# Nested stochastic block models applied to the analysis of single cell data

Leonardo Morelli[1,2], Valentina Giansanti[1, 3], Davide Cittaro[1*],

**1** Center for Omics Sciences, IRCCS San Raffaele Institute, Milan, Italy
**2** Università Vita-Salute San Raffaele, Milan, Italy
**3** Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy

* cittaro.davide@hsr.it

## Abstract

Single cell profiling has been proven to be a powerful tool in molecular biology to understand the complex behaviours of heterogeneous system. While properties of single cells is the primary endpoint of such analysis, these are typically clustered to underpin the common determinants that can be used to describe functional properties of the cell mixture under investigation. Several approaches have been proposed to identify cell clusters; while this is matter of active research, one popular approach is based on community detection in neighbourhood graphs by optimisation of modularity. In this paper we propose an alternative solution to this problem, based on nested Stochastic Block Models; we show a threefold advantage of our approach as it is able to correctly identify cell groups, it returns a meaningful hierarchical structure and, lastly, it provides a statistical measure of association between cells and the assigned clusters.

## Author summary

Identification of cell types is a key step of many single cell experiments. Many of the approaches to achieve this goal converge on the analysis of the graph representing cell-wise similarity by optimisation of modularity. While fast, this method does not provide a robust estimation of the number of groups, and it leaves large room of arbitrariness in the choice of appropriate resolution. To overcome these limitations, we propose a strategy based on nested stochastic block models. We show our method is accurate and provides a rich description of cell groups and their relationships.

## Introduction

Transcriptome analysis at single cell level by RNA sequencing (scRNA-seq) is a technology growing in popularity and applications [1]. It has been applied to study the biology of complex tissues [2, 3], tumor dynamics [4–7], development [8, 9] and to describe whole organisms [10, 11].

A key step in the analysis of scRNA-seq data and, more in general, of single cell data, is the identification of cell populations, groups of cells sharing similar properties. Several approaches have been proposed to achieve this task, based on well established clustering techniques [12, 13], consensus clustering [14, 15] and deep learning [16]; many more have been recently reviewed [17, 18] and benchmarked [19]. As the popularity of single cell analysis frameworks Seurat [20] and Scanpy [21] raised, methods based instead on graph partitioning became the *de facto* standards. Such methods require the construction of a cell neighbourhood graph (*e.g.* by $k$ Nearest Neighbours, $k$NN) which is then partitioned into communities; the latter step is typically performed using the Louvain method [22], a fast algorithm for optimisation of graph modularity. While fast, this method does not guarantee that small communities in large networks are well defined. To overcome its limits, a more recent approach, the Leiden algorithm [23], has been implemented and it has been quickly adopted in the analysis of single cell data, for example by Scanpy and PhenoGraph [24]. In addition to Newman's modularity [25], other definitions currently used in single cell analysis make use of a resolution parameter [26, 27] . In lay terms, resolution works as a threshold on the density within communities: lowering the resolution results in less and sparser communities and *viceversa*. Identification of an appropriate resolution has been recognised as a major issue [28], also because it requires the definition of a mathematical property (clusters) over biological entities (the cell groups), with little formal description of the latter. In addition, the larger the dataset, the harder is to identify small cell groups, as a consequence of the well-known resolution limit [29]. Moreover, it has been demonstrated that random networks can have modularity [30] and its optimisation is incapable of separating actual structure from those arising simply of statistical fluctuations of the null model. Additional solutions to cell group identification from neighbourhood graphs have been proposed, introducing resampling techniques [31] or clique analysis [32]. Lastly, it has been proposed that high resolution clustering, *e.g.* obtained with Leiden or Louvain methods, can be refined in agglomerative way using machine learning techniques [33].

