https://doi.org/10.1038/s41587-023-01733-8

The scverse project provides a computational ecosystem for single-cell omics data analysis

Single-cell omics technologies have enabled the creation of comprehensive cell atlases across tissues and species and delivered key insights into the biological mechanisms underlying development, homeostasis and disease. Deriving such insights relied on the development of a multitude of computational tools and data structures[1]. However, the explosive growth of tools led to incompatibilities in data formats, application programming interfaces (APIs) and user interfaces, posing growing challenges for both tool users and developers. Here, we present scverse (https://scverse.org), a multi-institution open-source software project to address storage and analysis needs of single-cell profiling data. Scverse will help to support an omics analysis ecosystem in Python by bringing together tools and analysts into a robust community. To deliver a sustainable solution, scverse provides well-maintained core functionality including interoperable data formats and community structures. Scverse is an open community initiated by the developers of some of the most popular tools in the Python single-cell analysis ecosystem, including scanpy[2], scvi-tools[3] and muon[4], which are successfully used in many downstream methods constructing various single-cell reference atlases of tissues, organs and organisms[5].

The rapid growth, broad adoption and substantial impact of single-cell genomics depends on computational tools, as reflected in the accelerated growth of the software ecosystem[1]. With this growth, common obstacles of academic open-source software start to arise, including scattered and overlapping functionality across tools, poor discoverability and documentation, and lack of testing and continuous integration for some tools. These issues are exacerbated by an incentive structure in academia that rewards novelty over maintenance of essential infrastructure[6]. Moreover, enhancement and maintenance of widely used academic software is often limited to the original authors and a small number of direct collaborators, which can severely limit their continued reliability. However, to ensure that all newly developed tools can interact with each other, a certain amount of centralization is required, especially concerning data structures.

At the core of the scverse ecosystem are the AnnData and MuData[4] classes for storing and handling unimodal and multimodal high dimensional data, respectively. These data

