Community-wide hackathons establish foundations for emerging single cell data integration

This manuscript (permalink) was automatically generated from BIRSBiointegration/whitePaper@1b91952 on September 1, 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

- **Vincent J. Carey**
  0000-0003-4046-0063 · vjcitn
  Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School · Funded by National Institutes of Health, National Human Genome Research Institute; National Institutes of Health, National Cancer Institute; Chan-Zuckerberg Initiative

- **Ricard Argelaguet**
  0000-0003-3199-3722 · rargelaguet · RArgelaguet
  European Bioinformatics Institute (EMBL-EBI) · Funded by EMBL PhD programme

- **Susan Holmes**
  0000-0002-2208-8168 · spholmes · SherlockpHolmes
  Department of Statistics, Stanford University, USA · Funded by National Institute of Health, NIAID

- **Michael I. Love**
  0000-0001-8401-0545 · mikelove · mikelove
  Department of Biostatistics, UNC-Chapel Hill; Department of Genetics, UNC-Chapel Hill · Funded by National Institutes of Health, National Human Genome Research Institute; National Institutes of Health, National Institute of Mental Health

- **Guo-Cheng Yuan**
  XXXX-XXXX-XXXX-XXXX · gcyuan · gc_yuan
  Department of Pediatric Oncology, Dana-Farber Cancer Institute; Department of Pediatrics, Harvard Medical School · Funded by National Institutes of Health, National Cancer Institute; National Institutes of Health, National Institute on Aging

- **Al J Abadi**
  0000-0002-4146-2848 · ajabadi · aljabadi
  Melbourne Integrative Genomics, School of Mathematics and Statistics, University of Melbourne, Australia · Funded by Chan Zuckerberg initiative (HCA2-A-1708-02277); Australian Research Council (DP200102903)

- **Casey S. Greene**
  0000-0001-8713-9213 · cgreene · greenescientist
  Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA · Funded by National Institutes of Health, National Cancer Institute (R01 CA237170); National Institutes of Health, National Human Genome Research Institute (R01 HG010067)

- **Ayshwarya Subramanian**
  0000-0002-4134-7612 · ayshwaryas · ayshwaryas
  Klarman Cell Observatory, Broad Institute; Kidney Disease Initiative, Broad Institute

- **Amrit Singh**
  0000-0002-7475-1646 · singha53 · asingh_22g
  Department of Pathology and Laboratory Medicine, University of British Columbia; PROOF Centre of Excellence · Funded by Michael Smith Foudation for Health Research; Mitacs

- **Alexis Coullomb**
  0000-0003-3164-6640 · alexcoul · AlexisCoullomb
  Toulouse Cancer Research Center (CRCT), INSERM · Funded by Pierre Fabre

- **Arshi Arora**
  0000-0002-4040-1787 · arorarshi · arorarshi
  Department of Epidemiology and Biostatistics, Memorial Sloan Ketterign Cancer Center

- **Joshua S. Sodicoff**
  0000-0001-5182-960X · jsodicoff
  Department of Computational Medicine and Bioinformatics, University of Michigan; Department of Biomedical Engineering, University of Michigan

- **Joshua D. Welch**
  [0000-0002-5869-2391](https://orcid.org/0000-0002-5869-2391) · [jw156605](https://github.com/jw156605) · [LabWelch](https://twitter.com/LabWelch)
  Department of Computational Medicine and Bioinformatics, University of Michigan; Department of Computer Science and Engineering, University of Michigan · Funded by National Human Genome Research Institute, NIH; National Institute of Allergy and Infectious Diseases, NIH; National Institute of Mental Health, NIH

- **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

List of Figures, Tables and online resources

Figures

Tables

Online resources (optional as referred above)

Abstract

Multi-omics hackathon studies illustrate standards and computational challenges in cell biology

Single-cell multimodal omics has claimed the title of method of the year only six years after single-cell sequencing [1], demonstrating the rapid pace of technological development in biology. Multi-omics technologies provide a unique opportunity to fully characterize biological systems at the spatial and molecular levels. While each data modality can uniquely resolve specific biological scales and biological questions, complementary data integration techniques can resolve multi-scale interactions. At the same time, multi-modal omics technological advances have coincided with the formation of tremendous new data resources and the formation of Atlas based initiatives to characterize biological systems. However, despite the potential of single cell multi-omics technologies, computational techniques and benchmark strategies to integrate datasets across high-throughput measurement technologies remain an active area of research. While these emerging data cna unravel novel biological processes, the central question is whether there are optimal analysis methods that can be carried forward, and what are the new developments needed for efficient and insightful statisticial and mathematical analyses.

The hackathon studies we selected for our Mathematical Frameworks for Integrative Analysis of Emerging Biological Data workshop tailored independent challenges oand raised topical data integration challenges. The first challenge included spatial molecular profiling. While this technology is rapidly emerging, it often provides lower molecular resolution than their non-spatial counterparts. Integration strategies that merge spatial and omics datasets have the promise to enhance the molecular resolution of spatially resolved profiling. To address this challenge, we designed a hackathon using spatially resolved transcriptional data from seqFISH with corresponding non-spatial single cell profiling data from the mouse visual cortex [2]. The second challenge dealt with the limited availability of tissue to obtain multiple measurements in samples from identical conditions, raising the question as to whether information can be transferred from datasets between disparate sample cohorts. Therefore, we designed the second hackathon to contain two triple negative breast cancer cohorts profiled with single cell proteomics profiling from mass cytometry (CyTOF) [4] and spatial in-situ proteomics from Multiplexed Ion Beam Imaging (MIBI) [5]. Contrary to the latter challenges, the third challenge presented data at different molecular scales but on the same cells, to investigate how genetic and epigenetic alterations to DNA further drive the transcriptional regulation that mediates intra- and inter-cellular signaling processes underlying cellular fate transitions and states. Thus, our third hackathon was designed with scNMT-seq data to obtain concurrent DNA methylation, chromatin accessibility, and RNA expression from the same cells to delineate the regulatory networks that underlie mouse gastrulation [6].

Altogether, the analysis approaches that were employed to address these disparate hackathons across biological contexts provided an unique opportunity to identify technology-specific challenges and unifying themes that are essential to effectively employ multi-omics datasets into new biological knowledge. This article presents the study-specific and common challenges faced during this workshop, provide guidelines and articulate needs on technologies, data, tools and methods ato model the multi-scale regulatory processes in biological systems in the area of computational biology.

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

Overview and biological question

The first hackathon aimed to leverage the complementary strengths of sequencing and imaging based single-cell transcriptomic profiling with computational techniques to integrate scRNA-seq and seqFISH data in the mouse visual cortext. While single cells are considered the smallest units and building blocks of each tissue, they still require proper spatial and structural three-dimensional organization in order to assemble into a functional tissue that can exert its physiological function. In the last decade, single-cell RNA-seq (scRNA-seq) has played a key role to capture single cell gene expression profiles, allowing us to to virtually map all the different cell types and states in whole organisms. Despite this remarkable achievement, this technology is based on cellular dissociation and hence does not maintain spatial relationships between single cells. Emerging technologies can now profile the transcriptome of single cells within their original environment, thus offering the possibility to examine how gene expression is influenced by cell-to-cell interactions, and how it is organized in a spatially coherent manner. One such approach is sequential single-molecule fluorescence in situ hybridization (seqFISH [7]), which can identify single molecules at (sub)cellular resolution with high sensitivity.

In contrast with scRNAseq, seqFISH and many other spatial transcriptomic technologies often pose significant technological challenges, resulting in a small number of profiled genes per cell (10-100s). The newer generation of seqFISH technology (called seqFISH+ [9]) has greatly enhanced its capacity to profile up to 10,000 genes, but this technology is more complex and costly than seqFISH.

New computational approaches are needed to effectively integrate scRNAseq and seqFISH data analyses. This first hackathon provided seqFISH and scRNAseq data corresponding to the mouse visual cortex ([3], [2]) and our participants were challenged to accurately identify cell types. The scRNA-seq data included transcriptional profiles at a high molecular resolution whereas the seqFISH data provided spatial characterization at a lower molecular resolution. Two key computational challenges were identified to enable high-resolution spatial molecular resolution. Firstly, we explored a number of strategies to identify the most likely cell types in the seqFISH dataset based on information obtained from the scRNAseq dataset. Secondly, we sought the opposite direction to transfer spatial information obtained from the seqFISH dataset to that of the scRNAseq dataset. Cell type labels were derived from scRNA-seq analysis [2] and previous seqFISH/scRNA-seq integration [3] were also provided as reference. **Could we have a more detailed description of the data chracteristics here please (number of cells, genes per data set, any filtering applied in the hackathon, is applicable**

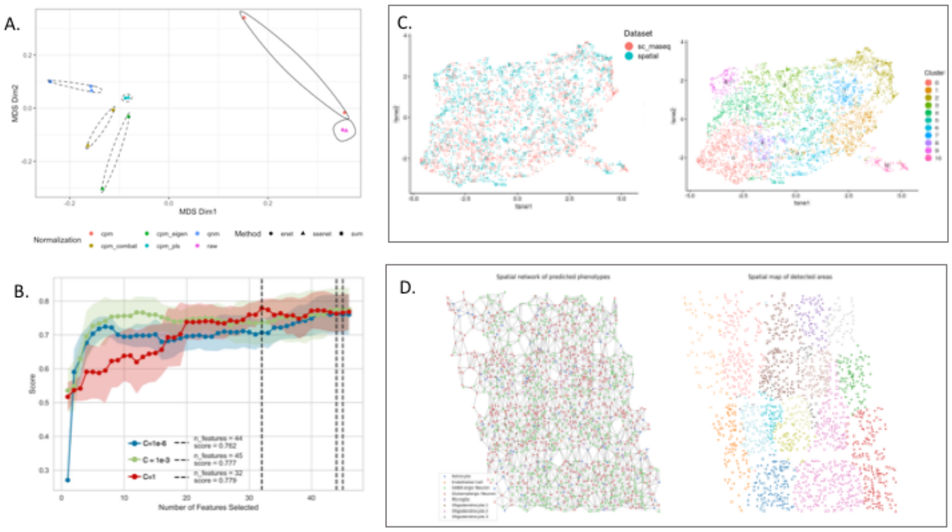


Figure 1:

Caption Figure: **Overview of seqFISH and scRNA-seq integration analysis**. **A** Assessment of cell type prediction using different data normalizations and classifiers. Normalization strategies included none (raw), counts per million (cpm), ComBat batch correction applied to cpm (cpm_combat), scRNAseq and seqFISH scaled using the first eigenvalue (cpm_eigen), latent variables retained for both datasets after applying Partial Least Squares regression to cpm_eigen normalized data (cpm_pls). Classifiers approaches included a supervised multinomial classifier with elastic net penalty (enet), a semi-supervised multinomial classifier with elastic net penalty (ssenet) and Support Vector Machine (SVM, supervised). Each classifier was trained using the scRNAseq data and the known (provided) cell type labels, then predicted the cell type labels in the seqFISH data; for the SVM we used the predictions from the original study (Challenge 1). The Gower distance between each method-normalization pair was computed and depicted on a multidimensional scaling plot. The first dimension (x-axis) separates methods that normalize the scRNAseq and seqFISH data together (dashed) and separately (solid), showing that normalization had a stronger impact on cell type predictions than the classification method used. **B** SVM classification models with different C parameters were trained with different number of genes in scRNAseq data using Recursive Feature Elimination (RFE) to evaluate the minimal number of genes required for data integration. The results show that a smaller gene list than what the original study proposed was sufficeint to identify cell types in both data types (Challenge 1). **C** LIGER was applied to combine spatial and single cell transcriptomic datasets. From the separate and integrative analyses, plots of identified and known clusters were generated and metrics of integration performance were compared, showing some loss of information as a result of the integration (Challenge 1). **D** Construction of a spatial network from cells' positions using Voronoi tessellation, where cell types were inferred from SVM trained on scRNA-seq data. Left: A neighbors aggregation method computes aggregation statistics on the seqFISH gene expression data for each node and its first order neighbors to address Challenge 2. Right: Identification of spatially coherent areas that can contain one or several cell types and can be used to detect genes whose expression is modulated by spatial factors rather than cell type.

