https://doi.org/10.1038/s41592-024-02371-x

WebAtlas pipeline for integrated single-cell and spatial transcriptomic data

Check for updates

Single-cell and spatial transcriptomics illuminate complementary features of tissues. Computational integration can synergize these technologies to resolve cell types and transcriptomes in situ. However, online dissemination and exploration of integrated datasets remains challenging. Here, we introduce the WebAtlas pipeline for user-friendly sharing and interactive navigation of integrated single-cell and spatial transcriptomic datasets (Fig. 1a; https://cellatlas.io/webatlas).

Multimodal tissue atlasing datasets pose two key challenges for online dissemination and equitable access. First, single-cell RNA-sequencing (scRNA-seq) and spatial transcriptomics data objects are often saved in non-unified sequencing and imaging file formats that perform poorly with web technologies. Second, existing software platforms do not readily support simultaneous browsing of multiple integrated data modalities.

To address these challenges, we provide 1) a new data ingestion pipeline to convert and unify datasets from multiple single-cell and spatial technologies into the cloud-ready Zarr format[1] (Fig. 1b) and 2) a front-end web client based on the Vitessce framework[2] for interactive exploration and cross-query of gene expression and cell types across modalities (Fig. 1d). WebAtlas allows bioinformaticians and software engineers to build public-facing data portals, as well as non-technical community members to access tissue atlases. (See Supplementary Note 1 for detailed comparison to other platforms.)

Within WebAtlas, single-cell and spatial datasets are linked by biomolecular metadata, such as shared cell-type or gene annotations. Linkage is performed before WebAtlas ingestion using existing data-integration methods like Cell2location[3] and StabMap[4] that map scRNA-seq cell-type references onto spatial transcriptomics datasets and impute unobserved gene expression in the latter (Fig. 1a).

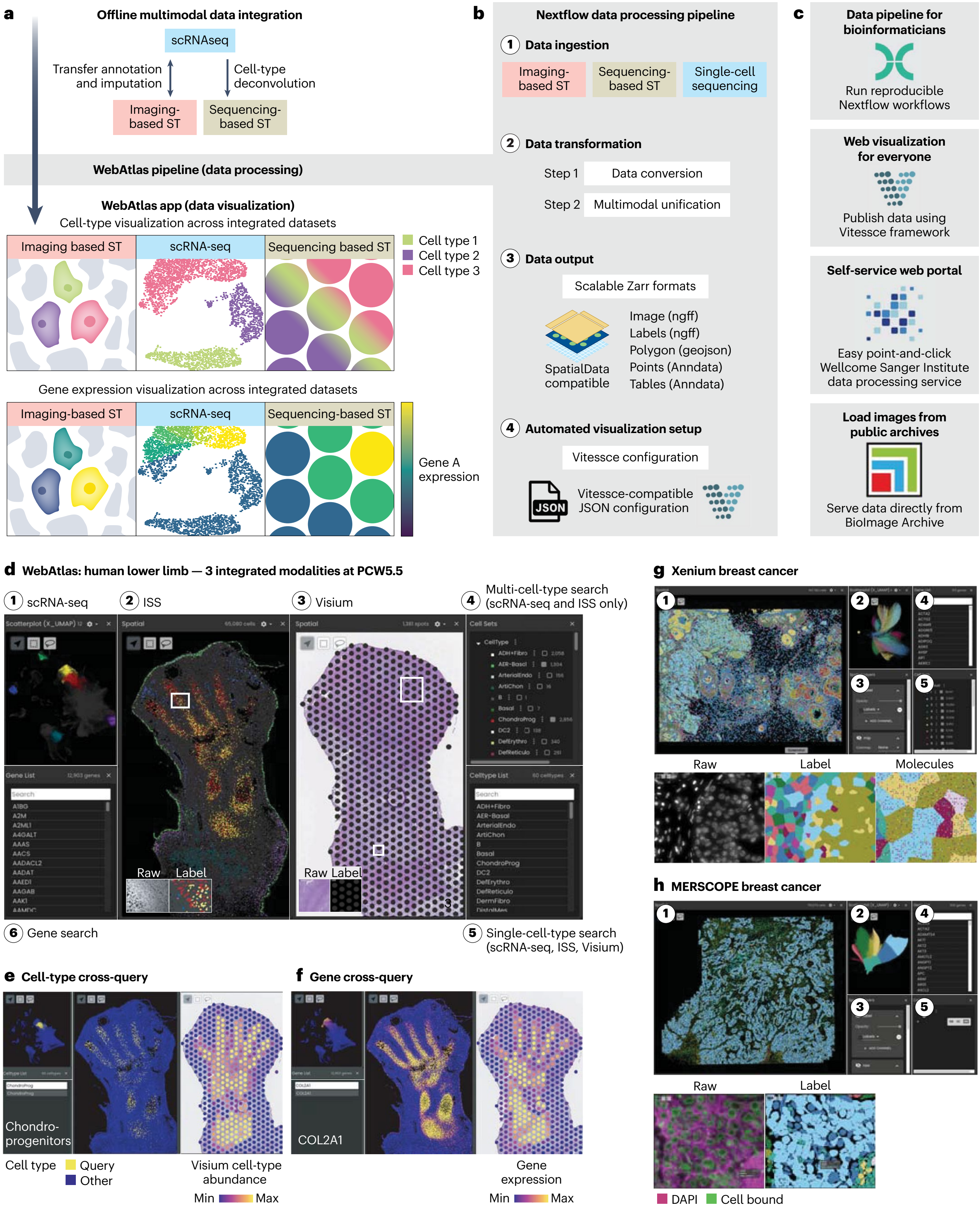
Our data ingestion pipeline performs the extract–transform–load steps for sequencing and imaging data objects generated from various technologies (Fig. 1b). To enable efficient browsing of tissue atlas datasets online, we produce an output using the array-chunked Zarr file format. Tabular gene expression and cell-type annotation files for scRNA-seq and spatial transcriptomics are converted into AnnData-Zarr format. Raw microscopy images and cell segmentation label images from spatial transcriptomics data are converted into OME-Zarr[5] format for multi-scale visualization. Other spatial transcriptomics data elements (for example, RNA molecules or points) are also stored in Zarr. Originally developed as a new format, the WebAtlas Zarr convention is now aligned with the recently released SpatialData[6] format, and WebAtlas can output into SpatialData format.

