


## METHOD

## Open Access



# PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells

F. Alexander Wolf<sup>1</sup> , Fiona K. Hamey<sup>2</sup>, Mireya Plass<sup>3</sup>, Jordi Solana<sup>3</sup>, Joakim S. Dahlin<sup>2,4</sup>, Berthold Göttgens<sup>2</sup>, Nikolaus Rajewsky<sup>3</sup>, Lukas Simon<sup>1</sup> and Fabian J. Theis<sup>1,5\*</sup>

## Abstract

Single-cell RNA-seq quantifies biological heterogeneity across both discrete cell types and continuous cell transitions. Partition-based graph abstraction (PAGA) provides an interpretable graph-like map of the arising data manifold, based on estimating connectivity of manifold partitions (<https://github.com/theislab/paga>). PAGA maps preserve the global topology of data, allow analyzing data at different resolutions, and result in much higher computational efficiency of the typical exploratory data analysis workflow. We demonstrate the method by inferring structure-rich cell maps with consistent topology across four hematopoietic datasets, adult planaria and the zebrafish embryo and benchmark computational performance on one million neurons.

## Background

Single-cell RNA-seq offers unparalleled opportunities for comprehensive molecular profiling of thousands of individual cells, with expected major impacts across a broad range of biomedical research. The resulting datasets are often discussed using the term transcriptional landscape. However, the algorithmic analysis of cellular heterogeneity and patterns across such landscapes still faces fundamental challenges, for instance, in how to explain cell-to-cell variation. Current computational approaches attempt to achieve this usually in one of two ways [1]. Clustering assumes that data is composed of biologically distinct groups such as discrete cell types or states and labels these with a discrete variable—the cluster index. By contrast, inferring pseudotemporal orderings or trajectories of cells [2–4] assumes that data lie on a connected manifold and labels cells with a continuous variable—the distance along the manifold. While the former approach is the basis for most analyses of single-cell data, the latter enables a better interpretation of continuous phenotypes

and processes such as development, dose response, and disease progression. Here, we unify both viewpoints.

A central example of dissecting heterogeneity in single-cell experiments concerns data that originate from complex cell differentiation processes. However, analyzing such data using pseudotemporal ordering [2, 5–9] faces the problem that biological processes are usually incompletely sampled. As a consequence, experimental data do not conform with a connected manifold and the modeling of data as a continuous tree structure, which is the basis for existing algorithms, has little meaning. This problem exists even in clustering-based algorithms for the inference of tree-like processes [10–12], which make the generally invalid assumption that clusters conform with a connected tree-like topology. Moreover, they rely on feature-space based inter-cluster distances, like the euclidean distance of cluster means. However, such distance measures quantify biological similarity of cells only at a local scale and are fraught with problems when used for larger-scale objects like clusters. Efforts for addressing the resulting high non-robustness of tree-fitting to distances between clusters [10] by sampling [11, 12] have only had limited success.

Partition-based graph abstraction (PAGA) resolves these fundamental problems by generating graph-like

\*Correspondence: [fabian.theis@helmholtz-muenchen.de](mailto:fabian.theis@helmholtz-muenchen.de)

<sup>1</sup>Helmholtz Center Munich – German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Munich, Germany

<sup>5</sup>Department of Mathematics, Technische Universität München, Munich, Germany

Full list of author information is available at the end of the article



maps of cells that preserve both continuous and disconnected structure in data at multiple resolutions. The data-driven formulation of PAGA allows to robustly reconstruct branching gene expression changes across different datasets and, for the first time, enabled reconstructing the lineage relations of a whole adult animal [13]. Furthermore, we show that PAGA-initialized manifold learning algorithms converge faster, produce embeddings that are more faithful to the global topology of high-dimensional data, and introduce an entropy-based measure for quantifying such faithfulness. Finally, we show how PAGA abstracts transition graphs, for instance, from RNA velocity and compare to previous trajectory-inference algorithms. With this, PAGA provides a graph abstraction method [14] that is suitable for deriving interpretable abstractions of the noisy kNN-like graphs that are typically used to represent the manifolds arising in scRNA-seq data.