An alternative solution to community detection is the Stochastic Block Model, a generative model for graphs organized into communities [34]. In this scenario, identification of cell groups requires the estimation of the proper parameters underlying the observed neighbourhood graph. According to the microcanonical formulation [35], the parameters are node partitions into groups and the matrix of edge counts between groups themselves. Under this model, nodes belonging to the same group have the same probability to be connected with other nodes. It is possible to include node degree among the model parameters [36], to account for heterogeneity of degree distribution of real-world graphs. A Bayesian approach to infer parameters has been developed [37] and implemented in the *graph-tool* python library (https://graph-tool.skewed.de). There, a generative model of network $\boldsymbol{A}$ has a probability $P(\boldsymbol{A}|\boldsymbol{\theta}, \boldsymbol{b})$ where $\boldsymbol{\theta}$ is the set of parameters and $\boldsymbol{b}$ is the set of partitions. The likelihood of the network being generated by a given partition can be measured by the posterior probability

$$P(\boldsymbol{b}|\boldsymbol{A}) = \frac{P(\boldsymbol{A}|\boldsymbol{\theta}, \boldsymbol{b})P(\boldsymbol{\theta}, \boldsymbol{b})}{P(\boldsymbol{A})} \tag{1}$$

and inference is performed by maximising the posterior probability. The numerator in this equation can be rewritten exponentiating the description length

$$\Sigma = -\ln P(\boldsymbol{A}|\boldsymbol{\theta}, \boldsymbol{b}) - \ln P(\boldsymbol{\theta}, \boldsymbol{b}) \tag{2}$$

so that inference is performed by minimizing the information required to describe the data (Occam's razor); *graph-tool* is able to efficiently do this by a Markov Chain Monte Carlo approach [38]. SBM itself may fail to identify small groups in large graphs, hence hierarchical formulation has been proposed [39]. Under this model, communities are agglomerated at a higher level in a block multigraph, also modelled using SBM. This process is repeated recursively until a graph with a single block is reached, creating a Nested Stochastic Block Model (nSBM).

In this work we propose nSBM for the analysis of single cell data, in particular scRNA-seq data. This approach identifies cell groups in a statistical robust way and, moreover, it is able to determine the likelihood of the grouping, thus allowing model selection. In addition, it is possible to measure the confidence of assignment to groups. We show that such information may be exploited to perfect the notion of cell groups and the identification of markers.

We developed *schist* (https://github.com/dawe/schist), a python library compatible with *scanpy*, to facilitate the adoption of nested stochastic block models in single-cell analysis.

# Materials and methods

## Analysis of cell mixtures

Data and metadata for five cell mixture profiled by Chromium 10x were downloaded from the sc-mixology repository (https://github.com/LuyiTian/sc_mixology). Data were analysed using scanpy v1.4.6 [21]. Cells with less than 200 genes were excluded, as genes detected in less than 3 cells. Cells with less than 5% of mitochondrial genes were retained for subsequent analysis. Data were normalised and log-transformed; number of genes and percentage of mitochondrial genes were regressed out. nSBM was initialised three times. Analysis was performed at level 2 of the nSBM hierarchy. Random Poisson noise was generated at given $\lambda$ (range: 0 - 1000) and added to the initial count matrix. Analysis was performed at level 2 of the nSBM hierarchy.

## Analysis of hematopoietic differentiation

Data were retrieved using scanpy's built-in functions and were processed as in [48], except for $k$NN graph built using 30 principal components, 30 neighbours and diffmap as embedding. Gene signatures were calculated using the following gene lists

- Erythroids: Gata1, Klf1, Epor, Gypa, Hba-a2, Hba-a1, Spi1
- Neutrophils, Elane, Cebpe, Ctsg, Mpo, Gfi1
- Monocytes, Irf8, Csf1r, Ctsg, Mpo

nSBM was completed with 3 initialisations

## Analysis of cluster consistency

Count matrices were downloaded from GEO using the following accession numbers: GSE133535 (Chromium 10Xv3), GSE133543 (Quartz-seq2), GSE133542 (MARS-seq) and GSE133541 (iCELL8). Data were processed according to the methods in the original paper [42]. Briefly, cells with less than 10,000 total number of reads as well as the cells having less than 65% of the reads mapped to their reference genome were discarded. Cells in the 95th percentile of the number of genes/cell and those having less than 25% mitochondrial gene content were included in the downstream analyses. Genes that were expressed in less than five cells were removed. Data were normalized and log-transformed, highly variable genes were detected at minimal dispersion equal to 0.5. Neighbourhood graph was built using 30 principal components and 20 neighbours. nSBM was completed with 3 initialisations.

