METHOD

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

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

*Correspondence:
cittaro.davide@hsr.it
[1]Center for Omics Sciences, IRCCS San Raffaele Hospital, Via Olgettina 58, 20132 Milan, Italy
Full list of author information is available at the end of the article

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 neighborhood 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.

Keywords: Single cell; Cluster analysis; Stochastic Block Models

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 [4–7], 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 neighborhood graph (*e.g.* by k Nearest Neighbors, 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 neighborhood 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 neighborhood 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 (MCMC) 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 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 neighborhood 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 neighbors 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 neighbors. In general, higher number of components and neighbors 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

neighbors. If fewer PCs are used, a smaller fraction of the total variance, hence less information, is used to build the kNN graph; if fewer neighbors 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).

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.

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 [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].

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 the underlying libraries, runtimes are proportional to the number of edges in the neighborhood graph and while it supports CPU-level parallelisation, a model minimisation is hundreds times slower than the extremely fast Leiden approach. Runtimes are further inflated if MCMC equilibration is performed. We are well aware that this will limit the adoption of any strategy based on nSBM, but we believe that the quality of the results greatly justifies the additional resources.

Materials and Methods

A detailed view of the parameters used for the analysis presented in this manuscript is available as jupyter notebook at https://github.com/dawe/schist-notebooks/tree/master/schist_paper.

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 neighbors and diffmap as embedding. nSBM was completed with 3 initialisations. 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

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. Neighborhood graph was built using 30 principal components and 30 neighbors. nSBM was completed with 3 initialisations.

Competing interests
The authors declare that they have no competing interests.

Author's contributions

LM performed the analysis, contributed to the code and wrote the manuscript. VG supervised the analysis and wrote the manuscript. DC conceived the project, performed the analysis, contributed to the code and wrote the manuscript.

Acknowledgements

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.

Author details

[1]Center for Omics Sciences, IRCCS San Raffaele Hospital, Via Olgettina 58, 20132 Milan, Italy. [2]Università Vita-Salute San Raffaele, Via Olgettina 58, 20132 Milan, Italy. [3]Dept. of Informatics, Systems and Communication, Univ. of Milan-Bicocca, Viale Sarca 336, 20126 Milan, Italy.