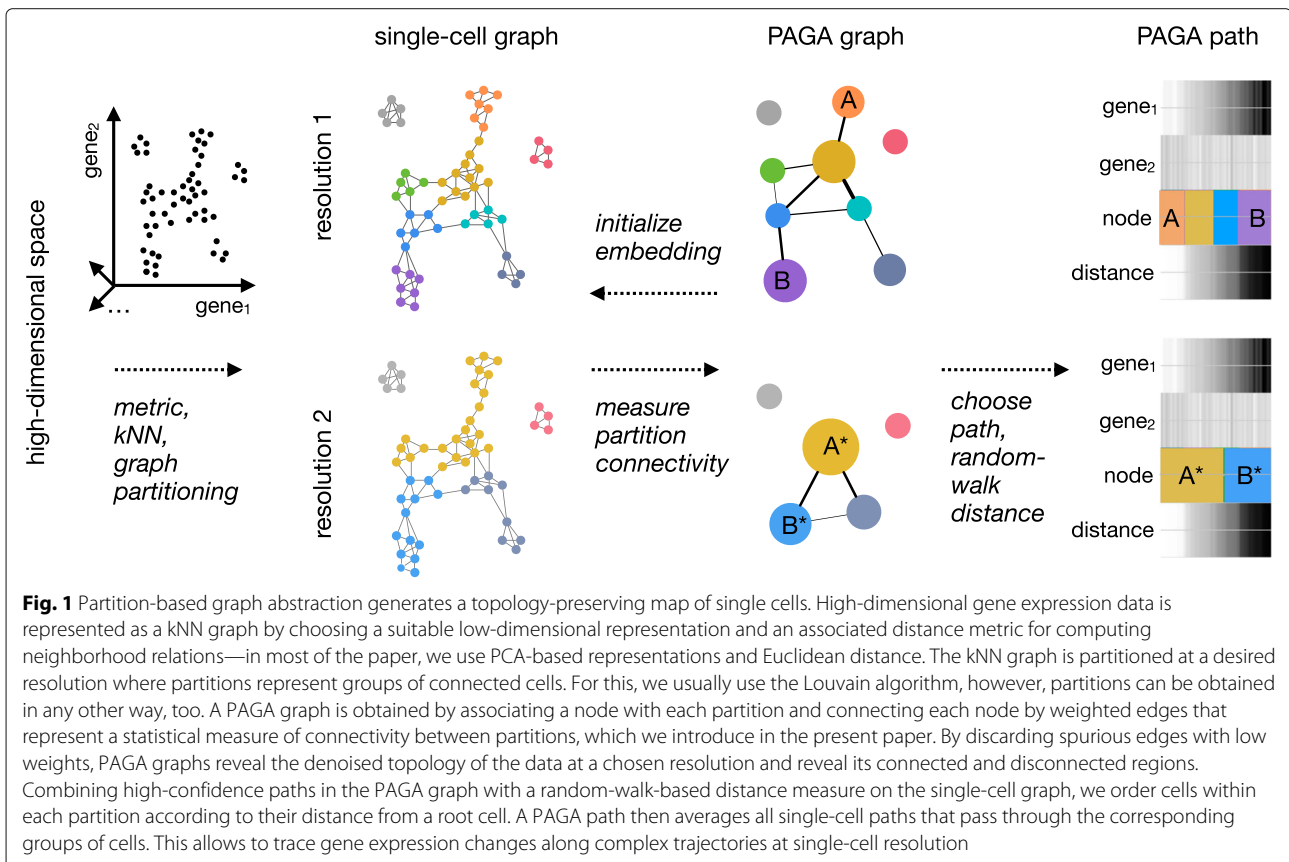
### Results

#### PAGA maps discrete disconnected and continuous connected cell-to-cell variation

Both established manifold learning techniques and single-cell data analysis techniques represent data as a neighborhood graph of single cells  $G = (V, E)$ , where each

node in  $V$  corresponds to a cell and each edge in  $E$  represents a neighborhood relation (Fig. 1) [3, 15–17]. However, the complexity of  $G$  and noise-related spurious edges make it both hard to trace a putative biological process from progenitor cells to different fates and to decide whether groups of cells are in fact connected or disconnected. Moreover, tracing isolated paths of single cells to make statements about a biological process comes with too little statistical power to achieve an acceptable confidence level. Gaining power by averaging over distributions of single-cell paths is hampered by the difficulty of fitting realistic models for the distribution of these paths.

We address these problems by developing a statistical model for the connectivity of groups of cells, which we typically determine through graph-partitioning [17–19] or alternatively through clustering or experimental annotation. This allows us to generate a simpler PAGA graph  $G^*$  (Fig. 1) whose nodes correspond to cell groups and whose edge weights quantify the connectivity between groups. Similar to modularity [20], the statistical model considers groups as connected if their number of inter-edges exceeds a fraction of the number of inter-edges expected under random assignment. The connection strength can be interpreted as confidence in



the presence of an actual connection and allows discarding spurious, noise-related connections (Additional file 1: Note 1). While  $G$  represents the connectivity structure of the data at single-cell resolution, the PAGA graph  $G^*$  represents the connectivity structure of the data at the chosen coarser resolution of the partitioning and allows to identify connected and disconnected regions of the data. Following paths along nodes in  $G^*$  means following an ensemble of single-cell paths that pass through the corresponding cell groups in  $G$ . By averaging over such an ensemble of single-cell paths, it becomes possible to trace a putative biological process from a progenitor to fates in a way that is robust to spurious edges, provides statistical power, and is consistent with basic assumptions on a biological trajectory of cells (Additional file 1: Note 2). Note that by varying the resolution of the partitioning, PAGA generates graphs at multiple resolutions, which enables a hierarchical exploration of data (Fig. 1, Additional file 1: Note 1.3).

To trace gene dynamics at single-cell resolution, we extended existing random-walk-based distance measures (Additional file 1: Note 2, Reference [7]) to the realistic case that accounts for disconnected graphs. By following high-confidence paths in the abstracted graph  $G^*$  and ordering cells within each group in the path according to their distance  $d$  from a progenitor cell, we trace gene changes at single-cell resolution (Fig. 1). Hence, PAGA covers both aspects of clustering and pseudotemporal ordering by providing a coordinate system  $(G^*, d)$  that allows us to explore variation in data while preserving its topology (Additional file 1: Note 1.6). PAGA can thus be viewed as an easily interpretable and robust way of performing topological data analysis [9, 21] (Additional file 1: Note 3).

#### PAGA-initialized manifold learning produces topology-preserving single-cell embeddings

The computationally almost cost-free coarse-resolution embeddings of PAGA can be used to initialize established manifold learning and graph drawing algorithms like UMAP [22] and ForceAtlas2 (FA) [23]. This strategy is used to generate the single-cell embeddings throughout this paper. In contrast to the results of previous algorithms, PAGA-initialized single-cell embeddings are faithful to the global topology, which greatly improves their interpretability. To quantify this claim, we took a classification perspective on embedding algorithms and developed a cost function  $KL_{\text{geo}}$  (Box 1 and Additional file 1: Note 4), which captures faithfulness to global topology by incorporating geodesic distance along the representations of data manifolds in both the high-dimensional and the embedding space, respectively. Independent of this, PAGA-initialized manifold learning converges about six times faster with respect to established cost functions in manifold learning (Additional file 1: Figure S10)

**Box 1.** Taking a classification view on embedding algorithms, we quantify how faithful an embedding is to the global topology of the high-dimensional data by comparing the distributions  $P$  and  $Q$  of edges in the high-dimensional and embedding spaces using a weighted Kullback-Leibler divergence

$$KL_{\text{geo}}(P||Q) = KL_{\text{geo}}^{\text{disc}}(P||Q) + KL_{\text{geo}}^{\text{overl}}(P||Q)$$

