High-resolution sample size enrichment of single-cell multi-modal low-throughput Patch-seq datasets

High-resolution sample size enrichment of single-cell multi-modal low-throughput Patch-seq datasets

Lorenzo Martini
Politecnico di Torino
Control and Computer
Engineering Department
Torino,10129, Italy
Email: lorenzo.martini@polito.it
0000-0002-7794-7791

Roberta Bardini
Politecnico di Torino
Control and Computer
Engineering Department
Torino,10129, Italy
Email: roberta.bardini@polito.it
0000-0002-1809-3212

Alessandro Savino
Politecnico di Torino
Control and Computer
Engineering Department
Torino,10129, Italy
Email: alessandro.savino@polito.it
0000-0003-0529-7950

Stefano Di Carlo
Politecnico di Torino
Control and Computer
Engineering Department
Torino,10129, Italy
Email: stefano.dicarlo@polito.it
0000-0002-7512-5356

***Abstract*—Single-cell multimodal technologies are becoming the hot topic of single-cell heterogeneity and function studies, promising to unravel the hidden relationship and functionalities of different aspects of the cells. Among the plethora of single-cell technologies, interesting is the patch-seq technology, which simultaneously performs Patch clamp measures and scRNA-seq on the same cells. However, given the experimental limitations of throughput of Patch clamp, the scRNA-seq analysis is challenging because it requires more samples to investigate cellular heterogeneity. Usually, the solution is associating the cells with the cell types in an existing scRNA-seq dataset. However, doing so loses part of the single cell resolution of the multimodal technique. Therefore, this work proposes a procedure leveraging the Seurat Integration process to find from a reference dataset the most similar cells to the ones from the patch-seq. The similarity is how much gene expression profiles are identical, and to evaluate that, this work defines various metrics based on Rand Index. In this way, one obtains a selection of suitable Reference cells to enrich the number of cells on which to perform multimodal investigation.**

***Index Terms*—Patch-seq, dataset enrichment, scRNA-seq, multimodal technologies**

I. Introduction

A brain is an incredible machine, composed of specific elements which define its highly specialized functions [1]. Even if the general brain cell categories are few (inhibitory/excitatory neurons and non-neuronal cells), there is a constellation of cellular subtypes differing in functionality and scope [2]. So far, a complete taxonomy of the cortical cell population is an almost unreachable goal in neuroscience, a relevant and impending obstacle in this field [3].

There are many differences, even considering only neuronal cells in the brain cortex. Historically, researchers investigate cells' dissimilarities based on morphology [4], location [5], electrophysiological characteristics [6], and molecular markers [7]. More recently, Next Generation Sequencing (NGS) technologies provide additional insight into the cells. In particular, single-cell profiling techniques, like single-cell RNA sequencing (scRNA-seq), help the cellular heterogeneity investigation [8]. However, these technologies often lead to cell-type characterizations that are not trivially overlapping [9].

For this reason, the research on cellular heterogeneity now focuses on multi-modal approaches, meaning a simultaneous analysis with two or more methodologies, allowing a direct comparison of the different results. In this regard, one of the latest technologies is Patch-sequencing (patch-seq) [10], which combines patch clamp, measuring electrophysiological properties, and scRNA-seq, which investigates gene expression. This technique can unravel the relationship between electrophysiological states and genome-wide transcriptomic profiles. However, these two technologies have different throughputs. scRNA-seq can profile tens of thousands of cells at once. Instead, a patch clamp rarely goes over 100 cells since it requires long and manual measurements on each cell.

The fact that the number of cells in a patch-seq dataset is bounded by the throughput of the patch clamp technique poses a *caveat* to the use of these datasets for transcriptomic data analysis, which usually requires a high amount of cells [11].

A common strategy to overcome this problem is to increase the number of cells for the gene expression analysis by

mapping the patch-seq cells to existing reference datasets of cell types [12]. Nonetheless, this method has the drawback of assigning cells to the general cell types of the reference dataset and generalizing their electrophysiological features to groups. This loses part of the multi-modal dataset resolution, particularly the direct and single-cell relation between the two modalities. Moreover, neuron electrophysiology studies aim to identify functional cell states (e.g., activity states or plasticity states [13]) not necessarily related to cell types. Therefore this bulk-fashion cell type-related method to increase the number of cells is not ideal.

As a first step to going beyond this limitation, this paper presents a single-cell data sample size enrichment tailored for patch-seq datasets with low cell numerosity. The proposed method is based on data integration analysis. It starts from the idea of integrating the patch-seq dataset with an appropriate reference dataset, finding and taking from it the most similar cells, and coupling them with the original ones. Based on the assumption that the connected cells are in the same state, the electrophysiological features and labels from the original dataset can be transferred to cells from the reference dataset (reference cells). The main advantage is maintaining the single-cell resolution in the multi-modal information, guaranteeing a more direct relation between electrophysiology and reference cells. The paper focuses on the methodology to find the correct cells from the reference dataset, the metrics to measure how similar they are to cells in the patch-seq dataset, and finally on the application in different case studies.

II. Background

Patch-sequencing data are an incredible source of biological information about neurons. However, their low throughput is incompatible with the sample size required for gene expression analysis. The general solution to enrich the dataset for gene expression analysis is to employ an external scRNA-seq dataset selecting cells with the same cell type as the patch-seq data. This usually relies on two available methodologies. The first is the implementation of a classifier, trained on the external dataset, to assign cell types labels to the patch-seq cells [12]. The second is label transfer through data integration, i.e., integrating the datasets to obtain a joint visualization and clustering of all the cells [14]. Integration can be implemented leveraging Satija lab's R package Seurat [15].