References

1. Svensson, V., Vento-Tormo, R., Teichmann, S.A.: Exponential scaling of single-cell RNA-seq in the past decade. Nature Protocols **13**(4), 599–604 (2018). doi:10.1038/nprot.2017.149
2. Guo, J., Grow, E.J., Mlcochova, H., Maher, G.J., Lindskog, C., Nie, X., Guo, Y., Takei, Y., Yun, J., Cai, L., Kim, R., Carrell, D.T., Goriely, A., Hotaling, J.M., Cairns, B.R.: The adult human testis transcriptional cell atlas. Cell Research **28**(12), 1141–1157 (2018). doi:10.1038/s41422-018-0099-2. Accessed 2019-04-27
3. Vento-Tormo, R., Efremova, M., Botting, R.A., Turco, M.Y., Vento-Tormo, M., Meyer, K.B., Park, J.-E., Stephenson, E., Polański, K., Goncalves, A., Gardner, L., Holmqvist, S., Henriksson, J., Zou, A., Sharkey, A.M., Millar, B., Innes, B., Wood, L., Wilbrey-Clark, A., Payne, R.P., Ivarsson, M.A., Lisgo, S., Filby, A., Rowitch, D.H., Bulmer, J.N., Wright, G.J., Stubbington, M.J.T., Haniffa, M., Moffett, A., Teichmann, S.A.: Single-cell reconstruction of the early maternal-fetal interface in humans. Nature **563**(7731), 347–353 (2018). doi:10.1038/s41586-018-0698-6. Accessed 2019-06-24
4. Rozenblatt-Rosen, O., Regev, A., Oberdoerffer, P., Nawy, T., Hupalowska, A., Rood, J.E., Ashenberg, O., Cerami, E., Coffey, R.J., Demir, E., Ding, L., Esplin, E.D., Ford, J.M., Goecks, J., Ghosh, S., Gray, J.W., Guinney, J., Hanlon, S.E., Hughes, S.K., Hwang, E.S., Iacobuzio-Donahue, C.A., Jané-Valbuena, J., Johnson, B.E., Lau, K.S., Lively, T., Mazzilli, S.A., Pe'er, D., Santagata, S., Shalek, A.K., Schapiro, D., Snyder, M.P., Sorger, P.K., Spira, A.E., Srivastava, S., Tan, K., West, R.B., Williams, E.H., Network, H.T.A.: The human tumor atlas network: Charting tumor transitions across space and time at single-cell resolution. Cell **181**(2), 236–249 (2020). doi:10.1016/j.cell.2020.03.053. Accessed 2020-06-30
5. Tirosh, I., Izar, B., Prakadan, S.M., Wadsworth, M.H., Treacy, D., Trombetta, J.J., Rotem, A., Rodman, C., Lian, C., Murphy, G., Fallahi-Sichani, M., Dutton-Regester, K., Lin, J.-R., Cohen, O., Shah, P., Lu, D., Genshaft, A.S., Hughes, T.K., Ziegler, C.G.K., Kazer, S.W., Gaillard, A., Kolb, K.E., Villani, A.-C., Johannessen, C.M., Andreev, A.Y., Van Allen, E.M., Bertagnolli, M., Sorger, P.K., Sullivan, R.J., Flaherty, K.T., Frederick, D.T., Jané-Valbuena, J., Yoon, C.H., Rozenblatt-Rosen, O., Shalek, A.K., Regev, A., Garraway, L.A.: Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science **352**(6282), 189–196 (2016). doi:10.1126/science.aad0501. Accessed 2019-04-28
6. Patel, A.P., Tirosh, I., Trombetta, J.J., Shalek, A.K., Gillespie, S.M., Wakimoto, H., Cahill, D.P., Nahed, B.V., Curry, W.T., Martuza, R.L., Louis, D.N., Rozenblatt-Rosen, O., Suvà, M.L., Regev, A., Bernstein, B.E.: Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science **344**(6190), 1396–1401 (2014). doi:10.1126/science.1254257. Accessed 2019-04-28
7. Neftel, C., Laffy, J., Filbin, M.G., Hara, T., Shore, M.E., Rahme, G.J., Richman, A.R., Silverbush, D., Shaw, M.L., Hebert, C.M., Dewitt, J., Gritsch, S., Perez, E.M., Gonzalez Castro, L.N., Lan, X., Druck, N., Rodman, C., Dionne, D., Kaplan, A., Bertalan, M.S., Small, J., Pelton, K., Becker, S., Bonal, D., Nguyen, Q.-D., Servis, R.L., Fung, J.M., Mylvaganam, R., Mayr, L., Gojo, J., Haberler, C., Geyeregger, R., Czech, T., Slavc, I., Nahed, B.V., Curry, W.T., Carter, B.S., Wakimoto, H., Brastianos, P.K., Batchelor, T.T., Stemmer-Rachamimov, A., Martinez-Lage, M., Frosch, M.P., Stamenkovic, I., Riggi, N., Rheinbay, E., Monje, M., Rozenblatt-Rosen, O., Cahill, D.P., Patel, A.P., Hunter, T., Verma, I.M., Ligon, K.L., Louis, D.N., Regev, A., Bernstein, B.E., Tirosh, I., Suvà, M.L.: An integrative model of cellular states, plasticity, and genetics for glioblastoma. Cell **178**(4), 835–84921 (2019). doi:10.1016/j.cell.2019.06.024. Accessed 2019-08-11
8. Rosenberg, A.B., Roco, C.M., Muscat, R.A., Kuchina, A., Sample, P., Yao, Z., Graybuck, L.T., Peeler, D.J., Mukherjee, S., Chen, W., Pun, S.H., Sellers, D.L., Tasic, B., Seelig, G.: Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science **360**(6385), 176–182 (2018). doi:10.1126/science.aam8999. Accessed 2019-04-28
9. Wagner, D.E., Weinreb, C., Collins, Z.M., Briggs, J.A., Megason, S.G., Klein, A.M.: Single-cell mapping of gene expression landscapes and lineage in the zebrafish embryo. Science **360**(6392), 981–987 (2018). doi:10.1126/science.aar4362. Accessed 2019-04-27
10. Plass, M., Solana, J., Wolf, F.A., Ayoub, S., Misios, A., Glažar, P., Obermayer, B., Theis, F.J., Kocks, C., Rajewsky, N.: Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. Science **360**(6391) (2018). doi:10.1126/science.aaq1723. Accessed 2018-04-19
11. Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., Clevers, H., Deplancke, B., Dunham, I., Eberwine, J., Eils, R., Enard, W., Farmer, A., Fugger, L., Göttgens, B., Hacohen, N., Haniffa, M., Hemberg, M., Kim, S., Klenerman, P., Kriegstein, A., Lein, E., Linnarsson, S., Lundberg, E., Lundeberg, J., Majumder, P., Marioni, J.C., Merad, M., Mhlanga, M., Nawijn, M., Netea, M., Nolan, G., Pe'er, D., Phillipakis, A., Ponting, C.P., Quake, S., Reik, W., Rozenblatt-Rosen, O., Sanes, J., Satija, R., Schumacher, T.N., Shalek, A., Shapiro, E., Sharma, P., Shin, J.W., Stegle, O., Stratton, M., Stubbington, M.J.T., Theis, F.J., Uhlen, M., van Oudenaarden, A., Wagner, A., Watt, F., Weissman, J., Wold, B., Xavier, R., Yosef, N., Participants, H.C.A.M.: The human cell atlas. eLife **6** (2017). doi:10.7554/eLife.270. Accessed 2020-01-16