The WebAtlas data ingestion pipeline is implemented on Nextflow and is configured through a YAML schema that defines input data files and visualization parameters. Input datasets are processed independently, and users can choose to process an individual dataset (for example, Visium), specific dataset components (for example, cell segmentation masks but not raw images) or any given combination of integrated datasets (for example, scRNA-seq and Visium, or scRNA-seq and imaging).

To visualize integrated datasets, we used Vitessce, which provides a serverless web framework for interactive exploration of multimodal data. Currently, preparing data and configuration files for Vitessce requires programmatic expertise and familiarity with multiple data models. WebAtlas automates these steps for coordinated viewing and querying of genes and cell types across integrated datasets (Supplementary Note 2). WebAtlas is comprehensively documented, including sample workflows for common technologies. We also provide a self-service web portal that allows users to upload data files through a web page (Supplementary Note 2).

To showcase WebAtlas, we applied it to a developing human lower limb tissue atlas that integrates public scRNA-seq, Visium spatial RNA-seq and in situ sequencing (ISS) datasets[7,8] (Supplementary Fig. 1a). WebAtlas enabled coordinated navigation of these datasets. We could readily cross-query cell types and compare spatial cell-type locations revealed by Visium data deconvolution to single-cell-resolution cell maps obtained by ISS imaging. We could also cross-query genes, comparing the expression patterns of genes

Fig. 1 | Overview of WebAtlas pipeline. a, WebAtlas incorporates integrated scRNA-seq, imaging- and sequencing-based spatial transcriptomics (ST) datasets for interactive web visualization, enabling cross-query of cell types and gene expression across modalities. **b**, The WebAtlas data pipeline processes diverse data objects from integrated single-cell and spatial technologies. (1) The pipeline can ingest data from image and/or sequencing-based ST technologies and single-cell sequencing methods such as scRNA-seq. (2) Data transformation is a two-step process. First, data are converted into a standard format so they can be handled more easily. Second, the data are filtered, reindexed and concatenated to unify integrated modalities and prepare for data visualization. (3) Data are output in scalable Zarr format, and optionally SpatialData. (4) The setup required to visualize and query shared gene and cell-type features across modalities is prepared, and a Vitessce compatible configuration file is automatically generated through the pipeline. **c**, Key features and intended audience for WebAtlas. **d**, A WebAtlas snapshot visualizes integrated scRNA-seq, ISS and Visium datasets of the human lower limb at post-conception week (PCW) 5.5. Components include the following: (1) UMAP cell-type representation with queried types highlighted, (2) Spatial map of segmented cells in ISS tissue, (3) Visium tissue section with spot label masks, (4) Cell-set search widget, (5) Cell-type search widget and (6) Gene search widget for cross-querying data. **e**, Cell-type cross-query snapshot. Selecting chondroprogenitors via the cell-type search console simultaneously highlights their cell cluster in scRNA-seq and their spatial locations in ISS and Visium datasets. On Visium data, the predicted abundance of chondroprogenitors per Visium spot is shown. **f**, Gene expression cross-query snapshot. Selecting the chondrocyte lineage marker COL2A1 via the gene search console returns its expression pattern in all three modalities, plotted per cell or Visium spot. **g**, WebAtlas app snapshots of the Xenium human breast cancer dataset. Inset panels show raw DAPI images, segmented cell masks and RNA molecules. **h**, WebAtlas app snapshots of MERSCOPE formalin-fixed, paraffin-embedded breast cancer dataset. Inset panels show raw microscopy images and segmented cell masks.


