nature methods

Brief Communication

https://doi.org/10.1038/s41592-024-02299-2

The tidyomics ecosystem: enhancing omic data analyses

Received: 24 August 2023

Accepted: 5 May 2024

Published online: 14 June 2024

Check for updates

William J. Hutchison[1,2,35], Timothy J. Keyes[3,4,35], The tidyomics Consortium*, Helena L. Crowell[5,6], Jacques Serizay[7], Charlotte Soneson[8,9], Eric S. Davis[10], Noriaki Sato[11], Lambda Moses[12], Boyd Tarlinton[13], Abdullah A. Nahid[14], Miha Kosmac[15], Quentin Clayssen[16], Victor Yuan[17], Wancen Mu[18], Ji-Eun Park[18], Izabela Mamede[19], Min Hyung Ryu[20,21], Pierre-Paul Axisa[22], Paulina Paiz[3], Chi-Lam Poon[23], Ming Tang[24], Raphael Gottardo[9,25,26], Martin Morgan[27], Stuart Lee[28], Michael Lawrence[28], Stephanie C. Hicks[29,30,31], Garry P. Nolan[32], Kara L. Davis[4], Anthony T. Papenfuss[1,2] ✉, Michael I. Love[18,33] ✉ & Stefano Mangiola[1,2,34] ✉

The growth of omic data presents evolving challenges in data manipulation, analysis and integration. Addressing these challenges, Bioconductor provides an extensive community-driven biological data analysis platform. Meanwhile, tidy R programming offers a revolutionary data organization and manipulation standard. Here we present the tidyomics software ecosystem, bridging Bioconductor to the tidy R paradigm. This ecosystem aims to streamline omic analysis, ease learning and encourage cross-disciplinary collaborations. We demonstrate the effectiveness of tidyomics by analyzing 7.5 million peripheral blood mononuclear cells from the Human Cell Atlas, spanning six data frameworks and ten analysis tools.

High-throughput technologies for genomics, epigenomics, transcriptomics, spatial analysis and multi-omics have revolutionized biomedical research, presenting opportunities and challenges for data manipulation, exploration, analysis, integration and interpretation[1]. To address these challenges, the scientific community has developed object-oriented frameworks for data organization and specialized operations.

In response to the complexity of the software landscape, Bioconductor[2] has emerged as a premier R software repository and platform for omic data analysis. Bioconductor provides international standardization and interoperability for data processing workflows and statistical analysis. With extensive annotation resources and standardized data formats that link metadata, Bioconductor promotes reproducibility and community-driven open-source development.

Recently, the tidy R paradigm and the tidyverse software ecosystem[3] have transformed R-based data science by prioritizing intuitive data representation and manipulation over complex data structures and syntax. This paradigm uses tables to represent data, with variables as columns and observations as rows. It simplifies data manipulation with operations connected in pipelines that use standardized and natural language vocabulary. The components of the tidyverse rank as the most frequently downloaded R packages[4] and are widely taught in Data Science and Bioinformatics programs worldwide[5].

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[6–8] and analysis[8] would enable researchers to shift their focus from technical challenges to biological questions. Also, leveraging a standard in data science education would lower the barrier to entry for analyzing diverse omic data.

In this Brief Communication, we present tidyomics, an interoperable software ecosystem that bridges Bioconductor and other omic analysis frameworks (for example, Seurat[7]) with the tidyverse. This ecosystem is installable with a single meta-package, available in

A full list of affiliations appears at the end of the paper. *A list of authors and their affiliations appears at the end of the paper.
✉e-mail: papenfuss@wehi.edu.au; michaelisaiahlove@gmail.com; stefano.mangiola@adelaide.edu.au

