The Journal of Open Source Software

scphylo-tools: A Python toolkit for single-cell tumor phylogenetic analysis

Farid Rashidi Mehrabadi [1]

1 Laboratory of Human Carcinogenesis, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, United States of America

DOI: 10.21105/joss.10589

Software

- Review
- Repository
- Archive

Editor: William Gearty

Reviewers:

- @rachelicr
- @shrishar

Submitted: 11 January 2026
Published: 01 September 2026

Summary

`scphylo-tools` is a Python library designed to unify single-cell tumor phylogeny inference methods. It addresses the lack of standardization in the field by providing a cohesive interface for data processing, tree reconstruction, visualization, and benchmarking. By streamlining these tasks, `scphylo-tools` empowers researchers with limited programming expertise to easily install and utilize complex inference methods, accelerates algorithm development, and ensures reproducible benchmarking of tumor phylogeny reconstruction methods.

Statement of Need

Cancer is a dynamic evolutionary process driven by the acquisition of somatic mutations and the competitive selection of clonal populations. Single-Cell Sequencing (SCS) technologies have enabled the profiling of genomic alterations at cellular resolution, offering unprecedented insights into intratumor heterogeneity (Navin et al., 2011; Wang et al., 2014). Elucidating the phylogenetic relationships between tumor cells is essential for understanding metastasis (Leung et al., 2017; Roper et al., 2020), drug resistance (Gopalan et al., 2021; Gruen et al., 2023; Kim & Simon, 2014), and clonal dynamics (Hirsch et al., 2025; Laks et al., 2019; Liu et al., 2025).

Reconstructing the evolutionary history of a tumor from SCS data presents unique computational challenges. These datasets are plagued by high noise levels, including Allele Drop-Out (ADO), false positives, missing data, and doublet artifacts (Malikić et al., 2021; Rashidi Mehrabadi, 2022). Consequently, a diverse array of computational tools has been developed to address these challenges, including stochastic methods like SCITE (Jahn et al., 2016), infSCITE (Kuipers et al., 2017), OncoNEM (Ross & Markowetz, 2016), SiFit (Zafar et al., 2017), and SiCloneFit (Zafar et al., 2019); combinatorial approaches such as PhISCS (Malikic et al., 2019), PhISCS-BnB (Sadeqi Azer et al., 2020), SPhyR (El-Kebir, 2018), ScisTree (Wu, 2019), gpps (Ciccolella, Soto Gomez, et al., 2020), and SASC (Ciccolella, Ricketts, et al., 2020); and specialized methods including HUNTRESS (Kızılkale et al., 2022), SCIPhI (Singer et al., 2018), and Scelestial (Foroughmand-Araabi et al., 2022).

However, the software landscape remains highly fragmented. Existing methods typically function as standalone binaries or scripts with inconsistent input/output formats, rendering comparative analysis difficult. Researchers attempting to utilize these tools face a laborious process of installation, data conversion, and script wrapping. Furthermore, integrating these tools into modern Python-based environments (e.g., alongside SCANPY (F. A. Wolf et al., 2018) or Biopython (Cock et al., 2009)) often requires custom development. `scphylo-tools` addresses these challenges by providing a single, cohesive Python API that wraps independently developed standalone applications, including Trisicell (Mehrabadi et al., 2021), SCITE, PhISCS, and HUNTRESS.

The target audience includes computational biologists, bioinformaticians, and cancer researchers who need to:

- Compare multiple phylogeny inference methods on the same dataset
- Benchmark new algorithms against established methods
- Integrate tumor phylogeny analysis into existing Python workflows
- Access curated single-cell cancer datasets for research

State of the Field

Several tools exist for tumor phylogeny inference, but none provide the unified framework that `scphylo-tools` offers. Individual tools like SCITE, PhISCS, and HUNTRESS each require separate installation, different input formats, and produce outputs that are not directly comparable. Workflow frameworks like Snakemake or Nextflow could orchestrate these tools, but would require significant custom development for format conversion and result standardization.