## Overview of *schist*

*schist* is a convenient wrapper to the *graph-tool* python library, written in python and designed to be used with *scanpy*. The most prominent function is *schist.inference.nested_model()* which takes a *AnnData* object as input and fits a nested stochastic block model on the $k$NN graph built with *scanpy* functions (*e.g.* *scanpy.tools.neighbors()*). When launched with default parameters, *schist* fits a model which maximises the posterior probability of having a set of cell groups (or blocks) given a graph. *schist* annotates cells in the data object with all the groups found at each level of a hierarchy. As there could be more model fits with similar entropy, *schist* could explore the space of solutions with a Markov Chain Monte Carlo (MCMC) algorithm, to perform model averaging; this step is performed until it converges, that is the difference in model entropy in $n$ continuous iterations remains under a specified threshold. Sampling from the posterior distribution can be used to study the distribution of the number of groups, at each level. This information (group marginals) could be studied to identify the most probable number of cell groups. Once *schist* has fitted a model, it evaluates the difference in entropy given by assigning every cell to every possible group. This step generates the matrix of *cell affinity*, that is the probability for a cell to belong to a specific group. A cell affinity matrix is generated and returned for every hierarchy level. We show that this information can be useful to evaluate cluster consistence.

## nSBM correctly identifies cell populations

To benchmark *schist*, we tested our approach on scRNA-seq mixology data [40], in particular on a mixture of 5 cell lines profiled with Chromium 10x platform. At a first evaluation of the UMAP embedding, all lines appear well separated. Only the lung cancer line H1975 shows a certain degree of heterogeneity with some cells being embedded in other cell groups (Fig. 1A). Inference on the neighbourhood graph is influenced by the graph structure itself, therefore we built multiple graphs changing the number of principal components used in PCA reduction and the number of neighbours in the $k$NN graph. We then calculated the Adjusted Rand Index ($ARI$) between the cell line assignments (ground truth) and the cell groups identified by nSBM at each level. We found a peak of $ARI = 0.977$ with 30 principal components (PC) and 30 neighbours. In general, higher number of components and neighbours has a positive impact on the performance (Fig. 1B). Conversely, if few PCs (10) or neighbors (5) are used, performances degrade, with a minimum $ARI = 0.669$ at 20 PCs and 5 neighbours. If fewer PCs are used, a smaller fraction of the total variance, hence less information, is used to build the $k$NN graph; if fewer neighbours are chosen, the graph is sparser and the model is fit from less edges (S1 Fig). Running MCMC algorithm recovers the performances of the majority of the configurations (S2 Fig).

Analysis of the nSBM hierarchy reveals that five levels are needed to describe the experiment (Fig. 1C, upper panel), with level 2 properly catching the cell identity ($ARI = 0.977$). In addition to the five major groups, observe two small groups, summing to 11 cells, that were merged to HCC827 and H838 at hierarchy level 3. Interestingly, these groups are enriched in cells whose identity was reassigned from H1975 to H838 or HCC827 in the original paper using Demuxlet [41], indicating that nSBM was able to recognise peculiar properties and isolate them. It may be worth mention that the second best ranked group by cell affinity for these 11 cells is the correct group assigned in the original paper, except for a single cell assigned to H2228. As a high separation between cell lines is observable, optimisation of modularity by Leiden algorithm is also able to identify cell identities with high precision, given that a proper resolution threshold is set (Fig. 1C, lower panel); we found that when resolution is set to 0.05 the cell lines are properly separated ($ARI = 0.975$), with the exception of the above mentioned cells.