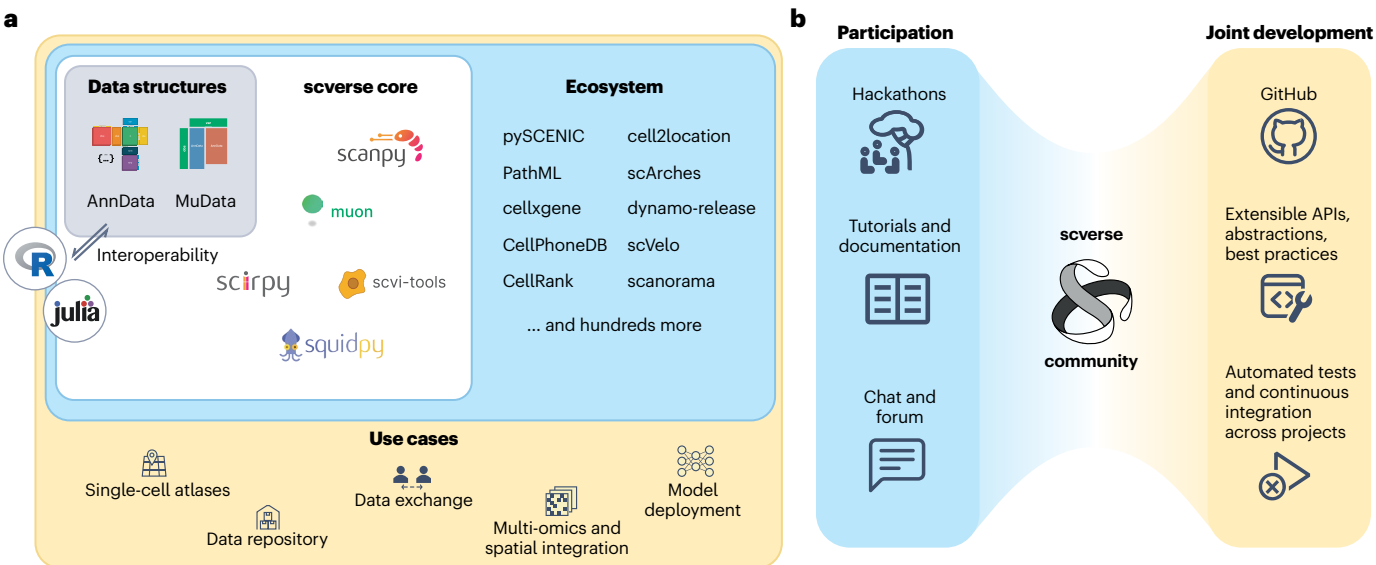

Fig. 1 | Scverse. a, The scverse ecosystem. Central to the scverse ecosystem are the AnnData and MuData data structures, which are containers for high-dimensional data with storage formats designed for interoperability with multiple languages and tools. The scverse core tools provide foundational support for single-cell data modalities and analysis paradigms. They are selected frameworks that scverse developers can build on and are guaranteed to be further maintained. On top of this core tooling, an ecosystem containing hundreds of dependent packages (Supplementary Tables 1 and 2) developed by members of the global research community has emerged. As a whole, this ecosystem has already provided functionality for powering a variety of use cases. **b**, Modes of participation for users and developers in scverse. Scverse is an open ecosystem, aiming to build resources and community structures that serve the needs of both analysts and ecosystem developers. Analysts will be supported with regular workshops for scverse core and ecosystem tools, comprehensive tutorials and community platforms for help and discussion. Developer-facing community structures include regular hackathons for networking and getting outside support, workshops for advertising their packages, and shared development infrastructure provided by the scverse core team.

structures, which resemble Bioconductor's SingleCellExperiment and MultiAssayExperiment classes[7], respectively, keep annotations and learned representations (such as a *k*-nearest-neighbors topology or feature attribution loadings) tightly associated with the measured data, facilitating an iterative approach to understanding the data. Datasets are stored using the language-agnostic high-performance array storage formats HDF5 and Zarr – allowing scalable out-of-core access, whereby datasets that do not fit into memory are efficiently accessed from local or cloud storage. Standardized storage has enabled use of AnnData and MuData by single-cell tools built in different programming languages, such as vitessce (JavaScript)[8], zellkonverter (R)[9], MuDataSeurat (R)[4], scVI.jl (Julia) and anndata-rs (Rust). Data interoperability is a high priority for the scverse team and will be furthered with improved metadata handling, further formalization of data schemas, and collaboration with standards like OME-NGFF[10].

On top of these data structures, scverse provides toolkits for fundamental tasks in single-cell genomics. Scanpy, a framework for single-cell data analysis in Python, is complemented by muon for integrating data from multiple modalities, scirpy[11] for T and B cell receptor repertoire analysis, squidpy[12] for spatial omics data analysis, and scvi-tools for building and deploying deep learning models[3] (Fig. 1a). The mutual development of these scverse core tools allows efficient progress with fewer overlaps, and more support and continuity from developers across the core packages. All scverse core tools regularly validate new development versions through automated tests, ensuring interoperability of new releases.

Beyond interoperability, performance and robustness have practical and tangible repercussions for users who work with finite computational resources and with datasets that are increasingly scaling to millions of cells and more. Hence, tools in scverse primarily build on Python and its scientific software stack (https://scientific-python.org), allowing support for sparse data representations and operations[13,14], just-in-time compilation[15] and lazy operations on large arrays[16,17], as well as deep learning libraries such as PyTorch[18] and Jax[19]. These advantages have led to broad adoption of scverse tools for data-intensive and computationally complex tasks such as atlas creation[5], infrastructure for large-scale methods[20–22] and hardware performance benchmarks.

