The Journal of Open Source Software

scmorph: a python package for analysing single-cell morphological profiles

Jesko Wagner [1], **Hugh Warden** [1], **Ava Khamseh** [1,2¶], **and Sjoerd Viktor Beentjes** [1,3¶]

1 MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK **2** School of Informatics, University of Edinburgh, Edinburgh, UK **3** School of Mathematics, University of Edinburgh, Edinburgh, UK ¶ Corresponding author

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

License
Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

Summary

scmorph is a Python package to analyse single-cell data from morphological profiling experiments which generate large tabular data. scmorph combines domain-specific methods such as single-cell hit calling and batch correction with the versatile and scalable scverse tools to offer feature selection, dimensionality reduction and more. Overall, scmorph brings together a host of single-cell morphological profiling methods, making it applicable for a wide range of experimental designs and workflows.

Statement of need

Morphological profiling has become an essential tool in biology and drug discovery, but there is a lack of open-source software tools for analysing single-cell morphological data. Existing tools are commercial, do not scale to large datasets, or do not offer single-cell specific tools (Omta et al., 2016; **?**). scmorph complements existing tools by providing a comprehensive set of methods for analysing single-cell morphological data, which do not require averaging of features. By integrating with the growing scverse of single-cell tools, scmorph also opens up advanced processing capabilities including access to deep learning tools (Wolf et al., 2018).

Briefly, scmorph provides five modules to analyze morphological profiles:

- Reading and writing (IO). scmorph allows reading data from csv, sql, sqlite, and h5ad files, including from the popular CellProfiler software (Stirling et al., 2021). Once converted, scmorph works with AnnData objects stored as h5ad, which track processing steps and can easily be written to disk (Virshup et al., 2024).
- Quality control. scmorph integrates two levels of quality control: image-level and single-cell level. To improve reusability of these methods, both approaches operature unsupervised. Image-level correction is performed with a kNN-based outlier detection method, whereas single-cell profiles that are outliers are detected via pyod (Li et al., 2022).
- Preprocessing. Provided functions perform feature selection, compute PCA coordinates, and optionally aggregate data. For the first time in the field, scmorph integrates scone as batch correction function, which retains interpretability of features (Cole et al., 2019). Additionally, the integrated feature selection methods can remove feature associated with known confounders or with high correlation structures, as is common in morphological profiling experiments (Kruskal & Wallis, 1952; Lin & Han, 2021).
- Plotting. scmorph uses scanpy for easy plotting of PCA and UMAP coordinates, either in 2D or as cumulative densities, which can be useful for identifying technical artifacts

DRAFT

such as batch effects (Wolf et al., 2018). It also provides methods for plotting features over known covariates, such as plates.

- Downstream analysis. For experiments focused on profiling non-dynamic responses, like to a small molecule library, `scmorph` integrates functions to perform hit calling from single-cell profiles using the KS statistic of single-cells to controls in PCA space. For dynamic systems like differentiating cells, `scmorph` incorporates differential trajectory inference modelling via `slingshot` and `condiments` through the `rpy2` translation layer (Roux de Bézieux et al., 2024; Street et al., 2018).

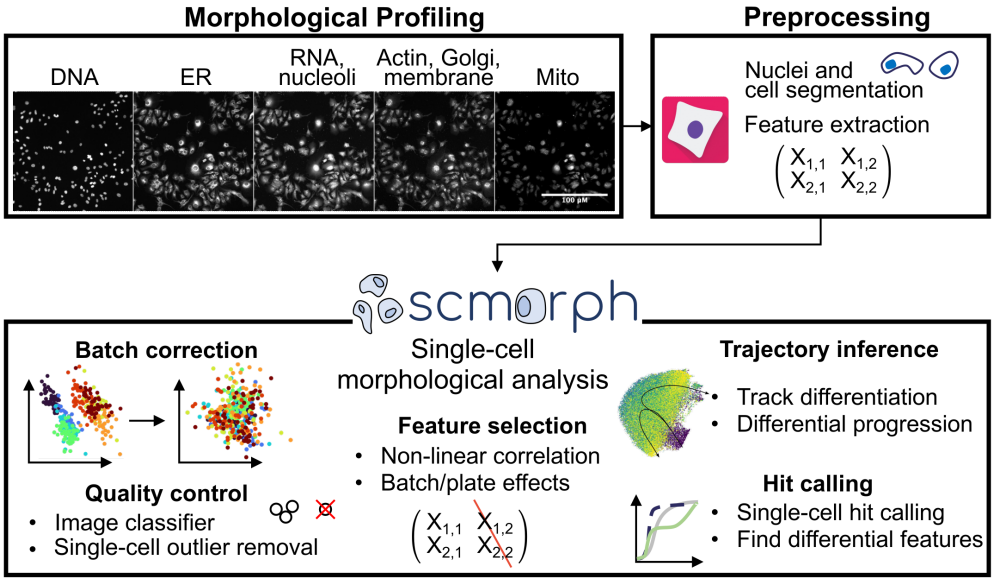

Figure 1: Overview of `scmorph` functionality. `scmorph` processes profiles generated with software such as CellProfiler to facilitate downstream analysis by performing batch correction, image- and single-cell QC, feature selection, hit calling and trajectory inference. All methods are built with single-cell analysis in mind and do not require subsampling.