`scphylo-tools` distinguishes itself by:

1. **Unified Interface**: A consistent Python API across all wrapped methods, eliminating format conversion overhead
2. **Comprehensive Metrics**: Built-in implementation of specialized tumor tree comparison metrics (MLTD (Karpov et al., 2019), CASet/DISC (DiNardo et al., 2019), MP3 (Ciccolella, Bernardini, et al., 2020))
3. **Curated Datasets**: Direct access to published SCS datasets from landmark cancer evolution studies (Gawad et al., 2014; Hou et al., 2012; Leung et al., 2017; Morita et al., 2020; Wang et al., 2014; Y. Wolf et al., 2019)
4. **Ecosystem Integration**: Seamless integration with the Python scientific stack (NumPy, NetworkX, Matplotlib)

Software Design

`scphylo-tools` follows a modular architecture inspired by SCANPY. Modules mirror the analysis workflow: `io` (genotype and tree I/O, VCF export), `pp` (filtering of binary and read-count matrices, node collapsing, subsampling), `tl` (solvers, tree-comparison scores, consensus trees, partition functions, ancestral-state reconstruction), `pl` (tree rendering and genotype heatmaps), `ul` (utilities), and `datasets`. A `scphylo` console script exposes the same operations to shell users and workflow managers.

A single data model spans the package: genotype matrices are pandas `DataFrames` encoding presence, absence, and missing entries; read counts are AnnData objects; trees are NetworkX `DiGraph` objects. Over twenty solvers share the call shape `scp.tl.<method>(df_input, ...)` and return a conflict-free matrix, so switching methods is a one-line change; serialization, execution, parsing, and logging are internal. The ten tree-comparison metrics take the same two arguments, reducing a benchmark to a short loop.

To remove the main installation barrier, performance-critical C++ codebases (SCITE, ScisTree, MLTD, and the partition-function sampler) are vendored, compiled as Cython extensions, and shipped as prebuilt wheels. Non-redistributable tools stay optional, located by one resolution layer (explicit path, `SCPHYLO_TOOLS_DIR`, `scphylo.settings.tools_dir`, `PATH`) raising typed exceptions that name the missing artifact and how to supply it. Heavy optional dependencies (Gurobi, rpy2, graph-tool, Graphviz) load lazily.

Adding a method requires one module under `tl/solver` and no changes elsewhere, as with the existing wrappers. For validation, `datasets` retrieves published SCS datasets by identifier and the simulation engine generates ground-truth genotypes under configurable false-positive, false-negative, missing-value, and doublet noise.

Research Impact Statement

`scphylo-tools` has enabled published research in three ways.

First, it provided the interface through which Trisicell and ScisTree were applied to build tumor phylogenies. These reconstructions underpinned the analysis of clonal sublines that uncovered heterogeneity-driven immunotherapy resistance in melanoma (Gruen et al., 2023), the comparison of genomic and epigenomic evolutionary trajectories in long-read sequencing of single cell-derived subclones (Liu et al., 2025), the stochastic modeling that quantified the non-genomic contribution to subclone evolution (Hirsch et al., 2025), and the recovery of metastatic migration histories from single-cell methylation sequencing (Liu et al., 2023).

Second, it supported the identification of expressed mutation profiles in single cells, which revealed subclonal expansion patterns and the therapeutic impact of intratumor heterogeneity (Mehrabadi et al., 2021).

Third, its uniform solver interface, simulation engine, and tree-comparison metrics were used to benchmark competing inference methods on common inputs during the evaluation of HUNTRESS (Kızılkale et al., 2022).

The package therefore serves as both the analysis layer for biological studies and the comparison harness for methods development. It is distributed on PyPI with source on GitHub and documentation on Read the Docs.

AI Usage Disclosure

No generative AI tools were used in the development of the `scphylo-tools` software, the writing of this manuscript, or the preparation of supporting documentation.

Acknowledgements

We acknowledge the contributions of the open-source community and the authors of the underlying algorithms whose work is wrapped within this toolkit. This work was supported by the Intramural Research Program of the National Institutes of Health, National Cancer Institute.

References

Ciccolella, S., Bernardini, G., Denti, L., Bonizzoni, P., Previtali, M., & Della Vedova, G. (2020). Triplet-based similarity score for fully multilabeled trees with poly-occurring labels. *Bioinformatics*, *37*(2), 178–184. https://doi.org/10.1093/bioinformatics/btaa676

Ciccolella, S., Ricketts, C., Soto Gomez, M., Patterson, M., Silverbush, D., Bonizzoni, P., Hajirasouliha, I., & Della Vedova, G. (2020). Inferring cancer progression from Single-Cell Sequencing while allowing mutation losses. *Bioinformatics*, *37*(3), 326–333. https://doi.org/10.1093/bioinformatics/btaa722

