Nested Stochastic Block Models Applied to the Analysis of Single Cell Data

Leonardo Morelli[1,2], Valentina Giansanti[1,3], and 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

July 6, 2020

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.

Background

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, 5, 6, 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, KNN) 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. Our approach identifies cell groups in a statistical robust way and, moreover, is able to determine the likelihood of the grouping, thus allowing model selection. In addition, our approach measures the confidence of assignment to groups; we show that this information may be exploited to perfect the notion of cell groups and the identification of markers.

Lastly, 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.

Results

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 kNN 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 evaluate 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, here including level 0. As we show below, cell affinities could be exploited as covariates in testing marker genes and, more in general, to define the stability of any cell group.

nSBM correctly identifies cell populations

We tested our approach on scRNA-seq mixology data [40], in particular on the 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 kNN 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 kNN graph; if fewer neighbours are chosen, the graph is sparser and the model is fit from less edges (Fig. S1). Running MCMC algorithm recovers the performances of the majority of the configurations (Fig. S2).

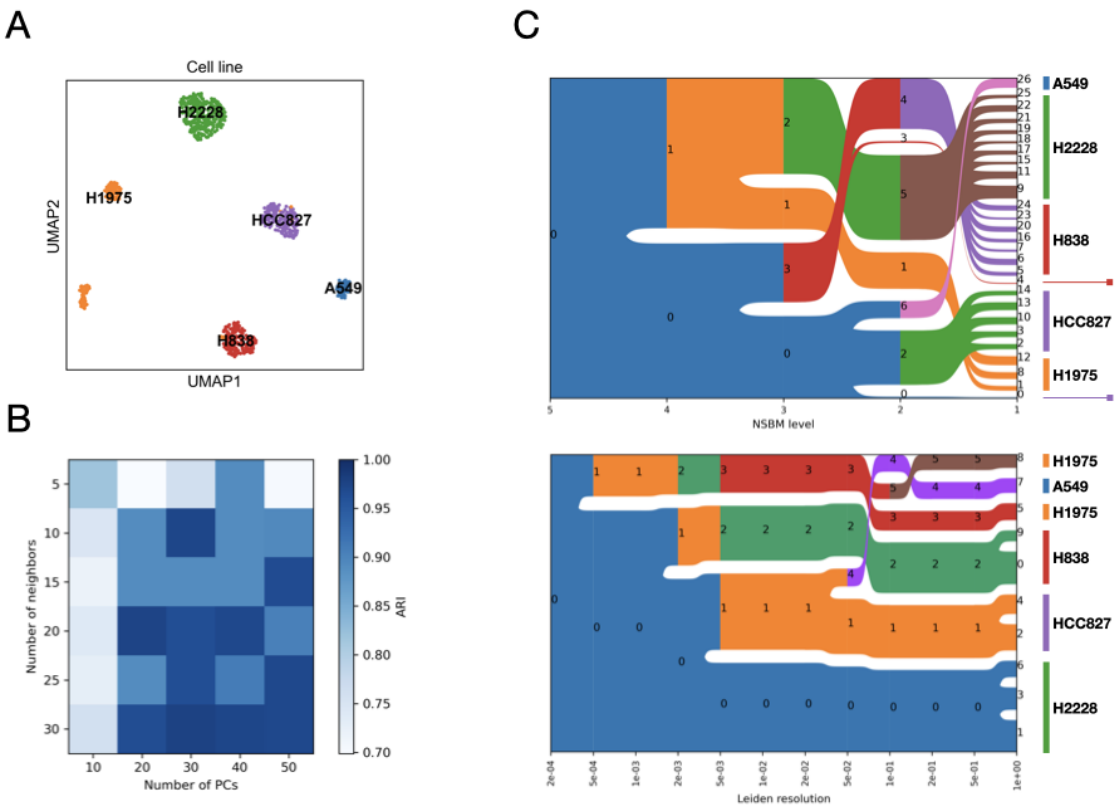

FIGURE 1: *schist* applied to scRNA-seq mixology data. (A) UMAP embedding of 10x Chromium data, cells are colored according to the given cell line in the original paper. A small number of H1975 cells are found in HCC827 and H838 clusters. (B) Heatmap showing the maximal Adjusted Rand Index for different kNN 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

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.

These observations show that nSBM is able to perform accurate identification of cell groups, without the need of an arbitrary threshold on the resolution parameter. These data also hint at the possibility to identify rare cell types in larger populations.

nSBM 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 [42], previously used to benchmark the consistency of cell grouping with differentiation trajectories by graph abstraction [43]. Standard processing of those data reveals three major branchings (Erythroids, Neutrophils and Monocytes) stemming from the progenitor cells (Fig. S3A). After applying nSBM, we identify 27 groups at level 3 of the hierarchy (Fig. S3B), compared to the 24 using Leiden method at default resolution (Fig. S4). We found that the hierarchy proposed by our model is consistent with the developmental model (Fig. 2). 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 (Fig. S4); 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.