Computational challenges

Challenge 1: overlay of scRNA-seq onto seqFISH for resolution enhancement

The mouse visual cortex consists of multiple complex cell types, however, the number of profiled genes in the seqFISH dataset was limited to 125. In addition, these genes were not prioritized based on their ability to discriminate between cell types. Assigning the correct cell identity presents an important challenge. In contrast, the scRNAseq dataset is transcriptome wide, and includes the 125 aforementioned genes.

This challenge proposed to use all genes to identify the cell type labels for each cell in the scRNA-seq data with high certainty. Next, we leveraged that information to build a classifier based on (a subset of) the 125 common genes only that was subsequently applied to the seqFISH dataset to assign cell types to each cell.

During the hackathon, participants aimed to test various machine learning or data integration models. However, preliminary analyses highlighted that normalization strategies had a significant impact on the final results (Figure 1A). In addition, although unique molecular identifier (UMI) based scRNA-seq and seqFISH can both be considered as count data, we observed dataset specific biases that could be attributed to either platform (imaging vs. sequencing batch effects) or sample specific sources of variation. In this hackathon, we opted to apply a quantile normalization approach which forces a similar expression distribution for each shared gene.

Two classifier approaches were considered: supervised and semi-supervised generalized linear model regularized with elastic net penalty (enet and ssenet) and support vector machines (SVM) that are supervised. The ssenet approach builds a model iteratively: it combines both datasets and initially only retains the highest confidence labels, then gradually adds more cell type labels until all cells are classified (Figure 1A). This type of self-training approach might be promising to generalize to other datasets. To further improve the SVM model, several combinations of kernels and optimal hyperparameters were assessed using a combination of randomized and zoomed search. In addition, different flavors of gene selection using recursive feature elimination were considered to identify the optimal or minimal number of genes needed to correctly classify the majority of the cells (Figure 1A). Finally, different classification accuracy metrics were considered to alleviate the major class imbalance in the dataset, since more than 90% of cells were excitatory or inhibitory neurons, using for example balanced classification error rate. Another approach, LIGER, based on integrative non-negative matrix factorization (NMF) was applied to integrate both datasets in a subspace based on shared factors, where cell type labels could be transferred using a nearest neighbor approach (Figure 1D).

could we have some elements of conclusion here?

Challenge 2: Identifying spatial expression patterns at the tissue level through integration of gene expression and spatial cellular coordinates

Most tools that were originally developed for scRNA-seq data can be adapted for spatial transcriptomic datasets (Section [**???**]{sec:common}), however, methods to extract sources of variation that originate from spatial factors are still lacking. Novel methods which can integrate the information obtained from gene expression with that of the spatial coordinates from each individual cell or transcript (for sub-cellular resolution) within a tissue of interest are needed.

To identify spatial expression patterns in the seqFISH dataset the participants first retained spatial information at the cellular level through the formation of a spatial network based on Voronoi tessellation ([10]). Within this spatial network the expression of each individual cell was spatially smoothed by calculating the average gene expression levels over all its neighboring cells. UMAP on the now smooth and aggregated data matrix was applied to identify cell clusters with a density based clustering approach (Figure 1D). Interestingly, these results showed that the obtained clusters themselves are spatially separated and do not necessarily overlap with specific cell types, suggesting that the spatial dimension cannot be simply captured from the expression data only.

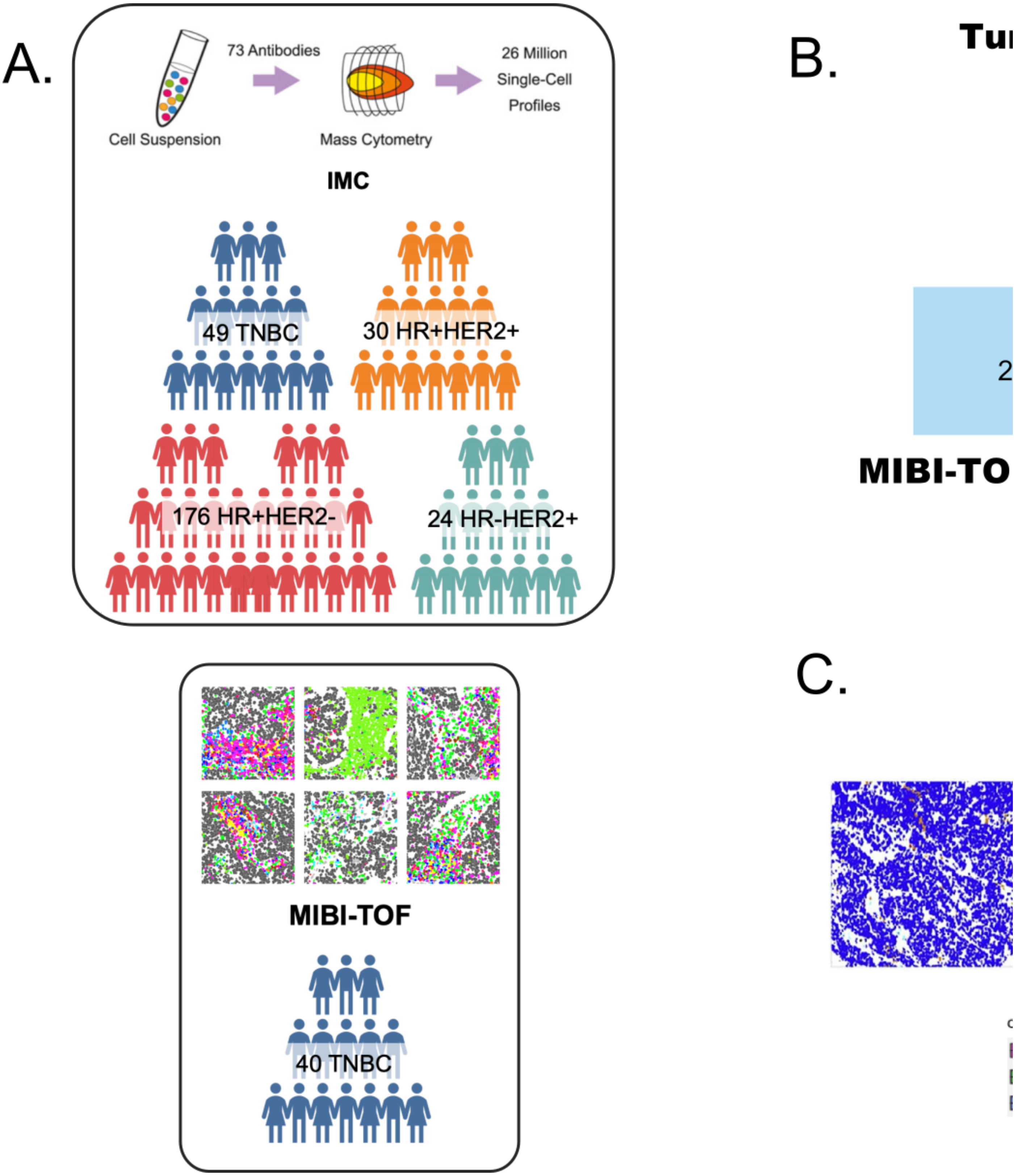
An unanswered question is whether the identified combinatorial spatial patterns can be extracted directly from a matching scRNA-seq dataset, similar to what was previously shown for individual gene expression profiles that can be mapped to their known spatial locations ([11], [**???**]). However, this still constitutes both a technological and analytical challenge that will require careful benchmarking in the near future (Section [**???**]{sec:benchmarking}).

Spatial proteomics as a case for cross-study and cross-platform analysis

Overview and biological question

Contrary to the first hackathon where seqFISH and and scRNA-seq data included samples from the same biological conditions, our second hackathon was obtained from different cohorts of patients primary breast cancer tissue using single cell targeted proteomics. The aim of this hackathon was to study tumor-immune microenvironment in primary breast cancer. Thus, challenges in this hackathon were considerable, including cross-study and cross-platform integrative analysis with a low number of features overlaping (Section @ref{sec:common}). Single cell proteomics data were generated on different antibody-based targeted proteomics technological platforms and in different laboratories. Mass cytometry (CyTOF) measured 73 proteins in two panels (immune and tumor) in 194 tissue samples from 143 subjects, of which 6 patients had triple-negative negative breast cancer [4], while the second dataset applied Multiplexed Ion Beam Imaging (MIBI) to quantify spatial in-situ expression of 36 proteins in 41 triple-negative breast cancer patients [5] (Figure ??AB). **what about the third study, can you describe in 1 sentence here**

This formidable data integration task included three challenges. The first challenge was to assess whether analytical methods could integrate partially-overlapping proteomic data collected on different patients with similar phenotypes, and whether measurements from one technology (MIBI spatial location and expression of proteins) could be transferred and used to predict information in the second technology (spatial expression patterns of proteins measured on mass-tag). The second challenge pertained to the specificity of spatial 'omics technologies and whether integrated analyses of single cell spatial technologies could capture unique information of the spatial location of immune cell populations in breast cancer beyond cell compositions. The third challenge posed the critical issue of lack of overlap in patients, and whether heterogeneous phenotype information could be used to integrate patient 'omics data with a low number of features and no tumor biological sample overlap.



Caption figure: **Overview of spatial proteomics cross-study and cross-platform integration analysis. A** Overview of single cell_targeted proteins hackathon challenge. **B** Challenge 1: Partial to no overlap between protein features across studies. **C** Challenge 2: Spatial analysis with Moran's index computed on Gabriel graph shown in boxplot according to tumor/immune status showing a significant difference between groups (Red asterisks indicate significance of an ANOVA of each group with all others with p-value from an overall ANOVA across the three groups reported).

Computational challenges

Challenge 1: Low overlap between protein features across studies

There were only 20 proteins that were assayed in both studies (Figure ??A-B **numbers dont match, specify which study**), which precluded integration of features at the level of gene set or pathways and required the use of surrogate measures for cross-study association. The majority of proteins were cell-type markers or biomarkers targets of breast cancer therapeutic intervention, providing the opportunity to perform cross-study integration of cell type proportions in tumor tissue samples.

Several semi-supervised and supervised algorithms were applied to transfer cell labels and cell compositions from one dataset to the second. Random forest was considered to capture the hierarchical structure of cell lineage and perform feature transfer learning of cell type labels, using an adaptation of the prediction strength approach [12] to assess model robustness: first, a model was trained on the labeled dataset, then used to predict labels in the unlabeled dataset; next, a second model was trained based on the second dataset with the newly predicted labels; finally, the ability of the second model to recover the correct original labels when making predictions on the labeled dataset was assessed. Mapping cells from CyTOF to imaging with spatial information was handled by solving an entropic regularization optimal transport problem [4, 11] **proper DOI ref please, in google doc**, using the cosine distance of the common proteins between the two datasets as transport cost. The constructed optimal transport plan can be considered as likelihood of cells from one modality mapped to cells from the other modality, which allows the prediction of protein expression measured only in CyTOF on imaging data. After cluster analysis of the resulting imputed expression matrix, sub tumour cell type could be identified that was not revealed in the original matrix.