Scverse is a community project, and as such thrives through joint development and communication (Fig. 1b). The broader community has developed tools built on scverse packages to address key challenges in data analysis pipelines. For example, cellxgene[23] and vitessce[8] enable the interactive visualization of large heterogeneous data, sfaira[24] orchestrates the management and harmonization of datasets, and PathML[25] tackles the analysis of large-scale pathology data. These packages, along with all others that depend on scverse core tools and data structures, are part of the scverse ecosystem.

Ecosystem packages that follow development best practices (for example, continuous testing, documentation, and availability through standard distribution tools) are highlighted on scverse.org to increase their visibility and encourage creation of high-quality tools. This set of packages is curated through a public GitHub repository (https://github.com/scverse/ecosystem-packages) that contains detailed requirements and instructions for adding new packages. Inspired by the documentation efforts of Bioconductor[26], we are launching a community effort to create a curated set of learning materials on https://scverse.org using materials from across the scverse ecosystem.

We are committed to public platforms and to making it easy to participate in the scverse community. Our work is done in public repositories on GitHub, and we discuss development in an open Zulip chat, as well as a Discourse forum (https://discourse.scverse.org) to facilitate communication between users and developers of both core and ecosystem packages. These open platforms explicitly bypass the barrier of individual research groups, with the goal of making more useful software available to more people. Like Bioconductor[7], we see our community as composed of overlapping groups of developers and users. For method developers, we provide communication channels and facilitate the creation of ecosystem packages by providing a best-practice code template (https://github.com/scverse/cookiecutter-scverse). For both users and developers, we aim to provide opportunities to participate in our community with workshops, hackathons and other hands-on events to foster tighter interactions between users and developers (Fig. 1b). In contrast to Bioconductor, we do not plan on expanding centralized infrastructure to include software distribution and servers for testing[26]. Instead, our packages and template rely on services like GitHub Actions, PyPI and Read the Docs that are commonly used in open-source software development.

As more powerful experimental techniques emerge, leading to increasingly large and complex datasets, it is crucial for the analysis software to keep up with this progress. By providing a core of interoperable, scalable and user-friendly tools for an ever-growing ecosystem of cutting-edge methods to build on, scverse is well set up to support the next decade of discoveries in single-cell genomics.

Code availability

The source code for the scverse core tools is publicly available at https://github.com/scverse. Code for determining the number of dependent packages and repositories is available from https://gist.github.com/ivirshup/4bff45c0a8d38b97c5289c8b2407dfcf.

Isaac Virshup[1,32] ✉, **Danila Bredikhin**[2,3,4,32] ✉, **Lukas Heumos**[1,5,6,32] ✉, **Giovanni Palla**[1,6,32], **Gregor Sturm**[7,32], **Adam Gayoso**[8,32], **Ilia Kats**[3], **Mikaela Koutrouli**[9], **Scverse Community***, **Bonnie Berger**[10,11,12], **Dana Pe'er**[13,14], **Aviv Regev**[15], **Sarah A. Teichmann**[16,17], **Francesca Finotello**[18,19,33], **F. Alexander Wolf**[1,20,33], **Nir Yosef**[8,21,22,23,33], **Oliver Stegle**[2,3,16,24,33] & **Fabian J. Theis**[1,6,16,25,33]

[1]Computational Health Center, Helmholtz Center Munich, Neuherberg, Germany. [2]European Molecular Biology Laboratory (EMBL), Genome Biology Unit, Heidelberg, Germany. [3]Division of Computational Genomics and Systems Genetics, German Cancer Research Center (DKFZ), Heidelberg, Germany. [4]Collaboration for joint PhD degree between EMBL and Heidelberg University, Faculty of Biosciences, Heidelberg, Germany. [5]Institute of Lung Health and Immunity and Comprehensive Pneumology Center with the CPC-M bioArchive; Helmholtz Zentrum Munich, Member of the German Center for Lung Research (DZL), Munich, Germany. [6]School of Life Sciences, Technical University of Munich, Munich, Germany. [7]Biocenter, Institute of Bioinformatics, Medical University of Innsbruck, Innsbruck, Austria. [8]Center for Computational Biology, University of California, Berkeley, Berkeley, CA, USA. [9]Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [10]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology,