https://doi.org/10.1038/s41592-024-02299-2

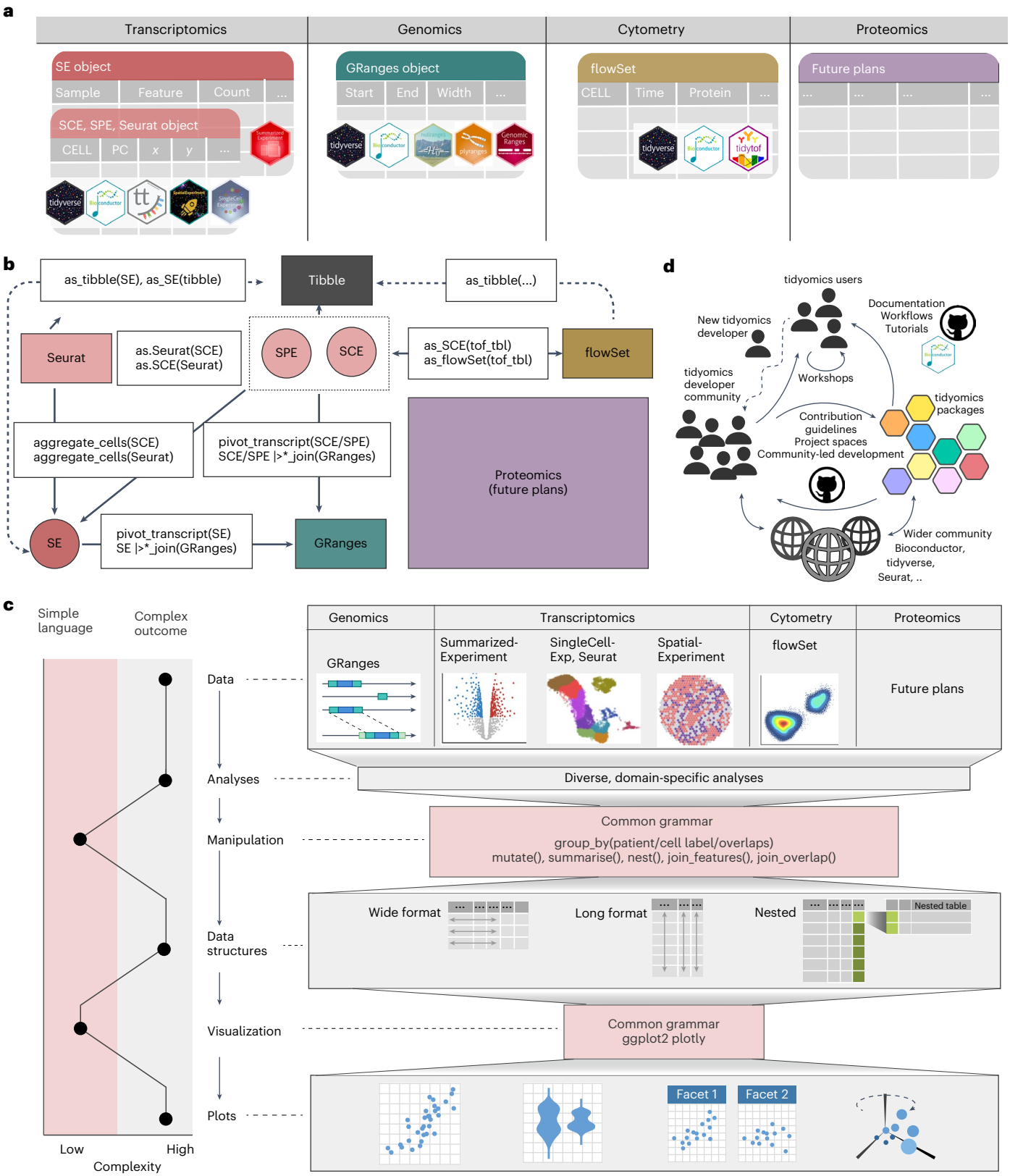
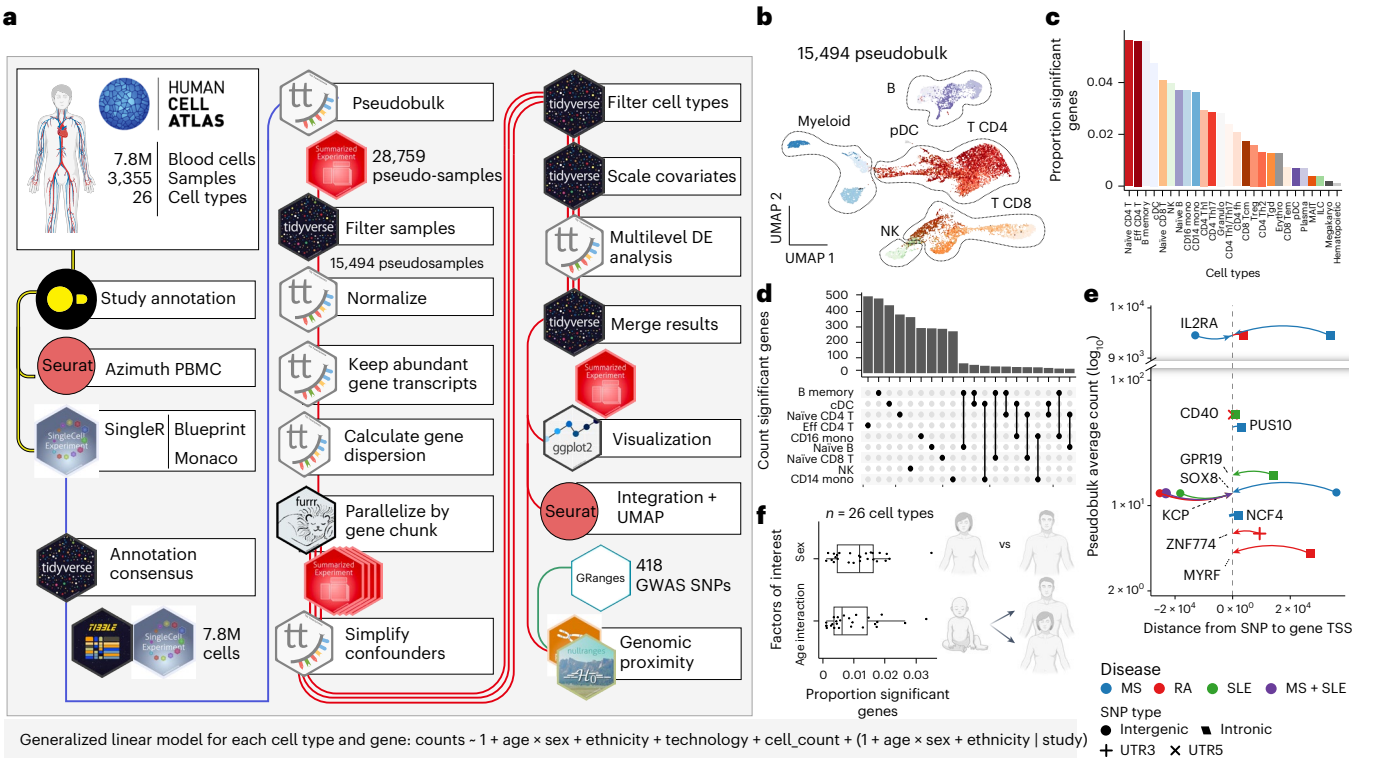
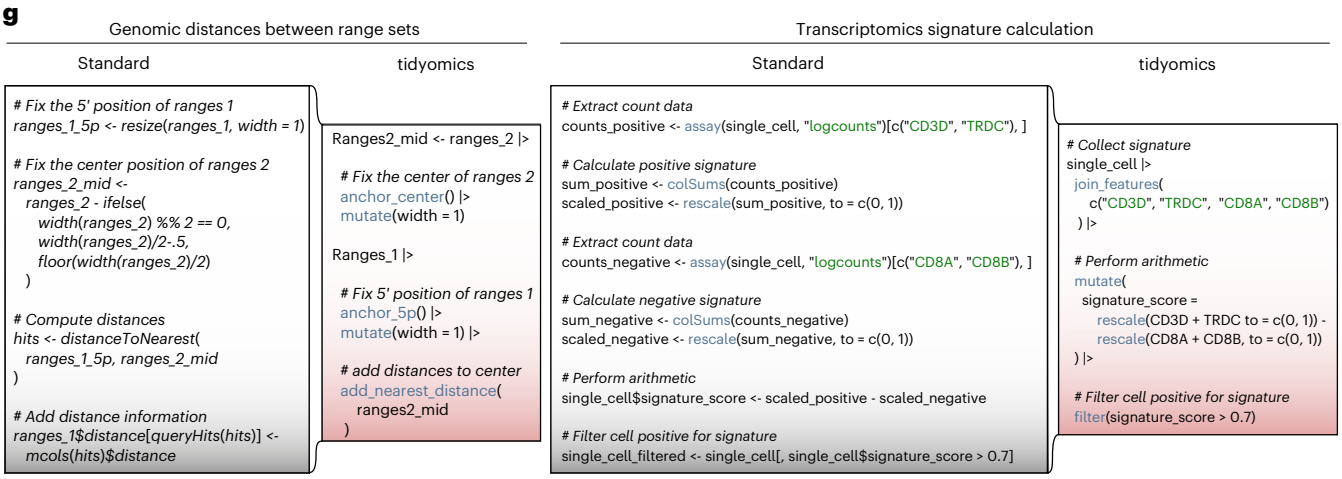
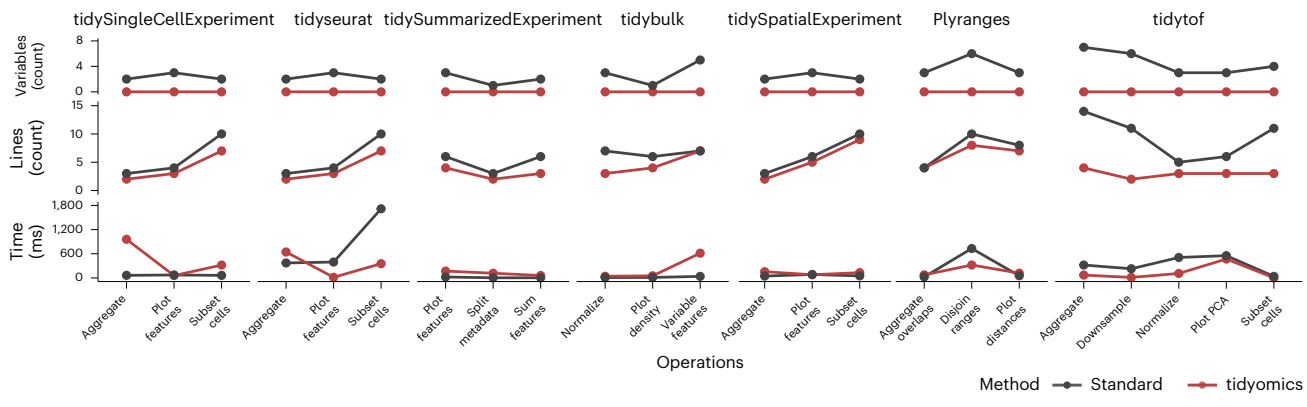