This challenge raised other topical issues. The scales of protein expression was a limiting factor to integrate cell compositions using correlation of the expression of protein markers, as some cell markers were expected on a range of cell types (e.g. CD45), while others were more specialized and represented only in a subset of those cells (e.g. CD4). Others challenges associated with cell compositions analysis of proteomics analysis included uncertainty over antibody specificity and consistency between studies, specific sensitivity and specificity of protein markers for cell types, tissue and disease heterogeneity. The assignment of cell type was also seen as a major challenge, as it relied on manually curated protein annotation, and was dependent on domain-specific knowledge (e.g. CD4 is expressed by T-cells). To date, methods for cell type assignment, classification or extraction of differentially expressed proteins cannot easily be applied to targeted proteomics. There is thus an urgent need for a unifying map between cells present in different datasets, and for annotation resources to provide quality metric or priors of protein cell type markers. The construction of protein expression atlases would support cell type classification, even if antibodies used and their performances might vary between labs.

Challenge 2: spatial protein expression analysis

CyToF mass spectrometry data provided protein expression and counts/composition of cells in breast tumor-immune environment, while the MIBI-TOF data provided spatial information that quantified cell attributes (shape/size/spatial coordinates) in addition to expression levels. these two data sets thus provide the opportunity to examine protein expression, cell microenvironment, and predict cell-cell interactions and the cellular community ecosystem.

Spatial information can be encoded as a set of XY coordinates (cell centroid), a line (e.g. tumor-immune boundary) or a polygon, which is a closed plane defined by a number of lines and can define complex shapes such as a cell or a community of cells. Spatial protein expression can be summarized using spatial descriptive statistics, such as the autocorrelation of the expression of a protein within a neighborhood of polygons, using techniques developed in geographical information science or ecology to assess whether a spatially measured variable has a random, dispersed or clustered pattern [7] **ref DOI**.

We investigated whether expression data could be used to predict spatial properties of tissue samples using a variety of approaches. A K-nearest neighbor graph was used to build spatial response variables and random forest model trained from expression data to predict spatial features. Topic modelling was trained on protein expression and cell compositions in the CyToF data to predict cell co-locations in a fraction of MIBI-ToF considered as test data (10%),or vice versa. Among the five topics identified, the first topic was dominated in most of the immune cells from CyTOF data and the other four dominated in all other cells. Prognostic performance of different higher level spatial metrics was also examined using Moran's Index with a sphere distance, cell type localisation using nearest neighbour correlation, or cell type interaction composition with Ripley's L-function. Cox models with fused lasso penalty and random forest survival models were thenb fitted based on clinical features such as tumour stage, tumour grade, age and tumour size, as well as cell type composition. The spatial metrics were found to be predictive, especially in triple negative breast cancer where clinical features such as grade are poorly prognostic. Further investigation in spatial metric using a graph-based neighborhood measure (Gabriel graph, based on Delaunay triangulation) found that Moran's I differed significantly between the three prognostic tumor scores described [5](Figure ??C). Thus, this challenge emphasizes the need to develop new spatial measures specific for single cell spatial proteomics data. **Note: we will need to link back to the vignettes**

Challenge 3: Fourth corner Integration of data at the level of phenotype

Cross-study integration also raises the challenge of non-overlapping biological samples but with similar phenotypes. Here the aim is to identify biomarkers from the different data types to predict phenotype, and, more importantly, to explore how the markers selected across multiple datasets are in agreement or distinct from each other. Such markers should enable to extend biological knowledge that is not available by single omics data. To solve this third challenge, phenotypical data (such as the cell attributes) are the critical factors that should be used to link the two datasets (Figure ??D).

Integrating patient phenotype measures such as grade, stage and overall survival is one first step that we were able to achieve **How can you say it was successful? on which basis?**. However, integrating proteins from data sets that used different approaches to cell type annotation and had 13 proteins in common was extremely challenging **numbers dont match, specify which study**. Borrowing from ecology and the French school of ordination, this problem can be described as a case of the fourth corner problem (or RLQ, Figure ??D). Briefly, given two omics data where both features and samples are non overlapping, and phenotypical data are available for each omics data, multiplying the two phenotypical factors should derive a bridging matrix that links the features of two omics data. This require the two phenotypical matrices to be multipliable, i.e. describing the same phenotypical factors. The fourth corner RLQ can be solved using matrix decomposition [13; doi:10.1111/ecog.02302]. However, this approach was not attempted in this hackathon **was it too hard or lack of time?**

scNMT-seq as a case-study for epigenetic regulation

Overview and biological question

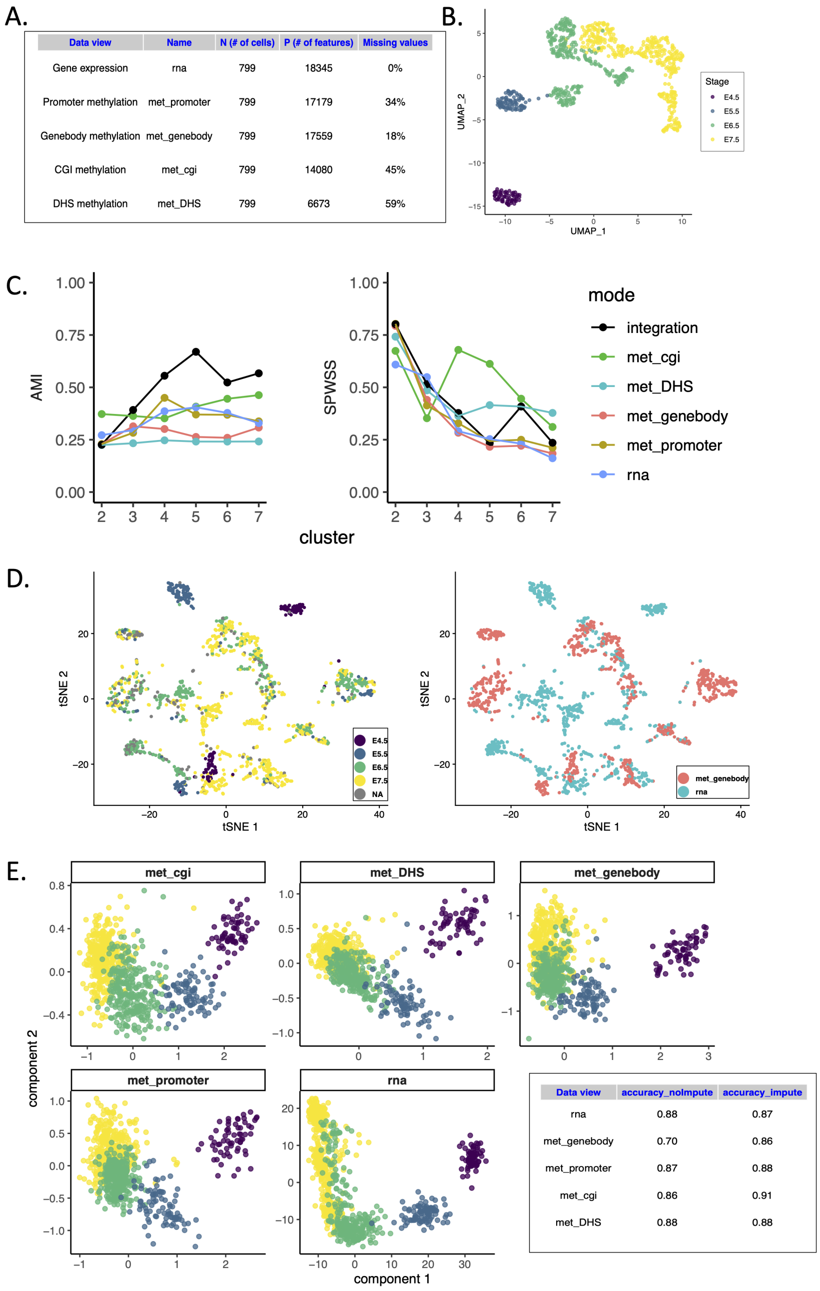
While the first two hackathons leveraged datasets from complementary technologies to enable high molecular and spatial resolution of biological systems, datasets spanning disparate molecular scales, such as DNA and RNA level measurements, enable to further resolve the regulatory networks that mediate cell fate decisions. The maturation of scRNA-seq technologies has enabled the identification of transcriptional profiles associated with lineage diversification and cell fate commitment [14], but the accompanying epigenetic changes and the role of epigenetic layers in driving cell fate decisions still remains poorly understood [15].

scNMT-seq is one of the first experimental protocols that enable simultaneous quantification of RNA expression and epigenetic information from individual cells [16]. Briefly, cells are incubated with a GpC methyltransferase enzyme that labels accessible GpC sites via DNA methylation. Thus, GpC methylation marks can be interpreted as direct read-outs for chromatin accessibility, whereas CpG methylation marks can be interpreted as endogenous DNA methylation. By physically separating the genomic DNA from the mRNA, scNMT-seq can profile RNA expression, DNA methylation and chromatin accessibility read-outs from the same cell. This third hackathon focused on data integration strategies to detect global covariation between RNA expression and DNA methylation variation from scNMT-seq data in a mouse gastrulation study [17].

Mouse gastrulation is a major lineage specification event in mammalian embryos that is accompanied by profound transcriptional rewiring and epigenetic remodelling [6]. In this study, four developmental stages were profiled, spanning exit from pluripotency and germ layer commitment (E4.5 to E7.5). For simplicity in this hackathon, we focused on the integration of RNA expression and DNA methylation, quantified over the following genomic contexts: gene bodies, promoters, CpG islands, and DHS open sites. A total of 799 cells passed quality control (Figure ??A). Preliminary analyses using dimensionality reduction methods confirmed that all four embryonic stages could be separated on the basis of RNA expression (Figure ??B). The main challenge was to leverage the multi-faceted nature of measurements to better resolve the single-cell subpopulations from distinct embyonic stages.

Computational challenges

Three computational strategies were considered by our participants: MOSAIC, a Multi-Omics Supervised Integrative Clustering algorithm inspired by `survClust` [18] that classifies samples by creating weighted distance matrices of effect sizes across data modalities with an outcome of interest. The weights are defined as the maximum of the ratio of Cluster specifc vs population log likelihoods. To facilitate integration, weighted distance matrices are standardized and Multidimensional scaling (MDS) is then used to map the subjects into an n-dimensional space that preserves between-subject distances for clustering. (Figure ??C); LIGER, an unsupervised non-negative matrix factorization model for manifold alignment that assumes a common feature space by aggregating DNA methylation over gene-centric elements (promoters or gene bodies) but allows cells to vary between data modalities [19] (Figure ??D); and Multi-block sparse Projection to Latent Structures (multiblock sPLS), a sparse generalization of canonical correlation analysis that maximizes paired covariances between the RNA data set and each of the other genomic context data sets [20 [21] (Figure ??E).



Caption Figure: **Overview of hackathon analyses for the scNMT-seq challenge.** **A** Summary of the data modalities analysed, including different putative regulatory regions. **B** UMAP of RNA measurements using 671 highly variable genes shows separation of the four embryonic stages.
C Supervised analysis using view-specific and integrative distance measures with MOSAIC: The integration identifies five clusters of cell populations based on Adjusted Mutual Information and Standardized Pooled Within Sum of Squares that outperforms individual (single omics) analyses.
D LIGER joint alignment using genebody methylation and RNA expression: cells are coloured by stage (left) or original data modality (right). **E** Unsupervised integration using multiblock sPLS: cells are projected into the space spanned by each data view components that are maximally correlated. For performance assessment, two types of analyses were considered, either by omitting the missing DNA methylation values or by incorporating imputed values. K-means clustering analysis based on the multiblock sPLS components was used to calculate balanced accuracy measures.

