Towards non-arbitrary reproducibility of omics data clustering through recursivity

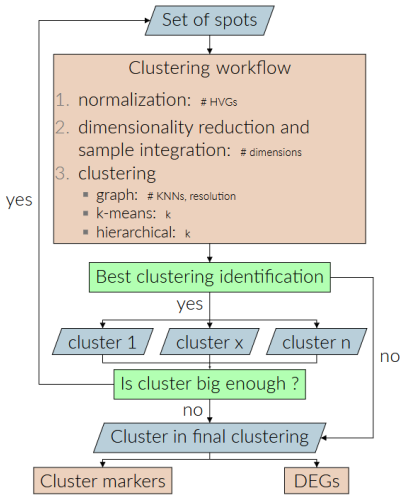

Fig. 1. Recursive clustering workflow.

Background Clustering samples with similar transcriptomic profiles into domains is one of the first and most important steps in the analysis of omics data. It is a process composed of several ordered highly parameterizable steps. Some parameters may drastically affect the final clustering in outcome. Besides, clustering algorithms are affected by randomness. To achieve reproducibility of the outcome, the seed, which controls the random number generator, has to be set to a constant value. However, the arbitrary way to handle reproducibility is rarely addressed in biological studies. Yet, it is an isssue of the utmost importance since downstream analyses, *e.g.* the identifcation of markers or differentially expressed genes, and the subsequent biological interpretation entirely rely on the quality of the clustering. In other words, the results of most biological studies that use clustering of omics data could be called into question simply by resetting the seed of the random number generator.
We have developed a recursive clustering workflow that iteratively processes an initial dataset in multiple independent rounds of sub-clustering (Fig. 1). At each round, the goal is to find a non-arbitrary reproducible partition of the dataset, *i.e.* without setting the seed. To achieve this goal, we hypothesize that the partition in a small number of clusters would be more likely to be reproducible in a non-arbitrary way and driven by the strongest source of biological variation in the dataset. Thus, the challenge consists in extracting the minimum amount of biological information necessary to make such a partition. We expect this approach to make clustering of omics data more robust and most biologically meaningful.

Methods After normalising raw counts, we identified highly variable genes (HVGs) with different residual variance threshold values and reduced the dimensionality of the dataset by PCA with 50 principal components (PCs). To extract the minimal biological information necessary to make a partition in at least two clusters, we used the elbow method to identify the optimal number of first PCs for the clustering. To go further in this direction, we also evaluated the extreme case of only considering the first PC. We then performed the clustering in the obtained low-dimensionsal space with a maximum number of clusters of 4 using different algorithms: graph-based clustering with a varying number of nearest neighbors and different resolution values to identify communities with the Louvain algorithm, k-means clustering and hierarchical clustering with euclidean distance. Before each clustering, we set the seed to the set to the date and time with a second accuracy before each clustering function call, guaranteeing a different seed for each clustering.

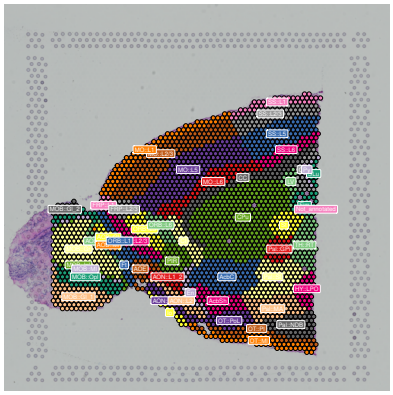

Fig. 2. Pathologist annotations of the mouse brain Visium dataset provided by 10x Genomics.

Only clusterings that are non-arbitrarily reproducible over 20 iterations, *i.e.* generating identical clusterings for all iterations, were retained and the clustering with the highest mean silhouette score was selected. This entire workflow, from the normalization to the clustering itself, was applied recursively to all the obtained clusters.
We also applied the classical clustering approach, which consists in analysing the whole dataset at once. To do so, we used the popular Seurat R package with the default parameter values at each each step of the clustering workflow, *i.e.* 3 000 HVGs, 30 first PCs for the integration of multiple samples, if any, 10 first PCs to build a 20-Nearest Neighbors graph. We also evaluated the 30 first PCs, since it is a common practice to capture all the biological heterogeneity. We identified the resolution value to use with the Louvain algorithm to generate a clustering with the same number of clusters as we obtained with our recursive approach.
For both approaches, we repeated the entire clustering workflow 10 times, compared all the obtained clusterings to each other and measured their similarity with the adjusted Rand Index (ARI). When available, we also used the pathologist annotations as ground truth to assess the quality of the clusterings using the normalized mutual information (NMI).

Results We analysed a Visium dataset from a mouse brain tissue section provided by 10x Genomics that was annotated by pathologists (Figure 2). To extract the minimal biological to partition the dataset in at least two clusters, we evaluated three different residual variance threshold, from 1.3 to 2, and two ways to reduce the dimensionality of the dataset: either the elbow method or only the first PC. The recursivity of our workflow performed multiple independent rounds of sub-clustering with different parameter settings (data not shown) and generated final recursive clustering composed of 48 clusters using the first PC and from 43 to 48 clusters with the elbow method. It appeared that the less highly variable genes and the less PCs are used, the better the reproducibility is (Figure 3.A and B).