Fig. 1 | Overview of the tidyomics ecosystem. a, Diagrams of data interfaces show consistent data representation for the diverse data containers. The hexagonal icons represent the compatible R packages for each data container. **b**, The landscape of rich data objects in R/Bioconductor, with tidyomics verbs as paths connecting these objects. The data containers are represented by rounded rectangles and functions that connect them as white boxes. SPE, SpatialExperiment; SCE, SingleCellExperiment; SE, SummarizedExperiment. **c**, Left: contrast between the simplicity of the tidy syntax/grammar and the complex outcome and input data containers. Right: example workflows include data, biological analysis, data/results manipulation and summarization, diverse data structures, visualization and resulting plots. The pink areas include the infrastructure that shares grammar across omics. **d**, Engagement within the tidyomics community is multifaceted, centering around a suite of R packages tailored for streamlined data analysis.

https://doi.org/10.1038/s41592-024-02299-2







Bioconductor. tidyomics includes three new packages: tidySummarizedExperiment, tidySingleCellExperiment and tidySpatialExperiment, and five publicly available R packages: plyranges[6], nullranges[9], tidyseurat[7], tidybulk[8] and tidytof[10]. Importantly, while tidyomics represents omic data in a tidy format (Fig. 1a), it leaves the original data containers and methods unaltered, ensuring compatibility with existing software, maintainability and long-term Bioconductor support.