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 **14**(4), 414–416 (2017). doi:10.1038/nmeth.4207. Accessed 2019-04-28
13. Lin, P., Troup, M., Ho, J.W.K.: CIDR: Ultrafast and accurate clustering through imputation for single-cell RNA-seq data. Genome Biology **18**(1), 59 (2017). doi:10.1186/s13059-017-1188-0. Accessed 2019-04-28
14. Huh, R., Yang, Y., Jiang, Y., Shen, Y., Li, Y.: SAME-clustering: Single-cell aggregated clustering via mixture model ensemble. Nucleic Acids Research **48**(1), 86–95 (2020). doi:10.1093/nar/gkz959. Accessed 2020-06-30
15. Kiselev, V.Y., Kirschner, K., Schaub, M.T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K.N., Reik, W., Barahona, M., Green, A.R., Hemberg, M.: SC3: consensus clustering of single-cell RNA-seq data. Nature Methods **14**(5), 483–486 (2017). doi:10.1038/nmeth.4236. Accessed 2019-01-14
16. Li, X., Wang, K., Lyu, Y., Pan, H., Zhang, J., Stambolian, D., Susztak, K., Reilly, M.P., Hu, G., Li, M.: Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. Nature Communications **11**(1), 2338 (2020). doi:10.1038/s41467-020-15851-3. Accessed 2020-05-21
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 **10**, 1253 (2019). doi:10.3389/fgene.2019.01253. Accessed 2020-05-21
18. Kiselev, V.Y., Andrews, T.S., Hemberg, M.: Challenges in unsupervised clustering of single-cell RNA-seq data. Nature Reviews. Genetics **20**(5), 273–282 (2019). doi:10.1038/s41576-018-0088-9. Accessed 2019-02-01
19. Duò, A., Robinson, M.D., Soneson, C.: A systematic performance evaluation of clustering methods for single-cell RNA-seq data. F1000Research **7**, 1141 (2018). doi:10.12688/f1000research.15666.2. Accessed 2019-03-27
20. Butler, A., Hoffman, P., Smibert, P., Papalexi, E., Satija, R.: Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology **36**(5), 411–420 (2018). doi:10.1038/nbt.4096. Accessed 2019-04-28
21. Wolf, F.A., Angerer, P., Theis, F.J.: SCANPY: large-scale single-cell gene expression data analysis. Genome Biology **19**(1), 15 (2018). doi:10.1186/s13059-017-1382-0. Accessed 2019-04-26
22. Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008**(10), 10008 (2008). doi:10.1088/1742-5468/2008/10/P10008. Accessed 2016-12-07
23. Traag, V.A., Waltman, L., van Eck, N.J.: From louvain to leiden: guaranteeing well-connected communities. Scientific Reports **9**(1), 5233 (2019). doi:10.1038/s41598-019-41695-z. Accessed 2019-07-23
24. Levine, J.H., Simonds, E.F., Bendall, S.C., Davis, K.L., Amir, E.-a.D., Tadmor, M.D., Litvin, O., Fienberg, H.G., Jager, A., Zunder, E.R., Finck, R., Gedman, A.L., Radtke, I., Downing, J.R., Pe'er, D., Nolan, G.P.: Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. Cell **162**(1), 184–197 (2015). doi:10.1016/j.cell.2015.05.047. Accessed 2017-04-12
25. Newman, M.E.J., Girvan, M.: Finding and evaluating community structure in networks. Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics **69**(2 Pt 2), 026113 (2004). doi:10.1103/PhysRevE.69.0261. Accessed 2019-06-26
26. Traag, V.A., Van Dooren, P., Nesterov, Y.: Narrow scope for resolution-limit-free community detection. Physical Review E **84**(1) (2011). doi:10.1103/PhysRevE.84.0161. Accessed 2020-06-30