Independently from the methodology, each patch-seq cell is associated with a whole cluster or cell-type group. Therefore, all the multi-modal downstream analyses associate the single-cell electrophysiological features with entire groups of cells, losing part of the single-cell resolution in favor of a bulk approach. This work aims to find an alternative method to enrich the patch-seq data conserving at least in part the single cell resolution.

III. Methods

This work proposes a procedure to enrich the number of cells of a low-throughput patch-seq dataset referred to as Query Dataset (QD) with cells from a high-quality and highly-characterized external scRNA-seq dataset referred to as Reference Dataset (RD), maintaining as much as possible the multi-modal single-cell properties. The procedure leverages the data integration approach to find the cells in RD (reference cells) that are most similar to the cells in QD (query cells) based on their gene expression profiles. Data integration leverages the functionality of Seurat, an R package from Satija lab becoming a staple in single-cell analysis [15].

Following the workflow presented in Fig. 1, Section III-A presents how to select, preprocess and prepare the datasets, Section III-B explains how to perform data integration, Section III-C illustrates how to choose the cells, and Section III-D defines the metrics to evaluate the similarity between cells. Section IV finally presents experimental results.

A. Dataset Selection and Processing

The proposed workflow requires finding an external and suitable scRNA-seq dataset to use as RD to enrich a low-throughput QD (Fig.1-A). RD must have high experimental quality, ensuring its cells are well-characterized (i.e., high sequencing depth and accurate cell types labels). Fortunately, many projects study the brain, its functionality, and its cell type landscape, and they provide top-quality and freely available scRNA-seq datasets [16].

Patch clamp electrophysiology data derive from measures taken on neurons, whether excitatory or inhibitory. However, neurons are not the only cell type the brain comprises. In general, scRNA-seq datasets from cortex samples contain cells like Astrocytes and Oligodendrocytes that are not part of the patch-clamp experiments. Therefore, it is advisable to filter RD from non-neuronal cells before integration, thus reducing the possibility of integration with wrong cell types.

Both QD and RD must be preprocessed before integration. One crucial step is normalization. While the common practice is to perform log-normalization based on pseudo counts, an interesting alternative is using the SCTrasform technique proposed by Satija Lab and implemented in Seurat. SCTrasform is a *"modeling framework for the normalization and variance stabilization of molecular count data from scRNA-seq experiments"* [17]. It is an alternative normalization and scaling method, particularly effective in case of high differences in sequencing depth. As suggested by Seurat, performing SCTrasform on both QD and RD can improve the integration, limiting the effect of technical differences between them.

B. Finding Anchors and Data integration

Once QD and RD are correctly processed, data integration takes place. Seurat relies on the concept of *anchors* to perform data integration. The integration workflow consists of selecting features, finding anchors (Fig.1-B), and integrating the data (Fig.1-C). An anchor is a couple of cells from two different datasets assigned to the same functional state. Finding anchors allows us to identify the cell states existing in both datasets. In particular, finding anchors allows inferring the query cells type from the paired reference cells.

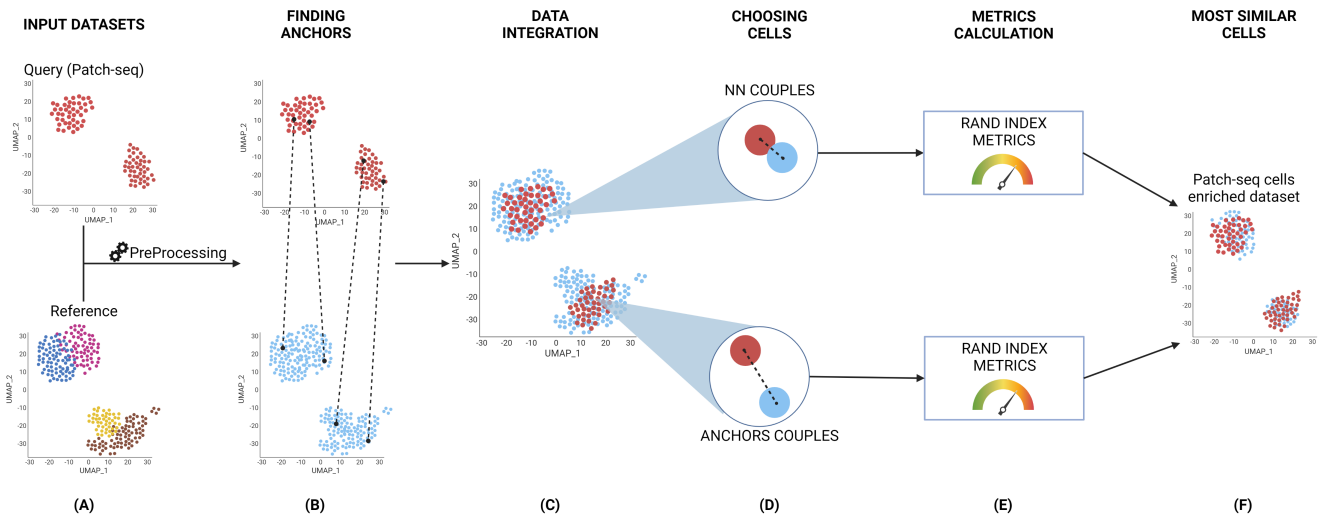

Fig. 1. Workflow. (A) QD and RD are first preprocessed in preparation for the following steps. (B) The procedure identifies the optimal anchors between QD (red) and RD (light blue), meaning couples of cells likely to share a cell state. (C) Integration of the data allows a joint 2D visualization of all the cells. (D) Selection of the reference cells related to the query cells, looking at Nearest Neighbors in the joint UMAP visualization and the integration anchors found in step B. (E) Calculate the similarity metrics based on the Rand Index of all identified couples. (F) Selection of the best reference cells (based on the Rand Index) to add to QD for downstream analyses.