It is worth noting that since nSBM reflects the amount of information contained in the data, it does not return clusters when data are noisy. To demonstrate this property, we added increasing levels of random Poisson noise to the count matrix (Fig.2). The number of groups reported by nSBM decreases with increasing amount of noise and it is 1 (*i.e.* no clusters) at $\lambda = 200$, whereas a standard approach overfits data.

These observations show that nSBM is suitable for accurate identification of cell groups, without the need of an arbitrary threshold on the resolution parameter and with the possibility to identify rare cell types in larger populations. Moreover, nSBM avoids excessive clustering of data in presence of noisy measurements.

**Fig 1.** *schist* applied to scRNA-seq mixology data. (A) UMAP embedding of 10x Chromium data, cells are colored according to the given cell line in the original paper. A small number of H1975 cells are found in HCC827 and H838 clusters. (B) Heatmap showing the maximal Adjusted Rand Index for different $k$NN graphs. We tested the impact of varying the number of Principal Components and the number of neighbors used in *sc.pp.neighbors()* function in *scanpy*. Adjusted Rand Index between the actual cell lines and the identified groups is shown. Darker blue indicates higher concordance between the model and the ground truth. (C) Alluvial plots showing the hierarchy of cell groups as identified by *schist* (above) or by Leiden method at different resolution thresholds (below). The bars on the right indicate the cell identity; two marks in the *schist* plot indicate two groups of cells discussed in the main text

**Fig 2.** Identification of cell groups at different level of random noise. When random Poisson noise is added to the gene count martix, at increasing values of $\lambda$, a strategy based on optimization of modularity does return cell groups, whereas nSBM does not cluster cells when the noise exceeds a certain threshold (solid lines). The behaviour of the former approach follows the trend of the modularity (dashed lines), which is sustained at high levels independently from the level of noise.

## Model hierarchy contains biological information

The hierarchical model of cell groups implies that a relationship exists between groups. We next wanted to explore if the hierarchy proposed by the nSBM had a biological interpretation. To this end, we analysed data for hematopoietic differentiation [47], previously used to benchmark the consistency of cell grouping with differentiation trajectories by graph abstraction [48]. Standard processing of those data reveals three major branchings (Erythroids, Neutrophils and Monocytes) stemming from the progenitor cells (S3 FigA). After applying nSBM, we identify 27 groups at level 3 of the hierarchy (S3 FigB), compared to the 24 using Leiden method at default resolution (S4 Fig). We found that the hierarchy proposed by our model is consistent with the developmental model (Fig. 3). Of note, we found that clustering with Leiden method produces cell groups that are mixed and split at different resolutions (0.1 - 1), in a non hierarchical manner (S4 Fig); we spotted several occurrences of such phenomenon, *e.g.* group 9 at resolution $r = 0.4$ splits into groups 0 and 6 at $r = 0.3$ or group 3 at $r = 0.6$ splits in groups 4, 8 and 12 at $r = 0.5$.

In all, these data suggest that not only nSBM is able to identify consistent cell groups at different scales, but also that the hierarchy proposed by the model has a direct biological interpretation.

**Fig 3.** Analysis of hematopoietic differentiation. Each panel presents a low dimensional embedding of single cells next to a radial tree representation of the nSBM hierarchy. Cells are colored according to groupings at level 5 of the hierarchy, group 0 marks the progenitor population (A). In subsequent panels, cells are colored using a signature of erythroid lineage (B), monocytes (C) or neutrophils (D).

## Cell affinities can be used to evaluate cluster purity

The computational framework underlying *schist* calculates the model entropy, that is the amount of information required to describe a block configuration. Given that minimisation of such quantity can be used to perform model selection, it can be also used to evaluate the impact of modifying the assignment of a cell to a cluster. Once a model is minimised, *schist* performs an exhaustive exploration of all model entropies resulting from moving all cells into all possible clusters. The differences in entropies could be interpreted as affinities of cells to given clusters. Such affinities are, in fact, probability values and could be used to evaluate the internal consistency of a given cell cluster.