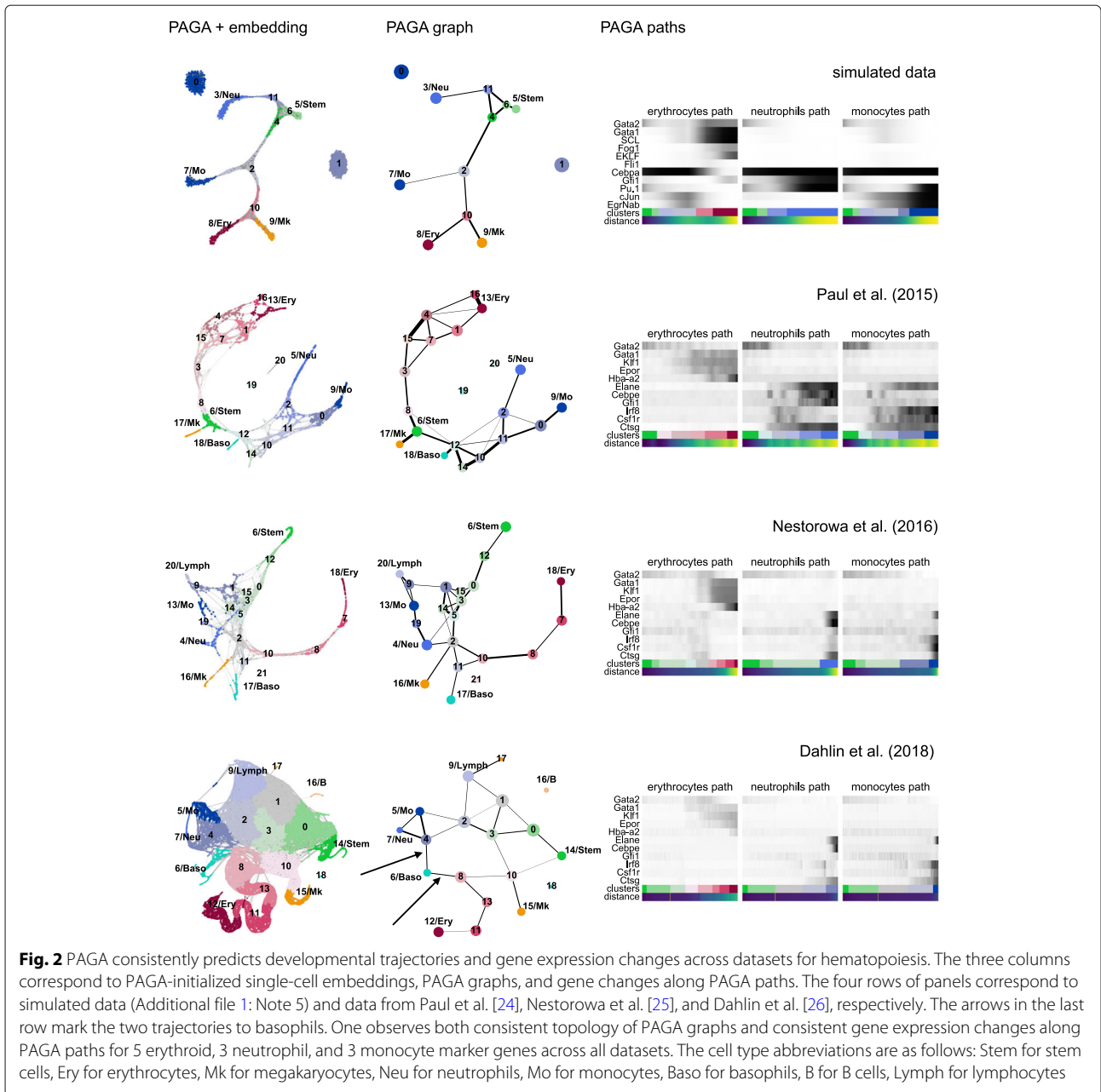
$$= \sum_{e \in E_{\text{fc}}} \underbrace{\frac{d_e^q}{d_e^p} p_e \log\left(\frac{p_e}{q_e}\right)}_{\text{disconnected cost}} + \underbrace{\frac{d_e^p}{d_e^q} (1-p_e) \log\left(\frac{1-p_e}{1-q_e}\right)}_{\text{overlapping cost}},$$

where  $p_e$  and  $q_e$  are the probabilities for an edge being present in the kNN graphs in the high-dimensional and embedding spaces, respectively. Analogously,  $d_e^p$  and  $d_e^q$  denote random-walk based estimators of geodesic distances on the manifolds in these spaces, respectively.  $E_{\text{fc}}$  denotes the edge set of the fully connected graph (Additional file 1: Note 4 and Figure S10).

#### PAGA consistently predicts developmental trajectories and gene expression changes in datasets related to hematopoiesis

Hematopoiesis represents one of the most extensively characterized systems involving stem cell differentiation towards multiple cell fates and hence provides an ideal scenario for applying PAGA to complex manifolds. We applied PAGA to simulated data (Additional file 1: Note 5) for this system and three experimental datasets: 2730 cells measured using MARS-seq [24], 1654 cells measured using Smart-seq2 [25], and 44,802 cells from a 10× Genomics protocol [26]. These data cover the differentiation from stem cells towards cell fates including erythrocytes, megakaryocytes, neutrophils, monocytes, basophils, and lymphocytes.