Finding anchors requires a set of features to pair cells. This work uses the genes with high expression variability in both datasets. These genes are the ones that mainly characterize the cellular heterogeneity of the gene expression data, and anchors should find cells with similar gene variability patterns.

Anchors are identified using the Weighted-Nearest Neighbor (WNN) analysis, an unsupervised framework to learn the relative utility of each data type in each cell, enabling an integrative analysis of multiple modalities [18]. In particular, this work exploits the Seurat function `FindIntegrationAnchors` rather than the other similar function `FindTransferAnchors` even if it is computationally more expensive. This is because the first procedure aims for full data integration and not just a transfer of information (e.g., metadata such as cell type labels) from RD to QD. Moreover, Seurat recommends the `FindTransferAnchors` function when datasets are similar, with conserved cell type populations, and not too different in the number of cells, which is not the case considered in this paper.

The `FindIntegrationAnchors` function takes as input the two datasets and the previously selected features (i.e., genes with high expression variability). It produces the anchors, i.e., the couples of cells from the two datasets (Fig.1-B). In the case of datasets with comparable numbers of cells, this process would find several anchors and filter them to retain the best ones based on an internal score. However, when the query cells are few, none are filtered out, and the result is a list of about five reference cells per query cell.

Anchors can be used to perform datasets integration. Integration refers to combining the datasets at the data level for a joint analysis. This work implements it through the Seurat `IntegrateData` function, which computes a centered and corrected version of the data, starting from the expression differences of the anchor cells and modifying the expression data for all the other cells. This produces an integrated data matrix, where the features are the same set of genes, the cells are the two combined groups, and the data are a jointly corrected version of the original ones. This matrix can be analyzed as a gene expression matrix [19], obtaining a joint 2D visualization that in this paper is a Uniform Manifold Approximation and Projection (UMAP) (Fig.1-C), which allows investigating the query cells' position compared to the reference cells. This can be used to assess the query cell types based on their neighbors, which is fair to assume of the same kind.

C. Choosing cells

After obtaining the integrated dataset, it is, in principle, possible to link the electrophysiology data from the query cells with the reference cell types. However, a direct relationship between the gene expression profile and electrophysiological features is inaccurate. The only possible correlation is between electrophysiology data and neuron cell types, under the non-obvious hypothesis that neurons classified with the same cell type are in the same electrophysiological cell state. Indeed, cell states and cell types do not trivially correlate.

Therefore, this work aims to select only a few specific cells from RD that are "similar" to cells in QD to employ in the actual multi-modal study (Fig. 1-D). The cells in question must have the gene expression profiles similar to the query cells, so you can transfer whatever single-cell analysis was performed on QD to the reference coupled cells.

The main question is: which are the best cells to choose from RD? To answer this question, leveraging on the previous

computations, there are two options: (1) using the Nearest Neighbors (NN) concept and (2) employing the anchors.

NN are the cells with a lower euclidean distance from the query cells in the joint 2D UMAP. In this approach, the shorter the distance, the higher the cellular similarity (i.e., near cells likely are in the same cell state). This paper only considers the first NN, obtaining one neighbor reference cell for each query cell.

The anchors are couples of cells from the two datasets, and their computation process (see Section III-B) is naturally linked to cells sharing the same state between datasets. Thus, they are perfect candidates. There are several anchors for each query cell, each characterized by a score. Therefore, it is possible to decide how many of them to take based on this score. This ends up with 1 up to 4-5 reference cells per query cell.

Even if these two possible cell subgroups seem similar, they tend to consist of very different cells. Therefore, it is advised to understand which are the best, meaning the ones sharing the most similar gene expression profiles to the query cells. The following section proposes a set of metrics to accomplish this task.

D. Similarity Metrics

This paper relies on quantitative metrics to define the optimal reference cells for each query cell. The goal is to investigate if the cells identified in Section III-C have gene expression profiles comparable with their related query cells (Fig. 1-E). This is not a trivial problem. First, gene expressions range in a continuum, meaning it is impossible to assess their similarity through a simple comparison. Second, the difference in sequencing depth makes comparisons even harder. The wide discrepancies in data sparsity lead to genes detected in one dataset and not in others due to the experimental setup rather than to the biological information [11].

To overcome these limitations, this paper proposes to employ the Rand Index (RI) as a similarity metric. RI is used in data clustering to measure the similarity of data partitions [20]. Given two partitions, X and Y, of the same group S of n elements, RI is defined as:

$$RI(X,Y) = \frac{C}{\binom{n}{2}} \tag{1}$$

where C is the number of agreements between X and Y. In our problem, S is the set of considered genes, while X and Y are the gene expression profiles of two cells. RI cannot be computed on continuum values. The employed solution is to binarize the expression profiles (i.e., a gene can be either expressed or not expressed). RI indicates if two cells express the same set of genes, evaluating how much their expression profiles match.

