The tidyomics ecosystem: Enhancing omic data analyses

Hutchison WJ[1,2,+], Keyes TJ[3,4,+], *The Tidyomics Consortium*, Crowell LH[5,6], Soneson C[7,8], Yuan V[9], Nahid AA[10], Mu W[11], Park J[11] Davis ES[12], Tang M[13], Axisa PP[14], Kitt JW[15], Poon CL[16], Kosmac M[17], Serizay J[18], Sato N[19], Gottardo R[20,8], Morgan M[21], Lee S[22], Lawrence M[22], Hicks SC[23,24], Nolan GP[25], Davis KL[4], Papenfuss AT[1,2,*], Love M[26,11,*], Mangiola S[1,2,*]

+ These authors contributed equally, * These authors contributed equally

Abstract

The exponential availability of omic data presents challenges in data manipulation, analysis, and integration. Addressing these challenges, Bioconductor offers an extensive community-driven data analysis platform, while R tidy programming offers a revolutionary standard for data organisation and manipulation. Bioconductor and tidy R have mostly remained independent; bridging them would streamline omic analysis and ease learning and cross-disciplinary collaborations. Here, we introduce the *tidyomics* software ecosystem that brings the R tidy toolkit to omic data analysis. We demonstrate its benefits by analysing 7.5 million PBMCs from the Human Cell Atlas, bridging six data frameworks and ten analysis tools.

[1] Walter and Eliza Hall Institute of Medical Research
[2] Department of Medical Biology, University of Melbourne, Parkville, VIC 3052, Australia
[3] Stanford University School of Medicine, Department of Biomedical Data Science
[4] Stanford University School of Medicine, Department of Pediatrics
[5] University of Zurich, Switzerland
[6] Centre for Genomic Regulation, Barcelona Institute of Science and Technology, Barcelona, Spain
[7] Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland
[8] SIB Swiss Institute of Bioinformatics, Basel, Switzerland
[9] Department of Statistics, The University of British Columbia
[10] Department of Biochemistry and Molecular Biology, Shahjalal University of Science and Technology, Sylhet, Bangladesh
[11] Biostatistics Department, University of North Carolina-Chapel Hill, Chapel Hill, NC, USA
[12] Bioinformatics and Computational Biology Program, University of North Carolina-Chapel Hill, Chapel Hill, NC, USA
[13] Immunitas Therapeutics
[14] Centre de Recherches en Cancerologie de Toulouse, Inserm, Toulouse, France
[15] Genetics, Diversity and Ecophysiology of Cereals, INRAE, Clermont-Ferrand, France
[16] Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA
[17] Achilles Therapeutics
[18] Institut Pasteur, CNRS UMR3525, Unité Régulation Spatiale des Génomes, F-75015, Paris, France
[19] Division of Health Medical Intelligence, Human Genome Center, The Institute of Medical Science, The University of Tokyo
[20] University of Lausanne
[21] Roswell Park Comprehensive Cancer Center
[22] Genentech, Department of Bioinformatics and Computational Biology
[23] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health
[24] Malone Center for Engineering in Healthcare, Johns Hopkins University, MD, USA
[25] Stanford University School of Medicine, Department of Pathology
[26] Genetics Department, University of North Carolina-Chapel Hill, Chapel Hill, NC, USA

Main

High-throughput genomic, transcriptomic, spatial, and multi-omic technologies have revolutionised biomedical research, presenting formidable opportunities and challenges for data manipulation, exploration, analysis, integration, and interpretation[1]. To address these challenges, the scientific community has embraced increasing abstract programming languages **(Figure 1A)**. Building upon the hardware-focused Fortran to higher-level instructions in C, scientists adopted modular data representations in Python, Perl, and Java, suited to genes and DNA sequences. Biological data analysis has gravitated toward higher-level languages oriented to data structures and statistical modelling, such as S, Matlab, R, and Julia[2].