To this end we calculate the entropy of the group-wise distribution of cell affinities, which is maximal when all cells have affinity equal to 1 for a given group. We tested this idea on four datasets recently published to benchmark single cell technologies in the Human Cell Atlas project [42]; in particular, we chose two technologies resulting in high quality data: Quartz-seq2 [43] and Chromium 10x v3 [44], and two technologies resulting in more noisy data: MARS-seq [45] and iCell8 [46] (Fig. 4).

Cluster consistency is not a measure of the data quality, in fact we identify low consistency groups in all datasets. High consistency, instead, appears to be linked to the biological purity of the cells and it is inverse to the diversity index, estimated using cell annotation from the original paper. Consequently, filtering low consistency groups increases concordance with biological groups, at the cost of a reduced number of cells ( S5 Fig).

Similarly, we can use cell affinities to derive a stability parameter, a measure of the tendency for a cell to be stably associated to given clusters at all levels of the hierarchy. To this end, we first calculate the cell-wise entropy $H_{i,h}$ of cell affinity at each hierarchy level $h$, then we define the stability as $S_i = 1 - \max(H_i)$. While we conceived this measure to identify and exclude cells with dubious assignment, we found that it may be more useful to assess the general data quality: the fraction of cells having $S > 0.95$ was 0.783, 0.795, 0.831 and 0.855 for the iCELL8, MARS-seq, Chromium 10x and Quartz-seq2 technology respectively, in line with the evaluation on increasing performances of those platforms in [42].

**Fig 4.** Analysis of cell cluster consistency. Every panel reports a UMAP embedding of a PBMC + HEK293 cells profiled on different platform. Cells are annotated by cell type and by consistency value, which is assigned to cell clusters at nSBM level 1. The charts next to UMAPs show the correlation between consistency and diversity index for each cell cluster. Technologies showed here are (A) Chromium 10x v3, (B) Quartz-seq 2, (C) MARS-seq and (D) iCELL8.

## Analysis of runtimes

Minimisation of the nSBM is a process that may require a large amount of computational resources. The analysis of a relatively small scRNA-seq dataset, such as the ones in [42], may require several minutes to be processed. This could be a serious limitation to the adoption of nSBM in the analysis of single cell data, especially because several parameters should be tested. To overcome this limitation, it was suggested to let a greedy merge-split MCMC algoritthm [49] to explore the solutions and stop iterations when the difference in entropy is below a defined threshold. We tested this approach on a commodity hardware (MacBook Air, dual core 1.6 GHz i5 processor, 16 GB RAM) and compared to the default approach. Results are reported in Table 1. The merge-split algorithm greatly reduces the time needed to propose the final model. In addition, the partitions found are largely overlapping the ones found by the default approach.

| **Dataset** | **Cells** | **Minimize** | **Merge-split MCMC** | **Overlap** |
|---|---|---|---|---|
| sc-mixology [40] | 860 | 01:11 | 00:03 | 0.884 |
| Quartzseq [42] | 1266 | 00:31 | 00:02 | 0.726 |
| MARS-seq [42] | 1401 | 00:29 | 00:08 | 0.834 |
| Chromium 10X [42] | 1523 | 00:40 | 00:04 | 0.695 |
| iCELL8 [42] | 1830 | 00:54 | 00:07 | 0.623 |
| Paul15 [47] | 2730 | 02:37 | 00:10 | 0.575 |
| Planaria [10] | 21612 | 13:12 | 03:29 | 0.589 |

**Table 1.** Time required to minimise the nSBM using the default minimization method compared to the greedy merge-split MCMC. Times are expressed in mm:ss. Partition overlap measures concordance between the two models over the full hierarchy. Timing is the average after 3 initialisations.