When applied to the specific problem presented in this paper, RI can be calculated in different flavors. Let us denote with $\mathbf{CQ}_{|G|\times|Q|}$ the binarized raw count matrix of QD and $\mathbf{CR}_{|G|\times|R|}$ the binarized raw count matrix of RD, where G is the set of genes, Q the set of query cells, and R the set of reference cells. The standard RI between two cells $q \in Q$ and $r \in R$ is computed using (1) between the two vectors $\mathbf{CQ}_{|G|\times q}$, and $\mathbf{CR}_{|G|\times r}$.

Due to the high variability of scRNA-seq data, considering all genes in the datasets makes it hard to identify identical profiles between cells. Therefore, it is compelling to restrict the analysis to a reduced set of cell-state-related genes (relevant for downstream analyses) and have a higher chance of being concordant between cells. The proposed solution is to look at the genes with high expression variability coming from the pre-integration selection in Section III-B and a list of synaptic protein-related genes employed by Fuzik et al. in [12]. These synaptic protein-related genes are specific for patch-seq downstream functional analysis. In general, the approach is to look for functionally peculiar genes related to the target investigation. This led to the definition of two specific RI versions defined as:

$$\begin{aligned} &RI_{var}(\mathbf{CQ}_{|V|\times q}, \mathbf{CR}_{|V|\times r}) \\ &RI_{eph}(\mathbf{CQ}_{|E|\times q}, \mathbf{CR}_{|E|\times r}) \end{aligned} \tag{2}$$

where $V \subseteq G$ is the subset of the variable genes and $E \subseteq G$ is the subset of electrophysiology-related genes.

In case QD has a significantly low depth, very few RNA molecules may be detected, and too many genes appear not expressed in the cells. A good solution, in this case, is to define the strict RI between two cells q and r computed on the subset of genes expressed in the query cells to ensure the comparison of something experimentally detected:

$$RI^s(\mathbf{CQ}_{|S_q|\times q}, \mathbf{CR}_{|S_q|\times r}) \tag{3}$$

where $S_q \subseteq G$ is the subset of genes such that $\mathbf{CQ}_{s_q\times q} = 1$ with $s_q \in S_q$, meaning all the genes expressed in cell q. Based on this definition, it is possible to define the strict version RI^s_{var} and RI^s_{eph} of the two metrics defined in (2).

Finally, it is possible to calculate RI for two cells q and r on the integrated data:

$$RI^i(\mathbf{ID}_{|G|\times q}, \mathbf{ID}_{|G|\times r}) \tag{4}$$

where $\mathbf{ID}_{|G|\times|I|}$ is the integrated data matrix, and $I = Q \cup R$ is the ensemble of query and reference cells. Again, it is possible to compute the integration version RI^i_{var} and RI^i_{eph} of the two metrics defined in (2).

The proposed method computes all these metrics between the query cells and their respective reference cells from Anchors and NN couples. Given the multitude of variations of the metric, it is fair, for explanation purposes, to investigate the mean values of all of them. In particular, the mean between the three different features subsets on the standard version RI_{mean}, on the strict version RI^s_{mean}, and the integration version RI^i_{mean}. This estimation clarifies which are the most similar cells to each query cell, to which it is fair to assign the labels and analysis results from electrophysiology. The following section presents the results of the proposed procedure and the metrics obtained from two examples.

IV. Results

The proposed methodology was tested on two use cases.

The first use case aimed at performing cross-validation using an *ad hoc* scRNA-seq dataset created from the Allen Brain Map database [16] consisting of 14,249 cells from the motor and visual cortex of a mouse brain sample [21]. The Allen Institute provides very high-quality experimental data for various brain analyses. In particular, single-cell data are high-throughput and have high sequencing depth, therefore are the perfect candidates to be used as RD. As discussed in Section III-A, non-neuronal cells were removed from the dataset using only cells labeled as "Glutamatergic" or "Gabaergic." The resulting dataset comprising 13,411 cells was used to create QD and QD.

The second use case consisted of a patch-seq dataset of 83 neurons from a mouse used as QD. The scRNA-seq part of the data is freely available on the NCBI's Gene Expression Omnibus (GEO) portal with the accession code GSE70844. The full dataset of 13,411 neurons from the first use case was used as RD.

A. Cross-validation: Allen Subset vs. Allen

n

QD was created by randomly selecting 80 cells from the Allen dataset (after removing the non-neuronal cells), creating a synthetic analogous of a patch-seq dataset.

As a preliminary cross-check, RD was created considering all 13,411 cells, including those in QD, to assess that each query cell had one anchor corresponding to its copy in RD, confirming the capability of the approach to identify identical cells correctly.

The remaining part of the cross-validation was conducted by building RD from the 13,411 available neurons and removing the 80 selected query cells.

Fig. 2 shows the embeddings of RD (left) and QD (right) after applying preprocessing (see Section III-A). From RD, one can appreciate the different neuron subtypes, which compose the excitatory and inhibitory neurons of the cortex. In the case of QD, the 80 cells are equally divided into two groups, likely identifying only the two general types of neurons and not distinguishing all the different subtypes. This highlights how low-throughput datasets do not have a sufficient sample size to detect cellular heterogeneity properly.

The next step was SCTrasform, which resulted in a 17,419x80 matrix for QD and a 34,401x13,331 matrix for RD. The selection of the shared variable genes resulted in 400 genes. Employing RD, the anchor function returned 400 anchor couples, corresponding to about five reference cells for each query cell.