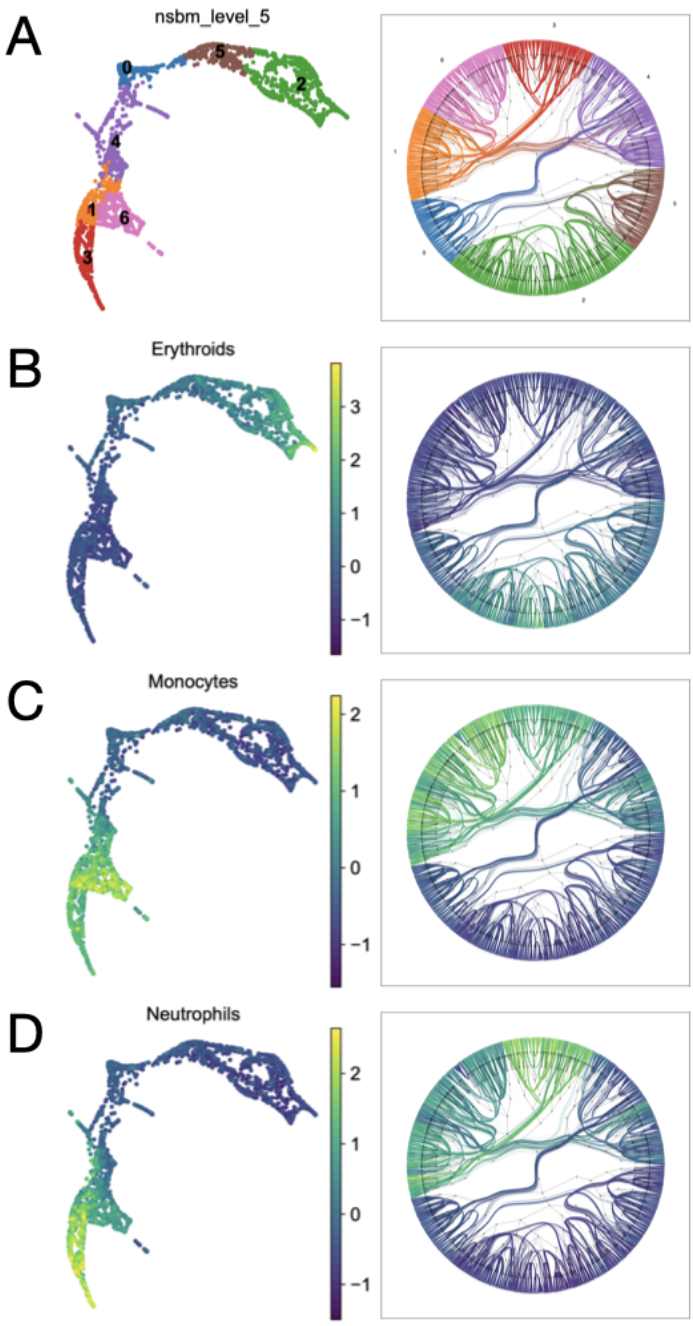

FIGURE 2: Analysis of hematopoietic differentiation. Each panel presents a low dimensional embedding of single cells next to a radial tree representation of the nSBM hierarchy. Cells are colored according to groupings at level 5 of the hierarchy, group 0 marks the progenitor population (A). In subsequent 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.