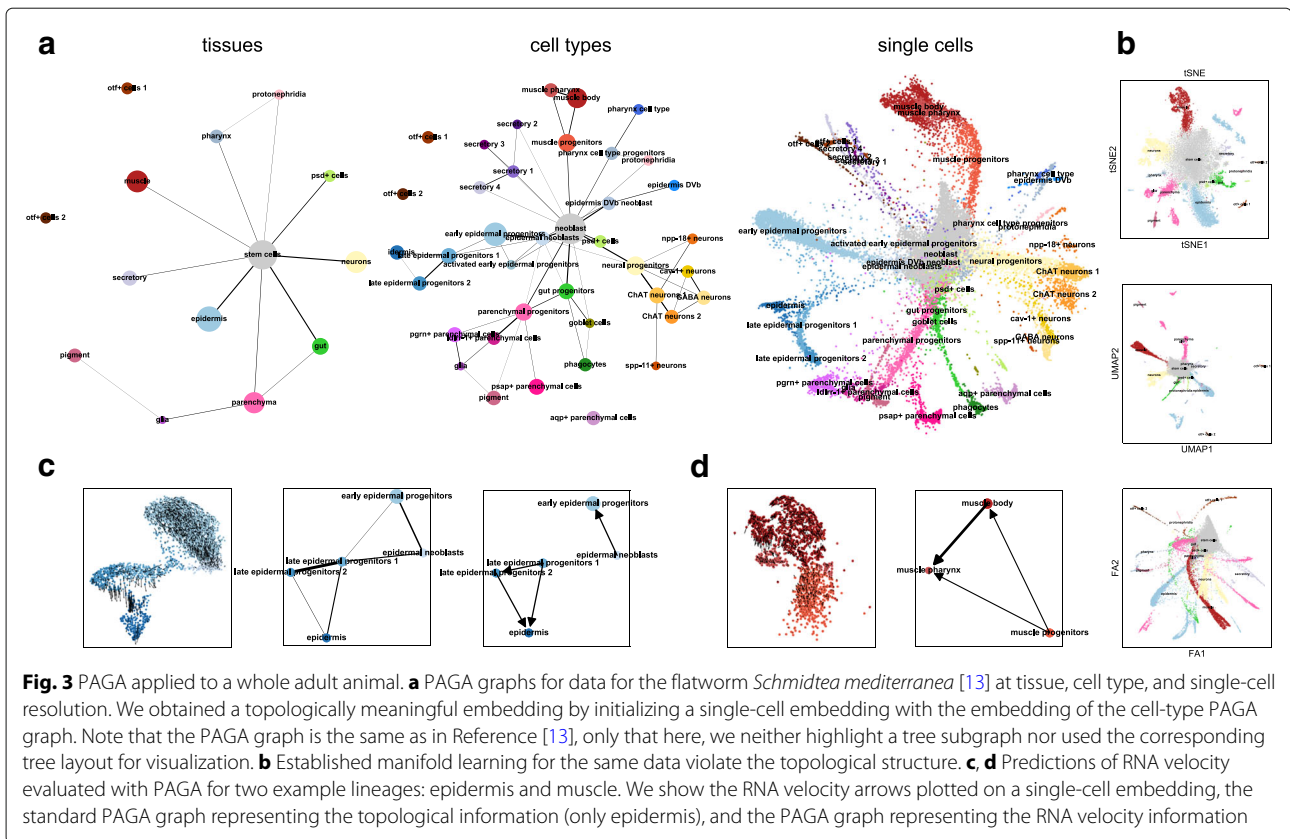
The PAGA graphs (Fig. 2) capture known features of hematopoiesis, such as the proximity of megakaryocyte and erythroid progenitors and strong connections between monocyte and neutrophil progenitors. Under debate is the origin of basophils. Studies have suggested both that basophils originate from a basophil-neutrophil-monocyte progenitor or, more recently, from a shared erythroid-megakaryocyte-basophil progenitor [27, 28]. The PAGA graphs of the three experimental datasets highlight this ambiguity. While the dataset of Paul et al. falls in the former category, Nestorowa et al. falls in the latter and Dahlin et al., which has by far the highest cell numbers and the densest sampling, allows us to see both trajectories. Aside from this ambiguity that can be explained by insufficient sampling in Paul et al. and Nestorowa et al., even with the very different experimental protocols and vastly different cell numbers the PAGA graphs show consistent topology between the three datasets. Beyond consistent topology between cell subgroups, we find consistent continuous gene expression



changes across all datasets—we observe changes of erythroid maturity marker genes (*Gata2*, *Gata1*, *Klf1*, *Epor*, and *Hba-a2*) along the erythroid trajectory through the PAGA graphs and observe sequential activation of these genes in agreement with known behavior. Activation of neutrophil markers (*Elane*, *Cepbe*, and *Gfi1*) and monocyte markers (*Irf8*, *Csf1r*, and *Ctsg*) are seen towards the end of the neutrophil and monocyte trajectories, respectively. While PAGA is able to capture the dynamic transcriptional processes underlying multilineage hematopoietic differentiation, previous algorithms often fail to robustly produce meaningful results (Additional file 1: Figures S8, S9, S10).

**PAGA maps single-cell data of whole animals at multiple resolutions**

Recently, Plass et al. [13] reconstructed the first cellular lineage tree of a whole adult animal, the flatworm *Schmidtea mediterranea*, using PAGA on scRNA-seq data from 21,612 cells. While Plass et al. focussed on the tree-like subgraph that maximizes overall connectivity—the minimum spanning tree of  $G^*$  weighted by inverse PAGA connectivity—here, we show how PAGA can be used to generate maps of data at multiple resolutions (Fig. 3a). Each map preserves the topology of data, in contrast to state-of-the-art manifold learning where connected tissue types appear as either disconnected or overlapping



(Fig. 3b). PAGA’s multi-resolution capabilities directly address the typical practice of exploratory data analysis, in particular for single-cell data: data is typically reclustered in certain regions where a higher level of detail is required.

