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[1], [1], [11], . . . [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

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

Submitted: 12 Jun 2024

Published by BioHackrXiv.org

Introduction

[Main goal of the hackathon and setting](Marconato et al., 2024)

Results

[Main outcomes]

Workgroup pipelines

- Nextflow:
  - nf-core/molkart template update
  - nf-core spotiflow module
  - nf-core stardist module
  - Spot2cell python+conda+docker+nf-core
- 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 outcomes]

Workgroup spatial transcriptomics

[Workgroup outcomes]

Workgroup spatial proteomics

[Workgroup outcomes]

Common issues in spatial proteomics analysis were found to be: support for reading data from specific platforms, with support for physical pixel size, cycle and exposure time. One topic would be to have a reader for MACSima.

- readers:
  - MACSima reader
    - stack of tiffs misaligned
    - specific channels not well aligned
  - Akoya reader
  - Lunaphore COMET
    - autofluorescence image
    - channel, is it already substracted?
  - MIBI, heavy metals
    - BIN files, toffy software
    - ark-analysis
    - Mesmer
    - retraing cellpose with tisuenet dataset
    - IMCDataAnalysis, also using, but also move to Python
  - Slidescanner

Some data loading issues were misaligned tiles, misaligned channels. Some favorite tools of the participants are: TissUUmaps, scimap, Hydra config, snakemake, Nextflow.

For segmentation, a problem was

common issues: normalization -

segmentation: - CLAHE - cellpose, fine-tune - random subset

feature calculation: - medians

batch problems: - correction: Harmony - check visualy - maybe can overcorrect if using LogNorm -

spatial spillover, double positive - Starling - REDSEA - Stellar

Weird cells tissue: heart, brain, difficult cell shapes

segmentation, intensities of cells, clustering get poor results solutions: - Ilastik object classifier, semi-supervised classifier - downside: get data back in https://github.com/orgs/saeyslab/projects/5/views/4?pane=issue&itemId=65964689 - use napari-ilastik:

spatial spillover of membrane bound markers: - just label until things look good - manually label all edge cases - downside: manual, reproduce to other sample

Workgroup spatial multi-omics

[Workgroup outcomes]

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) - 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, conda (runs ok)	rpy2 issues, all data should be in .h5ad	https://github.com/JinmiaoChenLab/SpatialGlue?tab=readme-ov-file	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							

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

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

[Workgroup outcomes]

Discussion

[Main general takeaways for the field and future outlook]

Acknowledgements

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

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.

Supplemental information

Tables and figures

References

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 **[cito:usesMethodIn]**