More recently, tidy R programming and the tidyverse software ecosystem[3] have transformed data science by prioritising intuitive data manipulation, pipelines, and statistical models over the intricacy of data structures. Tidy data organises variables as columns and observations as rows. Tidy programming leverages standardised verbs for data manipulation and visualisation, streamlining workflows and promoting transparency. This approach is now standard in data science educational programs.

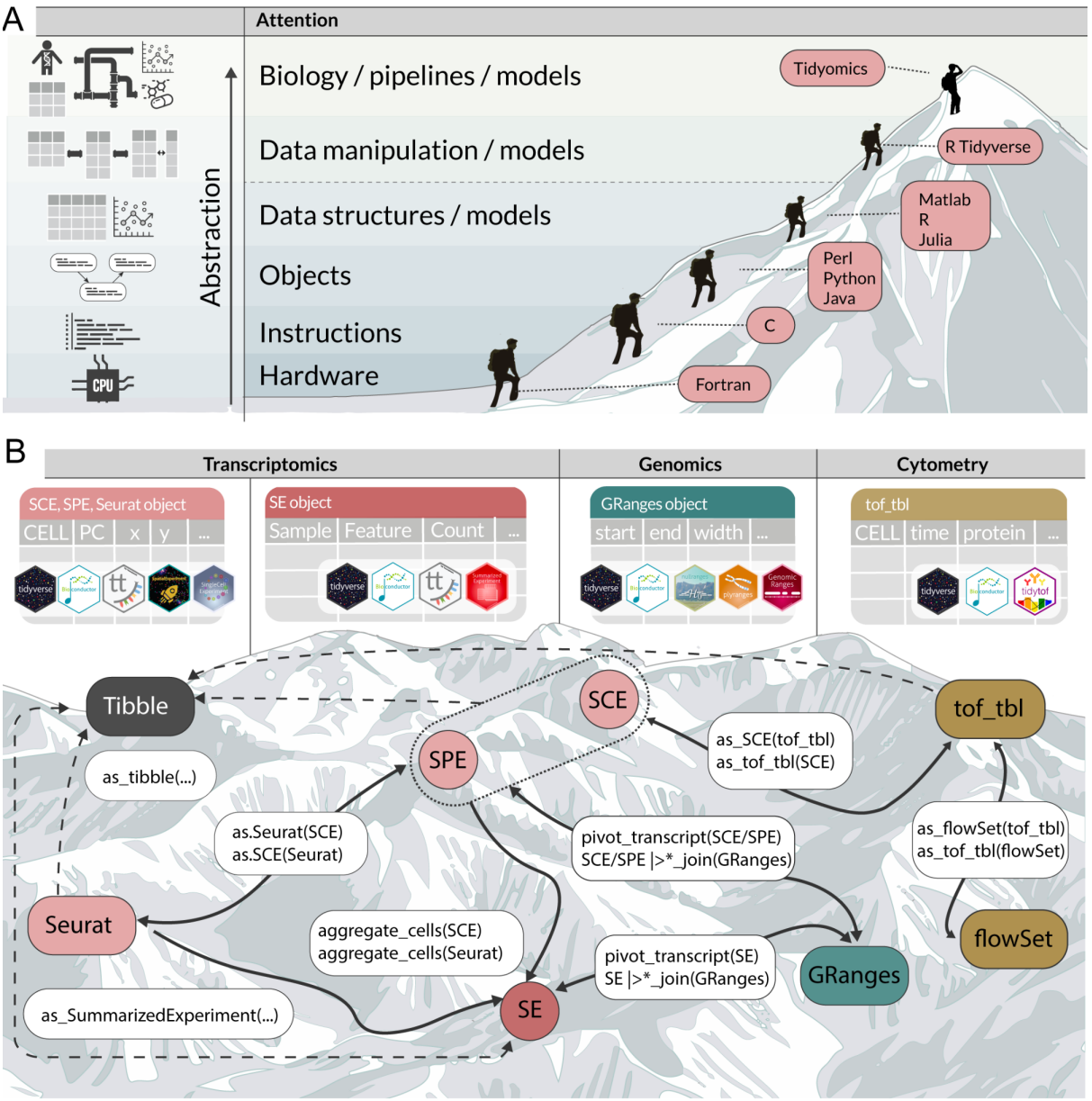

Figure 1: A: The landscape from low- to high-level languages for quantitative biology. The icons on the left represent the programmer's focus on coding languages. **B:** The landscape of interconnectivity within the tidyomics ecosystem. The top side shows the consistent data representation for the diverse data containers. The hexagonal icons represent the compatible R packages. On the bottom, the data containers are represented by rounded rectangles and functions that connect them as white boxes. SPE = SpatialExperiment, SCE = SingleCellExperiment, SE = SummarizedExperiment.

In response to the complexity of the software landscape, the Bioconductor Project[4] has emerged as a premier R software repository and platform for omic data analysis. Bioconductor provides international standardisation and interoperability for data processing workflows and statistical analysis. With extensive annotation resources

and standardised data formats that link metadata, Bioconductor promotes reproducibility and community-driven open-source development.

Bioconductor has remained largely independent of the tidyverse ecosystem. Creating a bridge between these two ecosystems by providing a tidy interface to standard data formats[5–7] and analysis[7] would shift the researcher's focus from technical challenges to biological questions. Also, introducing a standard established in data science education would lower the barrier to entry for analysing diverse omics data.

Here, we present *tidyomics*, an interoperable software ecosystem that bridges Bioconductor and the tidyverse. *tidyomics* is installable with a single homonymous meta-package. This ecosystem includes three new packages: tidySummarizedExperiment, tidySingleCellExperiment, and tidySpatialExperiment, and five publicly available R packages: plyranges[5], nullranges[8], tidyseurat[6], tidybulk[7], tidytof[9]. Importantly, *tidyomics* leaves the original data containers and methods unaltered, ensuring compatibility with existing software, maintainability and long-term Bioconductor support.

