BioHackrXiv

VIB Hackathon on spatial omics tools and methods

Benjamin Rombaut[1,2,3], Lotte Pollaris[1,2,3], Chananchida Sang-aram[1,2,3], Michiel Ver Cruysse[1,3], Robrecht Cannoodt[5,1,2], Frank Vernaillen[4], Arne Defauw[4], Julien Mortier[4], Luuk Harbers[8], Miguel A. Ibarra-Arellano[6], Kresimir Bestak[6], Aroj Hada[6,7], Vladislav Vlasov[9], Michele Bortolomeazzi[10], Paul Kiessling[11], Alexander Sudy[12], Wouter-Michiel Vierdag[13], Miray Cetin[14], Lotte Van de Vreken[15], Quentin Blampey[16], Anastasiia Okhtienko[17], Daniel Dimitrov[6], Mayar Ali[18,19], Francesca Drummer[18, 20], Benedetta Manzato[21], Carlos Ariel Pulido-Vicuna[22,8], Susmita Mandal[23], Giovanni Palla[18], Laurens Lehner[18], Lorenzo Giordani[24], Claudio Novella-Rausell[21], . . .[1], and Yvan Saeys[1,2,3]

1 Data Mining and Modelling for Biomedicine, VIB-UGent Center for Inflammation Research, Ghent, Belgium **2** Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium **3** VIB Center for AI and Computational Biology, Ghent, Belgium **4** VIB Spatial Catalyst **5** Data Intuitive, Lebbeke, Belgium **6** Institute for Computational Biomedicine, Faculty of Medicine, Heidelberg University Hospital, Heidelberg, Germany **7** AI-Health Innovation Cluster, Heidelberg, Germany **8** VIB-KU Leuven Center for Cancer Biology, Leuven, Belgium **9** Brain and Systems Immunology Lab, Brussels Center for Immunology, Vrije Universiteit Brussel **10** ScOpen Lab, German Cancer Research Center (DKFZ), Heidelberg, Germany **11** RWTH Aachen, University Hospital **12** Center of Digital Health, Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Germany **13** European Molecular Biology Laboratorium, Heidelberg, Germany **14** Systems Immunology and Single-Cell Biology, German Cancer Research Center (DKFZ), Heidelberg, Germany **15** VIB-UGent Center for Plant Systems Biology, Ghent, Belgium **16** MICS Laboratory, CentraleSupélec, Paris-Saclay University, Paris, France **17** Institute of Virology,Technical University of Munich, Munich, Germany **18** Institute of Computational Biology, Helmholtz Munich, Neuherberg, Germany **19** Institute for Tissue Engineering and Regenerative Medicine, Helmholtz Munich, Neuherberg, Germany **20** Institute for Stroke and Dementia Research, Klinikum Der Universität München, Ludwig-Maximilians-Universität, Munich, Germany **21** Department of Human Genetics, Leiden University Medical Center, Leiden 2333ZC, The Netherlands **22** Laboratory for Molecular Cancer Biology, Center for Cancer Biology, VIB, Leuven, Belgium; Department of Oncology, KU Leuven, Leuven, Belgium. **23** Institute of Pathology at Charité – Universitätsmedizin Berlin, Germany **24** Sorbonne Université, INSERM UMRS 974, Association Institut de Myologie, Centre de Recherche en Myologie, 75013 Paris, France.

BioHackathon series:
COVID-19 BioHackathon
Virtual conference 2020
Code repository

Submitted: 12 Jun 2024

License:
Authors retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC-BY).

Published by BioHackrXiv.org

Introduction

During a three-day hackathon, work was performed on various topics within the field of spatial omics data analysis. The topics were organized in five workgroups and included benchmarking, pipelines, spatial transcriptomics, spatial proteomics, spatial multi-omics and cell-cell communication. Most tools and methods were considered in the context of the analysis ecosystem in Python for spatial (Marconato et al., 2024) and single-cell (Virshup et al., 2023) data analysis.

Results

Results were summarized in a final slide deck. A project board collected all task items and GitHub Issues. Here we give a brief overview for each of the five workgroups.

Workgroup pipelines

Nextflow During this hackathon, we have worked on and finished the template update for nf-core/molkart, an nf-core pipeline for processing Molecular Cartography data, allowing for the next expansion that will include spot-based segmentation options. Additionally, we have finished adding Spotiflow, a FISH spot-detection tool into the nf-core framework.