The next step was to check that the anchor couples were consistent in cell type, meaning the query cells were linked to the right cell types guaranteeing at least some similarity. This was trivially done by checking the labels of query and reference cells (which, in this case, come from the same dataset). Then, the actual integration followed. The result was a 2,000x13,491 data matrix containing the corrected gene expression values for all cells. Proceeding with the processing of the integrated matrix yielded the joint embedding of the cells. In the visualization, the general structure of the embedding remained relatively unchanged from that of RD alone, given the small number of cells added from QD. The query cells were spread across all the different cell subtypes clusters. Based on this UMAP embedding, the search for the NN of the query cells that, as explained in Section III-C, are good candidates for integration identified 80 reference cells, one for each query cell. The NN cells did not correspond to the anchors; they were a separate subset of cells. This highlights how identifying the best candidate cells for the query dataset enrichment is far from trivial. Therefore, the RI metrics investigation is fundamental. The best case scenario is to find couples of cells with RI ideally between 0.9 and 1, meaning they make an almost perfect match. Since the calculated RI metrics are many, to jointly account for their contribution to the quality of cell couples matching, this step considered their mean values (already explained at the end of Section III-D).

Fig. 3 shows a scatterplot of the cells visualizing their RI_{mean} against their RI^{s}_{mean}. The plot includes all the NN and the top 80 anchors (based on mean RI). One can observe that, in general, the identified cells have pretty high metrics values (around 0.8). Moreover, it appears that the anchors (in red) tend to have better similarity with QD than the NN (in light blue). Fig. 4 plots cells using RI_{mean} against RI^{int}_{mean}. Here, cells appear to be more scattered, and the higher RI values of anchors are less evident yet still present. As last Fig. 5 shows $RI_{varmean}$(the mean between versions of RI on the variable features) against $RI_{ephmean}$ (the mean between versions of RI on the electrophysiology genes) which are relevant and characterizing versions for the similarity investigation between cells. Again, the cells represented have both high metrics (around 0.8-0.9).

B. Patch-seq vs. Allen

The second use case applied the same analysis presented for the cross-validation to a 83 cells QD obtained from a patch-seq dataset [12], and the 13,249 cells RD mouse brain cortex Allen dataset already used for cross-validation [21].

Analyzing QD alone results in Fig. 6 show a population of cells that were not clearly divided into subgroups, making it difficult to distinguish between excitatory and inhibitory neurons, hampering cell type investigation and thus highlighting the need to enrich the sample size. Applying the proposed method, the first step was the SCTrasform of QD, resulting in an 8,853x83 matrix. Using this matrix, we could identify 415 anchor couples, corresponding, similarly to the previous case, to about 5 reference cells per query cell. At this point, the first differences from the prior case appeared. The scores attributed to the anchor cells were lower (with a mean of 0.53 among all cells, against 0.6 in the previous case). This was probably due to the more significant difference between QD and RD and the consequent higher difficulty in finding the proper anchors. The following step, i.e., data integration, created a

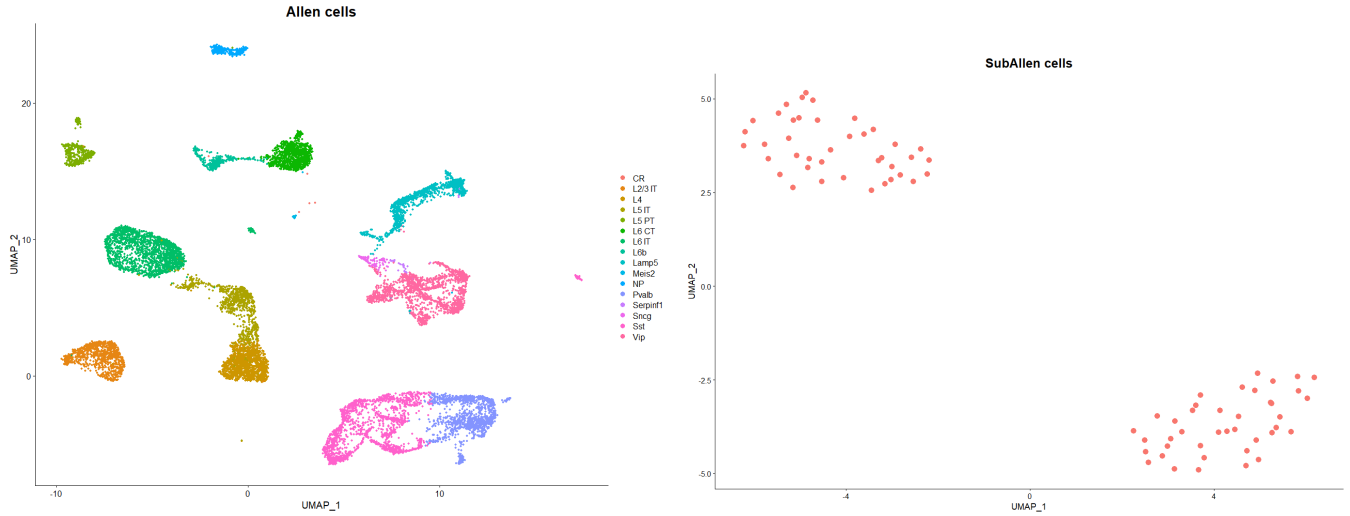

Fig. 2. UMAP visualization of the RD (left) and the 80 cells QD (right). The cell types indicated in the legend are all Excitatory and Inhibitory neuron subtypes. The analysis of the QD results in two groups, even if the low number of cells is insufficient for a more precise cellular heterogeneity analysis.