Within Bioconductor, GenomicRanges[10] organises genomic features as ranges (e.g. genes, exons, SNPs, CpGs) in rows, linked with variables (e.g. range width) as columns (like the BED[11] format). plyranges[5] extends dplyr verbs to GenomicRanges objects, facilitating ranges integration, overlap analysis, summarisation, and visualisation. Plyranges interacts with nullranges for matching[12] or bootstrapping[8] ranges to perform overlap enrichment analyses.

In Bioconductor, SummarizedExperiment and SingleCellExperiment[13] organise transcriptional abundance as a feature/gene-by-sample matrix linked with metadata. *tidyomics* generalises the concept of the *variable* providing a tabular interface with observations (e.g. gene/sample pair) as rows and variables (e.g. abundance, metadata) as columns. This approach enables complex filtering, summarisation, analysis, and visualisation using the tidyverse. tidybulk[7] offers a tidy and modular analysis pipeline for bulk and pseudobulk data.

Bioconductor's flowCore[14] package organises data from mass, flow and sequence-based cytometry in a cell-by-feature matrix and facilitates data manipulation. tidytof[9] interfaces flowCore with the tidyverse, tidySingleCellExperiment and tidySummarizedExperiment.

Bioconductor's SpatialExperiment[15] organises data from cell/pixel-based technologies[16], such as 10X Genomics Xenium, CosMX, Mibi, and MERSCOPE. tidySpatialExperiment offers a tidy interface for data with spatial coordinates and provides specialised operations such as filtering based on geometric shapes and merging multilayer or multiomic datasets.

tidyomics is a unified and interoperable software ecosystem for omic technologies that covers the Bioconductor and Seurat[6] frameworks. Through conversion and join operations, a network of functionalities connects all data containers **(Figure 1B)**. This harmonised approach facilitates seamless container switching, decreasing the dependence on a specific framework created by domain-specific syntax and effectively increasing the umbrella of used tools.