a
Offline multimodal data integration
scRNAseq
Transfer annotation and imputation
Cell-type deconvolution
Imaging-based ST
Sequencing-based ST
WebAtlas pipeline (data processing)
WebAtlas app (data visualization)
Cell-type visualization across integrated datasets
Imaging based ST
scRNA-seq
Sequencing based ST
Cell type 1
Cell type 2
Cell type 3
Gene expression visualization across integrated datasets
Imaging-based ST
scRNA-seq
Sequencing-based ST
Gene A expression
b
Nextflow data processing pipeline
1 Data ingestion
Imaging-based ST
Sequencing-based ST
Single-cell sequencing
2 Data transformation
Step 1 Data conversion
Step 2 Multimodal unification
3 Data output
Scalable Zarr formats
SpatialData compatible
Image (ngff)
Labels (ngff)
Polygon (geojson)
Points (Anndata)
Tables (Anndata)
4 Automated visualization setup
Vitessce configuration
JSON
Vitessce-compatible JSON configuration
c
Data pipeline for bioinformaticians
Run reproducible Nextflow workflows
Web visualization for everyone
Publish data using Vitessce framework
Self-service web portal
Easy point-and-click Wellcome Sanger Institute data processing service
Load images from public archives
Serve data directly from BioImage Archive
d WebAtlas: human lower limb — 3 integrated modalities at PCW5.5
1 scRNA-seq
2 ISS
3 Visium
4 Multi-cell-type search (scRNA-seq and ISS only)
Scatterplot (X_UMAP)
Spatial
Cell Sets
CellType
ADH+Fibro
AER-Basal
ArterialEndo
ArtiChon
B
Basal
ChondroProg
DC2
DefErythro
DefReticulo
Gene List
12,903 genes
Search
A1BG
A2M
A2ML1
A4GALT
AAAS
AACS
AADACL2
AADAT
AAED1
AAGAB
AAK1
Celltype List
60 celltypes
ADH+Fibro
AER-Basal
ArterialEndo
ArtiChon
B
Basal
ChondroProg
DC2
DefErythro
DefReticulo
DermFibro
Raw
Label
Raw
Label
6 Gene search
5 Single-cell-type search (scRNA-seq, ISS, Visium)
e Cell-type cross-query
Chondro-progenitors
Cell type
Query
Other
Visium cell-type abundance
Min
Max
f Gene cross-query
COL2A1
Gene expression
Min
Max
g Xenium breast cancer
1
2
3
4
5
Raw
Label
Molecules
h MERSCOPE breast cancer
1
2
3
4
5
Raw
Label
DAPI
Cell bound

imputed in ISS data to their direct measurements in Visium.

The WebAtlas Zarr file convention, optimized for web visualization, can also support navigation of reference tissue atlas datasets hosted in central repositories. To demonstrate this, we deposited WebAtlas Zarr outputs of scRNA-seq, Visium and ISS limb datasets to the BioImage Archive[9] and visualized them directly on the WebAtlas App (Supplementary Fig. 1a).

Finally, WebAtlas is technology agnostic and scalable. We applied WebAtlas to Xenium (Fig. 1g), MERSCOPE (Fig. 1h) and Visium CytAssist datasets, as well as a mouse embryonic atlas integrating scRNA-seq and seqFISH[10] (Supplementary Table 2 and Supplementary Fig. 2), scaling up to 900,000 cells and 1.1 million RNA molecules (Supplementary Table 2 and Supplementary Fig. 2). Public access to all datasets used here (Supplementary Table 2) is available via our portal.

WebAtlas provides an intuitive pipeline for online exploration of integrated single-cell and spatial transcriptomics, and facilitates the creation of rich and easily accessible tissue atlases for biologists. We envision diverse use cases, including anatomical and pathology annotation of spatial datasets, data dissemination alongside publications, and centralized reference atlases.

Code availability

All software code has been made publicly available on GitHub at https://github.com/haniffalab/webatlas-pipeline. Each software release is permanently archived on Zenodo at https://doi.org/10.5281/zenodo.7405818. Comprehensive documentation, tutorials and sample workflows are available at https://haniffalab.github.io/webatlas-pipeline.