Challenge 1: defining genomic features

The first challenge presented in this hackathon concerns the definition of the input data. The output of single-cell bisulfite sequencing are binary DNA methylation measurements for individual CpG sites. Integrative analysis at the CpG level is extremely challenging due to the sparsity levels, the binary nature of the read-outs, and the intricacy in interpretation of individual dinucleotides. To address these problems, DNA methylation measurements are typically aggregated over pre-defined sets of genomic elements (i.e. promoters, enhancers, etc.). This preprocessing step reduces sparsity, permits the calculation of binomial rates that are approximately continuous and can also improve interpretability of the model output.

We observed remarkable differences between genomic contexts on the integration performance. In MOSAIC, stages are better separated when using DNA methylation measurements on promoter regions and at least four clusters (AMI=0.45). Interestingly, this setting performed better than using RNA expression alone (AMI=0.40). Notably, when using an integrated solution across data modalities, stages were better classified (AMI = 0.68) (Figure ??C). LIGER, that was also applied in the first hackathon (Section [**???**]{sec:spatial}) requires a common feature space to perform alignment of cells when profiled for different data modalities. This hackathon provides unambiguous cell matching between the data modalities and thus represents a gold standard for testing this approach. LIGER was applied on gene expression and gene body methylation: the poor alignment suggested a complex coupling of gene expression and genebody methylation during gastrulation (Figure ??D). Finally, multiblock sPLS identified covarying components between RNA expression and DNA methylation that separated cell stages in all putative regulatory contexts considered (Figure ??E). Taken altogether, these results confirmed that the appropriate selection of the feature space is critical for a successful integration with RNA expression.

Challenge 2: Missing values in DNA methylation

Single-cell bisulfite sequencing protocols are limited by incomplete CpG coverage because of the low amounts of starting material. Nonetheless, in contrast to scRNA-seq, missing data can be distinguished from dropouts. Integrative methods can be divided into approaches that can handle missing values (e.g. MOSAIC, multiblock sPLS which omit the missing values during inference), or approaches that require *a priori* imputation (e.g. LIGER). In this hackathon, missing values were imputed using nearest neighbor averaging (as implemented in the `impute` package [22]) in the methylation data.

We compared the integration performance of multiblock sPLS either with original or with imputed data. The missing values were inferred using nearest neighbor averaging (as implemented in the `impute` package [22]) in the methylation data. The components associated to each data set showed varying degree of separation of the embryonic stages, depending on the genomic contexts (Figure ??E). Accuracy measures based on k-means clustering analysis on the multiblock sPLS components showed that genebody methylation components were better at characterizing embryonic stage after imputation (from 70% with original data to 86% after imputation).

Missing values in regulatory context data represent a topical challenge in data analysis, and further methodological developments are needed to either handle and accurately estimate missing values.

Challenge 3: Linking epigenetic features to gene expression

One of the main advantages of scNMT-seq is the ability to unbiasedly link epigenetic variation with gene expression. Transcriptional activation is associated with specific chromatin states near the gene of interest. This includes deposition of activatory histone marks such as H3K27ac, H3K4me3 and H3K36me3, binding of transcription factors, promoter and/or enhancer demethylation and chromatin remodelling. All these events are closely interconnected and leave a footprint across multiple molecular layers that can only be (partially) recovered by performing an association analysis between a specific chromatin read-out and mRNA expression. However, given the large amount of genes and regulatory regions, this task can become prohibitively large, with the associated multiple testing burden. In addition, some of our analyses have shown that the correlations between epigenetic layers and RNA expression calculated from individual genomic features can be generally weak, or spurious.

A simple and practical approach from a computational perspective involves considering only putative regulatory elements within each gene's genomic neighbourhood. Nonetheless, this might miss important links with regulatory elements located far away from the neighbourhood.

In recent years, chromosome conformation capture experiments, have uncovered a complex network of chromatin interactions inside the nucleus connecting regions separated by multiple megabases along the genome and potentially involved in gene regulation. Early genome-wide contact maps generated by HiC uncovered domains spanning on the order of 1 Mb (in humans) within which genes would be coordinately regulated. Thus, a second strategy to associate putative regulatory elements to genes is to build on existing promoter-centered chromatin contact networks to restrict the association analysis to putative regulatory elements that are in 3D contact with genes. Although this is a promising strategy to reduce the complexity of the association analysis, most of our 3D interaction datasets are produced in bulk samples and it is so far unclear how much of these structures are preserved across individual cells. Single-cell conformation capture experiments are still limited by data sparsity and high levels of technical noise, but we envision that technological advances in this area will deepen our understanding on the regulatory roles of chromatin states.

Commonalities between analytical multi-omics approaches for hackathons

Each hackathon study highlighted disparate challenges to multi-omics from different measurement technologies. Yet, these studies were unified by the underlying problem of data integration. We summarize the common main challenges faced across all hackathons and the common approaches participants adopted that highlight the critical computational challenges faced in multi-omics single cell data analysis.

The choice of methods mostly relied on the challenge or biological question to address: data integration was conducted using projection approaches, cell prediction required machine or statistical learning methods (SVM, Enet) and spatial analysis was conducted using Hidden Markov random field or Moran's Index. As computational methodologies span technologies, so do the central challenges highlighted in each hackathon. For example, the accuracy of the analysis critically depended on data pre-processing (normalization, upstream feature selection), differences in scale across data sets and overlap (or lack thereof) of features and cells (Figure ??). In many cases, preprocessing can yield data mapping to common molecular features, such as genes, that can be the focus of the integration task. However, the spatial proteomics challenge showed that many multi-omics questions may have limited shared features between studies. In cross-study and cross-platform analyses, methods that investigate hierarchical structure among the 'omics, cell and phenotype layers and apply measures of higher order concordance are critical. Even in cases with matching molecular features, such analyses can reveal novel aspects of the biology.

Table ?? summarizes the main methods that were applied across all hackathons. A large number of computational analysis methods that were applied derive from bulk RNA-seq literature, with the exception of projection methods such as tSNE, UMAP and LIGER (the latter two are based on the common techniques NMF and PCA that were further developed for single cell data). In this section, we briefly highlight the three common challenges faced across all hackathons.

Common challenge 1: Dependence on pre-processing method and/or variable selection

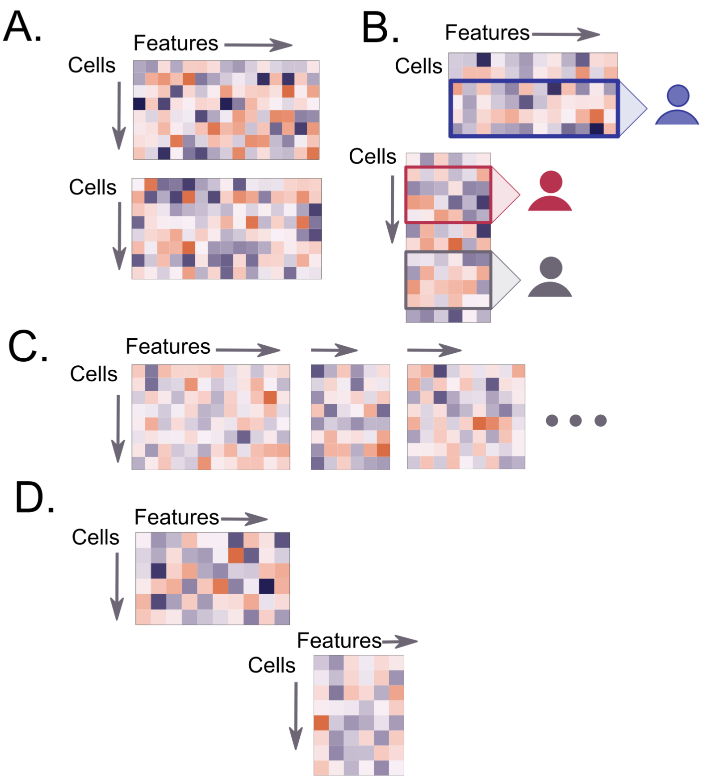
Pre-processing steps strongly affect downstream analyses. Our participants thoroughly assessed the effect of normalization and data transformation (e.g. spatial transcriptomics, Figure 1A), as well as preliminary feature selection (mostly on based on highly variable genes) or feature summarization (scNMT-seq study). Ease of comparisons between analyses was facilitated by providing processed input data ([**???**]{sec:software}), but even such step did not avoid facing reproducibility issues between the original published study and the new analyses. For example in the spatial transcriptomics study, 19 genes were selected in [COullomb vignette to ref] (*in scRNA-seq? or seqFISH?*) whereas the original paper selected 47 genes based on the same feature selection process [3]. No consensus was reached across participants analyses regarding the best way to process such emerging data, as no extensive benchmark, ground truth nor established biological results are yet available as we dicuss in [**???**]{sec:benchmark}.

Common challenge 2: Managing differences in scale and size across datasets

Different types of techniques were used to address the differences in scale or resolution across data sets. For spatial transcriptomics and proteomics, participants focused on a common set of genes (via feature selection in spatial transcriptomics) or proteins. The scNMT-seq study that included overlap between cells raised the issue of differences in data set size with a varying number of features per dataset ranging from 6,673 to 18,345 (Figure ??A). Some projection-based methods, such as MOFA [23], require a similar number of features in each data set, whilst others such as PLS / sGCCA [20] are not limited by such data setting and enable flexible analysis (Abadi vignette). Difference in data scale may result in one data set contributing to either too much variation or noise during data integration. Techniques such as re-scaling (Jenagan vignette), batch effect removal approaches, such as Combat [24] (Singh vignette), or weighting specific data sets (Arora, Abadi vignettes) were considered and all offered further improvement in the analyses.

Common challenge 3: Addressing partial overlap of information across cells or features

Degree of feature or cell overlap between datasets varied dramatically within each study. Intuitively, one requires at least one type of overlap (whether on the features, or on cells, Figure ??) in order to integrate information across modalities. The field has made progress in developing methods to integrate data sets across the same (bulk) samples of single cells, mostly based on dimension reduction techniques. Amongst them, NMF (LIGER), Projection to Latent Structures (sGCCA [20]) were used for the scNMT-seq study. When there was no cell overlap (spatial studies), imputation methods were used to predict gene, protein or spatial expression values based on nearest neighbors, latent variables or optimal transport, or to predict cell types (Hsu vignette). The most challenging study was the spatial proteomics, which raised the issue of no overlap between cells or features - the so called fourth corner that relies on phenotypes (Challenge 3 in [**???**]{sec:proteomics}). We anticipate that this scenario will be avoided once technological progress and increase in data sets availability is achieved [**???** 10.1186/s13059-020-1926-6].



Caption figure: **A.** Overlap of features (genes) but not cells (e.g. spatial transcriptomics where cell type prediction for seqFISH data was performed based on scRNAseq where cell types are known. **B.** Partial overlap of features (proteins) but no overlap of cells (e.g. spatial proteomics that required data imputation or cell type prediction). **C.** Overlap of cells across assays, but no overlap of features (e.g. scNMT-seq where data integration was performed). **D.** Lack of overlap between cells and features (the so-called fourth corner problem [**???**]{sec:proteomics}).