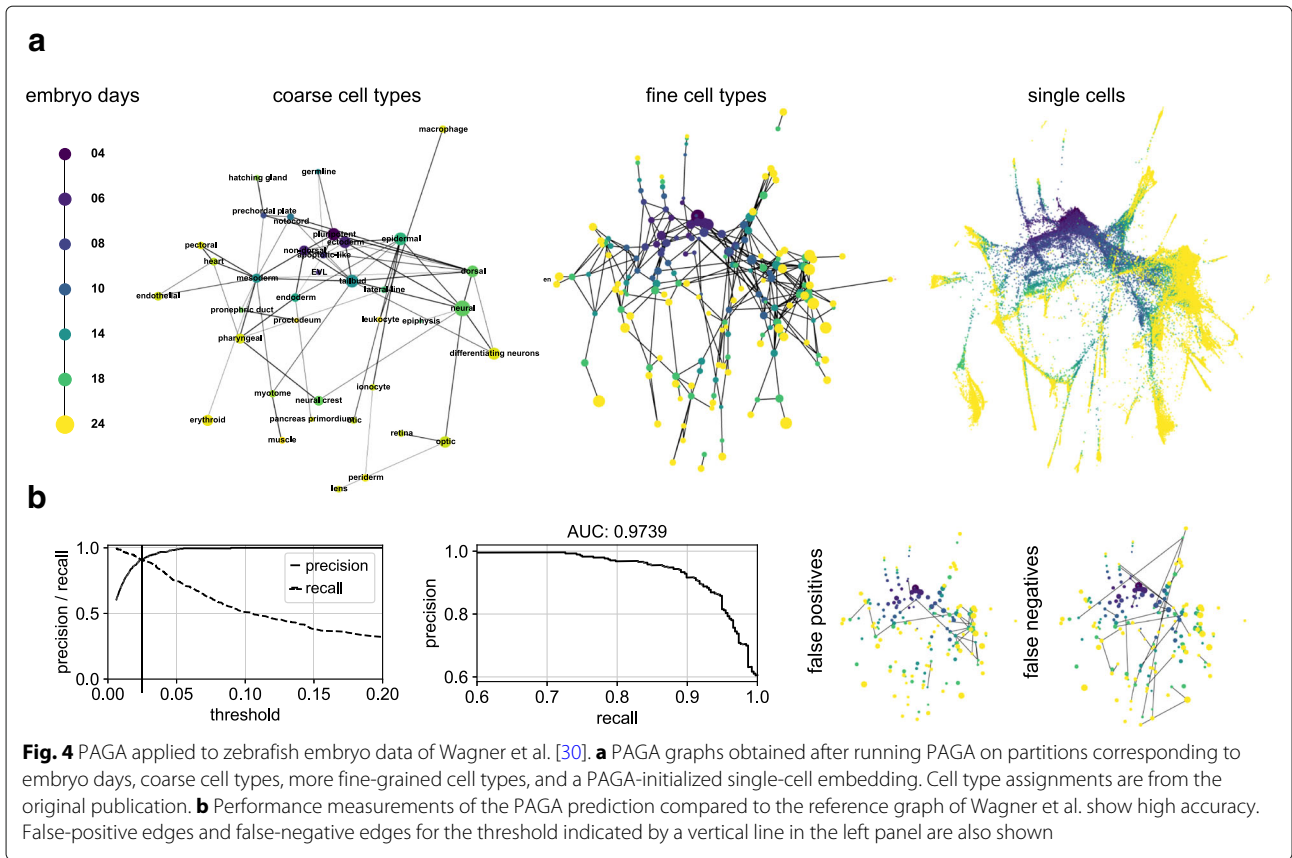
**PAGA abstracts information from RNA velocity**

Even though the connections in PAGA graphs often correspond to actual biological trajectories, this is not always the case. This is a consequence of PAGA being applied to kNN graphs, which solely contain information about the topology of data. Recently, it has been suggested to also consider directed graphs that store information about cellular transition based on RNA velocity [29]. To include this additional information, which can add further evidence for actual biological transitions, we extend the undirected PAGA connectivity measure to such directed graphs (Additional file 1: Note 1.2) and use it to orient edges in PAGA graphs (Fig. 3c). Due the relatively sparsely sampled, high-dimensional feature space of scRNA-seq data, both fitting and interpreting an RNA velocity vector without including information about topology—connectivity of neighborhoods—is practically impossible. PAGA provides a natural way of abstracting both topological information and information about RNA velocity.

Next, we applied PAGA to 53,181 cells collected at different developmental time points (embryo days) from the zebrafish embryo [30]. The PAGA graph for partitions corresponding to embryo days accurately recovers the chain topology of temporal progression, whereas the PAGA graph for cell types provides easily interpretable overviews of the lineage relations (Fig. 4a). Initializing a ForceAtlas2 layout with PAGA coordinates from fine cell types automatically produced a corresponding, interpretable single-cell embedding (Fig. 4a). Wagner et al. [30] both applied an independently developed computational approach with similarities to PAGA (Additional file 1: Note 3) to produce a coarse-grained graph and experimentally validated inferred lineage relations. Comparing the PAGA graph for the fine cell types to the coarse-grained graph of Wagner et al. reproduced their result with high accuracy (Fig. 4b).

**PAGA increases computational efficiency and interpretability in general exploratory data analysis and manifold learning**

Comparing the runtimes of PAGA with the state-of-the-art UMAP [22] for 1.3 million neuronal cells of 10× Genomics [31] we find a speedup of about 130, which enables interactive analysis of very large-scale data (90 s



versus 191 min on 3 cores of a small server, tSNE takes about 10 h). For complex and large data, the PAGA graph generally provides a more easily interpretable visualization of the clustering step in exploratory data analysis, where the limitations of two-dimensional representations become apparent (Additional file 1: Figure S12). PAGA graph visualizations can be colored by gene expression and covariates from annotation (Additional file 1: Figure S13) just as any conventional embedding method.

**PAGA is robust and qualitatively outperforms previous lineage reconstruction algorithms**

To assess how robustly graph and tree-inference algorithms recover a given topology, we developed a measure for comparing the topologies of two graphs by comparing the sets of possible paths on them (Additional file 1: Note 1.4, Figure S4). Sampling widely varying parameters, which leads to widely varying clusterings, we find that the inferred abstraction of topology of data within the PAGA graph is much more robust than the underlying graph clustering algorithm (Additional file 1: Figure S5). While graph clustering alone is, as any clustering method, an ill-posed problem in the sense that many highly degenerate quasi-optimal clusterings exist and

some knowledge about the scale of clusters is required, PAGA is not affected by this.

Several algorithms [5, 10–12] have been proposed for reconstructing lineage trees (Additional file 1: Note 3, [4]). The main caveat of these algorithms is that they, unlike PAGA, try to explain any variation in the data with a tree-like topology. In particular, any disconnected distribution of clusters is interpreted as originating from a tree. This produces qualitatively wrong results already for simple simulated data (Supplementary Figure 6) and only works well for data that clearly conforms with a tree-like manifold (Supplementary Figure 7). To establish a fair comparison on real data with the recent popular algorithm, Monocle 2, we reinvestigated the main example of Qiu et al. [5] for a complex differentiation tree. This example is based on the data of Paul et al. [24] (Fig. 2), but with cluster 19 removed. While PAGA identifies the cluster as disconnected with a result that is unaffected by its presence, the prediction of Monocle 2 changes qualitatively if the cluster is taken into account (Supplementary Figure 8). The example illustrates the general point that real data almost always consists of dense and sparse—connected and disconnected—regions, some tree-like, some with more complex topology.