# Conclusion

Identification of cells sharing similar properties in single cell experiments is of paramount importance. A large number of approaches have been described, although the standardisation of analysis pipelines converged to methods that are based on modularity optimisation. We tackled the biological problem using a different approach, nSBM, which has several advantages over existing techniques. The most important advantage is the hierarchical definition of cell groups which eliminates the choice of an arbitrary threshold on clustering resolution. In addition, we showed that the hierarchy itself could have a biological interpretation, implying that the hierarchical model is a valid representation of the cell ensemble. Our approach introduces the evaluation of cluster consistency, which can be used to isolate cells with heterogeneous identity. Lastly, a statistical way to evaluate models is made available, allowing for reliable model selection. This last capability has the obvious advantage that the choice of parameters, hence the definition of cell clusters, could be conditioned to an evaluation metric which is robust and easy to understand (*i.e.* the model entropy).

The major drawback of adopting this strategy is the substantial increase of runtimes. According to the developers of *graph-tool*, runtimes are proportional to the number of edges in the neighbourhood graph and while it supports CPU-level parallelisation, a model minimisation is hundreds times slower than the extremely fast Leiden approach. Nevertheless, we show that a greedy merge-split MCMC algorithm can overcome this limitation, achieving performances that allow the usage of *schist* on standard desktop hardware to analyse various single cell datasets.

# Supporting information

**S1 Fig. Degree distribution of $k$NN graphs.** Degree distribution of multiple $k$NN graphs derived from scRNA-seq mixology datasets using variable number of

Principal Components or number of neighbors. Each histogram shows the number of nodes (on y axis) within a specific degree bin (on x axis). Both the parameters influence the sparseness of the graph.

**S2 Fig. nSBM performance after MCMC.** Adjusted Rand Index for different $k$NN graphs after MCMC run. Maximal ARI over all hierarchy level is shown. Darker color indicates higher concordance with the ground truth.

**S3 Fig. Analysis of hematopoietic differentiation.** (A) Low dimensional embedding of single cells colored by original cell type and pseudotime. (B) Cells are colored according to the nSBM grouping at level 3 of the hierarchy, next to a radial tree representation of the same model.

**S4 Fig. Non hierarchical clustering at different resolution.** Low dimension embedding of single cells for hematopoietic differentiation colored according to Leiden clustering at decreasing resolution, from 1.0 to 0.1. Lowering the distribution does not grant that cells are grouped in a hierarchical way, *e.g.* group 3 at resolution $r$=0.3 splits in groups 2 and 0 at resolution $r$=0.2.

**S5 Fig. Cluster consistency and cell type identification.** Adjusted Rand Index between cell clusters and cell type annotation filtering data at different cutoffs of consistency. Dot size is proportional to the number of cells remaining after filtering.

## Acknowledgments

We would like to thank Tiago de Paula Peixoto (Central European University, ISI Foundation) and Giovanni Petri (ISI Foundation) for the discussions and the precious hints. We also would like to thank all people at COSR, in particular Giovanni Tonon and Paolo Provero.

## References

1. Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the past decade. Nature Protocols. 2018;13(4):599–604. doi:10.1038/nprot.2017.149.

2. Guo J, Grow EJ, Mlcochova H, Maher GJ, Lindskog C, Nie X, et al. The adult human testis transcriptional cell atlas. Cell Research. 2018;28(12):1141–1157. doi:10.1038/s41422-018-0099-2.

3. Vento-Tormo R, Efremova M, Botting RA, Turco MY, Vento-Tormo M, Meyer KB, et al. Single-cell reconstruction of the early maternal-fetal interface in humans. Nature. 2018;563(7731):347–353. doi:10.1038/s41586-018-0698-6.

4. Rozenblatt-Rosen O, Regev A, Oberdoerffer P, Nawy T, Hupalowska A, Rood JE, et al. The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. Cell. 2020;181(2):236–249. doi:10.1016/j.cell.2020.03.053.

5. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science. 2016;352(6282):189–196. doi:10.1126/science.aad0501.

6. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014;344(6190):1396–1401. doi:10.1126/science.1254257.