Ciccolella, S., Soto Gomez, M., Patterson, M. D., Della Vedova, G., Hajirasouliha, I., & Bonizzoni, P. (2020). gpps: an ILP-based approach for inferring cancer progression with mutation losses from single cell data. *BMC Bioinformatics*, *21*(S1). https://doi.org/10.1186/s12859-020-03736-7

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & Hoon, M. J. L. de. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

DiNardo, Z., Tomlinson, K., Ritz, A., & Oesper, L. (2019). Distance measures for tumor evolutionary trees. *Bioinformatics*, *36*(7), 2090–2097. https://doi.org/10.1093/bioinformatics/btz869

El-Kebir, M. (2018). SPhyR: tumor phylogeny estimation from single-cell sequencing data under loss and error. *Bioinformatics*, *34*(17), i671–i679. https://doi.org/10.1093/bioinformatics/bty589

Foroughmand-Araabi, M.-H., Goliaei, S., & McHardy, A. C. (2022). Scelestial: Fast and accurate single-cell lineage tree inference based on a Steiner tree approximation algorithm. *PLOS Computational Biology*, *18*(8), e1009100. https://doi.org/10.1371/journal.pcbi.1009100

Gawad, C., Koh, W., & Quake, S. R. (2014). Dissecting the clonal origins of childhood acute lymphoblastic leukemia by single-cell genomics. *Proceedings of the National Academy of Sciences*, *111*(50), 17947–17952. https://doi.org/10.1073/pnas.1420822111

Gopalan, V., Singh, A., Rashidi Mehrabadi, F., Wang, L., Ruppin, E., Arda, H. E., & Hannenhalli, S. (2021). A Transcriptionally Distinct Subpopulation of Healthy Acinar Cells Exhibit Features of Pancreatic Progenitors and PDAC. *Cancer Research*, *81*(15), 3958–3970. https://doi.org/10.1158/0008-5472.can-21-0427

Gruen, C., Yang, H. H., Sassano, A., Wu, E., Gopalan, V., Marie, K. L., Castro, A., Mehrabadi, F. R., Wu, C. H., Church, I., Needle, G. A., Smith, C., Chin, S., Ebersole, J., Marcelus, C., Fon, A., Liu, H., Malikic, S., Sahinalp, C., … Pérez-Guijarro, E. (2023). *Melanoma clonal subline analysis uncovers heterogeneity-driven immunotherapy resistance mechanisms*. https://doi.org/10.1101/2023.04.03.535074

Hirsch, M. G., Pal, S., Rashidi Mehrabadi, F., Malikic, S., Gruen, C., Sassano, A., Pérez-Guijarro, E., Merlino, G., Sahinalp, S. C., Molloy, E. K., Day, C.-P., & Przytycka, T. M. (2025). Stochastic modeling of single-cell gene expression adaptation reveals non-genomic contribution to evolution of tumor subclones. *Cell Systems*, *16*(1), 101156. https://doi.org/10.1016/j.cels.2024.11.013

Hou, Y., Song, L., Zhu, P., Zhang, B., Tao, Y., Xu, X., Li, F., Wu, K., Liang, J., Shao, D., Wu, H., Ye, X., Ye, C., Wu, R., Jian, M., Chen, Y., Xie, W., Zhang, R., Chen, L., … Wang, J. (2012). Single-Cell Exome Sequencing and Monoclonal Evolution of a JAK2-Negative Myeloproliferative Neoplasm. *Cell*, *148*(5), 873–885. https://doi.org/10.1016/j.cell.2012.02.028

Jahn, K., Kuipers, J., & Beerenwinkel, N. (2016). Tree inference for single-cell data. *Genome Biology*, *17*(1). https://doi.org/10.1186/s13059-016-0936-x

Karpov, N., Malikic, S., Rahman, Md. K., & Sahinalp, S. C. (2019). A multi-labeled tree dissimilarity measure for comparing "clonal trees" of tumor progression. *Algorithms for Molecular Biology*, *14*(1). https://doi.org/10.1186/s13015-019-0152-9

Kim, K. I., & Simon, R. (2014). Using single cell sequencing data to model the evolutionary history of a tumor. *BMC Bioinformatics*, *15*(1). https://doi.org/10.1186/1471-2105-15-27