## Conclusions

In view of an increasing number of large datasets and analyses for even larger merged datasets, PAGA fundamentally addresses the need for scalable and interpretable maps of high-dimensional data. In the context of the Human Cell Atlas [32] and comparable databases, methods for their hierarchical, multi-resolution exploration will be pivotal in order to provide interpretable accessibility to users. PAGA allows to present information about clusters or cell types in an unbiased, data-driven coordinate system by representing these in PAGA graphs. In the context of the recent advances of the study of simple biological processes that involve a single branching [6, 7], PAGA provides a similarly robust framework for arbitrarily complex topologies. In view of the fundamental challenges of single-cell resolution studies due to technical noise, transcriptional stochasticity, and computational burden, PAGA provides a general framework for extending studies of the relations among single cells to relations among noise-reduced and computationally tractable groups of cells. This could facilitate obtaining clearer pictures of underlying biology.

In closing, we note that PAGA not only works for scRNA-seq based on distance metrics that arise from a sequence of chosen preprocessing steps, but can also be applied to any learned distance metric. To illustrate this point, we used PAGA for single-cell imaging data when applied on the basis of a deep-learning-based distance metric. Eulenberg et al. [33] showed that a deep learning model can generate a feature space in which distances reflect the continuous progression of cell cycle. Using this, PAGA correctly identifies the biological trajectory through the interphases of cell cycle while ignoring a cluster of damaged and dead cells (Additional file 1: Figure S14).

## Methods

### Preprocessing scRNA-seq data

We preprocess scRNA-seq data as commonly done following steps mostly inspired by Seurat [34] in the implementation of Scanpy [35]. These steps consist in basic filtering of the data, total count normalization,  $\log_1p$  logarithmization, extraction of highly variable genes, a potential regression of confounding factors, and a scaling to  $z$ -scores. On this corrected and homogenized representation of the count data, we perform a PCA and represent the data within the reduced space of principal components. As an alternative to this “classical” procedure, which is built on the PCA representation of the data, one might consider using the latent space representation of neural network model such as scVI for scRNA-seq data [36], or as the classifier discussed in Additional file 1: Note 5.6. Detailed parameters used for the processing can be found in Additional file 1: Note 5 and at

<https://github.com/theislab/paga>. In the GitHub repository, each figure of the paper is reproduced in a dedicated notebook.

### Graph construction

Using the compressed and denoised representation of the data in the previous step, we construct a symmetrized kNN-like graph, typically using the approximate nearest neighbor search within UMAP [22]. While one might potentially choose different distance metrics, we always choose Euclidean distance. Depending on user choice, the graph is either weighed using adaptive Gaussian kernels [7] or the exponential kernel within UMAP [22]. For all results shown in the manuscript, we used the exponential kernel.

### Graph partitioning and abstraction

We consider all partitionings of interest of the kNN-like graph. To determine those, typically, we use the Louvain algorithm in the implementation of [37] at suitable resolutions, but PAGA works with any underlying clustering algorithm or experimentally generated groupings of observations. In the present work, we exclusively used the Louvain algorithm.

In the conventional undirected case, for each partitioning, we generate a PAGA graph using the “PAGA connectivity measure” defined in Additional file 1: Eq. (11). This measure is a test statistic quantifying the degree of connectivity of two partitions and has a close relation with modularity [20]. For each pair of clusters, PAGA connectivity is the ratio of the number of inter-edges between the clusters normalized with the number of inter-edges expected under random assignment of edges.

In the directed case, in which we typically abstract a “velocity graph” originating from RNA velocity [29], we consider the ratio of arrows Additional file 1: Eq. (14), which are in- and outgoing for each pair of partitions to quantify a tendency of transition between partitions.

### Pseudotime estimation

For estimating pseudotime, we use an extended version of diffusion pseudotime (DPT) Reference [7] that accounts for disconnected graphs. The extension consists in a simple modification of the original algorithm that accounts for disconnected Eigen-subspaces of the graph adjacency matrix, which results in multiple subspaces of Eigen value 1 of the graph transition matrix. Practically, we assign an infinite distance to cells that reside in disconnected clusters and compute distances among cells within connected regions in the graph as it would be done in DPT. See Additional file 1: Note 2, both for details and for a review of random-walk-based distances. For instance, we show the close relation of DPT to mean commute distance.

### Consistent embeddings across resolutions

PAGA achieves consistent (i.e., minimally displaced in the embedding space) and topology-preserving embeddings by initializing an embedding of a fine-grained graph using the coordinates of a coarse-grained graph. For this initialization, the positions of nodes of the fine-grained graph that belong to a group corresponding to a node in the coarse-grained graph are randomly distributed in a non-overlapping rectangular region around the position of that node. This procedure is repeated for all nodes of the coarse-grained graph. Non-overlapping regions are trivially ensured by choosing rectangles with half-edge lengths of half the distance to the nearest neighbor in the coarse-grained embedding.

Conversely, for a given fine-grained graph, we position nodes in the coarse-grained graph by placing them on the median coordinates of the positions of the corresponding nodes in the fine-grained graph.

### Additional file

**Additional file 1:** Supplementary figures and notes. (PDF 6324 kb)

### Acknowledgements

F.A.W. thanks N. Yosef and D. Wagner for stimulating discussions, S. Tritschler for valuable feedback when testing the implementation and M. Luecken and V. Traag for comments on “connectivity-preserving” graph partitioning. We thank reviewer 1 for pointing us to the review of [14].