Within Bioconductor, GenomicRanges[11] organizes genomic features as ranges (for example, genes, exons, single-nucleotide

Fig. 2 | Performance of the tidyomics ecosystem. a, tidyomics powers large-scale cross-framework analyses. The logos represent data and analysis frameworks. The connecting lines represent pipelines, colored by object type. Parallel lines represent parallel workflows. **b**, Pseudosample UMAP, colored by cell type. NK, natural killer. **c**, Rank of cell types from the most to the least changed across sexes, colored as per **b**. **d**, Large 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. MS, multiple sclerosis; RA, rheumatoid arthritis; SLE, systemic lupus erythematosus. **f**, Fraction of significant (FDR <0.05) genes for age-independent sex effect or sex-age interaction. The box plot center line represents the median value, and the lower and upper hinges represent the first and third quartiles. The lower whisker extends from the lower hinge to 1.5 times the interquartile range or the lowest value. The upper whisker extends from the upper hinge to 1.5 times the interquartile range or the highest value. **g**, Comparison of code readability between standard and tidyverse programming. Two tasks showcased are visualizing 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 our ecosystem compared with standard (non-tidy) coding. The operations include common manipulations and analysis for each package (Methods).

polymorphisms (SNPs) and CpGs) in rows, linked with variables (for example, range width) as columns (like the BED[12] format). plyranges[6] extends dplyr verbs to GenomicRanges objects, facilitating ranges integration, overlap analysis, summarization and visualization. plyranges interacts with nullranges for matching[13] or bootstrapping[9] ranges to perform overlap enrichment analyses.

In Bioconductor, SummarizedExperiment and SingleCellExperiment[14] organize transcriptional abundance as a feature/gene-by-sample matrix linked with metadata. tidyomics generalizes the concept of the variable by providing a tabular interface with observations (for example, gene–sample pair) as rows and variables (for example, abundance and metadata) as columns. This approach enables complex filtering, summarization, analysis and visualization using the tidyverse. tidybulk[8] offers a tidy and modular analysis pipeline for bulk and pseudobulk data.

Bioconductor's flowCore[15] package organizes data from mass, flow and sequence-based cytometry in a cell-by-feature matrix and facilitates data manipulation. tidytof[10] interfaces flowCore with the tidyverse, tidySingleCellExperiment and tidySummarizedExperiment.

Bioconductor's SpatialExperiment[16] organizes data from cell- or pixel-based technologies[17], such as 10x Genomics Xenium, CosMX, Mibi and MERSCOPE. tidySpatialExperiment offers a tidy interface for data with spatial coordinates and provides specialized operations such as gating based on geometric and hand-drawn shapes.

tidyomics is a unified and interoperable software ecosystem for omic technologies that covers several omic analysis frameworks. Through conversion and join operations, a network of functionalities connects all data containers (Fig. 1b). This harmonized 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 (Fig. 1c).

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 (Fig. 2a), showcasing the benefit of consistently using tidy R grammar instead of mixing the syntaxes of base R, DuckDB, Seurat, SingleCell- and SummarizedExperiment, DGEList and GenomicRanges. After preprocessing, we tested 15,494 pseudobulk samples across 26 immune cell types (Fig. 2b) with a multilevel differential expression model. We identified T CD4 naive cells, T effectors and B memory cells as the most changing between sexes (Fig. 2c). Most sex-related transcriptional changes (excluding sex chromosomes) were cell-type specific rather than shared (false discovery rate (FDR) <0.05) (Fig. 2d). We tested the proximity of genes with a significant effect (FDR <0.05) for sex or its interaction with age in CD4 naive cells to genome-wide association study (GWAS) SNPs for three immune-cell-related and sex-biased diseases: multiple sclerosis, rheumatoid arthritis and systemic lupus erythematosus (Fig. 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-aging trajectories (Fig. 2f).

This analysis shows that tidyomics allows code repurposing across diverse data sources. For example, complex manipulation and visualization of genome and transcriptome data can be performed using modular, consistent grammar assembled into a compact and legible pipeline (Fig. 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 (Fig. 2h). Tidy R favors functional programming (for example, vectorization rather than for-loops), which helps avoid bugs caused by variable updating, especially in interactive programming.

The Bioconductor coding standards and contribution guidelines adopted by our dedicated developer community set a solid ground for the long-term maintainability of the ecosystem (Fig. 1d). This ecosystem will grow, including R packages with compatible goals and standards from Bioconductor and CRAN. While tidyomics is currently focused on simplification and harmonization, novel analysis and manipulation tools are among its goals.

The tidyverse and Bioconductor ecosystems are transforming R-based 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 with new software. Ultimately, the tidyomics ecosystem, consisting of new and publicly available R packages, has the potential to greatly accelerate scientific discovery.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-024-02299-2.

References

1. Tarazona, S., Arzalluz-Luque, A. & Conesa, A. Undisclosed, unmet and neglected challenges in multi-omics studies. *Nat. Comput. Sci.* **1**, 395–402 (2021).
2. Gentleman, R. C. et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* **5**, R80 (2004).
3. Wickham, H. et al. Welcome to the Tidyverse. *J. Open Source Softw.* **4**, 1686 (2019).
4. Li, P. *Computation and Visualization of Package Download Counts and Percentiles [R package packageRank version 0.8.3]* (R Project, 2023).
5. Çetinkaya-Rundel, M. et al. An educator's perspective of the tidyverse. Preprint at https://doi.org/10.48550/arXiv.2108.03510 (2021).
6. Lee, S., Cook, D. & Lawrence, M. plyranges: a grammar of genomic data transformation. *Genome Biol.* **20**, 4 (2019).