Table: Different methods were used in the hackathon. * indicates the method was not applied on the hackathon data. For some common challenges, 'bulk' indicates the method was originally developed for bulk omics, 'sc' indicates the method was specifically developed specifically for single cell data **table will include links to vignettes rather than name of participants TO DO** {#tbl:common}

Common challenges	Tasks	sc Spatial	sc targeted proteomics	sc NMT-seq
Pre-processing	**Normalization & data transformation**	Data distribution checks (Coullomb, Singh) High Variable Genes selection (Xu)	Variance Stabilization Normalisation [25] (Meng) Arcsinh transformation (Jeganathan). Inverse transformation (Jenagan) Selection of patients (Jenagan)	Summaries of DNA measurements (input data provided in hackathon)
Managing differences in scale	**Data integration**	LIGER [26] (Sodicoff) (sc) ComBat (Singh) Projection methods MFA, sGCCA [20] (Singh*) (bulk) UMAP/tSNE (Sodicoff) (sc)	Multi-block PCA [27] Weighting matrices based on their similarities: STATIS, MFA (Chen*)(bulk) Scale MIBI-TOF to the range of CyTOF values (Jenagan)	LIGER [26] (Welch) (sc) Projection method sGCCA [20] (Abadi) (bulk) Multi Omics Supervised Integrative Clustering with weights (Arora) (bulk)

Common challenges	Tasks	sc Spatial	sc targeted proteomics	sc NMT-seq
Overlap	**Cell overlap** (features not matching)			**Dimension reduction and projection methods:** LIGER [26] (Welch) (sc) sGCCA [20] (Abadi) (bulk)
	Partial feature overlap (cells not matching)		**Imputation:** Direct inversion with latent variables (Sankaran) Optimal transport to predict protein expression (Lin) K Nearest Neighbor averaging (Jenathan*)* ***No imputation:*** *Biological Network Interaction (Foster*)	
	Partial cell overlap (features not matching)		Multi block PCA [27] (Meng*)	
	No cell overlap (complete feature overlap)	Averaging nearest neighbors in latent space to impute unmeasured expression values (Coullomb?)	Transfer cell type label with Random Forest (Hsu)	LIGER [26] (Welch)
	No cell overlap (partial feature overlap)		Topic modeling to predict cell spatial co-location or spatial expression (Jenathan, partial feature overlap)	
	No overlap		RLQ [28] (Chen*)	
Generic approaches	**Classification & feature selection**	Backward selection with SVM (Coullomb) self training ENet (Singh) Balanced error rate (Coullomb, Singh) Recursive Feature Elimination (Xu) (all bulk)		Multi Omics Supervised Integrative Clustering (Arora) (bulk) Lasso penalization in regression-type models (bulk)
	Cell type prediction	Projection with LIGER [26] (Sodicoff) SVM (Coullomb, Xu) ssEnet (Singh) (all bulk)		
	Spatial analysis	Hidden Markov random field Voronoi tesselation (Coullomb) (bulk)	Spatial autocorrelation with Moran's Index (Hsu, Lin) Selection of spatial discriminative features: Moran's Index, NN correlation, Cell type, interaction composition, L function (Lin) (all bulk?)	
	Inclusion of additional information		Survival prediction: Cox regression based on spatial features (Lin)	Include annotated hypersensitive sites index to anchor new/unseen data from DNase-seq, (sc)ATAC-seq, scNMT-seq, for *de novo* peak calling (Meuleman*) (bulk)

Challenges for interpretation

While the three hackathon analyses emphasized that regardless of the common challenges faced by our participants, there is no one method fits all multi-omics challenges. An equally important complement to the diverse computational methods used to solve multi-omics analysis problems rests in the biological interpretation of their solutions. One notabe challenge to interpretation is that the the integrated data resulting from these approaches are often even higher dimensional than each of the input datasets to capture the multi-scale resolution of biological systems. For example, even abstract lower dimensional representations of spatial coordinates are often interpreted in terms of their ability to capture higher level cellular structure or prognostics, requiring even further data than the high-throughput multi-omics data as input. These approaches also suggest that new measures of the tumor or cell ecosystems of interacting cells are needed because these interactions are fundamental to bioloigcal systems. Both this high-dimensionality and biologicaly complexity introduce further challenges in the understanding and communication of results from complex data sets and analyses. Thus, these efforts to interpret multi-omics data will require standardized vocabulary, benchmarked methods, and common abstracted variables that can be compared between studies.

Supervised versus unsupervised

Interpretation hinges on the analysis method selected for a given dataset. One simple delineation between methods used throughout the hackathons and summarized in Table {#tbl:common} is that some aim to predict a clearly defined outcome at the start of the project, such as recognizing the environment of tumor cells versus that of healthy cells [**???**](sec:scProteomics). The supervised setting often provides easier interpretations, one can easily rank the covariates and contiguous data in terms of their predictive potential.

On the other hand when data are collected using multiple different technologies the data integration needs to provide organizing patterns that enable interpretation. Clustering is often used as one unsupervised method and is a good example of the use of a latent variable, in this case a factor or categorical variable which was not directly measured on the data but is often used to enable simple interpretations.

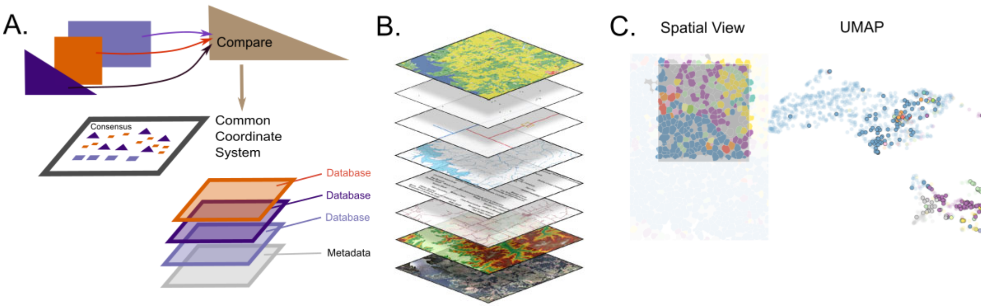
In cellular biology, a favorite such division into clusters is that involved in the definition of cell type [29].

Sometimes people get carried away in “clustering data” and manipulate the data, in cytometry one often sees cell gating done. The goal there is to eliminate cells in intermediary states to give clearly delineated inventories of cell types or cells in discrete states, this is a static description and will not enable researchers down the road to predict or understand transtitions between types.

Although a latent factor can be a useful first approximation, keep in mind that development of cells and their fate is a dynamic process and it can often be beneficial to keep data that enable interpretation of the cell trajectories: in that case, locally the underlying latent variable of interest is continuous along a gradient of development.

So far, we have seen two types of latent variables: clusters and a one dimensional continuous “gradient”, (pseduo-time, disease progression are two examples of such latent gradients). However the idea of latent variables is a rich anchor for many multimodal methods and can often be useful in highlighting what the modalities have in “common” and how they differ. The commonalities are well understood in the case of classical multivariate factor analyses where the data are decomposed into “commonalities” and uniqueness components [30].

A schematic summary of the different stages in interpretation is provided here:



Caption figure: **A** Schematic diagram of stages of interpretaion and integration of data sources. **B** Standards in Geographic Information Systems enable the integration of multiple layers of data. **C** Brushing an UMAP with a covariate can illustrate the dynamics of cell changes (**Kris: I have this feeling this is not the type of figure Susan intended to show**).

Multiple domains of knowledge can be combined easily if there is a common coordinate system, as in geospatial analyses.This is often a goal in multimodal or conjoint analyses, when the first step is to find a common compromise or consensus on which to project each of the individual modalities. Conjoint analyses also known as STATIS [31] was a very early multimodal method designed as PCA of PCAs where the first step in the analyses was to find what the different modalities had in common and define a consensus [32] onto which the individual tables were projected. This method can be seen as an extension of the class of matrix decomposition methods to data cubes. Many extensions to matrix decompositions have been designed for multimodal data, [33] offers an overview of the relations between many of them.

Reasonning by analogy with geospatial problems

In both the proteomics example [**???**](sec:scProteomics) and the [**???**](sec:scSpatial) examplary data, a spatial dimension is already naturally available. As in previous studies one can leverage extensive methods developed in spatial statistics to quantify spatial effects [34]. Contiguity and clustering can be tested and easily understood in the spatial context.

In these cases, layers of information can be mapped to the natural coordinate system in the same way a GIS system incorporates them (Figure ??B).

The spatial coordinate system analogy can be pursued further by finding a “consensus space” that provides a common coordinate system.

There are however pitfalls in using very sophisticated dimension reduction techniques which lead to over-interpretation or misinterpretation (size of clusters in tSNE related to sampling baselines rather than density, …)

Disparate sources of evidence are more compelling than more of the same.

Following Cardinal Newman’s principle[1] disparate sources of evidence, or in this case data from different technologies, are more compelling than many replicates of the same technology. Thus, if different technologies allow a consensus on underlying latent variables, this information is worth retaining.

Explaining results to biologists through generative models and simulations (ex: Factor Analysis, Hierarchical models).

Several difficulties arise when explaining summaries and conclusions, problems encountered include non-identifiability of models or non-sufficiency of summaries, simulations can often provide effective communication tools.

One can often generate data from different probabilistic models and show that the methods cannot differentiate between the generation processes, this is the identifiability problems that most overparametrized models lead to. Added constraints on the parameters can often be integrated into the analyses to make them more realistic and reduce if not eliminate the identifiability issues.

Meaningful Interpretation by linking in databases

In the right side of Figure ??A we show how connections to layers of information from outside databases can be incorporated into the final output. Real biological understanding is often subordinated to the integration of this contiguous information. Either from the metadata already available in the multiassay containers as for instance in the MultiAssayExperiment package or from exterior sources such as Gene Ontologies, Biomart [35], Kegg, Human Cell Atlas (HCA) or other sources often available through links provided within systems like bioconductor ().

Redundant biological knowledge is often enlightening, as many methods suffer from identifiability issues (ie in a gradient, the direction of the direction is unknown). By providing information on the extreme points in a map or brushing a map with known gene expression features one can delineate orientations and clusters.

For instance coloring by CD56 across time shows the dynamics of immune response [36] (Figure ??C).

Visualization tools for interpretation and communication to biologists

An example of effective visual interpretation tools is interactive brushing of UMAP plot, see Figure ??C by Kris Sankaran.

Interpretation for data scientists reading the methods sections requires a good understanding of the building blocks

Spanning all of these interpretation challenges is a further central communication barriers within the community of data scientists, computer scientists and computational biologists ie communicating about methods within a community of practitionners who do not have the same vocabulary or background.

Many tools are used as black boxes and users don't know or agree on what exactly the methods are doing (MOFA and tSNE are examples). The first step in unblinding these black boxes used as methodology shortcuts is to have a clear glossary of terms and how we are using them. Many synonyms for multimodal data exist and some have nuances, see the table we have compiled (ref: Table1). Understanding the relation between methods developed by different teams is essential and we often try to organize the methods first, thus it is useful to create a dichotomy of methods and their underlying properties.

A very useful tool for making methodological black boxes more transparent are simulated data. These can follow benchmark methods such as those presented in [**???**](sec:sec-benchmark) and use well defined generative processes to clarify what some complex methods do.

Visualization of the data, following the step by step transformations and optimizations of data representations also help clarify how certain methods fit models or compress and reduce data dimensionality. These visualizations are often very specialized (think for instance, correspondence analyses, goodness of fit plots like qqplots or rootograms or mean-variance fitting). These intermediary plots don't usually end up in the main text of final biological publications and serve as intermediary checks to unpack the black boxes.

Missing

- Validation through complementary data and sequential experimental design.
- Examples from other parts, references and commentary here missing until documents become availabe ([**???**])

References

Cell type definition: [29]

Factor Analysis: [30]

Statis, conjoint analysis: [31]

The French way: [32]

Overview and connections of methods: KS [33]

Kevin Murphy: Probabilistic Machine Learning, MIT Press [**???**] [37] [37]

GIS: reference https://www.usgs.gov/faqs/what-a-geographic-information-system-gis

Original: https://prd-wret.s3.us-west-2.amazonaws.com/assets/palladium/production/s3fs-public/styles/full_width/public/thumbnails/image/8BaseLayersofTheNationalMap.JPG

Biomart: [35]

UMAP: [38] https://arxiv.org/abs/1802.03426v2

Spatial tumor and immune cells: [34]

CD56 Immune cell coloring, paper with C. Blish: [36]

Footnote: Cardinal Newman wrote **The Grammar of Assent.** and cited in [Bruno de Finetti, Volume 1, 1974 Theory of Probability]:

Supposes a thesis (e.g. the guilt of an accused man) is supported by a great deal of circumstantial evidence of different forms, but in agreement with each other; then even if each piece of evidence is in itself insufficient to produce any strong belief, the thesis is decisively strengthened by their joint effect.