### Funding

F.A.W. acknowledges support by the Helmholtz Postdoc Programme, Initiative and Networking Fund of the Helmholtz Association. J.S.D. is supported by a grant from the Swedish Research Council. Work in B.G.'s laboratory is supported by grants from Wellcome, Bloodwise, Cancer Research UK, NIH-NIDDK, and core support grants by Wellcome to the Cambridge Institute for Medical Research and Wellcome-MRC Cambridge Stem Cell Institute. F.K.H. is the recipient of a Medical Research Council PhD Studentship. The work from M.P., J.S., and N.R. was funded by the German Center for Cardiovascular Research (DZHK BER 1.2 VD) and the DFG (grant RA 838/5-1). F.J.T. is supported by the German Research Foundation (DFG) within the Collaborative Research Centre 1243, Subproject A17.

### Availability of data and materials

PAGA as well as all processing steps used within the analyses are available within Scanpy [35]: <https://github.com/theislab/scanpy>. The analyses and results of the present paper are available from <https://github.com/theislab/paga> [38] together with all used data and have been obtained using Scanpy 1.2. PAGA is licensed under the BSD-3 license.

The Planaria dataset is available from NCBI GEO under accession number GSE103633 [13], the Zebrafish embryo dataset is available under GSE112294 [30].

### Authors' contributions

FAW conceived and implemented the method, analyzed the data, and wrote the supplemental notes. FKH analyzed the data of Dahlin et al. [26]. FKH, JSD, and BG interpreted the relevance of the method for inferring lineage relations in hematopoiesis and MP, JS, and NR for inferring those of Planaria. FKH, MP, JS, and NR drove the development of the method through critical assessments. LS and FJT contributed to the conception of the project. FJT supervised the project and wrote parts of Supplemental Note 1.3. FAW and FJT wrote the paper with contributions from all coauthors. All authors read and approved the final manuscript.

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

### Author details

<sup>1</sup>Helmholtz Center Munich – German Research Center for Environmental Health, Institute of Computational Biology, Neuherberg, Munich, Germany. <sup>2</sup>Department of Haematology and Wellcome and Medical Research Council Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK. <sup>3</sup>Berlin Institute for Medical Systems Biology, Max-Delbrück Center for Molecular Medicine, Berlin, Germany. <sup>4</sup>Department of Medicine, Karolinska Institutet and Karolinska University Hospital, Stockholm, Sweden. <sup>5</sup>Department of Mathematics, Technische Universität München, Munich, Germany.

Received: 5 November 2018 Accepted: 26 February 2019

Published online: 19 March 2019

### References

1. Wagner A, Regev A, Yosef N. Revealing the vectors of cellular identity with single-cell genomics. *Nat Biotechnol*. 2016;34(11):1145–60. <https://doi.org/10.1038/nbt.3711>.
2. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, Lennon NJ, Livak KJ, Mikkelsen T. S, Rinn JL. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat Biotechnol*. 2014;32(4):381–6. <https://doi.org/10.1038/nbt.2859>.
3. Bendall SC, Davis KL, Amir E-aD, Tadmor MD, Simonds EF, Chen TJ, Shenfeld DK, Nolan GP, Pe'er D. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell*. 2014;157(3):714–25. <https://doi.org/10.1016/j.cell.2014.04.005>.
4. Saelens W, Cannoodt R, Todorov H, Saey Y. A comparison of single-cell trajectory inference methods: towards more accurate and robust tools. *bioRxiv*. 2018;276907. <https://doi.org/10.1101/276907>.
5. Qiu X, Hill A, Packer J, Lin D, Ma YA, Trapnell C. Single-cell mRNA quantification and differential analysis with census. *Nat Methods*. 2017;14:309–15. <https://doi.org/10.1038/nmeth.4150>.
6. Setty M, Tadmor MD, Reich-Zeliger S, Angel O, Salame TM, Kathail P, Choi K, Bendall S, Friedman N, Pe'er D. Wishbone identifies bifurcating developmental trajectories from single-cell data. *Nat Biotechnol*. 2016;34:637–45. <https://doi.org/10.1038/nbt.3569>.
7. Haghverdi L, Büttner M, Wolf FA, Büttner F, Theis FJ. Diffusion pseudotime robustly reconstructs branching cellular lineages. *Nat Methods*. 2016;13:845–8. <https://doi.org/10.1038/nmeth.3971>.
8. Street K, Risso D, Fletcher RB, Das D, Ngai J, Yosef N, Purdom E, Dudoit S. Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*. 2018;19:477. <https://doi.org/10.1186/s12864-018-4772-0>.
9. Rizvi AH, Camara PG, Kandror EK, Roberts TJ, Schieren I, Maniatis T, Rabadan R. Single-cell topological rna-seq analysis reveals insights into cellular differentiation and development. *Nat Biotechnol*. 2017;35(6):551–60. <https://doi.org/10.1038/nbt.3854>.
10. Qiu P, Simonds EF, Bendall SC, Gibbs KD, Bruggner RV, Linderman M. D, Sachs K, Nolan GP, Plevritis SK. Extracting a cellular hierarchy from high-dimensional cytometry data with spade. *Nat Biotechnology*. 2011;29(10):886–91. <https://doi.org/10.1038/nbt.1991>.
11. Giecoold G, Marco E, Garcia SP, Trippa L, Yuan GC. Robust lineage reconstruction from high-dimensional single-cell data. *Nucleic Acids Res*. 2016;44(14):122. <https://doi.org/10.1093/nar/gkw452>.
12. Grün D, Muraro MJ, Boisset JC, Wiebrands K, Lyubimova A, Dharmadhikari G, van den Born M, van Es J, Jansen E, Clevers H, et al. De novo stem cell identity using single-cell transcriptome data. *Cell Stem Cell*. 2016;19(2):266–77. <https://doi.org/10.1016/j.stem.2016.05.010>.