7. Neftel C, Laffy J, Filbin MG, Hara T, Shore ME, Rahme GJ, et al. An integrative model of cellular states, plasticity, and genetics for glioblastoma. Cell. 2019;178(4):835–849.e21. doi:10.1016/j.cell.2019.06.024.

8. Rosenberg AB, Roco CM, Muscat RA, Kuchina A, Sample P, Yao Z, et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science. 2018;360(6385):176–182. doi:10.1126/science.aam8999.

9. Wagner DE, Weinreb C, Collins ZM, Briggs JA, Megason SG, Klein AM. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. Science. 2018;360(6392):981–987. doi:10.1126/science.aar4362.

10. Plass M, Solana J, Wolf FA, Ayoub S, Misios A, Glažar P, et al. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. Science. 2018;360(6391). doi:10.1126/science.aaq1723.

11. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The human cell atlas. eLife. 2017;6. doi:10.7554/eLife.27041.

12. Wang B, Zhu J, Pierson E, Ramazzotti D, Batzoglou S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nature Methods. 2017;14(4):414–416. doi:10.1038/nmeth.4207.

13. Lin P, Troup M, Ho JWK. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. Genome Biology. 2017;18(1):59. doi:10.1186/s13059-017-1188-0.

14. Huh R, Yang Y, Jiang Y, Shen Y, Li Y. SAME-clustering: Single-cell Aggregated Clustering via Mixture Model Ensemble. Nucleic Acids Research. 2020;48(1):86–95. doi:10.1093/nar/gkz959.

15. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: consensus clustering of single-cell RNA-seq data. Nature Methods. 2017;14(5):483–486. doi:10.1038/nmeth.4236.

16. Li X, Wang K, Lyu Y, Pan H, Zhang J, Stambolian D, et al. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. Nature Communications. 2020;11(1):2338. doi:10.1038/s41467-020-15851-3.

17. Krzak M, Raykov Y, Boukouvalas A, Cutillo L, Angelini C. Benchmark and Parameter Sensitivity Analysis of Single-Cell RNA Sequencing Clustering Methods. Frontiers in genetics. 2019;10:1253. doi:10.3389/fgene.2019.01253.

18. Kiselev VY, Andrews TS, Hemberg M. Challenges in unsupervised clustering of single-cell RNA-seq data. Nature Reviews Genetics. 2019;20(5):273–282. doi:10.1038/s41576-018-0088-9.

19. Duò A, Robinson MD, Soneson C. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. F1000Research. 2018;7:1141. doi:10.12688/f1000research.15666.2.

20. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology. 2018;36(5):411–420. doi:10.1038/nbt.4096.

21. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biology. 2018;19(1):15. doi:10.1186/s13059-017-1382-0.

22. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment. 2008;2008(10):P10008. doi:10.1088/1742-5468/2008/10/P10008.

23. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Scientific Reports. 2019;9(1):5233. doi:10.1038/s41598-019-41695-z.

24. Levine JH, Simonds EF, Bendall SC, Davis KL, Amir EaD, Tadmor MD, et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell. 2015;162(1):184–197. doi:10.1016/j.cell.2015.05.047.

25. Newman MEJ, Girvan M. Finding and evaluating community structure in networks. Physical Review E, Statistical, Nonlinear, and Soft Matter Physics. 2004;69(2 Pt 2):026113. doi:10.1103/PhysRevE.69.026113.

26. Traag VA, Van Dooren P, Nesterov Y. Narrow scope for resolution-limit-free community detection. Physical Review E. 2011;84(1). doi:10.1103/PhysRevE.84.016114.

27. Reichardt J, Bornholdt S. Statistical mechanics of community detection. Physical Review E. 2006;74(1). doi:10.1103/PhysRevE.74.016110.

28. Lähnemann D, Köster J, Szczurek E, McCarthy DJ, Hicks SC, Robinson MD, et al. Eleven grand challenges in single-cell data science. Genome Biology. 2020;21(1):31. doi:10.1186/s13059-020-1926-6.