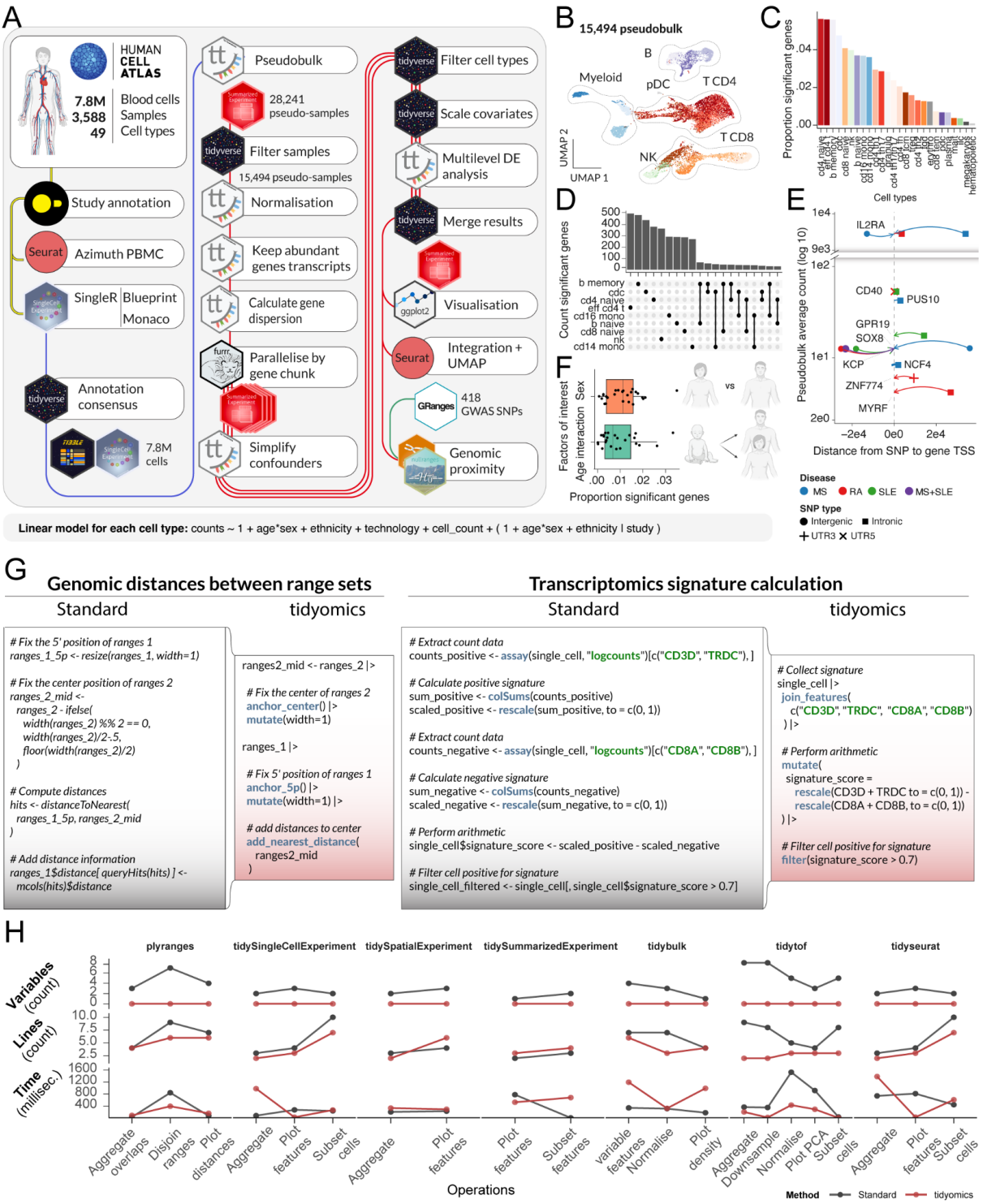

Figure 2: A: Tidyomics powers large-scale cross-framework analyses. We compared peripheral blood mononuclear cells between sexes, at the pseudobulk level. The logos represent data and analysis frameworks. The connecting lines represent pipelines, coloured by the object type. Parallel lines represent parallel workflows. **B:** Pseudosample UMAP, coloured by cell type. **C:** Rank of cell types from the most to the least changed across sexes, coloured as per the B panel. **D:** Significant gene overlap across the top nine cell types for sex effect or its interaction with age. **E:** Overlap of sex-related genes in CD4 naive cells with GWAS SNPs for multiple sclerosis, rheumatoid arthritis, and systemic

lupus erythematosus. **F:** Fraction of sex-related genes significant as a main effect or interaction with age. **G:** Comparison of code readability between standard and tidyverse programming. Two tasks showcased are visualising a histogram of genomic distances (left) and calculating a multi-gene signature from single-cell data (right). **H:** The benchmark of variables, lines of code, and time efficiency of the tidyomics ecosystem compared to standard (non-tidy) coding. The operations include common manipulations and analysis for each package (Methods).

To demonstrate *tidyomics'* utility and scalability, we tested sex transcriptomic differences of the peripheral immune system across 7.5 million blood cells. Our ecosystem seamlessly bridged six data and analysis frameworks **(Figure 2A)**. After preprocessing, we tested 15,494 pseudobulk samples across 26 immune cell types **(Figure 2B)** with a multilevel differential expression model. We identified T CD4 naive cells, T effector and B memory cells as the most changing between sexes **(Figure 2C)**. Most sex-related transcriptional changes (excluding sex chromosomes) were cell-type specific rather than shared (FDR < 0.05) **(Figure 2D)**. We tested the proximity of genes with a significant effect for sex or its interaction with