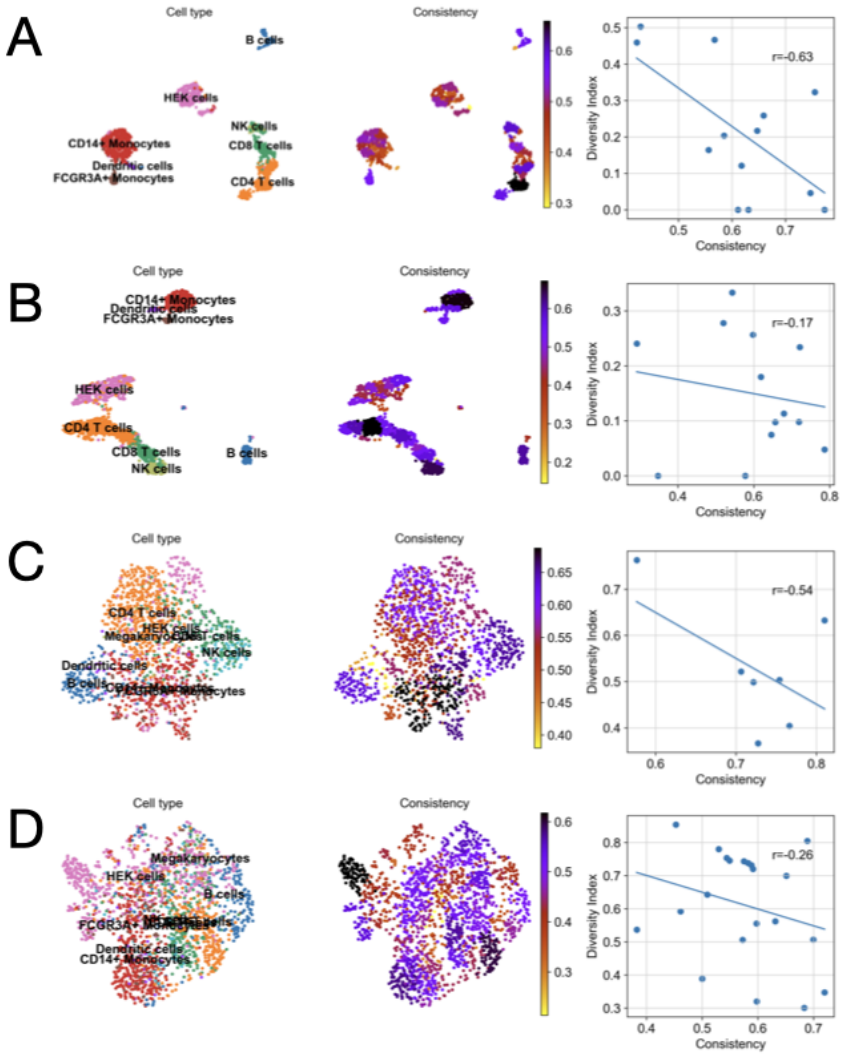

FIGURE 3: Analysis of cell cluster consistency. Every panel reports a UMAP embedding of a PBMC + HEK293 cells profiled on different platform. Cells are annotated by cell type and by consistency value, which is assigned to cell clusters at nSBM level 1. The charts next to UMAPs show the correlation between 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.

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 [44]; in particular, we chose two technologies resulting in high quality data: Quartz-seq2 [45] and Chromium 10x v3 [46], and two technologies resulting in more noisy data: MARS-seq [47] and iCell8 [48] (Fig. 3).

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 (Fig. S5).

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 [44].

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 [44], 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 [44]	1266	00:31	00:02	0.726
MARS-seq [44]	1401	00:29	00:08	0.834
Chromium 10X [44]	1523	00:40	00:04	0.695
iCELL8 [44]	1830	00:54	00:07	0.623
Paul15 [42]	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.

Conclusions

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.

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 of hematopoietic differentiation

Data were retrieved using scanpy's built-in functions and were processed as in [43], except for kNN 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 [44]. 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.

Acknowledgements

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. This work has been supported by Accelerator Award: A26815 entitled: "Single-cell cancer evolution in the clinic" funded through a partnership between Cancer Research UK and Fondazione AIRC

References

[1] Valentine Svensson, Roser Vento-Tormo, and Sarah A Teichmann. Exponential scaling of single-cell RNA-seq in the past decade. *Nature Protocols*, 13(4):599–604, mar 2018. ISSN 1754-2189. doi: 10.1038/nprot.2017.149.

[2] Jingtao Guo, Edward J Grow, Hana Mlcochova, Geoffrey J Maher, Cecilia Lindskog, Xichen Nie, Yixuan Guo, Yodai Takei, Jina Yun, Long Cai, Robin Kim, Douglas T Carrell, Anne Goriely, James M Hotaling, and Bradley R Cairns. The adult human testis transcriptional cell atlas. *Cell Research*, 28(12):1141–1157, oct 2018. doi: 10.1038/s41422-018-0099-2. URL `http://dx.doi.org/10.1038/s41422-018-0099-2`.

[3] Roser Vento-Tormo, Mirjana Efremova, Rachel A Botting, Margherita Y Turco, Miquel Vento-Tormo, Kerstin B Meyer, Jong-Eun Park, Emily Stephenson, Krzysztof Polański, Angela Goncalves, Lucy Gardner, Staffan Holmqvist, Johan Henriksson, Angela Zou, Andrew M Sharkey, Ben Millar, Barbara Innes, Laura Wood, Anna Wilbrey-Clark, Rebecca P Payne, Martin A Ivarsson, Steve Lisgo, Andrew Filby,

David H Rowitch, Judith N Bulmer, Gavin J Wright, Michael J T Stubbington, Muzlifah Haniffa, Ashley Moffett, and Sarah A Teichmann. Single-cell reconstruction of the early maternal-fetal interface in humans. *Nature*, 563(7731):347–353, nov 2018. ISSN 0028-0836. doi: 10.1038/s41586-018-0698-6. URL http://www.nature.com/articles/s41586-018-0698-6.