29. Fortunato S, Barthélemy M. Resolution limit in community detection. Proceedings of the National Academy of Sciences of the United States of America. 2007;104(1):36–41. doi:10.1073/pnas.0605965104.

30. Guimerà R, Sales-Pardo M, Amaral LAN. Modularity from fluctuations in random graphs and complex networks. Physical Review E. 2004;70(2). doi:10.1103/PhysRevE.70.025101.

31. Baran Y, Bercovich A, Sebe-Pedros A, Lubling Y, Giladi A, Chomsky E, et al. MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. Genome Biology. 2019;20(1):206. doi:10.1186/s13059-019-1812-2.

32. Xu C, Su Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. Bioinformatics. 2015;31(12):1974–1980. doi:10.1093/bioinformatics/btv088.

33. Miao Z, Moreno P, Huang N, Papatheodorou I, Brazma A, Teichmann SA. Putative cell type discovery from single-cell gene expression data. Nature Methods. 2020;17(6):621–628. doi:10.1038/s41592-020-0825-9.

34. Holland PW, Laskey KB, Leinhardt S. Stochastic blockmodels: First steps. Social networks. 1983;5(2):109–137. doi:10.1016/0378-8733(83)90021-7.

35. Peixoto TP. Nonparametric Bayesian inference of the microcanonical stochastic block model. Physical review E. 2017;95(1-1):012317. doi:10.1103/PhysRevE.95.012317.

36. Karrer B, Newman MEJ. Stochastic blockmodels and community structure in networks. Physical Review E, Statistical, Nonlinear, and Soft Matter Physics. 2011;83(1 Pt 2):016107. doi:10.1103/PhysRevE.83.016107.

37. Peixoto TP. Parsimonious module inference in large networks. Physical Review Letters. 2013;110(14):148701. doi:10.1103/PhysRevLett.110.148701.

38. Peixoto TP. Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models. Physical Review E, Statistical, Nonlinear, and Soft Matter Physics. 2014;89(1):012804. doi:10.1103/PhysRevE.89.012804.

39. Peixoto TP. Hierarchical Block Structures and High-Resolution Model Selection in Large Networks. Physical Review X. 2014;4(1):011047. doi:10.1103/PhysRevX.4.011047.

40. Tian L, Dong X, Freytag S, Lê Cao KA, Su S, JalalAbadi A, et al. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. Nature Methods. 2019;16(6):479–487. doi:10.1038/s41592-019-0425-8.

41. Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nature Biotechnology. 2018;36(1):89–94. doi:10.1038/nbt.4042.

42. Mereu E, Lafzi A, Moutinho C, Ziegenhain C, McCarthy DJ, Álvarez-Varela A, et al. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. Nature Biotechnology. 2020;38(6):747–755. doi:10.1038/s41587-020-0469-4.

43. Sasagawa Y, Danno H, Takada H, Ebisawa M, Tanaka K, Hayashi T, et al. Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. Genome Biology. 2018;19(1):29. doi:10.1186/s13059-018-1407-3.

44. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nature Communications. 2017;8:14049. doi:10.1038/ncomms14049.

45. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, et al. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science. 2014;343(6172):776–779. doi:10.1126/science.1247651.

46. Goldstein LD, Chen YJJ, Dunne J, Mir A, Hubschle H, Guillory J, et al. Massively parallel nanowell-based single-cell gene expression profiling. BMC Genomics. 2017;18(1):519. doi:10.1186/s12864-017-3893-1.

47. Paul F, Arkin Y, Giladi A, Jaitin DA, Kenigsberg E, Keren-Shaul H, et al. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. Cell. 2015;163(7):1663–1677. doi:10.1016/j.cell.2015.11.013.

48. Wolf FA, Hamey FK, Plass M, Solana J, Dahlin JS, Göttgens B, et al. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biology. 2019;20(1):59. doi:10.1186/s13059-019-1663-x.

49. Peixoto TP. Merge-split Markov chain Monte Carlo for community detection. arXiv. 2020;doi:arXiv:2003.07070.