Cambridge, MA, USA. [11]Harvard-MIT Health Sciences and Technology Program, Cambridge, MA, USA. [12]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [13]Computational and Systems Biology Program, Sloan Kettering Institute, New York, NY, USA. [14]Howard Hughes Medical Institute, Chevy Chase, MD, USA. [15]Genentech Research and Early Development, Genentech Inc, South San Francisco, CA, USA. [16]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, UK. [17]Cavendish Laboratory, University of Cambridge, Cambridge, UK. [18]Institute of Molecular Biology, University of Innsbruck, Innsbruck, Austria. [19]Digital Science Center (DiSC), University of Innsbruck, Innsbruck, Austria. [20]Lamin Labs, Munich, Germany. [21]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA, USA. [22]Ragon Institute of MGH, MIT and Harvard, Cambridge, MA, USA. [23]Chan Zuckerberg Biohub, San Francisco, CA, USA. [24]Faculty of Biosciences, Heidelberg University, Heidelberg, Germany. [25]School of Computation, Information and Technology, Technical University of Munich, Munich, Germany. [32]These authors contributed equally: Isaac Virshup, Danila Bredikhin, Lukas Heumos, Giovanni Palla, Gregor Sturm, Adam Gayoso. [33]These authors jointly supervised this work: Francesca Finotello, F. Alexander Wolf, Nir Yosef, Oliver Stegle, Fabian J. Theis. *A list of authors and their affiliations appears at the end of the paper.
✉e-mail: steering-council@scverse.org

Published online: 10 April 2023

References

1. Zappia, L. & Theis, F. J. *Genome Biol.* **22**, 301 (2021).
2. Wolf, F. A., Angerer, P. & Theis, F. J. *Genome Biol.* **19**, 15 (2018).
3. Gayoso, A. et al. *Nat. Biotechnol.* **40**, 163–166 (2022).
4. Bredikhin, D., Kats, I. & Stegle, O. *Genome Biol.* **23**, 42 (2022).
5. Liu, Z. & Zhang, Z. *Science* **376**, 695–696 (2022).
6. Woolston, C. Why science needs more research software engineers. *Nature* https://doi.org/10.1038/d41586-022-01516-2 (2022).
7. Huber, W. et al. *Nat. Methods* **12**, 115–121 (2015).
8. Keller, M. S. et al. Preprint at https://doi.org/10.31219/osf.io/y8thv (2021).
9. Zappia, L. & Lun, A. zellkonverter: Conversion Between scRNA-seq Objects. R package version 1.8.0, https://github.com/theislab/zellkonverter (2022).
10. Moore, J. et al. *Nat. Methods* **18**, 1496–1498 (2021).
11. Sturm, G. et al. *Bioinformatics* **36**, 4817–4818 (2020).
12. Palla, G. et al. *Nat. Methods* **19**, 171–178 (2022).
13. Virtanen, P. et al. *Nat. Methods* **17**, 261–272 (2020).
14. Harris, C. R. et al. *Nature* **585**, 357–362 (2020).
15. Lam, S. K., Pitrou, A. & Seibert, S. Numba: a LLVM-based Python JIT compiler. in *Proc. Second Workshop on the LLVM Compiler Infrastructure in HPC* 1–6 (Association for Computing Machinery, 2015).
16. Rocklin, M. Dask: parallel computation with blocked algorithms and task scheduling. in *Proc. 14th Python in Science Conference* (SciPy, 2015); https://doi.org/10.25080/majora-7b98e3ed-013
17. Hoyer, S. & Hamman, J. J. *J. Open Res. Softw.* **5**, 10 (2017).
18. Paszke, A. et al. in *Advances in Neural Information Processing Systems* vol. 32 (eds. Wallach, H. et al.) 8024–8035 (Curran Associates, 2019).
19. Bradbury, J. et al. *JAX: Composable Transformations of Python+NumPy Programs*, http://github.com/google/jax (2018).
20. Lance, C. et al. Preprint at https://doi.org/10.1101/2022.04.11.487796 (2022).
21. Li, B. et al. *Nat. Methods* **19**, 662–670 (2022).
22. Luecken, M. D. et al. *Nat. Methods* **19**, 41–50 (2022).
23. Megill, C. et al. Preprint at https://doi.org/10.1101/2021.04.05.438318 (2021).
24. Fischer, D. S. et al. *Genome Biol.* **22**, 248 (2021).
25. Rosenthal, J. et al. *Mol. Cancer Res.* **20**, 202–206 (2022).
26. Gentleman, R. C. et al. *Genome Biol.* **5**, R80 (2004).