- Infrastructure for pipelines:
  - Support for incremental IO (partial read/write) in SpatialData
  - Support for apply function in SpatialData
  - Use Viash to create a Nextflow job to view spatial omics datasets
- Specific issues:
  - improve performance of isoquant for large spatial omics datasets
  - Build a computational benchmark for spatial omics data
    - identify datasets
    - identify first becnmarks
- Accessing remote datasets:
  - Upload spatial omics datasets to S3
  - Support for private remove object storage in SpatialData

Workgroup spatial transcriptomics

Napari plugin

Napari is a scalable interactive viewer for multi-dimensional data. It works natively in python. Within this hackathon, we worked on adding functionality to napari-spatialdata, a SpatialData plugin for napari. Firstly, we worked on reusing colors previously defined in the SpatialData object. Secondly, progress had been made to only visualize subsets of the cells. This would allow to plot a certain celltype colored by gene expression of gene x and another celltype colored by gene expression y. Thirdly, work on the annotation wdget has been performed and checked.

Annotation workflows

We discussed user stories for a workflow that entails drawing annotations interactively with Napari and using the annotations in downstream analysis steps. To this end, we identified the following tasks that would enable such workflow: - napari-spatialdata widget that would enable: 1. Drawing annotations on a specific image or coordinate system. 2. Rename the annotations, specifying various metadata to the annotation, such as the identity of the annotator, labels for the annotations and others. 3. Save the annotations back to the spatialdata

Visium HD on-the-fly rasterization

As mentionned in this SpatialData issue, Visium HD data can't be rasterized in memory. Still, for visualization and analyses purposes, rasterization is needed. Therefore, we opened a new PR for bins rasterization, on which we support two modes: - rasterization of one or multiple channels (in-memory). It uses the indices of the sparse table in CSC format for efficiency. - lazy rasterization of the full data with dask (in particular, using map_blocks). The data is therefore rasterized when needed, for instance to display one or a few channels in napari-spatialdata.

Visium HD and Xenium * Available Xenium and Visium HD dataset: https://www.10xgenomics.com/products/visium-hd-spatial-gene-expression/dataset-human-crc from https://www.biorxiv.org/content/10.1101/2024.06.04.597233v1 * Aligning the Xenium and Visium HD dataset * Label transfer from scRNA-seq data to the spatial data * Merging spatialdata objects of Xenium and Viisum HD * Microenvironment detection using Banksy (https://github.com/prabhakarlab/Banksy_py)

Cellular niches Multiple unsupervised metrics have been added in this Squidpy PR to evaluate niches detection methods. Notably: - a niche continuity metric - a cross-slide homogeneity metric - DE tests to compare max gene expression across niches - ARI, NMI and Fowlkes-Mallows Index for niche result comparison (agreement)

Workgroup spatial proteomics

Group members had most experience with analysis of Miltenyi MACSima, Akoya Phenocycler, Lunaphore COMET and MIBI data.

Some common issues in spatial proteomics analysis were discussed. Reading in datasets in the SpatialData format still lacks for some platforms. Some interesting metadata is also included always included, such as physical pixel size, autofluorescence subtraction, imaging cycles and exposure time. The need in some datasets to detect misalignment and co-register the channel images, either all of them or specific ones. For segmentation, applying CLAHE and using cellpose was found to be sufficient for most cells. For exceptional cell shapes in tissues such as the heart and brain there is additional difficulty and need for fine-tuning the segmentation model with enough training data. This manual labeling is time-consuming and difficult to reproduce.

There was a lack of consensus on available normalization techniques and batch effect correction and their usefullness.

Four work items were selected:

1. Support for exporting cells in SpatialData and interactively annotating them using a classifier with Ilastik software (Berg et al., 2019).
2. Normalization facilitates the integration and comparison of data from different experiments, which is essential for large-scale studies and meta-analyses such as spatial omics data. Therefore, creation of an overview of normalization methods for downstream analysis of spatial proteomics datasets and a comparison between them is crucial. While evaluation & benchmarking would require a gold standard cell type dataset which is beyond the scope of this hackathon, a repository was created at https://github.com/SchapiroLabor/norm_methods/. that contains a summary of 9 methods adapted from published literature. All codes for each method are also available. A visualization of results obtained from these different methods on a MIBI dataset (not publicly available) is provided as well. Among the different methods, a visual qualitative comaprison provides evidence that a combined method (Shaban et al. + Greenbaum et al.) may yield more promising results. We plan to extend the work from this hackathon with a quantitative comparison in the future.
3. An alternative to `spatialdata.to_polygons()` label vectorization function, which features improved performance, resolution of the invalid geometries, and `shapely.MultiPolygon` geom filtering based on the area.
4. Creating a new reader for MACSima datasets in spatialdata-io.

