Mathematical Frameworks for Integrative Analysis of Multi-omics Biological Data

This manuscript (permalink) was automatically generated from BIRSBiointegration/whitePaper@12bcac0 on June 19, 2020.

Authors

- **Kim-Anh Lê Cao**
  0000-0003-3923-1116 · mixOmicsTeam · mixOmicsTeam
  Melbourne Integrative Genomics, School of Mathematics and Statistics, University of Melbourne, Australia · Funded by Grant National Health and Medical Research Council Career Development fellowship (GNT1159458)

- **Aedin C Culhane**
  0000-0002-1395-9734 · aedin · AedinCulhane
  Data Sciences, Dana-Farber Cancer Institute, Boston, MA, USA; Biostatsitics, Harvard TH Chan School of Public Health, Boston, MA, USA · Funded by Chan Zuckerberg Initative, NIH, DoD (need to get grant IDs)

- **Elana Fertig**
  0000-0003-3204-342X · ejfertig · FertigLab
  Department of Oncology, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University School of Medicine, Baltimore, MD, USA; Department of Biomedical Engineering, Johns Hopkins University School of Medicine, Baltimore, MD, USA; Department of Applied Mathematics and Statistics, Johns Hopkins University Whiting School of Engineering, Baltimore, MD, USA · Funded by National Institute of Health, National Cancer Institute; National Institute of Health, National Institute of Dental and Craniofacial Research; Lustgarten Foundation; Emerson Foundation; Allegheny Health Network

- **Jane Roe**
  XXXX-XXXX-XXXX-XXXX · janeroe (PLEASE COPY/PASTE DO NOT EDIT THIS ONE) · XXX
  Department of Something, University of Whatever; Department of Whatever, University of Something · Funded by XX

Abstract

Introduction to single cell and imaging multi-omics

Current multi-omic technologies

Challenges for interpretation

Need for technology-specific questions and analysis methods vs one size fits all data blender

Case studies

scNMT-seq as a case-study for epigenetic regulation

Overview and biological question

Computational challenges

- Identification of multi-omics signatures that characterise lineage, stage or both.
- Handling missing values
- Do epigenetic changes in some genomic contexts affect cell fate decision more than others? If so, how?

Methods for stats/maths analyses and results summary

scRNA-seq + FISH as a case study for spatial transcriptomics

Overview and biological question

Computational challenges

- Can scRNA-seq data be overlaid onto seqFISH for resolution enhancement
- What is the minimal number of genes needed for data integration?
- Are there signatures of cellular co-localization or spatial coordinates in the non-spatial scRNA-seq data?

Methods for stats/maths analyses and results summary

Spatial proteomics and cross-study analysis

Overview and biological question

Computational challenges

- Integrating partially-overlapping proteomic data collected on different patients with similar phenotypes
- Integration of spatial x-y coordinate co-location and co-expression
- Integration with other 'omics datasets (e.g., scRNA-seq) to support the results of these proteomic analyses
- Can we predict the spatial expression patterns of proteins measured on mass-tag but not measured in the MIBI-TOF data?
- What additional information can we learn about the different macrophage and immune populations in breast cancer by conducting integrated analyses of these datasets?

Methods for stats/maths analyses and results summary

Overview of common analytical methods spanning technologies / case studies

- matrix factorization
- neural network / autoencoders

Data structures and software packages to enable analyses

Basic objectives:

- Support development of efficient data containers/data services, balancing abstraction with concrete responsiveness to new biotechnologies
- Support software developers in achievement of scalability in their tools
- Support end-users in adoption and successful use

A software ecosystem for multimodal single cell genomics

- Defining the ecosystem; scope and principles of evolution
- Current content targets
  - spatial transcriptomics
  - scNMT-seq ...
- Governance concepts

Containers for developers and users

Details of working components

- Bioc/multiAssayExperiment for single cell
- new classes for proteomics
- PyTorch

Techniques and challenges for benchmarking methods

- realistic simulation studies
- cross-validation, issues in matching dimensions of latent space across folds
- cross-study validation
- benchmark datasets

Relevant citations to include as literature review on benchmarking multi-modal methods:

- Fertig 2012 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3460736/
- Haibe-Kains 2012 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3283537/
- Meng 2019 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6692785/
- Pratapa 2020 https://www.ncbi.nlm.nih.gov/pubmed/31907445

Other links:

- Levi Waldron's benchmarking repo https://github.com/waldronlab/awesome-bioinformatics-benchmarks
- Mike's reviews / evaluation section of `awesome-multi-omics` https://github.com/mikelove/awesome-multi-omics#multi-omics-reviews--evaluations

Discussion

Emerging analytical methods and technologies

Community needs for data structures, analysis methods, etc

References