27. Reichardt, J., Bornholdt, S.: Statistical mechanics of community detection. Physical Review E **74**(1) (2006). doi:10.1103/PhysRevE.74.0161. Accessed 2020-06-30
28. Lähnemann, D., Köster, J., Szczurek, E., McCarthy, D.J., Hicks, S.C., Robinson, M.D., Vallejos, C.A., Campbell, K.R., Beerenwinkel, N., Mahfouz, A., Pinello, L., Skums, P., Stamatakis, A., Attolini, C.S.-O., Aparicio, S., Baaijens, J., Balvert, M., Barbanson, B.d., Cappuccio, A., Corleone, G., Dutilh, B.E., Florescu, M., Guryev, V., Holmer, R., Jahn, K., Lobo, T.J., Keizer, E.M., Khatri, I., Kielbasa, S.M., Korbel, J.O., Kozlov, A.M., Kuo, T.-H., Lelieveldt, B.P.F., Mandoiu, I.I., Marioni, J.C., Marschall, T., Mölder, F., Niknejad, A., Raczkowski, L., Reinders, M., Ridder, J.d., Saliba, A.-E., Somarakis, A., Stegle, O., Theis, F.J., Yang, H., Zelikovsky, A., McHardy, A.C., Raphael, B.J., Shah, S.P., Schönhuth, A.: Eleven grand challenges in single-cell data science. Genome Biology **21**(1), 31 (2020). doi:10.1186/s13059-020-1926-6. Accessed 2020-02-12
29. Fortunato, S., Barthélemy, M.: Resolution limit in community detection. Proceedings of the National Academy of Sciences of the United States of America **104**(1), 36–41 (2007). doi:10.1073/pnas.0605965104. Accessed 2016-12-07
30. Guimerà, R., Sales-Pardo, M., Amaral, L.A.N.: Modularity from fluctuations in random graphs and complex networks. Physical Review E **70**(2) (2004). doi:10.1103/PhysRevE.70.0251. Accessed 2020-06-30
31. Baran, Y., Bercovich, A., Sebe-Pedros, A., Lubling, Y., Giladi, A., Chomsky, E., Meir, Z., Hoichman, M., Lifshitz, A., Tanay, A.: MetaCell: analysis of single-cell RNA-seq data using k-nn graph partitions. Genome Biology **20**(1), 206 (2019). doi:10.1186/s13059-019-1812-2. Accessed 2019-12-04
32. Xu, C., Su, Z.: Identification of cell types from single-cell transcriptomes using a novel clustering method. Bioinformatics **31**(12), 1974–1980 (2015). doi:10.1093/bioinformatics/btv088. Accessed 2019-04-28
33. Miao, Z., Moreno, P., Huang, N., Papatheodorou, I., Brazma, A., Teichmann, S.A.: Putative cell type discovery from single-cell gene expression data. Nature Methods **17**(6), 621–628 (2020). doi:10.1038/s41592-020-0825-9. Accessed 2020-06-21
34. Holland, P.W., Laskey, K.B., Leinhardt, S.: Stochastic blockmodels: First steps. Social networks **5**(2), 109–137 (1983). doi:10.1016/0378-8733(83)90021-7. Accessed 2020-06-30
35. Peixoto, T.P.: Nonparametric bayesian inference of the microcanonical stochastic block model. Physical review. E **95**(1-1), 012317 (2017). doi:10.1103/PhysRevE.95.0123. Accessed 2020-02-06
36. Karrer, B., Newman, M.E.J.: Stochastic blockmodels and community structure in networks. Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics **83**(1 Pt 2), 016107 (2011). doi:10.1103/PhysRevE.83.0161. Accessed 2020-06-30
37. Peixoto, T.P.: Parsimonious module inference in large networks. Physical Review Letters **110**(14), 148701 (2013). doi:10.1103/PhysRevLett.110.1487. Accessed 2020-06-30