Kızılkale, C., Rashidi Mehrabadi, F., Sadeqi Azer, E., Pérez-Guijarro, E., Marie, K. L., Lee, M. P., Day, C.-P., Merlino, G., Ergün, F., Buluç, A., Sahinalp, S. C., & Malikić, S. (2022). Fast intratumor heterogeneity inference from single-cell sequencing data. *Nature Computational Science*, *2*(9), 577–583. https://doi.org/10.1038/s43588-022-00298-x

Kuipers, J., Jahn, K., Raphael, B. J., & Beerenwinkel, N. (2017). Single-cell sequencing data reveal widespread recurrence and loss of mutational hits in the life histories of tumors. *Genome Research*, *27*(11), 1885–1894. https://doi.org/10.1101/gr.220707.117

Laks, E., McPherson, A., Zahn, H., Lai, D., Steif, A., Brimhall, J., Biele, J., Wang, B., Masud, T., Ting, J., Grewal, D., Nielsen, C., Leung, S., Bojilova, V., Smith, M., Golovko,

O., Poon, S., Eirew, P., Kabeer, F., … Aparicio, S. (2019). Clonal Decomposition and DNA Replication States Defined by Scaled Single-Cell Genome Sequencing. *Cell*, *179*(5), 1207–1221.e22. https://doi.org/10.1016/j.cell.2019.10.026

Leung, M. L., Davis, A., Gao, R., Casasent, A., Wang, Y., Sei, E., Vilar, E., Maru, D., Kopetz, S., & Navin, N. E. (2017). Single-cell DNA sequencing reveals a late-dissemination model in metastatic colorectal cancer. *Genome Research*, *27*(8), 1287–1299. https://doi.org/10.1101/gr.209973.116

Liu, Y., Goretsky, A., Keskus, A. G., Malikic, S., Ahmad, T., Gertz, E. M., Mehrabadi, F. R., Kelly, M., Hernandez, M., Seibert, C., Caravaca, J. M., Kline, K., Zhao, Y., Wu, Y., Shrestha, B., Tran, B., Ghosh, A., Cui, X., Sassano, A., … Kolmogorov, M. (2025). *Long-read sequencing of single cell-derived melanoma subclones reveals divergent and parallel genomic and epigenomic evolutionary trajectories*. https://doi.org/10.1101/2025.08.28.672865

Liu, Y., Li, X. C., Rashidi Mehrabadi, F., Schäffer, A. A., Pratt, D., Crawford, D. R., Malikić, S., Molloy, E. K., Gopalan, V., Mount, S. M., Ruppin, E., Aldape, K. D., & Sahinalp, S. C. (2023). Single-cell methylation sequencing data reveal succinct metastatic migration histories and tumor progression models. *Genome Research*, gr.277608.122. https://doi.org/10.1101/gr.277608.122

Malikic, S., Mehrabadi, F. R., Ciccolella, S., Rahman, Md. K., Ricketts, C., Haghshenas, E., Seidman, D., Hach, F., Hajirasouliha, I., & Sahinalp, S. C. (2019). PhISCS: a combinatorial approach for subperfect tumor phylogeny reconstruction via integrative use of single-cell and bulk sequencing data. *Genome Research*, *29*(11), 1860–1877. https://doi.org/10.1101/gr.234435.118

Malikić, S., Mehrabadi, F. R., Azer, E. S., Ebrahimabadi, M. H., & Sahinalp, S. C. (2021). Studying the History of Tumor Evolution from Single-Cell Sequencing Data by Exploring the Space of Binary Matrices. *Journal of Computational Biology*, *28*(9), 857–879. https://doi.org/10.1089/cmb.2020.0595

Mehrabadi, F. R., Marie, K. L., Pérez-Guijarro, E., Malikić, S., Azer, E. S., Yang, H. H., Kızılkale, C., Gruen, C., Robinson, W., Liu, H., Kelly, M. C., Marcelus, C., Burkett, S., Buluç, A., Ergün, F., Lee, M. P., Merlino, G., Day, C.-P., & Sahinalp, S. C. (2021). *Profiles of expressed mutations in single cells reveal subclonal expansion patterns and therapeutic impact of intratumor heterogeneity*. https://doi.org/10.1101/2021.03.26.437185