7. Mangiola, S., Doyle, M. A. & Papenfuss, A. T. Interfacing Seurat with the R tidy universe. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btab404 (2021).
8. 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).
9. Mu, W. et al. bootRanges: flexible generation of null sets of genomic ranges for hypothesis testing. *Bioinformatics* **39**, btad190 (2023).
10. 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).
11. Lawrence, M. et al. Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
12. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
13. Davis, E. S. et al. matchRanges: generating null hypothesis genomic ranges via covariate-matched sampling. *Bioinformatics* **39**, btad197 (2023).
14. Amezquita, R. A. et al. Orchestrating single-cell analysis with Bioconductor. *Nat. Methods* **17**, 137–145 (2020).
15. 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).
16. Righelli, D. et al. SpatialExperiment: infrastructure for spatially-resolved transcriptomics data in R using Bioconductor. *Bioinformatics* **38**, 3128–3131 (2022).
17. Wang, Y. et al. Spatial transcriptomics: technologies, applications and experimental considerations. *Genomics* **115**, 110671 (2023).

[1]The Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia. [2]Department of Medical Biology, University of Melbourne, Parkville, Victoria, Australia. [3]Department of Biomedical Data Science, Stanford University School of Medicine, Stanford, CA, USA. [4]Department of Pediatrics, Stanford University School of Medicine, Stanford, CA, USA. [5]University of Zurich, Zurich, Switzerland. [6]Centro Nacional de Análisis Genómico (CNAG), Barcelona, Spain. [7]Unité Régulation Spatiale des Génomes, Institut Pasteur, CNRS UMR3525, Paris, France. [8]Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland. [9]SIB Swiss Institute of Bioinformatics, Basel, Switzerland. [10]Bioinformatics and Computational Biology Program, University of North Carolina-Chapel Hill, Chapel Hill, NC, USA. [11]Division of Health Medical Intelligence, Human Genome Center, The Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [12]California Institute of Technology, Pasadena, CA, USA. [13]Queensland Department of Agriculture and Fisheries, Brisbane, Queensland, Australia. [14]Department of Biochemistry and Molecular Biology, Shahjalal University of Science and Technology, Sylhet, Bangladesh. [15]Achilles Therapeutics, London, UK. [16]DNA Script, Le Kremlin-Bicêtre, France. [17]Department of Statistics, The University of British Columbia, Vancouver, British Columbia, Canada. [18]Biostatistics Department, University of North Carolina-Chapel Hill, Chapel Hill, NC, USA. [19]Department of Biochemistry and Immunology, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil. [20]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. [21]Department of Medicine, Harvard Medical School, Boston, MA, USA. [22]Centre de Recherches en Cancérologie de Toulouse, Université de Toulouse, Inserm, CNRS, Université Toulouse III-Paul Sabatier, Toulouse, France. [23]Department of Physiology and Biophysics, Weill Cornell Medicine, New York, NY, USA. [24]Immunitas Therapeutics, Waltham, MA, USA. [25]University of Lausanne, Lausanne, Switzerland. [26]Lausanne University Hospital, Lausanne, Switzerland. [27]Roswell Park Comprehensive Cancer Center, Buffalo, NY, USA. [28]Department of Bioinformatics and Computational Biology, Genentech, South San Francisco, CA, USA. [29]Department of Biostatistics, Johns Hopkins University, Baltimore, MD, USA. [30]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA. [31]Malone Center for Engineering in Healthcare, Johns Hopkins University, Baltimore, MD, USA. [32]Department of Pathology, Stanford University School of Medicine, Stanford, CA, USA. [33]Genetics Department, University of North Carolina-Chapel Hill, Chapel Hill, NC, USA. [34]South Australian immunoGENomics Cancer Institute, The University of Adelaide, Adelaide, South Australia, Australia. [35]These authors contributed equally: William J. Hutchison, Timothy J. Keyes. ✉e-mail: papenfuss@wehi.edu.au; michaelisaiahlove@gmail.com; stefano.mangiola@adelaide.edu.au

The tidyomics Consortium

William J. Hutchison[1,2,35], Timothy J. Keyes[3,4,35], Helena L. Crowell[5,6], Jacques Serizay[7], Charlotte Soneson[8,9], Eric S. Davis[10], Noriaki Sato[11], Lambda Moses[12], Boyd Tarlinton[13], Abdullah A. Nahid[14], Miha Kosmac[15], Quentin Clayssen[16], Victor Yuan[17], Wancen Mu[18], Ji-Eun Park[18], Izabela Mamede[19], Min Hyung Ryu[20,21], Pierre-Paul Axisa[22], Paulina Paiz[3], Chi-Lam Poon[23], Ming Tang[24], Raphael Gottardo[9,25,26], Martin Morgan[27], Stuart Lee[28], Michael Lawrence[28], Stephanie C. Hicks[29,30,31], Garry P. Nolan[32], Kara L. Davis[4], Anthony T. Papenfuss[1,2], Michael I. Love[18,33] & Stefano Mangiola[1,2,34]

Methods

tidySummarizedExperiment

tidySummarizedExperiment introduces an innovative and versatile approach for representing and managing bulk data, offering an alternative to and complementing the conventional methodologies commonly employed in SummarizedExperiment. This novel approach incorporates adaptors tailored to widespread data manipulation and visualization packages, including dplyr, tidyr, ggplot2 and plotly. Crucially, the core structure of the SummarizedExperiment object remains unaltered, ensuring seamless compatibility with existing analytical workflows.