38. Peixoto, T.P.: 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 (2014). doi:10.1103/PhysRevE.89.0128. Accessed 2020-03-11
39. Peixoto, T.P.: Hierarchical block structures and high-resolution model selection in large networks. Physical Review X **4**(1), 011047 (2014). doi:10.1103/PhysRevX.4.0110. Accessed 2019-06-28
40. Tian, L., Dong, X., Freytag, S., Lê Cao, K.-A., Su, S., JalalAbadi, A., Amann-Zalcenstein, D., Weber, T.S., Seidi, A., Jabbari, J.S., Naik, S.H., Ritchie, M.E.: Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. Nature Methods **16**(6), 479–487 (2019). doi:10.1038/s41592-019-0425-8. Accessed 2019-06-05
41. Kang, H.M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., Wan, E., Wong, S., Byrnes, L., Lanata, C.M., Gate, R.E., Mostafavi, S., Marson, A., Zaitlen, N., Criswell, L.A., Ye, C.J.: Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nature Biotechnology **36**(1), 89–94 (2018). doi:10.1038/nbt.4042. Accessed 2019-04-28
42. Paul, F., Arkin, Y., Giladi, A., Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Winter, D., Lara-Astiaso, D., Gury, M., Weiner, A., David, E., Cohen, N., Lauridsen, F.K.B., Haas, S., Schlitzer, A., Mildner, A., Ginhoux, F., Jung, S., Trumpp, A., Porse, B.T., Tanay, A., Amit, I.: Transcriptional heterogeneity and lineage commitment in myeloid progenitors. Cell **163**(7), 1663–1677 (2015). doi:10.1016/j.cell.2015.11.013. Accessed 2017-06-15
43. Wolf, F.A., Hamey, F.K., Plass, M., Solana, J., Dahlin, J.S., Göttgens, B., Rajewsky, N., Simon, L., Theis, F.J.: PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome Biology **20**(1), 59 (2019). doi:10.1186/s13059-019-1663-x. Accessed 2019-04-09
44. Mereu, E., Lafzi, A., Moutinho, C., Ziegenhain, C., McCarthy, D.J., Álvarez-Varela, A., Batlle, E., Sagar, Grün, D., Lau, J.K., Boutet, S.C., Sanada, C., Ooi, A., Jones, R.C., Kaihara, K., Brampton, C., Talaga, Y., Sasagawa, Y., Tanaka, K., Hayashi, T., Braeuning, C., Fischer, C., Sauer, S., Trefzer, T., Conrad, C., Adiconis, X., Nguyen, L.T., Regev, A., Levin, J.Z., Parekh, S., Janjic, A., Wange, L.E., Bagnoli, J.W., Enard, W., Gut, M., Sandberg, R., Nikaido, I., Gut, I., Stegle, O., Heyn, H.: Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. Nature Biotechnology **38**(6), 747–755 (2020). doi:10.1038/s41587-020-0469-4. Accessed 2020-04-07
45. Sasagawa, Y., Danno, H., Takada, H., Ebisawa, M., Tanaka, K., Hayashi, T., Kurisaki, A., Nikaido, I.: Quartz-seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. Genome Biology **19**(1), 29 (2018). doi:10.1186/s13059-018-1407-3. Accessed 2019-04-28
46. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., Gregory, M.T., Shuga, J., Montesclaros, L., Underwood, J.G., Masquelier, D.A., Nishimura, S.Y., Schnall-Levin, M., Wyatt, P.W., Hindson, C.M., Bharadwaj, R., Wong, A., Ness, K.D., Beppu, L.W., Deeg, H.J., McFarland, C., Loeb, K.R., Valente, W.J., Ericson, N.G., Stevens, E.A., Radich, J.P., Mikkelsen, T.S., Hindson, B.J., Bielas, J.H.: Massively parallel digital transcriptional profiling of single cells. Nature Communications **8**, 14049 (2017). doi:10.1038/ncomms14049. Accessed 2019-06-24
47. Jaitin, D.A., Kenigsberg, E., Keren-Shaul, H., Elefant, N., Paul, F., Zaretsky, I., Mildner, A., Cohen, N., Jung, S., Tanay, A., Amit, I.: Massively parallel single-cell RNA-seq for marker-free decomposition of tissues into cell types. Science **343**(6172), 776–779 (2014). doi:10.1126/science.1247651. Accessed 2019-04-28
48. Goldstein, L.D., Chen, Y.-J.J., Dunne, J., Mir, A., Hubschle, H., Guillory, J., Yuan, W., Zhang, J., Stinson, J., Jaiswal, B., Pahuja, K.B., Mann, I., Schaal, T., Chan, L., Anandakrishnan, S., Lin, C.-W., Espinoza, P., Husain, S., Shapiro, H., Swaminathan, K., Wei, S., Srinivasan, M., Seshagiri, S., Modrusan, Z.: Massively parallel nanowell-based single-cell gene expression profiling. BMC Genomics **18**(1), 519 (2017). doi:10.1186/s12864-017-3893-1. Accessed 2019-04-28

Figures

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 *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.

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).

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.

Additional Files

Additional file 1 — Supplementary figures

The file supplementary.pdf includes supplementary figures from Figure S1 to Figure S5