Morita, K., Wang, F., Jahn, K., Hu, T., Tanaka, T., Sasaki, Y., Kuipers, J., Loghavi, S., Wang, S. A., Yan, Y., Furudate, K., Matthews, J., Little, L., Gumbs, C., Zhang, J., Song, X., Thompson, E., Patel, K. P., Bueso-Ramos, C. E., … Takahashi, K. (2020). Clonal evolution of acute myeloid leukemia revealed by high-throughput single-cell genomics. *Nature Communications*, *11*(1). https://doi.org/10.1038/s41467-020-19119-8

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., Muthuswamy, L., Krasnitz, A., McCombie, W. R., Hicks, J., & Wigler, M. (2011). Tumour evolution inferred by single-cell sequencing. *Nature*, *472*(7341), 90–94. https://doi.org/10.1038/nature09807

Rashidi Mehrabadi, F. (2022). *Scalable tumor phylogeny inference and validation from single-cell RNA or DNA data* [PhD thesis, Indiana University]. https://www.proquest.com/docview/2701106234/

Roper, N., Brown, A.-L., Wei, J. S., Pack, S., Trindade, C., Kim, C., Restifo, O., Gao, S., Sindiri, S., Mehrabadi, F., El Meskini, R., Ohler, Z. W., Maity, T. K., Venugopalan, A., Cultraro, C. M., Akoth, E., Padiernos, E., Chen, H., Kesarwala, A., … Guha, U. (2020). Clonal Evolution and Heterogeneity of Osimertinib Acquired Resistance Mechanisms in EGFR Mutant Lung Cancer. *Cell Reports Medicine*, *1*(1), 100007. https://doi.org/10.1016/j.xcrm.2020.100007

Ross, E. M., & Markowetz, F. (2016). OncoNEM: inferring tumor evolution from single-cell sequencing data. *Genome Biology*, *17*(1). https://doi.org/10.1186/s13059-016-0929-9

Sadeqi Azer, E., Rashidi Mehrabadi, F., Malikić, S., Li, X. C., Bartok, O., Litchfield, K., Levy, R., Samuels, Y., Schäffer, A. A., Gertz, E. M., Day, C.-P., Pérez-Guijarro, E., Marie, K., Lee, M. P., Merlino, G., Ergun, F., & Sahinalp, S. C. (2020). PhISCS-BnB: a fast branch and bound algorithm for the perfect tumor phylogeny reconstruction problem. *Bioinformatics*, *36*(Supplement_1), i169–i176. https://doi.org/10.1093/bioinformatics/btaa464

Singer, J., Kuipers, J., Jahn, K., & Beerenwinkel, N. (2018). Single-cell mutation identification via phylogenetic inference. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-018-07627-7

Wang, Y., Waters, J., Leung, M. L., Unruh, A., Roh, W., Shi, X., Chen, K., Scheet, P., Vattathil, S., Liang, H., Multani, A., Zhang, H., Zhao, R., Michor, F., Meric-Bernstam, F., & Navin, N. E. (2014). Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature*, *512*(7513), 155–160. https://doi.org/10.1038/nature13600

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

Wolf, Y., Bartok, O., Patkar, S., Eli, G. B., Cohen, S., Litchfield, K., Levy, R., Jiménez-Sánchez, A., Trabish, S., Lee, J. S., Karathia, H., Barnea, E., Day, C.-P., Cinnamon, E., Stein, I., Solomon, A., Bitton, L., Pérez-Guijarro, E., Dubovik, T., … Samuels, Y. (2019). UVB-Induced Tumor Heterogeneity Diminishes Immune Response in Melanoma. *Cell*, *179*(1), 219–235.e21. https://doi.org/10.1016/j.cell.2019.08.032

Wu, Y. (2019). Accurate and efficient cell lineage tree inference from noisy single cell data: the maximum likelihood perfect phylogeny approach. *Bioinformatics*, *36*(3), 742–750. https://doi.org/10.1093/bioinformatics/btz676

Zafar, H., Navin, N., Chen, K., & Nakhleh, L. (2019). SiCloneFit: Bayesian inference of population structure, genotype, and phylogeny of tumor clones from single-cell genome sequencing data. *Genome Research*, *29*(11), 1847–1859. https://doi.org/10.1101/gr.243121.118

Zafar, H., Tzen, A., Navin, N., Chen, K., & Nakhleh, L. (2017). SiFit: inferring tumor trees from single-cell sequencing data under finite-sites models. *Genome Biology*, *18*(1). https://doi.org/10.1186/s13059-017-1311-2