13. Plass M, Solana J, Wolf FA, Ayoub S, Misios A, Glažar P, Obermayer B, Theis FJ, Kocks C, Rajewsky N. Cell type atlas and lineage tree of a whole complex animal by single-cell transcriptomics. *Science*. 2018;360(6391):1723. <https://doi.org/10.1126/science.aaa1723>.
14. Hu Y, Shi L. Visualizing large graphs. *Wiley Interdiscip Rev Comput Stat*. 2015;7(2):115–36. <https://doi.org/10.1002/wics.1343>.
15. van der Maaten L, Hinton G. Visualizing data using t-sne. *J Mach Learn Res*. 2008;9(Nov):2579–605.
16. Islam S, Kjallquist U, Moliner A, Zajac P, Fan JB, Lonnerberg P, Linnarsson S. Characterization of the single-cell transcriptional landscape by highly multiplex rna-seq. *Genome Res*. 2011;21(7):1160–7. <https://doi.org/10.1101/gr.110882.110>.
17. Levine JH, Simonds EF, Bendall SC, Davis KL, Amir E-aD, Tadmor MD, Litvin O, Fienberg HG, Jager A, Zunder ER, Finck R, Gedman AL, Radtke I, Downing JR, Pe'er D, Nolan GP. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. *Cell*. 2015;162(1):184–97. <https://doi.org/10.1016/j.cell.2015.05.047>.
18. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *J Stat Mech*. 2008;2008:10008. <https://doi.org/10.1088/1742-5468/2008/10/P10008.0803.0476v2>.
19. Xu C, Su Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*. 2015;31(12):1974–80. <https://doi.org/10.1093/bioinformatics/btv088>.
20. Newman MEJ. Modularity and community structure in networks. *Proc Natl Acad Sci*. 2007;103(23):8577–82. <https://doi.org/10.1073/pnas.0601602103>.
21. Singh G, Mémoli F, Carlsson GE. Topological methods for the analysis of high dimensional data sets and 3d object recognition. In: *Eurographics Symposium on Point-Based Graphics*; 2007. p. 91–100. <http://cs233.stanford.edu/ReferencedPapers/mapperPBG.pdf>.
22. McInnes L, Healy J. Umap: Uniform manifold approximation and projection for dimension reduction. 2018;1802–03426. arXiv:1802.03426.
23. Jacomy M, Venturini T, Heymann S, Bastian M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the gephi software. *PLoS ONE*. 2014;9(6):98679. <https://doi.org/10.1371/journal.pone.0098679>.
24. Paul F, Arkin Y, Giladi A, Jaitin DA, Kenigsberg E, Keren-Shaul H, Winter D, Lara-Astiaso D, Gury M, Weiner A, David E, Cohen N, Lauridsen FKB, Haas S, Schlitzer A, Mildner A, Ginhoux F, Jung S, Trumpp A, Porse BT, Tanay A, Amit I. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*. 2015;163:1663–77. <https://doi.org/10.1016/j.cell.2015.11.013>.
25. Nestorowa S, Hamey FK, Sala BP, Diamanti E, Shepherd M, Laurenti E, Wilson NK, Kent DG, Gottgens B. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood*. 2016;128(8):20–31. <https://doi.org/10.1182/blood-2016-05-716480>.
26. Dahlin JS, Hamey FK, Pijuan-Sala B, Shepherd M, Lau WY, Nestorowa S, Weinreb C, Wolock S, Hannah R, Diamanti E, Kent DG, Göttgens B, Wilson NK. A single cell hematopoietic landscape resolves eight lineage trajectories and defects in kit mutant mice. *Blood*. 2018;131:1–11. <https://doi.org/10.1182/blood-2017-12-821413>.
27. Görgens A, Ludwig AK, Möllmann M, Krawczyk A, Dürig J, Hanenberg H, Horn PA, Giebel B. Multipotent hematopoietic progenitors divide asymmetrically to create progenitors of the lymphomyeloid and erythromyeloid lineages. *Stem Cell Rep*. 2014;3:1058–72. <https://doi.org/10.1016/j.stemcr.2014.09.016>.
28. Tusi BK, Wolock SL, Weinreb C, Hwang Y, Hidalgo D, Zilionis R, Waisman A, Huh JR, Klein AM, Socolovsky M. Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature*. 2018;555(7694):54–60. <https://doi.org/10.1038/nature25741>.
29. La Manno G, Soldatov R, Zeisel A, Braun E, Hochgerner H, Petukhov V, Lidschreiber K, Kastriiti ME, Lönnerberg P, Furlan A, et al. RNA velocity of single cells. *Nature*. 2018;560(7719):494. <https://doi.org/10.1038/s41586-018-0414-6>.
30. 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;362. <https://doi.org/10.1126/science.aar4362>.
31. 10X Genomics. 1.3 Million Brain Cells from E18 Mice. [https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.3.0/1M\\_neurons](https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.3.0/1M_neurons). Accessed 5 Apr 2017.
32. Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, Bodenmiller B, Campbell PJ, 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 SK, Klenerman P, Kriegstein A, Lein E, Linnarsson S, Lundberg E, Lundeberg J, Majumder P, Marioni JC, Merad M, Mhlanga M, Nawijn M, Netea M, Nolan G, Pe'er D, Phillipakis A, Ponting CP, Quake SR, Reik W, Rozenblatt-Rosen O, Sanes JR, Satija R, Schumacher TN, Shalek AK, Shapiro E, Sharma P, Shin JW, Stegle O, Stratton MR, Stubbington MJT, Theis FJ, Uhlen M, van Oudenaarden A, Wagner A, Watt FM, Weissman JS, Wold BJ, Xavier RJ, NY. Science forum: The human cell atlas. *eLife*. 2017;6:27041. <https://doi.org/10.7554/elife.27041>.
33. Eulenberg P, Köhler N, Blasi T, Filby A, Carpenter AE, Rees P, Theis FJ, Wolf FA. Reconstructing cell cycle and disease progression using deep learning. *Nat Commun*. 2017;8:463. <https://doi.org/10.1038/s41467-017-00623-3>.
34. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. *Nat Biotechnol*. 2015;33:495–502. <https://doi.org/10.1038/nbt.3192>.
35. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol*. 2018;19(1):15. <https://doi.org/10.1186/s13059-017-1382-0>.
36. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. *Nat Methods*. 2018;15(12):1053. <https://doi.org/10.1038/s41592-018-0229-2>.
37. Traag V. Louvain. GitHub repository. 2017. <https://doi.org/10.5281/zenodo.35117>.
38. Wolf FA, Hamey F, Plass M, Solana J, Dahlin JS, Göttgens B, Rajewsky N, Simon L, Theis FJ. PAGA: Graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. GitHub repository. 2019. <https://github.com/theislab/paga>.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