Data are represented as a long table, wherein observations are defined by the abundance of a feature–sample pair, while variables encompass feature and sample-related metadata. The fundamental columns composing this representation consist of the feature and sample columns. Importantly, when any of these essential columns are absent from the output of a given operation (for example, select) or when a summarized version of these columns is generated, an independent table is returned. This table adheres to the structure of the SummarizedExperiment tidy representation, facilitating separate analysis or visualization as needed. Furthermore, when the returned subset of observations does not represent a valid SummarizedExperiment (for example, it does not correspond to a rectangular slice of the feature–sample matrix), a tibble is returned for independent analyses.

With the tidySummarizedExperiment approach, newly created or joined columns, such as those obtained from a metadata table, are automatically incorporated into the colData, rowData or assays based on their alignment with feature or sample identifiers. This versatile mechanism extends the 'variable' concept, enabling a manipulation, displaying or visualization of sample, feature and abundance information with consistent grammar. Notably, the columns for the sample and feature identifiers and genomic ranges are designated as read-only to preserve data integrity.

tidySingleCellExperiment

tidySingleCellExperiment presents a novel approach for representing and manipulating single-cell data, providing an alternative to and complementing the conventional methods commonly used in SingleCellExperiment. The main goal and property of the API are consistent with tidySummarizedExperiment.

However, as the central analysis unit of single-cell data is cells, rather than genes for bulk data, tidySingleCellExperiment favors the cell-wise (metadata and reduced dimensions, for example, principal components and Uniform Manifold Approximation and Projection (UMAP) dimensions) and sample-wise information rather than gene-wise and sample-wise information for tidySummarizedExperiment. This design choice keeps the single-cell data representation highly interpretable and practically useful at the cost of a partial lack of consistency to the bulk data. The emphasis on cell-wise information over transcript-wise information is driven by the priority to facilitate ease of use, data summarization, information integration and data visualization in the context of cell analysis and by explicit feature-wise operation not being as common as cell-wise operation. By focusing on cell-wise information, the abstraction avoids unnecessary complexity from including feature-level information (for example, genes, proteins and genomic regions) by default, especially when performing cell-wise information subsetting.

To access transcript-level information, users can utilize the 'join_features' function. This function enriches the metadata by incorporating transcript identifiers, transcript abundance and transcript-wise annotations, including gene length, genomic coordinates and functional annotations, as additional columns of the cell metadata.

The tibble abstraction employed in tidySingleCellExperiment consists of two column types: editable columns, which allow user interaction and modification, and view-only columns, which encompass data-derived variables, such as reduced dimensions. Integrating all cell-wise information, including reduced dimensions, within a single tibble representation enables seamless data visualization, filtering and manipulation. Importantly, this design ensures compatibility with the tidyverse ecosystem, enabling data manipulation and plotting using routines from dplyr, tidyr, ggplot2 and plotly. Furthermore, adopting this abstraction allows users to operate on the data as if it were a standard tibble while preserving compatibility with any other algorithms or tools that utilize the SingleCellExperiment framework. This approach ensures full backward compatibility and facilitates seamless integration into existing workflows.

tidySingleCellExperiment shares the same grammar and data representation as tidyseurat, allowing users to use tidy code with SingleCellExperiment and Seurat data containers.

tidySpatialExperiment

tidySpatialExperiment provides a tidy R abstraction (tibble) of SpatialExperiment objects. Similarly to tidySingleCellExperiment, it provides cell-wise information, including cell metadata, spatial coordinates and metadata, and reduced dimensions. All information can be processed with tools provided by dplyr, tidyr, ggplot2 and plotly. tidySpatialExperiment provides the 'join_features' function to append the specified features to the cell-wise information and, consequently, the tibble abstraction. The 'aggregate_cells' function is provided to combine cells by shared variables and aggregate feature counts.

Benchmarking

We benchmarked standard and tidyomics workflows for common data analysis tasks. Briefly, the aggregate-cells-by-sample operation aggregates the feature counts of cells within each sample. The plot features per cell operation plots the distribution of summed feature counts for each cell within each sample. The subset cells by feature operation subsets cells by feature signature threshold. The subset-features-by-mean-count operation subsets features by mean count threshold. The plot-features-per-sample operation plots the distribution of feature counts for each sample. The normalise-features operation normalizes feature counts across samples. The plot-normalised-feature-density operation plots the density of normalized features for each sample. The identify-variable-features operation identifies the most variable features for each cell type. The aggregate-overlaps operation identifies and aggregates overlapping regions. The plot-feature-set-distances operation calculates and plots the distance between feature sets. The group-disjoin-ranges operation finds disjoint regions within groups of features and subsets overlaps. The downsample-cells operation randomly subsets cells from each sample. The plot-PCA operation calculates and plots principal components. The benchmarking operations were run using R v4.3.1, plyranges v1.22.0, tidySingleCellExperiment v1.13.3, tidySpatialExperiment v0.99.13, tidySummarizedExperiment v1.12.0, tidybulk v1.15.4, tidytof v0.0.0 and tidyseurat v0.8.0. Each benchmarking operation was executed 50 times using the microbenchmark package, and the mean time elapsed was recorded. The variable count was calculated as the number of times a new variable was created to store data. The line count was calculated as the number of lines required for each operation while following indentation best practice.

Transcriptional analyses

We collected all human peripheral blood mononuclear cells from the Human Cell Atlas[18] initiative using the CELLxGENE database. We downloaded the metadata and gene-transcript abundance through the R package CuratedAtlasQuery[30] (v1.1.3). We consistently represented age as days (named 'age_days' in our database). Where a categorical value was provided, we converted it into equivalent days using publicly available references. For example, the 'adolescent stage' was converted to 15 years old (= 5,475 days).