[4] Orit Rozenblatt-Rosen, Aviv Regev, Philipp Oberdoerffer, Tal Nawy, Anna Hupalowska, Jennifer E Rood, Orr Ashenberg, Ethan Cerami, Robert J Coffey, Emek Demir, Li Ding, Edward D Esplin, James M Ford, Jeremy Goecks, Sharmistha Ghosh, Joe W Gray, Justin Guinney, Sean E Hanlon, Shannon K Hughes, E Shelley Hwang, Christine A Iacobuzio-Donahue, Judit Jané-Valbuena, Bruce E Johnson, Ken S Lau, Tracy Lively, Sarah A Mazzilli, Dana Pe'er, Sandro Santagata, Alex K Shalek, Denis Schapiro, Michael P Snyder, Peter K Sorger, Avrum E Spira, Sudhir Srivastava, Kai Tan, Robert B West, Elizabeth H Williams, and Human Tumor Atlas Network. The human tumor atlas network: Charting tumor transitions across space and time at single-cell resolution. *Cell*, 181(2):236–249, apr 2020. ISSN 00928674. doi: 10.1016/j.cell.2020.03.053. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867420303469.

[5] Itay Tirosh, Benjamin Izar, Sanjay M Prakadan, Marc H Wadsworth, Daniel Treacy, John J Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, Mohammad Fallahi-Sichani, Ken Dutton-Regester, Jia-Ren Lin, Ofir Cohen, Parin Shah, Diana Lu, Alex S Genshaft, Travis K Hughes, Carly G K Ziegler, Samuel W Kazer, Aleth Gaillard, Kellie E Kolb, Alexandra-Chloé Villani, Cory M Johannessen, Aleksandr Y Andreev, Eliezer M Van Allen, Monica Bertagnolli, Peter K Sorger, Ryan J Sullivan, Keith T Flaherty, Dennie T Frederick, Judit Jané-Valbuena, Charles H Yoon, Orit Rozenblatt-Rosen, Alex K Shalek, Aviv Regev, and Levi A Garraway. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science*, 352(6282):189–196, apr 2016. ISSN 0036-8075. doi: 10.1126/science.aad0501. URL http://www.sciencemag.org/cgi/doi/10.1126/science.aad0501.

[6] Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, David N Louis, Orit Rozenblatt-Rosen, Mario L Suvà, Aviv Regev, and Bradley E Bernstein. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, jun 2014. doi: 10.1126/science.1254257. URL http://dx.doi.org/10.1126/science.1254257.

[7] Cyril Neftel, Julie Laffy, Mariella G Filbin, Toshiro Hara, Marni E Shore, Gilbert J Rahme, Alyssa R Richman, Dana Silverbush, McKenzie L Shaw, Christine M Hebert, John Dewitt, Simon Gritsch, Elizabeth M Perez, L Nicolas Gonzalez Castro, Xiaoyang Lan, Nicholas Druck, Christopher Rodman, Danielle Dionne, Alexander Kaplan, Mia S Bertalan, Julia Small, Kristine Pelton, Sarah Becker, Dennis Bonal, Quang-De Nguyen, Rachel L Servis, Jeremy M Fung, Ravindra Mylvaganam, Lisa Mayr, Johannes Gojo, Christine Haberler, Rene Geyeregger, Thomas Czech, Irene Slavc, Brian V Nahed, William T Curry, Bob S Carter, Hiroaki Wakimoto, Priscilla K Brastianos, Tracy T Batchelor, Anat Stemmer-Rachamimov, Maria Martinez-Lage, Matthew P Frosch, Ivan Stamenkovic, Nicolo Riggi, Esther Rheinbay, Michelle Monje, Orit Rozenblatt-Rosen, Daniel P Cahill, Anoop P Patel, Tony Hunter, Inder M Verma, Keith L Ligon, David N Louis, Aviv Regev, Bradley E Bernstein, Itay Tirosh, and Mario L Suvà. An integrative model of cellular states, plasticity, and genetics for glioblastoma. *Cell*, 178(4):835–849.e21, aug 2019. ISSN 00928674. doi: 10.1016/j.cell.2019.06.024. URL https://linkinghub.elsevier.com/retrieve/pii/S0092867419306877.