Workgroup spatial multi-omics

Day 1: introduction

Multi-omics often requires doing manual/automated image registration as a first step - find open datasets - same / consecutive section - same / different omics modality: - try out and compare existing registration tools

Morphological features: - Do they present bigger/smaller batch effects between slides compared to molecular features? - Do they correlate with molecular features / how well? - Can they be used as anchors for diagonal integration?

Day 1: hacking

Put data here: /dodrio/scratch/projects/starting_2024_011/multi-omic/datasets/

Potential methods for morphology extraction:

- HEIP
- UNI
- Resnet50 example
- ScDino (Immuno fluorescence)

Spatial transcriptomics + Morphology:

- Visium HD Cancer Colon: Raw data, Nuclei Segmentation + Domains,Preprint
- Xenium Lung Cancer: Spatialdata,Raw data
- Xenium Breast Cancer: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE243168
- Merfish RNA + IF How to dowload
- List of Visium, Xenium human cancer datasets: https://spatialdata.scverse.org/en/latest/tutorials/notebooks/datasets/README.html
- Morphology features tutorial squidpy (tensorflow) https://squidpy.readthedocs.io/en/stable/notebooks/tutorials/tutorial_tf.html

Multi-omics datasets (same/different slides):

- SPOTS with the 10x Visium technology capturing whole transcriptomes and extracellular proteins https://doi.org/10.1038/s41587-022-01536-3, GSE198353. High-resolution images (https://figshare.com/account/home#/projects/143019)
- Stereo-CITE-seq spatial transcriptomics + proteomics (https://doi.org/10.1101/2023.04.28.538364)
- spatial transcriptomics + DVP proteomics (https://doi.org/10.1038/s41593-022-01097-3)
- Spatial-ATAC-RNA-seq (https://doi.org/10.1038/s41586-023-05795-1)
- Cite-seq, proteogenomics (https://doi.org/10.1016/j.cell.2021.12.018)
- spatial CITE-seq transcriptomics+proteomics (https://doi.org/10.1038/s41587-023-01676-0)
- Benchmark datasets for 3D mass spec imaging (=2D Mass spec imaging on adjacent sections) (https://academic.oup.com/gigascience/article/4/1/s13742-015-0059-4/2707545)
- https://doi.org/10.1038/s41467-023-43105-5 (suppl table 1, collection of publicly available datasets from different studies)
- spatial-ATAC and the spatial RNA-seq (MISAR-seq, https://doi.org/10.1038/s41592-023-01884-1)
- Mass spec imaging + spatial transcriptomics (Visium): https://www.nature.com/articles/s41587-023-01937-y (see data availability, e.g. https://data.mendeley.com/datasets/w7nw4km7xd/1, sma zip file)

Data integration

Challenges: - number of detected features (e.g. RNA-seq VS proteomics) - different feature counts, statistical distributions - differences in resolution (imaging-based) - image alignment/overlay (imaging-based) - batch effect - technical (heavy data)

Horizontal

merging the same omic across different datasets Reasons: - 3D maps - technical replicates, integrating batches - integrating across different technologies

not true multi-omics integration

Examples: - STAGATE (spatial transcriptomics, consecutive sections, adaptive graph attention auto-encoder, https://doi.org/10.1038/s41467-022-29439-6) - STAligner (spatial transcriptomics datasets, batch effect-corrected embeddings, 3D reconstruction, https://

doi.org/10.1038/s43588-023-00543-x) - SpaGCN (spatial transcriptomics, graph convolutional network approach that integrates gene expression, spatial location and histology, https://doi.org/10.1038/s41592-021-01255-8) - PASTE (align and integrate ST data from multiple adjacent tissue sections) https://www.nature.com/articles/s41592-022-01459-6 - SpaceFlow (embedding is continuous both in space and time, Deep Graph Infomax (DGI) framework with spatial regularization, https://doi.org/10.1038/s41467-022-31739-w)