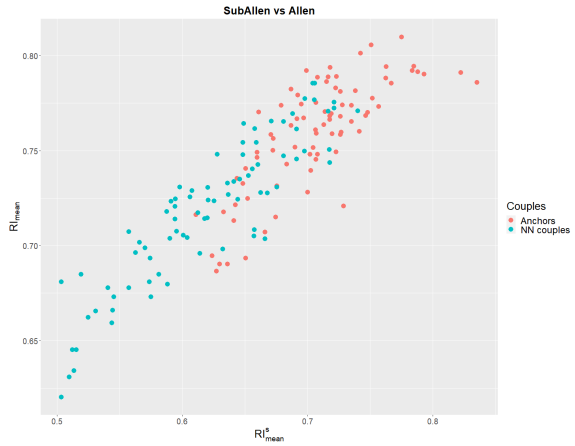

Fig. 3. Graph of the RI_{mean} against RI^{s}_{mean}. It highlights how the anchors appear to have overall better performances, especially for RI^{s}_{mean}.

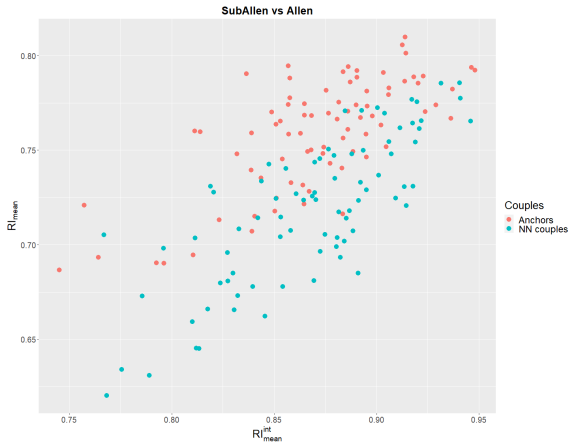

Fig. 4. Graph of the RI_{mean} against RI^{int}_{mean}. On the integrated version, the anchors are not better than the NN couples, but the latter has lower metrics in some isolated cases (bottom left).

2,000x13,574 matrix with the corrected integrated data. After processing it, we obtained a 2D joint visualization (Fig. 7), where again, there are not many differences with the RD alone, given the prevalence of reference cells compared to query cells. Interestingly, the query cells were not homogeneously divided across all cell types. There was an unbalance between query cells in the Vip/Lamp5 inhibitory neuron clusters to Pvalb/Sst inhibitory neuron clusters, which is in line with what was found by [12] on the same dataset. Again, it was trivial to find the 83 NN of the query cells based on this UMAP embedding. After identifying both the anchors and the NN, the procedure investigated their similarity with the query cells. Given the high discrepancy in sequencing depth and the consequent difference in the sparsity of the two datasets, cells could diverge in gene expression profile due to batch effects and not for actual discrepancies. For this reason, the strict version of RI, which considers only genes expressed in the query cells, is the best fit to measure the similarity. In this regard, Fig. 8 represents the NN and the top 83 anchors (by mean RI) in a graph of RI_{mean} against RI^{s}_{mean}.

Again, the anchors were the reference cells with higher performances, showing that they are the preferable cells to consider for integration. The RI_{mean} against RI^{int}_{mean} graph is not much different, showing high values for the RI metrics. Likewise, the RI metrics of $RI_{varmean}$ and $RI_{ephmean}$ were optimal (around 0.8), highlighting that these cells have gene expression similarities with the query cells on these relevant features, crucial for downstream analyses.

In both case studies, it was evident that the proposed method identified cells from RD with very similar gene expression profiles to QD, properly handling the differences in experimental quality between the two datasets. In particular, the anchors found by the integration process were the highest performing

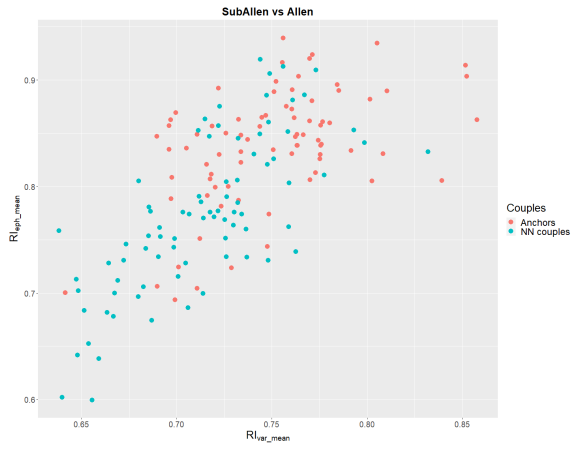

Fig. 5. Graph of the RI_{var_mean} against RI_{eph_mean}. These metrics are precious to investigating the relevant features for downstream analyses. The performances are quite high (especially for the anchors), ensuring the similarity of the couples of cells on these genes.

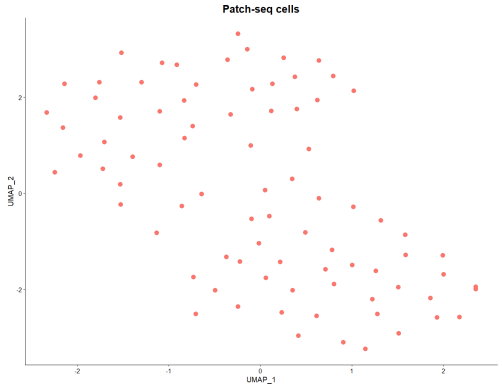

Fig. 6. patch-seq cells after processing. Even if not separated, one can identify two groups of cells, identifying Excitatory and Inhibitory neurons, but the study of the heterogeneity is limited.

ones, even if the NN found in the integrated visualization were also good candidates (Fig. 10).