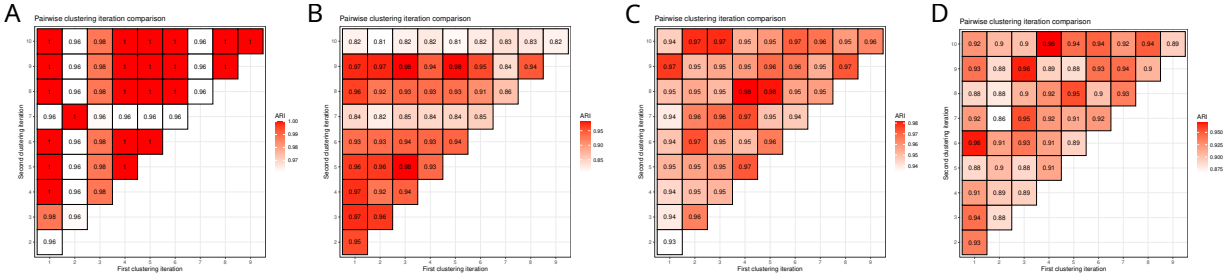

Fig. 3. Non-arbitrary reproducibility of the clustering of the 10x Genomics dataset: recursive approach using the first PC (A) or the elbow method (B) and the classical approach with the 10 (C) or 30 (D) first PCs.

Using a highest residual variance threshold of 2 and only the first PC, the 10 repeats of the recursive clustering were indeed very close to be identical. Using a higher number of dimensions with the elbow method reduces the reproducibilty but increases the biological quality compared to the ground truth annotations (NMI=0.642 with the elbow method, NMI=0.609). The classical approach yielded better reproducibility but worse results with a low number of dimensions (NMI=0.613 with 10 PCs, NMI=0.651 with 30 PCs). Overall, these results suggest that accurately identifying the minimum set of HVGs and PCs at each round of clustering with our approach could lead to non-arbitrarily reproducible and biologically relevant clustering.

We applied the same type of analysis to an in-house Visium dataset composed of brain tissue slices from two preeclamptic and two control mouse dams. Using a residual variance threshold of 1.5 to identify HVGs and the elbow method to reduce the dimensionality of the dataset at each round of clustering, we obtained final recursive clusterings composed of 29 or 30 clusters over 10 repetitions of the whole process. These replicates were almost perfectly identical with ARI beteween 0.96 and 1 (Figure 4.A).

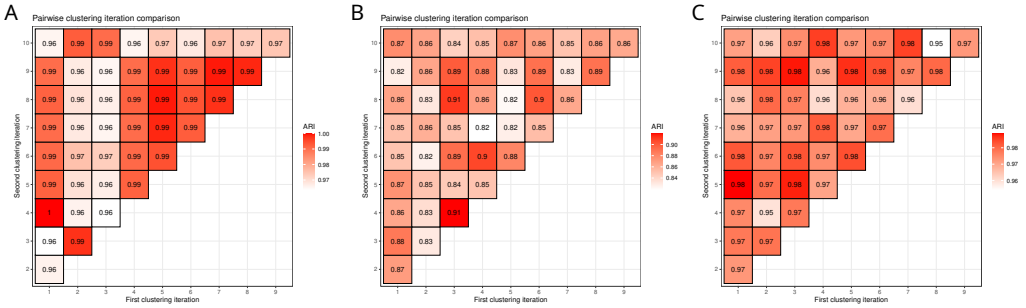

Fig. 4. Non-arbitrary reproducibility of the clustering of the preeclampsia dataset: recursive approach using the elbow method (A) and the classical approach with the 10 (B) or 30 (C) first PCs.

The classical clustering approach were much less reproducible with 10 dimensions (Figure 4.B) and provided slightly less reproducible clustering iterations in comparison with our recursive approach when increasing the number of dimensions to 30 (Figure 4.C). In addition to improving non-arbitrary reproducibility, our recursive clustering approach idenitfied well spatially defined cortical layers and hippocampal subfields (results not shown).

Perspectives The aim of our ongoing project is to find reproducible omics data clustering in a non-arbitrary way. Breaking down the challenging task of clustering a highly diverse high-dimensional dataset into successive simpler steps is key to achieving non-arbitrary reproducibility of omics data clustering. We hypothesise that identifying the minimum amount of biological information, *i.e.* the smallest set of HVGs and fewest PCs, necessary to reliably partition the dataset into at least two clusters in each step may enable to reach this goal. We will pursue of the methodological improvement of our approach in this direction and enhance its evaluation by using simulated datasets. We will also apply our approach to other Visium datasets from different organs, species, and with different sample sizes, as well as to single-cell RNA-seq datasets.