Techniques and challenges for benchmarking methods

Visualizations and biological assessment of marker gene lists resulting from multi-omics analyses provide critical interpretation of integrative analysis of high-throughput data, but additional quantitative metrics are necessary to delineate biologically-relevant features from features arising from either computational or technical artifacts. Beyond interpretation, quantitative benchmarks are also essential to enable unbiased comparisons between analytical methods. For example, the goal of multi-platform single cell data analysis is often the recovery of known cell types through computational methods, where metrics such as the adjusted Rand Index (ARI) enable direct assessment of the clustering results with respect to known cell types. When cell types or biological features are not known *a priori*, benchmark methods can also be used to discover known relationships between data modalities, for example *cis* gene regulatory mechanisms observed between chromatin accessibility and gene expression. Our hackathons highlighted that many of these relationships are not fully understood at the single cell level, and that benchmark standards are critically needed for validation (Figure 2A).

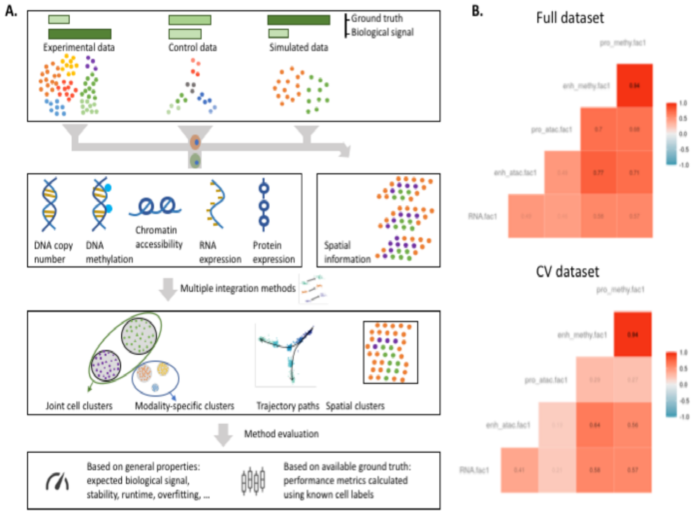


Figure 2:

Caption figure: **A** Systematic benchmarking of single cell multi-omic analysis methods can involve experimental data (as per our hackathons), custom control datasets, where known structure is imposed through the experimental design, or simulated data. The amount of biological signal and ground truth available varies considerably between these types of data. The resulting multi-omics datasets are analysed by competing methods and compared using metrics that have general purpose or take account of the ground truth (e.g. cell type labels or number of cell types simulated). **B** scNMT-seq study: correlations with linear projections (MOFA+) evaluated with cross-validation.

Challenges and strategies for benchmarking

Benchmarking multi-modal methods is inherently difficult, as ground truth is rarely known. Ground truth can be introduced through simulating high-throughput data *in silico*, but in the context of data integration, the simulation of a realistic covariance structure across features and across data modalities is challenging [39] and must rely on an underlying generative model that may introduce further biases into the benchmarking analysis. Another strategy is to use cross-validation within a study, or conduct cross-study validation to assess whether solutions found by multi-modal methods generalize to held-out observations or held-out studies. The latter was somewhat attempted in the spatial proteomics cross-study hackathon, but where ground truth was unknown ([**???**]{sec:proteomics}).

Challenge 1: creating benchmarking datasets

Benchmark datasets serve two main purposes: to provide ground truth for the intended effect of exposure in a proposed study design, and to provide validation for an analytic task for which a new computational method may be proposed (e.g. data integration in our hackathons), Figure 2A.

For single cell studies, benchmark datasets have largely focused on measuring sequencing depth and diversity of cell types derived from a single assay of interest (e.g. scRNAseq). Common experimental designs involve creating artificial samples through the mixing of cells in known proportions [40,41,42] or creating dilution series to simulate variation in cell size [40,43]. Simulating data is also popular and made more convenient through software such as the `splatter` R package [44].

For multi-modal assays, while the intended effects can vary based on the leading biological questions, one may abstract out common data integration tasks such as co-embedding, mapping or correlation, and inferring causal relationships. We distinguish data integration from further downstream analyses that may occur on integrated samples such as differential analysis of both assays with regard to a certain exposure. Both the intended effects and data integration task rely on study design that takes into account the biological and technical variability via replicates, block design, and randomization, the power analysis for the intended effect or data integration task, and the dependencies between modalities. For example, gene expression depends on gene regulatory element activity and thus requires that experiment design must also account for spatial and temporal elements in sampling for a given observation.

As such, no universal benchmark data scheme may suit every combination of modality (e.g. mising cells design does not generalise to the spatial context), and benchmark datasets may be established for commonly used combinations of modalities or technologies towards specific data integration tasks.

Challenge 2: cross-validation within study

Cross-validation within a representative multi-modal study is one possible approach for quantitative assessment for unbiased comparison of methods. We note that the approach of cross-validation – in which observations are split into folds or left out individually for assessing model fit – has been used often for parameter tuning within methods, or for other aspects of model selection [21,39,45,46,47,48,49,50,51,52,53,54].
Similarly, permutation has been used to create null datasets, either as demonstration that a particular method is not overfitting, or for parameter tuning, where the optimal parameter setting should result in a model score that is far from the null distribution of model scores [55,56,57]. Cross-validation is particularly useful as a quantitative assessment of a method's self-consistency, even though it cannot determine the *accuracy* of a method in a completely unbiased way if we do not have access to an external test data set for further confirmation.

A cross-validation analysis of the scNMT-seq dataset using MOFA+ was performed as part of the third hackathon, and demonstrated that strong relationships found among pairs of modalities in single cells used for training the model were likewise often found equally strong in held out cells (Figure 2B). This analysis in this hackathon revealed how we could reliably match dimensions of latent space across cross-validation folds. Previous evaluations of multi-modal methods have focused only on the top 'latent factor' [58], however, we showed in our analyses, many latent factors can be reliably discovered in held out cells in studies of complex biological processes such as differentiation of embryonic cells.

For clustering assessment, several studies have used resampling or data-splitting strategies to determine prediction strength [12,59,60,61]. These techniques could be further extended in a multi-modal setting for clustering of cells into putative cell types or cell states. Community-based benchmarking efforts in the area of multi-modal data analysis could follow the paradigm of the DREAM Challenges, with multi-modal training data provided and test samples held out, in order to evaluate the method submissions from participating groups.

Challenge 3: cross-validation between studies

Our benchmarking hackathons have emphasized the need to access external studies for methods assessment and validation, where either the ground truth is based on biological knowledge of the system being studied, or via high-quality control experiments where the ground truth (e.g. cell type labels) are known (Figure 2A). To take advantage of all data and technologies available, cross-study validation could also extend to cross-platform, to assess whether relationships discovered in one dataset are present in other datasets, looking across single cell and bulk as was recently proposed in [62].

Software strategies to enable analyses of multimodal single cell experiments

Open-source software are essential in bioinformatics and computational biology. Benchmark datasets, analysis pipelines, and development of multimodal genome-scale experiments are all enabled through community-developed, open source software and data sharing platforms. A wide array of genomics frameworks for multi-platform single cell data have been developed in R and Python. Along with other software, these frameworks use standardized licensing in Creative Commons, Artistic, or GNU so that all components are accessible for full vetting by the community. Our hackathons hinged on the central challenges such as widescale adoption, extension, and collaboration to enable inference and visualization of the multimodal single-cell experiments in our analytic frameworks. We designed each case study to leverage and build on these open frameworks to further develop and evaluate robust benchmark strategies. Easy to use data packages to distribute the multi-omics data and reproducible vignettes were key outputs from our workshop.

Collaboration enabled through continuous integration

Open-source software efforts facilitate a community-level coordinated approach to support collaboration rather than duplication of effort between groups working on similar problems. Real-time improvements to the tool-set should be feasible, respecting needs for stability, reliability, and continuity of access to evolving components. To that end, exploration and engagement with all these tools is richly enabled through code sharing resources. Our hackathons directly leveraged through GitHub with our reproducible analyses reports to enable continuous integration of changes to source codes (using Github Action), and containerized snapshots of the analyses environments. The hackathons analyses conducted in R were assembled into R packages to facilitate libraries loading, while those conducted in Python enabled automatic installation and deployment

Useability and adoption by the community

Robust software ecosystems are required to build broad user bases [63,64,65]. Bioconductor is one example of such ecosystem, that provides multiplatform and continuous delivery of contributed software, while assisting a wide range of users with standardised documentation, tests, community forums and workshops [66,67,68]. In the case of the hackathons, the R/Bioconductor ecosystem for multi-omics enabled data structures and vignettes to support reproducible, open-source, open development analysis. During this workshop we identified key software goals needed to advance the methods and interpretation of multi-omics.

Challenge 1: data accessibility

Providing data to the scientific community is a long-standing issue. A particular challeng in our hackathons was that each data modality was characterized by a different collection of features from possibly non-overlapping collections of samples [**???**]{sec:common}. thus, common data structures are needed to store and operate on these data collections, and support data dissemination with robust metadata and implementation of analysis frameworks.

The `MultiAssayExperiment` integrative data class from Bioconductor was our class of choice to enable the collation of standard data formats, easy data access and processing. It uses the S4 object-oriented structure in R [69,70] and includes several features to support multi-platform genomics data analysis, to store features from multiple data modalities (e.g. gene expression units from scRNA-seq and protein units in sc-proteomics) from either same or distinct cells, biological specimen of origin or from multiple dimensions (e.g. spatial coordinates, locations of eQTLs). This class also enables to store sample metadata (e.g. study, center, phenotype, perturbation) and provides a map between the datasets from different assays for downstream analysis.

In our hackathons, pre-processing steps applied on the raw data were fully documented. The input data were stored as `MultiAssayExperiment` objects that were centrally managed and hosted on `ExperimentHub` [71] as a starting point for all analyses. The `SingleCellMultiModal` package was used to query the relevant datasets for each analysis [doi:10.18129/B9.bioc.SingleCellMultiModal] (Figure 3). Text-based machine readable data were also made available for non-R users, and also to facilitate alternative data preprocessing for participants.

Besides efficient data storage, several hackathon contributors used the `MultiAssayExperiment` class to implement further data processing and extraction of spatial information from raster objects in their analyses. This infrastructure was readily suitable for the spatial and scNMT-seq hackathons but the lack of overlap between samples in the spatial proteomics hackathon revealed an important area of future work to link biologically related datasets without direct feature or sample mappings for multi-omics analysis. Further, our hackathons highlighted the need for scalability of storing and efficiently retrieving single cell data datasets [72,73]. New algorithms are emerging, that allow for data to be stored in memory or on disk (e.g. [74,75] in R or [76] in Python).