Immune cells were labeled using Seurat (v5.0.1) Azimuth mapping to the peripheral blood mononuclear cell reference[19] (using tidyseurat[7]) and SingleR[20] (v2.4.0) with the Blueprint[21] and Monaco[22] references. To identify a consensus, we compared and contrasted the high-resolution labels 'predicted.celltype.l2' for Seurat Azimuth and 'label.fine' Blueprint and Monaco references. When possible, the reference-specific cell-type labels were standardized under a common ontology[2]. Where the resolution of transcriptionally similar cell types was uncertain with the given tools, cell types were labeled with a coarser resolution. For example, innate lymphoid and natural killer cells were grouped under 'innate lymphoid'. The cell type curation was performed to obtain a high-confidence, meaningful representation of the immune system's heterogeneity, allowing data-rich cell types whose tissue composition can be modeled probabilistically rather than aiming for the finest resolution possible. The original annotation provided by the studies was integrated with the three new annotation sources to identify a total or partial consensus. Cell types without complete or partial consensus were filtered out.

We selected primary (no reanalyzed data or collections) physiological samples (that is, no disease). We also excluded samples with fewer than 30 cells. We excluded erythrocytes and platelets from the analyses as they were not of interest. To increase the sample size per demographic group, we merged Asian descendants (labeled in CELLxGENE as Asian and Chinese). We excluded samples whose ages were unknown.

tidySingleCellExperiment aggregated cells across samples and cell types in pseudobulk transcript counts. Pseudobulk samples with fewer than 5,000 genes or 10 cells were filtered out using tidySummarizedExperiment. Quantile normalization in limma[23] (v3.58.1) was used through tidybulk[8], as across the 28,241 pseudo samples the data distribution was heterogeneous and noncontrollable. Low-abundance gene transcripts were filtered out using edgeR[24] (v4.0.16) through tidybulk, using sex and ethnicity as factors of interest, minimum counts of 500 and minimum proportion of 0.9.

Gene-transcript abundance for each cell type was modeled using the following formulation, with age as a centered and scaled continuous variable (mean age of ~47 years):

$$\text{counts} \sim \text{age} \times \text{sex} + \text{ethnicity} + \text{technology} + \text{cell\_count} + (1 + \text{age} \times \text{sex} + \text{ethnicity}|\text{study}).$$

Given the complexity of the model, tidybulk was also used to identify data subsets that included complete covariate confounders. Tidybulk was used to fit the multilevel model through glmmSeq[25] (v0.5.5) and test hypotheses (FDR <0.05). Sex-related transcriptional changes (Fig. 2c–f) were defined as genes significant for the main effect of sex or its interaction with age, excluding genes on sex chromosomes.

Seurat[26] and tidyseurat[7] were used to remove the study effect across cell types from the pseudobulk data and calculate the UMAP dimensions. Ggplot2 (v3.5.0) was used for visualization[27].

To overlap genes with significant effect (FDR <0.05) for sex or its interaction with age in CD4 naive cells with GWAS lead SNPs, we used the tidySummarizedExperiment and plyranges packages to harmonize summary statistics from pseudobulk DE analysis and three GWAS for multiple sclerosis[28], rheumatoid arthritis and systemic lupus erythematosus[29,30]. As the GWAS data were provided in all cases for the hg19 genome build, the gene locations were loaded from the Bioconductor package TxDb.Hsapiens.UCSC.hg19.knownGene. The overlap analysis used genes with either significant main effect (FDR <0.05) for sex or sex × age, as estimated from the pseudobulk multilevel model, with FDR <0.05. Overlap was calculated as GWAS SNPs within 50 kb from the gene body, and distance was calculated from the GWAS lead SNP to the gene's transcription start site (TSS).

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Human Cell Atlas peripheral blood mononuclear single-cell data were downloaded from the CELLxGENE database. The relative weblink for each sample is listed in Supplementary Table 1. The samples analyzed are accessible at the Human Cell Atlas. Metadata and gene-transcript abundance for these datasets from the CuratedAtlasQuery database is accessible at sample_metadata.0.2.3.parquet. CELLxGENE sample accession codes are available in Supplementary Table 1. Source data are provided with this paper.

Code availability

The tidyomics homepage is https://github.com/tidyomics[31], which provides links to the constituent packages. The tidyomics meta-package is available at Bioconductor bioconductor.org/packages/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 Bioconductor bioconductor.org/packages/tidySpatialExperiment/. The code used to benchmark workflow efficiency and analyze peripheral blood mononuclear cells from the Human Cell Atlas is available at github.com/tidyomics/tidyomics_paper. Source data for Fig. 2h are available at github.com/tidyomics/tidyomics_paper.