Vertical

merges data from different omics within the same set of samples (matched integration) Anchor - cell Examples: - archr (https://doi.org/10.1038/s41588-021-00790-6, https://doi.org/10.1073/pnas.211002511) - MaxFuse (fuzzy smoothed embedding for weaky-linked modalities, proteomics, transcriptomics and epigenomics at single-cell resolution on the same tissue section https://doi.org/10.1038/s41587-023-01935-0) - MultiMAP (nonlinear manifold learning algorithm that recovers a single manifold on which several datasets reside and then projects the data into a single low-dimensional space so as to preserve the manifold structure, https://doi.org/10.1186/s13059-021-02565-y) - Seurat5

Diagonal

Different cells/consecutive slides/different studies (unmatched integration) Examples:

- SpatialGlue (https://doi.org/10.1101/2023.04.26.538404)
  - graph neural network with dual-attention mechanism
  - 2 separate graphs to encode data into common embedding space: a spatial proximity graph and a feature graph
  - Spatial-epigenome-transcriptome, Stereo-CITE-seq, SPOTS, and 10x Visium (to be continued)
  - python script and a set of jupyter notebooks with examples
  - need all data in adata .h5ad format (using scanpy)
  - calling R from Python
- MEFISTO (https://doi.org/10.1038/s41592-021-01343-9)
  - factor analysis + flexible non-parametric framework of Gaussian processes
  - spatio-temporally informed dimensionality reduction, interpolation, and separation of smooth from non-smooth patterns of variation.
  - different omics, multiple sets of samples (different experimental conditions, species or individuals)
  - each sample is characterized by "view", "group", and by a continuous covariate such as a one-dimensional temporal or two-dimensional spatial coordinate
  - no examples of real spatial multi-omics integration
  - integrated into the MOFA framework (in R)
- SLAT (https://doi.org/10.1038/s41467-023-43105-5)
  - aligning heterogenous spatial data across distinct technologies and modalities (is it so?)
  - single-cell spatial datasets
  - graph adversarial matching
  - benchmarked on 10× Visium, MERFISH, and Stereo-seq
- https://doi.org/10.1038/s41467-024-47883-4

Tool	Method	Data compatible/ benchmarked	Type of integration	Installation	Details on usage	Link to Github	other
SpatialGlue	GNN	Stereo-CITE-seq, SPOTS, 10x Visium + protein co-profiling, transcriptome-epigenome, generated data	linked data	PyPI (runs ok in conda)	rpy2 issues in env, all data should be in .h5ad	https://github.com/JinmiaoChenLab/SpatialGlue	returns attention weights for modalities
MEFISTO	factor analysis	generated data, 10x Visium, no examples of real integration	-	part of MOFA	-	https://biofam.github.io/MOFA2/MEFISTO.html	weights for factors (genes)
SLAT	GNN	aligning 2 Stereo-seq slices, 3D reconstruction from 4 Stereo-seq slices, 10x Xenium and 10x Visium alignment	cross-technology alignment, different slices	docker, PyPI	all data should be in .h5ad, requires manual preprocessing of the data	https://github.com/gao-lab/SLAT	notebooks with options for downstream analysis

General issue: gene-based, challenges with proteomics (and even more issues with metabolomics). Direct comparison of the tools is not possible due to different tasks and working principles. ### *In silico* datasets generation Experimental design planning; sampling strategy; statistics; tools benchmarking - https://www.nature.com/articles/s41592-023-01766-6 - tissue scaffold: random-circle-packing algorithm to generate a planar graph - attributes on nodes represent cell type assignments - the labeling is based on two data-driven parameters (prior knowledge) for a tissue type: the proportions of the k unique cell types, and the pairwise probabilities of each possible cell type pair being adjacent (a k × k matrix) - by changing these 2 params one should be able to obtain simulations for different tissues and technologies - ! Quite buggy in installation & running - scDesign3 https://www.nature.com/articles/s41587-023-01772-1 - SRTsim (transcriptomics only) https://doi.org/10.1186/s13059-023-02879-z

Image Registration

Spatial landmark detection and tissue registration with deep learning. Paper: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11009106/ Code: https://github.com/ekvall93/ELD

Misc:

Data used in STalign paper: https://www.nature.com/articles/s41467-023-43915-7#data-availability

Data used in CAST. Link to data doesn't work.