V. Conclusions

In conclusion, this work proposed a procedure to enrich the sample size of low-throughput scRNA-seq experiments, and in particular multi-modal experiments like patch-seq, where manual experimental procedures limit the throughput. The method aimed to go beyond the standard practices based on cell-type transfer that employ an external reference dataset and associate query cells to clusters of reference cells through their cell-type label. This bulk-fashion analysis is not optimal for specific electrophysiology investigations since electrophysiological cell states are not trivially correlated with cell types. Therefore, this paper presented a procedure to find the most similar reference cells to the query cells, providing a single-cell sample enrichment. The method employed data integration analysis (through Seurat). Throughout this process, it was possible to find two types of cell couples, the anchors and the Nearest Neighbors. Moreover, the paper proposed

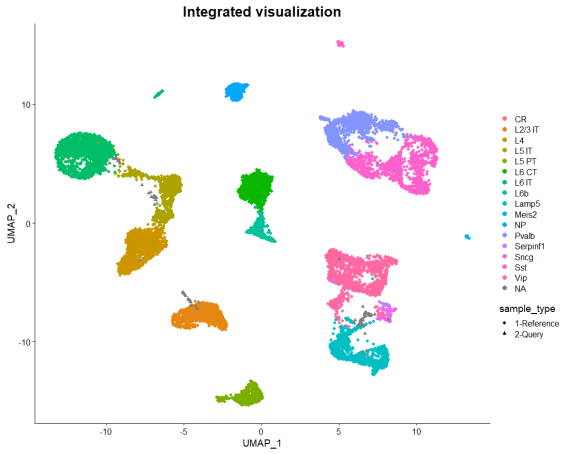

Fig. 7. Joint visualization of Query and Reference. The Query cells are not homogeneously divided between all cell types (as in the previous case), but there is a prevalence of them on Vip/Lamp5 Inhibitory neurons, which is in line with what was detected in [12]

a series of metrics based on the Rand Index to assess the similarities of the gene expression profiles of the coupled cells. These metrics were suitable to investigate cell similarities in case of high experimental differences between QD and RD, e.g., differences in sequencing depth and subsequent sparsity discrepancies. In particular, some of these metrics focused on specific sets of genes related to general cellular heterogeneity (such as variable features) and function (such as electrophysiology-related genes). The proposed method applied for cross-validation and to a patch-seq dataset led to interesting results. The technique generally identified optimal couples between query and reference cells, which showed overall high values on all the metrics. Interestingly, the anchors performed better than NN couples, highlighting a higher-performance type of couple.

For future works, it would be interesting to test this method between two patch-seq datasets, thanks to new strategies which

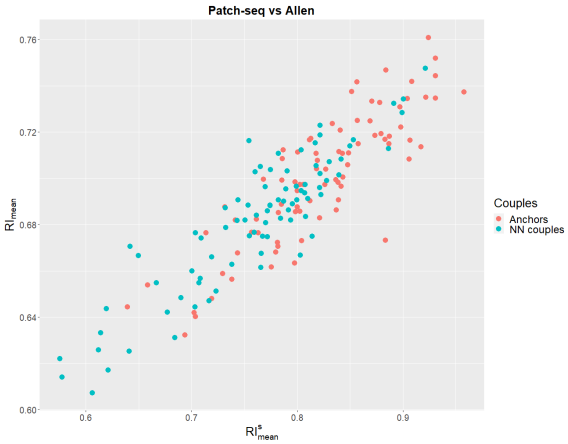

Fig. 8. Graph of the RI_{mean} against RI^s_{mean}. It highlights, again, how the anchors appear to have overall better performances, especially for RI^s_{mean}.

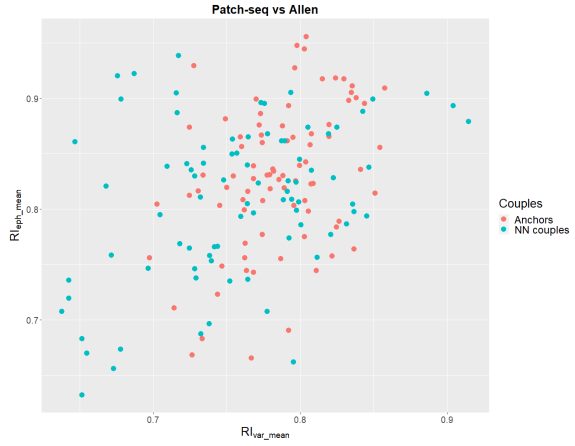

Fig. 9. Graph of the RI_{var_mean} against RI_{eph_mean}. These metrics are precious to investigating the relevant features for downstream analyses. The performances are quite high (especially for the anchors), ensuring the similarity of the couples of cells on these genes.

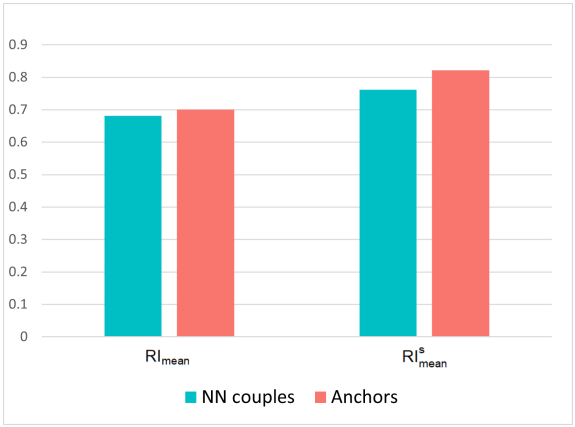

Fig. 10. Bar plots of the mean of RI_{mean} and RI^{int}_{mean} on all the Anchors and NN couples. It shows how the Anchors generally have better metrics than NN couples. It also shows how the strict version better uncovers the similarities of the cells.

promise to allow higher throughput for this technology [22]. Therefore, one could test that the chosen reference cells also have the same electrophysiological characteristics.