age in CD4 naive cells to GWAS SNPs for three immune-cell-related and sex-biased diseases: multiple sclerosis, rheumatoid arthritis, and systemic lupus erythematosus **(Figure 2E)**. We found nine genes overlapping or near SNPs associated with these diseases (*IL2RA*, *CD40*, and *KCP* associated with two or more). A large proportion of sex-related genes, 41%, define divergent immune-ageing trajectories **(Figure 2F)**.

As shown in this analysis, *tidyomics* allows code repurposing across diverse data sources. For example, complex manipulation and visualisation of genome and transcriptome data can be performed using modular, consistent grammar assembled into a compact and legible pipeline **(Figure 2G)**. Legibility and coding simplicity are promoted by fewer intermediate variables and lines of code compared with standard counterparts while incurring no major execution-time overhead **(Figure 2H)**. Tidyomics favours functional programming (e.g., vectorisation rather than for-loops), which helps avoid bugs caused by variable updating, especially in interactive programming.

The tidyverse and Bioconductor ecosystems are transforming data science and biological data analysis. Tidyomics bridges the gap between these ecosystems, enabling analysts to leverage the power of tidy data principles in omic analyses. This

integration fosters cross-disciplinary collaborations, reduces barriers to entry for new users, and enhances code readability, reproducibility, and transparency. The tidy standard applied to biological software creates an extensible development ecosystem where independent researchers can interface new software. Ultimately, the tidyomics ecosystem, consisting of new and publicly available R packages, has the potential to significantly accelerate scientific discovery.

Data availability

Human Cell Atlas PBMC single-cell data was downloaded from the CELLxGENE database (https://cellxgene.cziscience.com/). Metadata and gene-transcript abundance for these datasets were downloaded via the CuratedAtlasQuery R package.