[8] Alexander B Rosenberg, Charles M Roco, Richard A Muscat, Anna Kuchina, Paul Sample, Zizhen Yao, Lucas T Graybuck, David J Peeler, Sumit Mukherjee, Wei Chen, Suzie H Pun, Drew L Sellers, Bosiljka Tasic, and Georg Seelig. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. *Science*, 360(6385):176–182, apr 2018. ISSN 0036-8075. doi: 10.1126/science.aam8999. URL http://www.sciencemag.org/lookup/doi/10.1126/science.aam8999.

[9] Daniel E Wagner, Caleb Weinreb, Zach M Collins, James A Briggs, Sean G Megason, and Allon M Klein. Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. *Science*, 360(6392):981–987, jun 2018. ISSN 0036-8075. doi: 10.1126/science.aar4362. URL http://www.sciencemag.org/lookup/doi/10.1126/science.aar4362.

[10] Mireya Plass, Jordi Solana, F Alexander Wolf, Salah Ayoub, Aristotelis Misios, Petar Glažar, Benedikt Obermayer, Fabian J Theis, Christine Kocks, and Nikolaus Rajewsky. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*, 360(6391), may 2018. ISSN 0036-8075. doi: 10.1126/science.aaq1723. URL http://www.sciencemag.org/lookup/doi/10.1126/science.aaq1723.

[11] Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, Hans Clevers, Bart Deplancke, Ian Dunham, James Eberwine, Roland Eils, Wolfgang Enard, Andrew Farmer, Lars Fugger, Berthold Göttgens, Nir Hacohen, Muzlifah Haniffa, Martin Hemberg, Seung Kim, Paul Klenerman, Arnold Kriegstein, Ed Lein, Sten Linnarsson, Emma Lundberg, Joakim Lundeberg, Partha Majumder, John C Marioni, Miriam Merad,

Musa Mhlanga, Martijn Nawijn, Mihai Netea, Garry Nolan, Dana Pe'er, Anthony Phillipakis, Chris P Ponting, Stephen Quake, Wolf Reik, Orit Rozenblatt-Rosen, Joshua Sanes, Rahul Satija, Ton N Schumacher, Alex Shalek, Ehud Shapiro, Padmanee Sharma, Jay W Shin, Oliver Stegle, Michael Stratton, Michael J T Stubbington, Fabian J Theis, Matthias Uhlen, Alexander van Oudenaarden, Allon Wagner, Fiona Watt, Jonathan Weissman, Barbara Wold, Ramnik Xavier, Nir Yosef, and Human Cell Atlas Meeting Participants. The human cell atlas. *eLife*, 6, dec 2017. doi: 10.7554/{eLife}.27041. URL `http://dx.doi.org/10.7554/{eLife}.27041`.

[12] Bo Wang, Junjie Zhu, Emma Pierson, Daniele Ramazzotti, and Serafim Batzoglou. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. *Nature Methods*, 14(4):414–416, apr 2017. doi: 10.1038/nmeth.4207. URL `http://dx.doi.org/10.1038/nmeth.4207`.

[13] Peijie Lin, Michael Troup, and Joshua W K Ho. CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. *Genome Biology*, 18(1):59, mar 2017. doi: 10.1186/s13059-017-1188-0. URL `http://dx.doi.org/10.1186/s13059-017-1188-0`.

[14] Ruth Huh, Yuchen Yang, Yuchao Jiang, Yin Shen, and Yun Li. SAME-clustering: Single-cell aggregated clustering via mixture model ensemble. *Nucleic Acids Research*, 48(1):86–95, jan 2020. doi: 10.1093/nar/gkz959. URL `http://dx.doi.org/10.1093/nar/gkz959`.

[15] Vladimir Yu Kiselev, Kristina Kirschner, Michael T Schaub, Tallulah Andrews, Andrew Yiu, Tamir Chandra, Kedar N Natarajan, Wolf Reik, Mauricio Barahona, Anthony R Green, and Martin Hemberg. SC3: consensus clustering of single-cell RNA-seq data. *Nature Methods*, 14(5):483–486, may 2017. doi: 10.1038/nmeth.4236. URL `http://dx.doi.org/10.1038/nmeth.4236`.

[16] Xiangjie Li, Kui Wang, Yafei Lyu, Huize Pan, Jingxiao Zhang, Dwight Stambolian, Katalin Susztak, Muredach P Reilly, Gang Hu, and Mingyao Li. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. *Nature Communications*, 11(1):2338, may 2020. doi: 10.1038/s41467-020-15851-3. URL `http://dx.doi.org/10.1038/s41467-020-15851-3`.