Acknowledgements

The authors would like to thank all scverse contributors who opened issues, contributed code or answered questions on scverse forums to guide novice and advanced users. D.B. acknowledges funding by the EMBL International PhD Programme and Darwin Trust Fellowship. G.S. was supported by a DOC fellowship from the Austrian Academy of Sciences. F.F. was supported by the Austrian Science Fund (FWF) (T 974-B30). G.P. is supported by the Helmholtz Association under the joint research school Munich School for Data Science.

Author contributions

B.B., D.P., A.R. and S.A.T. are members of the scverse Advisory Committee. F.F., F.A.W., N.Y., O.S. and F.J.T. are members of the scverse Management Committee. Members of the scverse community are listed in alphabetical order.

Competing interests

F.J.T. consults for Immunai Inc., Singularity Bio B.V., CytoReason Ltd and Omniscope Ltd, and has ownership interest in Dermagnostix GmbH and Cellarity. A.R. is a co-founder of and equity holder in Celsius Therapeutics, an equity holder in Immunitas, and until 31 July 2020 was a scientific advisory board member of Thermo Fisher Scientific, Syros Pharmaceuticals, Neogene Therapeutics and Asimov. From 1 August 2020, A.R. has been an employee of Genentech and a member of the Roche Group, and has equity in Roche. M.D.L is a part-time contractor for the Chan Zuckerberg Initiative. V.B. reports being employed by and having ownership interest in Cellarity. G.E. has been an employee of Genentech since 4 April 2022. R.L. has been an employee of Genentech since 31 August 2021. F.A.W. holds equity in Lamin Labs, Cellarity, Retro Biosciences and Doloromics. N.Y. is an advisor to and/or has equity in Cellarity, Celsius Therapeutics and Rheos Medicines. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-023-01733-8.

Peer review information *Nature Methods* thanks Martin Hemberg and Wolfgang Huber for their contribution to the peer review of this work.

Scverse Community

Philipp Angerer[26], Volker Bergen[1,26], Pierre Boyeau[21], Maren Büttner[1,27], Gokcen Eraslan[15], David Fischer[1,6], Max Frank[2], Justin Hong[8,21], Michal Klein[1], Marius Lange[1,25], Romain Lopez[15,28], Mohammad Lotfollahi[1], Malte D. Luecken[1], Fidel Ramirez[29], Jeffrey Regier[30], Sergei Rybakov[1,25], Anna C. Schaar[1,25], Valeh Valiollah Pour Amiri[8,21], Philipp Weiler[1,25] & Galen Xing[8,31]

[26]Cellarity, Somerville, MA, USA. [27]Genomics and Immunoregulation, Life & Medical Sciences (LIMES) Institute, University of Bonn, Bonn, Germany. [28]Department of Genetics, Stanford University, Stanford, CA, USA. [29]Department of Global Computational Biology and Digital Sciences, Boehringer Ingelheim Pharma GmbH & Co. KG, Biberach an der Riss, Germany. [30]Department of Statistics, University of Michigan, Ann Arbor, MI, USA. [31]Gladstone-UCSF Institute of Genomic Immunology, San Francisco, CA, USA.