In contrast to the commonly used `pycytominer` package (**?**) and `SPACe`(Stossi et al., 2024), `scmorph` offers (i) interpretable batch correction techniques compatible with single-cell profiles, (ii) enhanced feature selection with an adapted Chatterjee correlation coefficient or Kruskal-Wallis test (Kruskal & Wallis, 1952; Lin & Han, 2021), (iii) lineage trajectory inference (Bézieux et al., 2021; Street et al., 2018), and (iv) the option to analyse multi-nucleated cells. Compared to `pycytominer`, `scmorph` also performs single-cell based hit calling. And unlike `SPACe`, `scmorph` is agnostic to the segmentation and feature extraction methods used upstream and therefore compatible with CellProfiler. `scmorph` also benefits from improvements of `AnnData` and `scanpy`, which `scmorph` is based on, which, going forward, will enable out-of-core computation crucial to big data analysis (Virshup et al., 2024; Wolf et al., 2018).

Already, `scmorph` has been used to quality control morphological profiling experiments involving differentiating liver cells (**?**). `scmorph` is also involved in three projects involving small compound and microRNA perturbations in the domains of drug discovery and fundamental research, spanning datasets of >20M cells. Going forward, we envision that `scmorph` will enable analysis of complex and large morphological profiling experiments.

Acknowledgements

JW and HW are funded by an MRC Unit Award.

DRAFT

References

Bézieux, H. R. de, Berge, K. V. den, Street, K., & Dudoit, S. (2021). *Trajectory inference across multiple conditions with condiments: Differential topology, progression, differentiation, and expression* (p. 2021.03.09.433671). bioRxiv. https://doi.org/10.1101/2021.03.09.433671

Cole, M. B., Risso, D., Wagner, A., DeTomaso, D., Ngai, J., Purdom, E., Dudoit, S., & Yosef, N. (2019). Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-Seq. *Cell Systems*, *8*(4), 315–328.e8. https://doi.org/10.1016/j.cels.2019.03.010

Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, *47*(260), 583–621. https://doi.org/10.1080/01621459.1952.10483441

Li, Z., Zhao, Y., Hu, X., Botta, N., Ionescu, C., & Chen, G. H. (2022). ECOD: Unsupervised Outlier Detection Using Empirical Cumulative Distribution Functions. *IEEE Transactions on Knowledge and Data Engineering*, 1–1. https://doi.org/10.1109/tkde.2022.3159580

Lin, Z., & Han, F. (2021). On boosting the power of Chatterjee's rank correlation. *arXiv*. https://doi.org/10.48550/arxiv.2108.06828

Omta, W. A., van Heesbeen, R. G., Pagliero, R. J., van der Velden, L. M., & Lelieveld, D. (2016). HC StratoMineR: A web-based tool for the rapid analysis of high-content datasets. *ASSAY and Drug Development Technologies*, *14*(8), 439–452. https://doi.org/gt2g9c

Roux de Bézieux, H., Van den Berge, K., Street, K., & Dudoit, S. (2024). Trajectory inference across multiple conditions with condiments. *Nature Communications*, *15*(1), 833. https://doi.org/10.1038/s41467-024-44823-0

Stirling, D. R., Swain-Bowden, M. J., Lucas, A. M., Carpenter, A. E., Cimini, B. A., & Goodman, A. (2021). CellProfiler 4: Improvements in speed, utility and usability. *BMC Bioinformatics*, *22*(1), 433. https://doi.org/10.1186/s12859-021-04344-9

Stossi, F., Singh, P. K., Marini, M., Safari, K., Szafran, A. T., Rivera Tostado, A., Candler, C. D., Mancini, M. G., Mosa, E. A., Bolt, M. J., Labate, D., & Mancini, M. A. (2024). SPACe: An open-source, single-cell analysis of Cell Painting data. *Nature Communications*, *15*(1), 10170. https://doi.org/10.1038/s41467-024-54264-4

Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., Purdom, E., & Dudoit, S. (2018). Slingshot: Cell lineage and pseudotime inference for single-cell transcriptomics. *BMC Genomics*, *19*(1), 477. https://doi.org/10.1186/s12864-018-4772-0

Virshup, I., Rybakov, S., Theis, F. J., Angerer, P., & Wolf, F. A. (2024). Anndata: Access and store annotated data matrices. *Journal of Open Source Software*, *9*(101), 4371. https://doi.org/10.21105/joss.04371

Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, *19*(1), 15. https://doi.org/10.1186/s13059-017-1382-0

DRAFT