[17] Monika Krzak, Yordan Raykov, Alexis Boukouvalas, Luisa Cutillo, and Claudia Angelini. Benchmark and parameter sensitivity analysis of single-cell RNA sequencing clustering methods. *Frontiers in genetics*, 10: 1253, dec 2019. doi: 10.3389/fgene.2019.01253. URL `http://dx.doi.org/10.3389/fgene.2019.01253`.

[18] Vladimir Yu Kiselev, Tallulah S Andrews, and Martin Hemberg. Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews. Genetics*, 20(5):273–282, 2019. ISSN 1471-0056. doi: 10.1038/s41576-018-0088-9. URL `http://www.nature.com/articles/s41576-018-0088-9`.

[19] Angelo Duò, Mark D Robinson, and Charlotte Soneson. A systematic performance evaluation of clustering methods for single-cell RNA-seq data. *F1000Research*, 7:1141, jul 2018. doi: 10.12688/f1000research.15666.2. URL `http://dx.doi.org/10.12688/f1000research.15666.2`.

[20] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature Biotechnology*, 36(5): 411–420, apr 2018. ISSN 1087-0156. doi: 10.1038/nbt.4096. URL `http://www.nature.com/doifinder/10.1038/nbt.4096`.

[21] F Alexander Wolf, Philipp Angerer, and Fabian J Theis. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):15, feb 2018. doi: 10.1186/s13059-017-1382-0. URL `http://dx.doi.org/10.1186/s13059-017-1382-0`.

[22] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008, oct 2008. ISSN 1742-5468. doi: 10.1088/1742-5468/2008/10/P10008. URL `http://stacks.iop.org/1742-5468/2008/i=10/a=P10008?key=crossref.46968f6ec61eb8f907a760be1c5ace52`.

[23] V A Traag, L Waltman, and N J van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1):5233, mar 2019. doi: 10.1038/s41598-019-41695-z. URL `http://dx.doi.org/10.1038/s41598-019-41695-z`.

[24] Jacob H Levine, Erin F Simonds, Sean C Bendall, Kara L Davis, El-ad D Amir, Michelle D Tadmor, Oren Litvin, Harris G Fienberg, Astraea Jager, Eli R Zunder, Rachel Finck, Amanda L Gedman, Ina Radtke, James R Downing, Dana Pe'er, and Garry P Nolan. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184–197, jul 2015. doi: 10.1016/j.cell.2015.05.047. URL `http://dx.doi.org/10.1016/j.cell.2015.05.047`.

[25] M E J Newman and M Girvan. Finding and evaluating community structure in networks. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, 69(2 Pt 2):026113, feb 2004. ISSN 1539-3755. doi: 10.1103/{PhysRevE}.69.026113. URL http://dx.doi.org/10.1103/{PhysRevE}.69.026113.

[26] V. A. Traag, P. Van Dooren, and Y. Nesterov. Narrow scope for resolution-limit-free community detection. *Physical Review E*, 84(1), jul 2011. ISSN 1539-3755. doi: 10.1103/{PhysRevE}.84.016114. URL http://link.aps.org/doi/10.1103/{PhysRevE}.84.016114.

[27] Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74(1), jul 2006. ISSN 1539-3755. doi: 10.1103/{PhysRevE}.74.016110. URL http://link.aps.org/doi/10.1103/{PhysRevE}.74.016110.

[28] David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, Luca Pinello, Pavel Skums, Alexandros Stamatakis, Camille Stephan-Otto Attolini, Samuel Aparicio, Jasmijn Baaijens, Marleen Balvert, Buys de Barbanson, Antonio Cappuccio, Giacomo Corleone, Bas E Dutilh, Maria Florescu, Victor Guryev, Rens Holmer, Katharina Jahn, Thamar Jessurun Lobo, Emma M Keizer, Indu Khatri, Szymon M Kielbasa, Jan O Korbel, Alexey M Kozlov, Tzu-Hao Kuo, Boudewijn P F Lelieveldt, Ion I Mandoiu, John C Marioni, Tobias Marschall, Felix Mölder, Amir Niknejad, Lukasz Raczkowski, Marcel Reinders, Jeroen de Ridder, Antoine-Emmanuel Saliba, Antonios Somarakis, Oliver Stegle, Fabian J Theis, Huan Yang, Alex Zelikovsky, Alice C McHardy, Benjamin J Raphael, Sohrab P Shah, and Alexander Schönhuth. Eleven grand challenges in single-cell data science. *Genome Biology*, 21(1):31, feb 2020. doi: 10.1186/s13059-020-1926-6. URL http://dx.doi.org/10.1186/s13059-020-1926-6.

[29] Santo Fortunato and Marc Barthélemy. Resolution limit in community detection. *Proceedings of the National Academy of Sciences of the United States of America*, 104(1):36–41, jan 2007. doi: 10.1073/pnas.0605965104. URL http://dx.doi.org/10.1073/pnas.0605965104.