Code availability

The *tidyomics* meta-package is available at github.com/tidyomics/tidyomics. The tidySummarizedExperiment package is available at Bioconductor bioconductor.org/packages/tidySummarizedExperiment. The tidySingleCellExperiment package is available at Bioconductor bioconductor.org/packages/tidySingleCellExperiment. The tidySpatialExperiment package is available at github.com/william-hutchison/tidySpatialExperiment. The code used to benchmark the efficiency of standard and tidyomics workflows is available as an R vignette on github.com/tidyomics/tidyomics/blob/main/vignettes/benchmarking.Rmd

Acknowledgements

We acknowledge Bioconductor and tidyverse communities, whose software and coding paradigms this work is based on and would not be possible without. We also thank the tidyomics community for their feedback and contribution. We thank Vincent Carey for his support and feedback on the project. Also, we thank Matt Richie for his continuous support and feedback.

Author information

Authors and Affiliations

Contributions

SM proposed the study, SM and MIL designed the study, HW and SM developed the novel tidy adapters for transcriptomics, and HW, KT, SM and MIL performed the analyses. The tidyomics Consortium contributed to the ecosystem development and ongoing improvement. SM, MIL, ATP, DK, SCH, LM, MM, and GR acted as supervisory team. All authors contributed to the writing of the manuscript.

Corresponding author

Mangiola Stefano mangiola.s@wehi.edu.au

Michael I Love michaelisaiahlove@gmail.com

Anthony T Papenfuss papenfuss@wehi.edu.au

Ethics declarations

The authors do not have any conflict of interest.

References

1. Tarazona, S., Arzalluz-Luque, A. & Conesa, A. Undisclosed, unmet and neglected challenges in multi-omics studies. *Nature Computational Science* **1**, 395–402 (2021).
2. Roesch, E. *et al.* Julia for biologists. *Nat. Methods* **20**, 655–664 (2023).
3. Wickham, H. *et al.* Welcome to the Tidyverse. *Journal of Open Source Software* **4**, 1686 (2019).
4. Gentleman, R. C. *et al.* Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
5. Lee, S., Cook, D. & Lawrence, M. plyranges: a grammar of genomic data transformation. *Genome Biol.* **20**, 4 (2019).
6. Mangiola, S., Doyle, M. A. & Papenfuss, A. T. Interfacing Seurat with the R tidy universe.

Bioinformatics (2021) doi:10.1093/bioinformatics/btab404.

7. Mangiola, S., Molania, R., Dong, R., Doyle, M. A. & Papenfuss, A. T. tidybulk: an R tidy framework for modular transcriptomic data analysis. *Genome Biol.* **22**, 42 (2021).
8. Mu, W. *et al.* bootRanges: flexible generation of null sets of genomic ranges for hypothesis testing. *Bioinformatics* **39**, (2023).
9. Keyes, T. J., Koladiya, A., Lo, Y.-C., Nolan, G. P. & Davis, K. L. tidytof: a user-friendly framework for scalable and reproducible high-dimensional cytometry data analysis. *Bioinform Adv* **3**, vbad071 (2023).
10. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
11. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
12. Davis, E. S. *et al.* matchRanges: generating null hypothesis genomic ranges via covariate-matched sampling. *Bioinformatics* **39**, (2023).
13. Amezquita, R. A. *et al.* Orchestrating single-cell analysis with Bioconductor. *Nat. Methods* **17**, 137–145 (2020).
14. Ko, M. E. *et al.* FLOW-MAP: a graph-based, force-directed layout algorithm for trajectory mapping in single-cell time course datasets. *Nat. Protoc.* **15**, 398–420 (2020).
15. Righelli, D. *et al.* SpatialExperiment: infrastructure for spatially-resolved transcriptomics data in R using Bioconductor. *Bioinformatics* **38**, 3128–3131 (2022).
16. Wang, Y. *et al.* Spatial transcriptomics: Technologies, applications and experimental considerations. *Genomics* 110671 (2023).