Papers

- Integration of Multiple Spatial Omics Modalities Reveals Unique Insights into Molecular Heterogeneity of Prostate Cancer Spatial transcriptomics and Mass spec imaging were performed on adjacent sections, and registered via their respective H&E images. The datasets are not publically available.
- Search and Match across Spatial Omics Samples at Single-cell Resolution
- https://frontlinegenomics.com/a-guide-to-multi-omics-integration-strategies/

Workgroup cell-cell communication

The goal of the group was to run multiple spatial CCC methods, compare evaluations/visualizations and results. We selected the methods from Armingol et al., 2024. Table is at

Results:

Methods were implemented and tested on a subset of the MERFISH whole mouse brain data (slice 80) from the Allen Brain Institute.

We obtained results for CCC for the following methods: COMMOT, SpatialDM, MEBOCOST, CellPhoneDB. SpatialDM, CellPhoneDB was run with LIANA+. Because MEBOCOST does not use Ligand-Recpetor interactions as the other tools we could not compare the results directly. We also ran SpaTalk but found no LR pairs, as the tool requires that the entire ligand-receptor-tf-target pathway is expressed for a LR pair to be considered, and this was likely not the case in a dataset with 1122 genes. For the other three tools we selected specific LR pairs to compare the results.

1. Comparision on cell type level

Q: Do the tools identify the same sender and receiver cells that participate at communication?

2. Comparison on spatial level

Where do the tools predict the communication to take place in tissue space? Do spatial methods benefit from the additional modality?

Discussion: - Comparison of results is difficult because i) there is no ground thruth regarding CCC, ii) output formats of methods vary, for example SpatialDM returns a $NxLR$ matrix with a score for each cell indicating the potential strength of a ligand or receptor and COMMOT returns a NxN matrix for each LR interaction, iii) different score metric - Different input databases on which communication analysis is based (metabolic vs ligand-receptor) but also within LR interactions it might use the CellPhoneDB or CellChat database

Papers:

- Armingol,E.,Baghdassarian,H.M.&Lewis,N.E.Thediversificationofmethodsforstudyingcell–cellinterac NatRevGenet1–20(2024)doi:10.1038/s41576-023-00685-8.
- Screeningcell–cellcommunicationinspatialtranscriptomicsviacollectiveoptimaltransport

Discussion

[Main general takeaways for the field and future outlook]

Links

Status updates and results were summarized in a slide deck. A project board collected all task items and a Zulip stream was used for communication. Code to use the computational resources was made available in a git repository.

Acknowledgements

[For every participant: sponsors, (travel) grants, infrastructure used. . .]

The hackathon was organized by the Saeys Lab and supported by Data Intuitive, the VIB Spatial Catalyst and the VIB Center for AI and Computational Biology.

The computational resources and services used in this work were provided by the VIB Data Core and the VSC (Flemish Super-computer Center), funded by the Research Foundation – Flanders (FWO) and the Flemish Government. B.R is supported by the Flanders AI Research Program.

References

Berg, S., Kutra, D., Kroeger, T., Straehle, C. N., Kausler, B. X., Haubold, C., Schiegg, M., Ales, J., Beier, T., Rudy, M., Eren, K., Cervantes, J. I., Xu, B., Beuttenmueller, F., Wolny, A., Zhang, C., Koethe, U., Hamprecht, F. A., & Kreshuk, A. (2019). Ilastik: Interactive machine learning for (bio)image analysis. *Nature Methods*, *16*(12), 1226–1232. https://doi.org/10.1038/s41592-019-0582-9

Marconato, L., Palla, G., Yamauchi, K. A., Virshup, I., Heidari, E., Treis, T., Vierdag, W.-M., Toth, M., Stockhaus, S., Shrestha, R. B., Rombaut, B., Pollaris, L., Lehner, L., Vöhringer, H., Kats, I., Saeys, Y., Saka, S. K., Huber, W., Gerstung, M., . . . Stegle, O. (2024). SpatialData: An open and universal data framework for spatial omics. *Nature Methods*, 1–5. https://doi.org/10.1038/s41592-024-02212-x

Virshup, I., Bredikhin, D., Heumos, L., Palla, G., Sturm, G., Gayoso, A., Kats, I., Koutrouli, M., Berger, B., Pe'er, D., Regev, A., Teichmann, S. A., Finotello, F., Wolf, F. A., Yosef, N., Stegle, O., & Theis, F. J. (2023). The scverse project provides a computational ecosystem for single-cell omics data analysis. *Nature Biotechnology*, *41*(5), 604–606. https://doi.org/10.1038/s41587-023-01733-8