[30] Roger Guimerà, Marta Sales-Pardo, and Luís A. Nunes Amaral. Modularity from fluctuations in random graphs and complex networks. *Physical Review E*, 70(2), aug 2004. ISSN 1539-3755. doi: 10.1103/{PhysRevE}.70.025101. URL http://link.aps.org/doi/10.1103/{PhysRevE}.70.025101.

[31] Yael Baran, Akhiad Bercovich, Arnau Sebe-Pedros, Yaniv Lubling, Amir Giladi, Elad Chomsky, Zohar Meir, Michael Hoichman, Aviezer Lifshitz, and Amos Tanay. MetaCell: analysis of single-cell RNA-seq data using k-nn graph partitions. *Genome Biology*, 20(1):206, oct 2019. ISSN 1474-760X. doi: 10.1186/s13059-019-1812-2. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1812-2.

[32] Chen Xu and Zhengchang Su. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, 31(12):1974–1980, jun 2015. doi: 10.1093/bioinformatics/btv088. URL http://dx.doi.org/10.1093/bioinformatics/btv088.

[33] Zhichao Miao, Pablo Moreno, Ni Huang, Irene Papatheodorou, Alvis Brazma, and Sarah A Teichmann. Putative cell type discovery from single-cell gene expression data. *Nature Methods*, 17(6):621–628, jun 2020. ISSN 1548-7091. doi: 10.1038/s41592-020-0825-9. URL http://www.nature.com/articles/s41592-020-0825-9.

[34] Paul W. Holland, Kathryn Blackmond Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, jun 1983. ISSN 03788733. doi: 10.1016/0378-8733(83)90021-7. URL http://linkinghub.elsevier.com/retrieve/pii/0378873383900217.

[35] Tiago P Peixoto. Nonparametric bayesian inference of the microcanonical stochastic block model. *Physical review. E*, 95(1-1):012317, jan 2017. doi: 10.1103/{PhysRevE}.95.012317. URL http://dx.doi.org/10.1103/{PhysRevE}.95.012317.

[36] Brian Karrer and M E J Newman. Stochastic blockmodels and community structure in networks. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, 83(1 Pt 2):016107, jan 2011. doi: 10.1103/{PhysRevE}.83.016107. URL http://dx.doi.org/10.1103/{PhysRevE}.83.016107.

[37] Tiago P Peixoto. Parsimonious module inference in large networks. *Physical Review Letters*, 110(14):148701, apr 2013. doi: 10.1103/{PhysRevLett}.110.148701. URL http://dx.doi.org/10.1103/{PhysRevLett}.110.148701.

[38] Tiago P Peixoto. Efficient monte carlo and greedy heuristic for the inference of stochastic block models. *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics*, 89(1):012804, jan 2014. doi: 10.1103/{PhysRevE}.89.012804. URL http://dx.doi.org/10.1103/{PhysRevE}.89.012804.

[39] Tiago P. Peixoto. Hierarchical block structures and high-resolution model selection in large networks. *Physical Review X*, 4(1):011047, mar 2014. ISSN 2160-3308. doi: 10.1103/{PhysRevX}.4.011047. URL https://link.aps.org/doi/10.1103/{PhysRevX}.4.011047.

[40] Luyi Tian, Xueyi Dong, Saskia Freytag, Kim-Anh Lê Cao, Shian Su, Abolfazl JalalAbadi, Daniela Amann-Zalcenstein, Tom S Weber, Azadeh Seidi, Jafar S Jabbari, Shalin H Naik, and Matthew E Ritchie. Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. *Nature Methods*, 16(6):479–487, may 2019. ISSN 1548-7091. doi: 10.1038/s41592-019-0425-8. URL http://www.nature.com/articles/s41592-019-0425-8.

[41] Hyun Min Kang, Meena Subramaniam, Sasha Targ, Michelle Nguyen, Lenka Maliskova, Elizabeth McCarthy, Eunice Wan, Simon Wong, Lauren Byrnes, Cristina M Lanata, Rachel E Gate, Sara Mostafavi, Alexander Marson, Noah Zaitlen, Lindsey A Criswell, and Chun Jimmie Ye. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nature Biotechnology*, 36(1):89–94, 2018. ISSN 1087-0156. doi: 10.1038/nbt.4042. URL http://www.nature.com/doifinder/10.1038/nbt.4042.