To conclude, this preliminary work aimed to highlight a different, more specific way to enrich low-throughput experiments as an alternative to the standard bulk-fashion integration methods. The results emphasized how this procedure is feasible and a promising approach for solving this problem.

VI. Data and code availability

The source code and all data required to reproduce the results presented in this paper and to apply the same procedure to other datasets are available on an Open Access repository: https://github.com/smilies-polito/patch-seq-sample-enrichment.

References

[1] S. Lodato and P. Arlotta, "Generating neuronal diversity in the mammalian cerebral cortex," *Annu Rev Cell Dev Biol.*, no. 31, pp. 699–720, 2015.

[2] A. Zeisel, A. B. Muñoz-Manchado, S. Codeluppi *et al.*, "Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq," *Science*, vol. 347, no. 6226, pp. 1138–1142, 2015. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.aaa1934

[3] J. A. Miller, N. W. Gouwens, B. Tasic *et al.*, "Common cell type nomenclature for the mammalian brain," *eLife*, vol. 9, p. e59928, dec 2020. [Online]. Available: https://doi.org/10.7554/eLife.59928

[4] H. Peng, P. Xie, L. Liu *et al.*, "Morphological diversity of single neurons in molecularly defined cell types." *Nature*, vol. 598, p. 174–181, 2021.

[5] T. P. O'Leary, K. E. Sullivan, L. Wang *et al.*, "Extensive and spatially variable within-cell-type heterogeneity across the basolateral amygdala," *eLife*, vol. 9, p. e59003, sep 2020.

[6] M. Carter and J. C. Shieh, "Chapter 4 - electrophysiology," in *Guide to Research Techniques in Neuroscience*, M. Carter and J. C. Shieh, Eds., 2010, pp. 91–118.

[7] L. Lyck, I. Dalmau, J. Chemnitz, B. Finsen, and H. Schrøder, "Immunohistochemical markers for quantitative studies of neurons and glia in human neocortex." *J Histochem Cytochem*, vol. 56, pp. 201–21, mar 2008.

[8] H. Byungjin, J. H. Lee, and B. Duhee, "Single-cell rna sequencing technologies and bioinformatics pipelines," *Experimental & Molecular Medicine*, vol. 50, no. 8, pp. 1–14, 2018.

[9] F. Scala, D. Kobak, M. Bernabucci *et al.*, "Phenotypic variation of transcriptomic cell types in mouse motor cortex." *Nature*, vol. 598, p. 144–150, 2021.

[10] M. Lipovsek, C. Bardy, C. R. Cadwell *et al.*, "Patch-seq: Past, present, and future," *Journal of Neuroscience*, vol. 41, no. 5, pp. 937–946, 2021.

[11] A. Haque, J. Engel, S. Teichmann *et al.*, "A practical guide to single-cell rna-sequencing for biomedical research and clinical applications." *Genome Med*, vol. 9, p. 75, 2017.

[12] J. Fuzik, A. Zeisel, Z. Máté *et al.*, "Integration of electrophysiological recordings with single-cell rna-seq data identifies neuronal subtypes." *Nat Biotechnol*, vol. 34, p. 175–183, 2016.

[13] C. Bardy, M. van den Hurk, B. Kakaradov *et al.*, "Predicting the functional states of human ipsc-derived neurons with single-cell rna-seq and electrophysiology." *Mol Psychiatry*, vol. 21, p. 1573–1588, 2016.

[14] M. Efremova and S. Teichmann, "Computational methods for single-cell omics across modalities." *Nat Methods*, vol. 17, pp. 14–17, 2020.

[15] T. Stuart, A. Butler, P. Hoffman *et al.*, "Comprehensive integration of single-cell data," *Cell*, vol. 177, no. 7, pp. 1888–1902, 2019.

[16] Allen Istitute for brain science, "© 2010 allen cell types database," Available at https://portal.brain-map.org/atlases-and-data/rnaseq.

[17] C. Hafemeister and R. Satija, "Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression," *Genome biology*, vol. 20, no. 1, pp. 1–15, 2019.

[18] Y. Hao, S. Hao, E. Andersen-Nissen *et al.*, "Integrated analysis of multimodal single-cell data," *Cell*, vol. 184, no. 13, pp. 3573–3587.e29, 2021.

[19] M. D. Luecken and F. J. Theis, "Current best practices in single-cell rna-seq analysis: a tutorial," *Molecular systems biology*, vol. 15, no. 6, p. e8746, 2019.

[20] J. R. Talburt, "3 - entity resolution models," in *Entity Resolution and Information Quality*, J. R. Talburt, Ed. Boston: Morgan Kaufmann, 2011, pp. 63–101.

[21] A. I. for brain science, "Mouse v1 and alm smart-seq," Available at https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-v1-and-alm-smart-seq.

[22] B. R. Lee, A. Budzillo, K. Hadley *et al.*, "Scaled, high fidelity electrophysiological, morphological, and transcriptomic cell characterization," *eLife*, vol. 10, p. e65482, aug 2021.