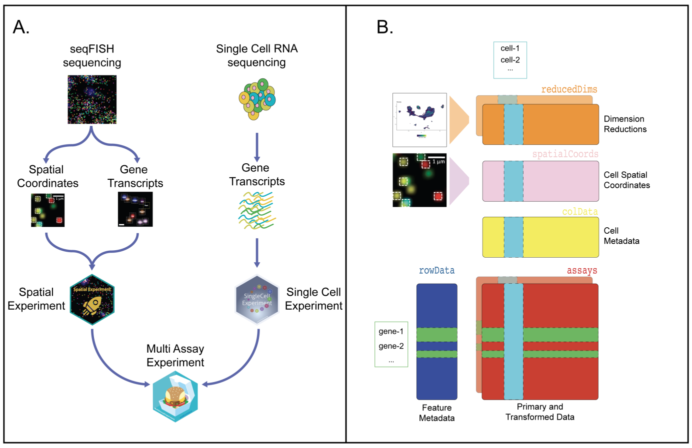


Figure 3:

Caption figure: **A** Software infrastructure using Bioconductor for the first hackathon to combine seqFISH-based `SpatialExperiment` and `SingleCellExperiment` instances into a `MultiAssayExperiment`. **B** To combine these two different experiments, the seqFISH data were stored into a `SpatialExperiment` S4 class object, while the scRNA-seq data were stored into a `SingleCellExperiment` class object [77]. These objects were then stored into a `MultiAssayExperiment` class object and released with the `SingleCellMultiModal` Bioconductor package [**???**].

Challenge 2: software infrastructure to handle assay-specific features

The hackathons further highlighted emerging challenges to handle different data modalities.

RNA-seq has well-defined units and IDs (e.g., transcript names), but other assays need to be summarized at different genomic scales (e.g., gene promoters, exons, introns, or gene bodies), as was highlighted in the scNMT-seq hackathon. Tools such as the `GenomicRanges` R package [78] have been proposed to compute summaries at different scales and overlaps between signal (e.g., ATAC-seq peaks) and genomic annotation.

Further, the observations of different modalities may not be directly comparable: for instance, gene expression may be measured from individual cells in single-cell RNA-seq but spatial transcriptomics may have a finer (sub-cellular) or coarser (multi-cellular) resolution. Methods such as SPOTlight [79] can be used to deconvolute multi-cellular spots signal.

Finally, in the absence of universal standards, the metadata available may vary from modalities, or independent studies (e.g. spatial proteomics), thus urging the need from the computational biology community to define the minimum set of metadata variables necessary for each assay, as well as for pairs of assays to be comparable for common analyses.

Challenge 3: accessible vizualization

Our brainstorm discussions on the Data Interpretation Challenge (@ref{sec:interp}) highlighted the importance of novel data visualization strategies to make sens of multi-modal data analyses. Often, these visualization strategies rely on heatmaps or reduced dimension plots, and utilize color to represent the different dimensions. These colors and low dimensional plots facilitate pattern detection and interpretation of increasingly complex and rich data. However, relying on color for interpretation leads to difficulties in perceiving patterns for a substantial proportion of the population with color vision deficiencies and can result in different data interpretations between individuals.

Presenting accessible scientific information requires the inclusion of colorblind friendly visualizations [80,81] standardized as default setting with using palettes such as R/viridis [82] and dittoSeq [83] with a limit of 10 colors. Additional visual cues to differentiate regions or cells can also reduce the dependence on colors using hatched areas or point shapes. The inclusion an "accessibility caption" accompanying figures which to guide the reader's perception of the images would also greatly benefit broader data accessibility. Thus, implementing community standards for accessible visualizations is essential for bioinformatics software communities to ensure standardized interpretation of multi-platform single cell data.

Type	Brief name (link)	Description
Matlab package	CytoMAP	CytoMAP: A Spatial Analysis Toolbox Reveals Features of Myeloid Cell Organization in Lymphoid Tissues
Matlab package	histoCAT	histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data
Python library	PyTorch	General framework for deep learning
Python package	SpaCell	SpaCell: integrating tissue morphology and spatial gene expression to predict disease cells
Python package	Scanpy	Python package for single cell analysis
R data class	MultiAssayExperiment	unify multiple experiments
R data class	SpatialExperiment	SpatialExperiment: a collection of S4 classes for Spatial Data
R package	Giotto	Spatial transcriptomics
R package	cytomapper	cytomapper: Visualization of highly multiplexed imaging cytometry data in R
R package	Spaniel	Spaniel: analysis and interactive sharing of Spatial Transcriptomics data
R package	Seurat	R toolkit for single cell genomics
R package	SpatialLIBD	Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex

Type	Brief name (link)	Description
R package	Cardinal	Cardinal: an R package for statistical analysis of mass spectrometry-based imaging experiments
R package	CoGAPS	scCoGAPS learns biologically meaningful latent spaces from sparse scRNA-Seq data
R package	projectR	ProjectR is a transfer learning framework to rapidly explore latent spaces across independent datasets
R package	SingleCellMultiModal	Serves multiple datasets obtained from GEO and other sources and represents them as MultiAssayExperiment objects
R scripts	SpatialAnalysis	Scripts for SpatialExperiment usage
Self-contained GUI	ST viewer	ST viewer: a tool for analysis and visualization of spatial transcriptomics datasets
Shiny app	Dynverse	A comparison of single-cell trajectory inference methods: towards more accurate and robust tools
R package	mixOmics	R toolkit for multivariate analysis of multi-modal data
Python package	totalVI	A variational autoencoder (deep learning model) to integrate RNA and protein data from CITE-seq experiments
Python web application		ImJoy
Python package	napari	Interactive big multi-dimensional 3D image viewer
Software	QuPath	Multiplex whole slide image analysis
Python package	Cytokit	Multiplex whole slide image analysis
Python package	cmIF	Multiplex whole slide image analysis
Software	Facetto	Multiplex whole slide image analysis, not available yet
Software, Python based	CellProfiler	Image analysis

Discussion

The Mathematical Frameworks for Integrative Analysis of Emerging Biological Data Workshop demonstrated the power of hackathons to both inform and develop new analysis methods to capture the complex, multi-scale nature of biological datasets from high-throughput data modalities. Notably, the hackathon studies of the workshop were specifically designed to span state-of-the-art multi-omics challenges to map the epigenetic, molecular, and cellular interaction across time and sample populations. Coupling single cell measurements can inherently simplify the challenge of linking disparate biological scales, but layering new sets of molecular measurements increases complexity of the analyses to interpret these data. The computational needs hinge on the underlying biological question being asked as well as the characteristics of the data themselves. In our workshop, different modelling considerations had to be made for multi-modal integration, as higlighted in the seqFISH and scNMT-seq challenges (matching on the same genes, or cells) and the scProteomics challenge (partially unmatched measurements). Regardless, through these hackathons we identified several common analysis themes spanning algorithmic advances, interpretation, benchmarking, and software infrastructure that are all necessary for biological interpretation. All hackathons required methods for dealing with data quality, data loss from summarization, timing variances both between and within omics layers, and batch effects. These represent the necessary challenges to overcome in the coming years, along with efficient and insightful data visualization strategies to infer regulatory relationships between different omics.

Technologies to profile biological systems at single cell resolution and across molecular scales are advancing at an unprecedented pace. Analytically, these advances require the computational analysis community to pursue research that can first enable robust, data set specific analyses tailored to a specific biology or measurement technology and, second, that can scale and adapt to these rapid advances. Our hackathons highlighted current technologies for spatial molecular profiling, which also require profiling samples with multiple other technologies to balance high resolution spatial and molecular profiling across data modalities. The scNMT-seq challenge did not include spatially resolved data, but highlighted the potential of further inference of gene regulation through concurrent profiling of RNA, methylation, and chromatin state. Technological advances for multi-omics spatial data and epigenetics data are rapidly advancing and becoming increasingly available through companies such as Nanostring, 10X Genomics, Akoya Biosciences, and others. Our workshop keynote Bernd Bodenmiller presented new research-level technological advances that enable three dimensional spatial molecular profiling [84]. Other technologies are currently expanding to allow for temporally resolved profiling [85]. Integration strategies aware of these future directions and of the mathematical analysis challenges that span technologies will be most adept to advance biological knowledge: this was the primary aim of this workshop.

The implementation of novel analysis tools requires further robust software ecosystems, including for example Bioconductor [86], Biopython, and toolkits such as Scanpy [76], Seurat [87], or Giotto [10], in which users can create their analysis approaches and while anticipating stable and adaptive data structures robust for these emerging technologies. The size of these emerging datasets, particularly in the context of their application to atlas projects (e.g. the Human Tumor Atlas Network [88], Human Cell Atlas [89], Allen Brain Initiative or ENCODE, to cite a few) are key examples that computational efficiency and scalability of these implementations are becoming ever more critical. <!–are there others or citations I should be using for this?–!>

In addition to new technologies, we wish to emphasize that arising multi-omics analysis methods can support the generation of new data sources to resolve the multi-scale nature of biological systems. For example, while the workshop posed the scNMT-seq data and spatial molecular datasets as distinct challenges for data integration, integration of matched datasets between these spatial and epigenetic profiling techniques could further resolve the dependence of cell-type and cellular-interactions of regulatory networks. By further embedding prior biological knowledge as rules in the analysis approaches, additional source of data from can generate a new representation of a biological system. For example, curated regulatory networks from databases such as KEGG, Biocarta, GO, or MSigDB <!– are there more resources and add citations–!>

provide commonly used frameworks for this prior knowledge, but must be extended to map context-specific networks in light of emerging single cell atlases. The regulatory networks and dynamic features captured in single cell data also provide the potential for future techniques to predict molecular and cellular states. Our hackathons and workshop have shown that merging single cell data with mathematical models are essential to forecast future states using rules derived from only prior biological knowledge. **to fix**

Glossary

Table 1: Glossary of interchangeable terms in the field of single-cell and bulk multi-omics (multi-source) data analysis.

Consensus Term	Related Terms	Description	Citation
network	graph, adjacency matrix	A set of *nodes*, representing objects of interest, linked by *edges*, representing specific relationships between nodes.	?
node	vertex	Element of interest in a network and linked to other nodes. For example: people, cells, proteins or genes. Nodes can have several properties called *attributes* like cell type or position.	?
edge	link	The relationship between 2 nodes in a network. For example: friendship in social networks, cells in contact in a spatial network, or gene-gene interactions in a gene regulatory network.	?
concordant	concordant, coherent, consistent	?	58
contributions	variable weights, loadings, eigenvector, axis, direction, dimension, coefficients, slopes	Contributions of the original variables in constructing the components.	21, 91
latent factors	variates, scores, projections, components, latent/hidden/unobserved variables/factors	Weighted linear combinations of the original variables.	21, 91
multimodal	Multiview, multiway arrays, multimodal, multidomain, multiblock, multitable, multi-omics, multi-source data analysis methods, N-integration	Methods pertaining to the analysis of multiple data matrices for the same set of observations.	21, 33, 92
conjoint analysis	conjoint analysis, P-integration, meta-analysis, multigroup data analysis	Methods pertaining to the analysis of multiple data matrices for the same set of variables.	21, 91, 93
variable	feature, variable	A measurable quantity that describes an observation's attributes. Variables from different modalities include age, sex, gene or protein abundance, single nucleotide variants, operational taxonomic units, pixel intensity *etc.*	?
biomarker	marker, biomarker	A variable that is associated with normal or disease processes, or responses to exposures, or interventions. Any change in this variable is also associated with a change in the associated clinical outcome. These variables may be used for diagnostic, monitoring, Pharmacodynamic responses. Examples include LDL cholesterol, CD4 counts, hemoglobin A1C.	94
panel	biomarker panel, biomarker signature	A subset of the originally measured variables that are determined to be associated with the outcome or response variable. This may be determined using statistical inference, feature selection methods, or machine/statistical learning.	95, 96
observation	sample, observation, array	A single entity belonging to a larger grouping. Examples include patients, subjects, participants, cells, biological sample, usually the unit of observation on which the variables are measured *etc.*	?