[42] Franziska Paul, Ya'ara Arkin, Amir Giladi, Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Deborah Winter, David Lara-Astiaso, Meital Gury, Assaf Weiner, Eyal David, Nadav Cohen, Felicia Kathrine Bratt Lauridsen, Simon Haas, Andreas Schlitzer, Alexander Mildner, Florent Ginhoux, Steffen Jung, Andreas Trumpp, Bo Torben Porse, Amos Tanay, and Ido Amit. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, 163(7):1663–1677, dec 2015. ISSN 00928674. doi: 10.1016/j.cell.2015.11.013. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867415014932.

[43] F Alexander Wolf, Fiona K Hamey, Mireya Plass, Jordi Solana, Joakim S Dahlin, Berthold Göttgens, Nikolaus Rajewsky, Lukas Simon, and Fabian J Theis. PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biology*, 20(1):59, mar 2019. doi: 10.1186/s13059-019-1663-x. URL http://dx.doi.org/10.1186/s13059-019-1663-x.

[44] Elisabetta Mereu, Atefeh Lafzi, Catia Moutinho, Christoph Ziegenhain, Davis J McCarthy, Adrián Álvarez-Varela, Eduard Batlle, Sagar, Dominic Grün, Julia K Lau, Stéphane C Boutet, Chad Sanada, Aik Ooi, Robert C Jones, Kelly Kaihara, Chris Brampton, Yasha Talaga, Yohei Sasagawa, Kaori Tanaka, Tetsutaro Hayashi, Caroline Braeuning, Cornelius Fischer, Sascha Sauer, Timo Trefzer, Christian Conrad, Xian Adiconis, Lan T Nguyen, Aviv Regev, Joshua Z Levin, Swati Parekh, Aleksandar Janjic, Lucas E Wange, Johannes W Bagnoli, Wolfgang Enard, Marta Gut, Rickard Sandberg, Itoshi Nikaido, Ivo Gut, Oliver Stegle, and Holger Heyn. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. *Nature Biotechnology*, 38(6):747–755, jun 2020. ISSN 1087-0156. doi: 10.1038/s41587-020-0469-4. URL http://www.nature.com/articles/s41587-020-0469-4.

[45] Yohei Sasagawa, Hiroki Danno, Hitomi Takada, Masashi Ebisawa, Kaori Tanaka, Tetsutaro Hayashi, Akira Kurisaki, and Itoshi Nikaido. Quartz-seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. *Genome Biology*, 19(1):29, mar 2018. doi: 10.1186/s13059-018-1407-3. URL http://dx.doi.org/10.1186/s13059-018-1407-3.

[46] Grace X Y Zheng, Jessica M Terry, Phillip Belgrader, Paul Ryvkin, Zachary W Bent, Ryan Wilson, Solongo B Ziraldo, Tobias D Wheeler, Geoff P McDermott, Junjie Zhu, Mark T Gregory, Joe Shuga, Luz Montesclaros, Jason G Underwood, Donald A Masquelier, Stefanie Y Nishimura, Michael Schnall-Levin, Paul W Wyatt, Christopher M Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D Ness, Lan W Beppu, H Joachim Deeg, Christopher McFarland, Keith R Loeb, William J Valente, Nolan G Ericson, Emily A Stevens, Jerald P Radich, Tarjei S Mikkelsen, Benjamin J Hindson, and Jason H Bielas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8:14049, jan 2017. doi: 10.1038/ncomms14049. URL http://dx.doi.org/10.1038/ncomms14049.

[47] Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Naama Elefant, Franziska Paul, Irina Zaretsky, Alexander Mildner, Nadav Cohen, Steffen Jung, Amos Tanay, and Ido Amit. Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. *Science*, 343(6172):776–779, feb 2014. doi: 10.1126/science.1247651. URL http://dx.doi.org/10.1126/science.1247651.

[48] Leonard D Goldstein, Ying-Jiun Jasmine Chen, Jude Dunne, Alain Mir, Hermann Hubschle, Joseph Guillory, Wenlin Yuan, Jingli Zhang, Jeremy Stinson, Bijay Jaiswal, Kanika Bajaj Pahuja, Ishminder Mann, Thomas Schaal, Leo Chan, Sangeetha Anandakrishnan, Chun-Wah Lin, Patricio Espinoza, Syed Husain, Harris Shapiro, Karthikeyan Swaminathan, Sherry Wei, Maithreyan Srinivasan, Somasekar Seshagiri, and Zora Modrusan. Massively parallel nanowell-based single-cell gene expression profiling. *BMC Genomics*, 18(1):519, jul 2017. doi: 10.1186/s12864-017-3893-1. URL http://dx.doi.org/10.1186/s12864-017-3893-1.

[49] Tiago P Peixoto. Merge-split markov chain monte carlo for community detection. *arXiv*, mar 2020. doi: arXiv:2003.07070. URL https://arxiv.org/abs/2003.07070.