Tong Li[1,8], **David Horsfall**[1,2,8], **Daniela Basurto-Lozada**[1,2,8], **Kenny Roberts**[1], **Martin Prete**[1], **John E. G. Lawrence**[1], **Peng He**[1], **Elisabeth Tuck**[1], **Josh Moore**[3], **Aybuke Kupcu Yoldas**[4], **Kolawole Babalola**[4], **Matthew Hartley**[4], **Shila Ghazanfar**[5,6], **Sarah A. Teichmann**[1,7], **Muzlifah Haniffa**[1,2,9] ✉ & **Omer Ali Bayraktar**[1,9] ✉

[1]Wellcome Sanger Institute, Hinxton, UK. [2]Biosciences Institute, Newcastle University, Newcastle upon Tyne, UK. [3]German BioImaging – Gesellschaft für Mikroskopie und Bildanalyse e.V., Konstanz, Germany. [4]European Bioinformatics Institute, Hinxton, UK. [5]School of Mathematics and Statistics, The University of Sydney, Sydney, New South Wales, Australia. [6]Charles Perkins Centre, The University of Sydney, Sydney, New South Wales, Australia. [7]University of Cambridge, Cambridge, UK. [8]These authors contributed equally: Tong Li, David Horsfall, Daniela Basurto-Lozada. [9]These authors jointly supervised this work: Muzlifah Haniffa, Omer Ali Bayraktar.
✉e-mail: mh32@sanger.ac.uk; ob5@sanger.ac.uk

Published online: 19 August 2024

References

1. zarr-developers/zarr-python: v2.17.1. https://doi.org/10.5281/zenodo.10790679 (2024).
2. Keller, M. S. et al. Preprint at *OSF Preprints* https://doi.org/10.31219/osf.io/y8thv (2021).
3. Kleshchevnikov, V. et al. *Nat. Biotechnol.* **40**, 661–671 (2022).
4. Ghazanfar, S., Guibentif, C. & Marioni, J. C. *Nat. Biotechnol.* **42**, 284–292 (2024).
5. Moore, J. et al. *Histochem. Cell Biol.* **160**, 223–251 (2023).
6. Marconato, L. et al. *Nat. Methods* https://doi.org/10.1038/s41592-024-02212-x (2024).
7. Zhang, B. et al. *Nature* https://doi.org/10.1038/s41586-023-06806-x (2023).
8. Lawrence, J. E. G. et al. Preprint at *bioRxiv* https://doi.org/10.1101/2023.12.20.572425 (2023).
9. Hartley, M. et al. *J. Mol. Biol.* **434**, 167505 (2022).
10. Lohoff, T. et al. *Nat. Biotechnol.* **40**, 74–85 (2022).

Acknowledgements

We thank the authors of the human lower limb atlas study for sharing data, Sébastien Besson for consultation on use of OME-NGFF as a common format, Pavel Mazin for help with Cell2location analysis of the lower limb Visium data, Sanger Institute's Informatics and Digital Solutions team for infrastructure support for hosting the datasets and website, and Jason Swedlow for comments on the manuscript. This work was funded by Wellcome Trust core funding (206194 and 220540/Z/20/A) and Strategic Science Award (221052/Z/20/Z, 221052/A/20/Z, 221052/B/20/Z, 221052/C/20/Z and 221052/E/20/Z) to M.H. and O.A.B., and a Wellcome Trust Senior Research Fellowship Award (223092/Z/21/Z) to M.H. J.M. was supported for work on OME-NGFF by grant numbers 2019-207272 and 2022-310144 and on Zarr by grant numbers 2019-207338 and 2021-237467 from the Chan Zuckerberg Initiative DAF, an advised fund of Silicon Valley Community Foundation, and was funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) 501864659 as part of NFDI4BIOIMAGE. S.G was supported for work on StabMap imputation by an Australian Research Council Discovery Early Career Researcher Award (DE220100964) funded by the Australian Government and a Chan Zuckerberg Initiative Single Cell Biology Data Insights grant (2022-249319).

Author contributions

T.L., D.H., D.B.L., M.H. and O.A.B. conceived the study. T.L., D.H. and D.B.L developed the WebAtlas pipeline and processed the sample datasets. M.P. supported WebAtlas implementation. J.E.G.L., P.H. and S.T. shared limb scRNA-seq and Visium data and contributed to limb data interpretation. K.R., E.T. and S.G. contributed to limb ISS data interpretation. A.K.Y., K.B. and M.H. assisted with BioImage Archive submission of limb datasets. J.M. consulted on the OME-Zarr specifications and reviewed and edited the manuscript. T.L., D.H., D.B.L., M.H. and O.A.B. wrote the manuscript with feedback from all authors.

Competing interests

In the past three years, S.A.T. has consulted for or been a member of scientific advisory boards at Qiagen, Sanofi, GlaxoSmithKline and ForeSite Labs. She is a consultant and equity holder for TransitionBio and EnsoCell. J.M. holds equity in Glencoe Software, which builds products based on OME-NGFF. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-024-02371-x.

Peer review information *Nature Methods* thanks Nils Gehlenborg and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.