References

18. Rozenblatt-Rosen, O. et al. Building a high-quality Human Cell Atlas. *Nat. Biotechnol.* **39**, 149–153 (2021).
19. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587.e29 (2021).
20. Aran, D. et al. Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. *Nat. Immunol.* **20**, 163–172 (2019).
21. Fernández, J. M. et al. The BLUEPRINT Data Analysis Portal. *Cell Syst* **3**, 491–495.e5 (2016).
22. Xu, W. et al. Mapping of γ/δ T cells reveals Vδ2$^+$ T cells resistance to senescence. *EBioMedicine* **39**, 44–58 (2019).
23. Law, C. W. et al. RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. *F1000Res* https://doi.org/10.12688/f1000research.9005.3 (2016).
24. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
25. Lewis, M., Goldmann, K., Sciacca, E., Cubut, C. & Surace, A. *glmmSeq: General Linear Mixed Models for Gene-Level Differential Expression* (glmmSeq: General Linear, 2022).
26. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
27. Wickham, H. *Ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).
28. International Multiple Sclerosis Genetics Consortium. Multiple sclerosis genomic map implicates peripheral immune cells and microglia in susceptibility. *Science* **365**, eaav7188 (2019).
29. Wang, Y.-F. et al. Identification of 38 novel loci for systemic lupus erythematosus and genetic heterogeneity between ancestral groups. *Nat. Commun.* **12**, 772 (2021).
30. Mangiola, S. et al. A multi-organ map of the human immune system across age, sex and ethnicity. Preprint at *bioRxiv* https://doi.org/10.1101/2023.06.08.542671 (2023).
31. tidyomics. *GitHub* https://github.com/tidyomics (2024).

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 V. Carey for his support and feedback on the project. Also, we thank M. Ritchie for his continuous support and feedback. Human illustrations were created with BioRender.com. S.M. was supported by the Victorian Cancer Agency Early Career Research Fellowship (ECRF21036). M.I.L. was supported by the Chan Zuckerberg Initiative (EOSS3-0000000057). A.T.P. was supported by the National Health and Medical Research Council (NHMRC) Senior Research Fellowship (1116955) and Investigator Grant (2026643). A.T.P., S.M. and W.H. were supported by the Lorenzo and Pamela Galli Medical Research Trust and the Galli Next Generation Discoveries Initiative. K.L.D. is the Anne T. and Robert M. Bass Endowed Faculty Scholar in Pediatric Cancer and Blood Diseases of the Stanford Maternal Child Health Research Institute and the Harriet and Mary Zelencik Endowed Faculty in Children's Cancer and Blood Diseases. P.-P.A. was supported by the Cancéropole GSO and Intergroupe Français du Myélome. R.G. was funded by a project grant from the Swiss National Foundation. M.M. was supported by the NHGRI and NCI of the National Institutes of Health under award numbers U41HG004059 and U24CA180996. This work was supported by an ASPIRE award from the Mark Foundation for Cancer Research and the B+ Foundation. The research benefited from support from the Victorian State Government Operational Infrastructure Support and Australian Government NHMRC Independent Research Institute Infrastructure Support. The funders had no role in study design, data collection and analysis, or decision to publish or prepare the manuscript.

Author contributions

S.M. proposed the study, and S.M. and M.I.L. designed the study. W.J.H. and S.M. developed the novel tidy adapters for transcriptomics, W.J.H., T.J.K., S.M. and M.I.L. performed the analyses. W.J.H., T.J.K., H.L.C., J.S., C.S., E.S.D., N.S., L.M., B.T., A.A.N., M.K., Q.C., V.Y., W.M., J.-E.P., I.M., M.H.R., P.-P.A., P.P., C.-L.P., M.T., R.G., M.M., S.L., M.L., S.C.H., G.P.N., K.L.D., A.T.P., M.I.L. and S.M. contributed to the ecosystem's development and ongoing improvement. S.M., M.I.L., A.T.P., K.L.D., S.C.H., M.L., M.M. and R.G. acted as the supervisory team. S.M., M.I.L. and A.T.P. contributed equally and jointly led the study. W.J.H. and T.J.K. contributed equally. All authors contributed to the manuscript's writing.

Competing interests

R.G. has received consulting income from Takeda and Sanofi, and declares ownership in Ozette Technologies. M.K. is an employee of and declares ownership in Achilles Therapeutics. The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-024-02299-2.

Correspondence and requests for materials should be addressed to Anthony T. Papenfuss, Michael I. Love or Stefano Mangiola.

Peer review information *Nature Methods* thanks Bo Li and Judith Zaugg for their contribution to the peer review of this work. Primary Handling Editor: Lei Tang, in collaboration with the *Nature Methods* team.

nature portfolio

Corresponding author(s): PapenFuss, Love, Mangiola

Last updated by author(s): May 1, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	https://github.com/tidyomics/tidyomics_paper/ CuratedAtlasQuery v1.1.3

Data analysis	https://github.com/tidyomics/tidyomics_paper/ Seurat v5.0.1, SingleR v2.4.0, Limma v3.58.1, edgeR v4.0.16, glmmSeq v0.5.5, Ggplot2 v3.5.0, R v 4.3.1, plyranges v1.22.0, tidySingleCellExperiment v1.13.3, tidySpatialExperiment v0.99.13, tidySummarizedExperiment v1.12.0, tidybulk v1.15.4, tidytof v0.0.0, tidyseurat v0.8.0

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Human Celi Atlas PBMC single-cell data was downloaded from the CELLxGENE database (https://cellxgene.cziscience.com/). Metadata and gene-transcript abundance far these datasets were downloaded via the CuratedAtlasQuery R package (available in Bioconductor). The samples analysed are accessible at the Human Cell Atlas portal at the address

https://explore.data.humancellatlas.org/projects?filter=%5B%7B%22categoryKey%22%3A%22specimenOrgan%22%2C%22value%22%3A%5B%22blood%22%2C%22peripheral+blood+mononuclear+cell%22%5D%7D%5D

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	3356
Data exclusions	Cell types without complete or partial consensus were filtered out.

Replication	The biological replication is represented by the samples included in the included datasets.
Randomization	N/A this study is observational.
Blinding	N/A this study is observational.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	N/A
Novel plant genotypes	N/A
Authentication	N/A