References

1. **Method of the Year 2019: Single-cell multimodal omics**
   Nature Methods
   (2020-01) https://www.nature.com/articles/s41592-019-0703-5
   DOI: 10.1038/s41592-019-0703-5

2. **Adult mouse cortical cell taxonomy revealed by single cell transcriptomics**
   Bosiljka Tasic, Vilas Menon, Thuc Nghi Nguyen, Tae Kyung Kim, Tim Jarsky, Zizhen Yao, Boaz Levi, Lucas T Gray, Staci A Sorensen, Tim Dolbeare, … Hongkui Zeng
   Nature Neuroscience (2016-01-04) https://doi.org/f778w5
   DOI: 10.1038/nn.4216 · PMID: 26727548 · PMCID: PMC4985242

3. **Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data**
   Qian Zhu, Sheel Shah, Ruben Dries, Long Cai, Guo-Cheng Yuan
   Nature Biotechnology (2018-10-29) https://doi.org/gfgn8x
   DOI: 10.1038/nbt.4260 · PMID: 30371680 · PMCID: PMC6488461

4. **A Single-Cell Atlas of the Tumor and Immune Ecosystem of Human Breast Cancer**
   Johanna Wagner, Maria Anna Rapsomaniki, Stéphane Chevrier, Tobias Anzeneder, Claus Langwieder, August Dykgers, Martin Rees, Annette Ramaswamy, Simone Muenst, Savas Deniz Soysal, … Bernd Bodenmiller
   Cell (2019-05) https://doi.org/gfzbz7
   DOI: 10.1016/j.cell.2019.03.005 · PMID: 30982598 · PMCID: PMC6526772

5. **A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging**
   Leeat Keren, Marc Bosse, Diana Marquez, Roshan Angoshtari, Samir Jain, Sushama Varma, Soo-Ryum Yang, Allison Kurian, David Van Valen, Robert West, … Michael Angelo
   Cell (2018-09) https://doi.org/gd4wms
   DOI: 10.1016/j.cell.2018.08.039 · PMID: 30193111 · PMCID: PMC6132072

6. **Epigenetic regulation in development: is the mouse a good model for the human?**
   Courtney W Hanna, Hannah Demond, Gavin Kelsey
   Human Reproduction Update (2018-09) https://doi.org/gd3d4z
   DOI: 10.1093/humupd/dmy021 · PMID: 29992283 · PMCID: PMC6093373

7. **Single-cell in situ RNA profiling by sequential hybridization**
   Eric Lubeck, Ahmet F Coskun, Timur Zhiyentayev, Mubhij Ahmad, Long Cai
   Nature Methods (2014-03-28) https://doi.org/ggrr5b
   DOI: 10.1038/nmeth.2892 · PMID: 24681720 · PMCID: PMC4085791

8. **In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus**
   Sheel Shah, Eric Lubeck, Wen Zhou, Long Cai
   Neuron (2016-10) https://doi.org/f8875g
   DOI: 10.1016/j.neuron.2016.10.001 · PMID: 27764670 · PMCID: PMC5087994

9. **Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+**
   Chee-Huat Linus Eng, Michael Lawson, Qian Zhu, Ruben Dries, Noushin Koulena, Yodai Takei, Jina Yun, Christopher Cronin, Christoph Karp, Guo-Cheng Yuan, Long Cai
   Nature (2019-03-25) https://doi.org/gfxgqx
   DOI: 10.1038/s41586-019-1049-y · PMID: 30911168 · PMCID: PMC6544023

10. **Giotto, a toolbox for integrative analysis and visualization of spatial expression data**
    Ruben Dries, Qian Zhu, Rui Dong, Chee-Huat Linus Eng, Huipeng Li, Kan Liu, Yuntian Fu, Tianxiao Zhao, Arpan Sarkar, Feng Bao, … Guo-Cheng Yuan
    bioRxiv (2020-05-30) https://doi.org/gg84qf
    DOI: 10.1101/701680

11. **Spatial reconstruction of single-cell gene expression data**
    Rahul Satija, Jeffrey A Farrell, David Gennert, Alexander F Schier, Aviv Regev
    Nature Biotechnology (2015-04-13) https://doi.org/f7bmck
    DOI: 10.1038/nbt.3192 · PMID: 25867923 · PMCID: PMC4430369

12. **Cluster Validation by Prediction Strength**
    Robert Tibshirani, Guenther Walther
    Journal of Computational and Graphical Statistics (2005-09) https://doi.org/fvtcf4
    DOI: 10.1198/106186005x59243

13. **Matching species traits to environmental variables: a new three-table ordination method**
    S. Dolédec, D. Chessel, C. J. F. ter Braak, S. Champely
    Environmental and Ecological Statistics (1996-06) https://doi.org/fhwz55
    DOI: 10.1007/bf02427859

14. **Using single-cell genomics to understand developmental processes and cell fate decisions**
    Jonathan A Griffiths, Antonio Scialdone, John C Marioni
    Molecular Systems Biology (2018-04-16) https://doi.org/gdgbtq
    DOI: 10.15252/msb.20178046 · PMID: 29661792 · PMCID: PMC5900446

15. **Reprogramming the Methylome: Erasing Memory and Creating Diversity**
    Heather J. Lee, Timothy A. Hore, Wolf Reik

Cell Stem Cell (2014-06) https://doi.org/f6f83c
DOI: 10.1016/j.stem.2014.05.008 · PMID: 24905162 · PMCID: PMC4051243

16. **scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells**
Stephen J. Clark, Ricard Argelaguet, ChantrioInt-Andreas Kapourani, Thomas M. Stubbs, Heather J. Lee, Celia Alda-Catalinas, Felix Krueger, Guido Sanguinetti, Gavin Kelsey, John C. Marioni, … Wolf Reik
Nature Communications (2018-02-22) https://doi.org/gc4q72
DOI: 10.1038/s41467-018-03149-4 · PMID: 29472610 · PMCID: PMC5823944

17. **Multi-omics profiling of mouse gastrulation at single-cell resolution**
Ricard Argelaguet, Stephen J. Clark, Hisham Mohammed, L. Carine Stapel, Christel Krueger, ChantrioInt-Andreas Kapourani, Ivan Imaz-Rosshandler, Tim Lohoff, Yunlong Xiang, Courtney W. Hanna, … Wolf Reik
Nature (2019-12-11) https://doi.org/ggfrnn
DOI: 10.1038/s41586-019-1825-8 · PMID: 31827285 · PMCID: PMC6924995

18. **Pan-cancer identification of clinically relevant genomic subtypes using outcome-weighted integrative clustering**
Arshi Arora, Adam B. Olshen, Venkatraman E. Seshan, Ronglai Shen
bioRxiv (2020-05-12) https://doi.org/gg8np9
DOI: 10.1101/2020.05.11.084798

19. **Single-Cell Multi-omic Integration Compares and Contrasts Features of Brain Cell Identity**
Joshua D. Welch, Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, Evan Z. Macosko
Cell (2019-06) https://doi.org/gf3m3v
DOI: 10.1016/j.cell.2019.05.006 · PMID: 31178122 · PMCID: PMC6716797

20. **Variable selection for generalized canonical correlation analysis**
A. Tenenhaus, C. Philippe, V. Guillemot, K.-A. Le Cao, J. Grill, V. Frouin
Biostatistics (2014-02-17) https://doi.org/gg583d
DOI: 10.1093/biostatistics/kxu001 · PMID: 24550197

21. **mixOmics: An R package for ‘omics feature selection and multiple data integration**
Florian Rohart, Benoît Gautier, Amrit Singh, Kim-Anh Lê Cao
PLOS Computational Biology (2017-11-03) https://doi.org/gcj84s
DOI: 10.1371/journal.pcbi.1005752 · PMID: 29099853 · PMCID: PMC5687754

22. **impute**
Robert Tibshirani Trevor Hastie
Bioconductor (2017) https://doi.org/gg9dds
DOI: 10.18129/b9.bioc.impute

23. **MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data**
Ricard Argelaguet, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C. Marioni, Oliver Stegle
Genome Biology (2020-05-11) https://doi.org/ggvwsr
DOI: https://doi.org/10.1186/s13059-020-02015-1

24. **Adjusting batch effects in microarray expression data using empirical Bayes methods**
W. Evan Johnson, Cheng Li, Ariel Rabinovic
Biostatistics (2007-01) https://doi.org/dsf386
DOI: 10.1093/biostatistics/kxj037 · PMID: 16632515

25. **Variance stabilization applied to microarray data calibration and to the quantification of differential expression**
W. Huber, A. von Heydebreck, H. Sultmann, A. Poustka, M. Vingron
Bioinformatics (2002-07-01) https://doi.org/dbb6xx
DOI: 10.1093/bioinformatics/18.suppl_1.s96 · PMID: 12169536

26. **Integrative inference of brain cell similarities and differences from single-cell genomics**
Joshua Welch, Velina Kozareva, Ashley Ferreira, Charles Vanderburg, Carly Martin, Evan Macosko
bioRxiv (2018-11-02) https://doi.org/gfgr7b
DOI: 10.1101/459891

27. **mogsa**
Chen Meng
Bioconductor (2017) https://doi.org/gg583f
DOI: 10.18129/b9.bioc.mogsa

28. **Combining the fourth-corner and the RLQ methods for assessing trait responses to environmental variation**
Stéphane Dray, Philippe Choler, Sylvain Dolédec, Pedro R. Peres-Neto, Wilfried Thuiller, Sandrine Pavoine, Cajo J. F. ter Braak
Ecology (2014-01) https://doi.org/gdsf9z
DOI: 10.1890/13-0196.1 · PMID: 24649641

29. **What Is Your Conceptual Definition of “Cell Type” in the Context of a Mature Organism?**
Cell Systems
(2017-03) https://doi.org/d38b
DOI: 10.1016/j.cels.2017.03.006 · PMID: 28334573

30. **Multiple factor analysis.**
L. L. Thurstone
Psychological Review (1931) https://doi.org/dq4k9p
DOI: 10.1037/h0069792

31. **The ACT (STATIS method)**
Christine Lavit, Yves Escoufier, Robert Sabatier, Pierre Traissac
Computational Statistics & Data Analysis (1994-08) https://doi.org/c8xttz
DOI: 10.1016/0167-9473(94)90134-1

32. **Multivariate data analysis: The French way**
Susan Holmes
Institute of Mathematical Statistics (2008) https://doi.org/cmnf7j
DOI: 10.1214/193940307000000455

33. **Multitable Methods for Microbiome Data Integration**
Kris Sankaran, Susan P. Holmes
Frontiers in Genetics (2019-08-28) https://doi.org/gf8dqn
DOI: 10.3389/fgene.2019.00627 · PMID: 31555316 · PMCID: PMC6724662

34. **Quantitative, Architectural Analysis of Immune Cell Subsets in Tumor-Draining Lymph Nodes from Breast Cancer Patients and Healthy Lymph Nodes**
A. Francesca Setiadi, Nelson C. Ray, Holbrook E. Kohrt, Adam Kapelner, Valeria Carcamo-Cavazos, Edina B. Levic, Sina Yadegarynia, Chris M. van der Loos, Erich J. Schwartz, Susan Holmes, Peter P. Lee
PLoS ONE (2010-08-25) https://doi.org/bp4qj5
DOI: 10.1371/journal.pone.0012420 · PMID: 20811638 · PMCID: PMC2928294

35. **Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt**
Steffen Durinck, Paul T Spellman, Ewan Birney, Wolfgang Huber
Nature Protocols (2009-07-23) https://doi.org/c4b7dd
DOI: 10.1038/nprot.2009.97 · PMID: 19617889 · PMCID: PMC3159387

36. **Characterization of the Impact of Daclizumab Beta on Circulating Natural Killer Cells by Mass Cytometry**
Thanmayi Ranganath, Laura J. Simpson, Anne-Maud Ferreira, Christof Seiler, Elena Vendrame, Nancy Zhao, Jason D. Fontenot, Susan Holmes, Catherine A. Blish
Frontiers in Immunology (2020-04-24) https://doi.org/gg5jcr
DOI: 10.3389/fimmu.2020.00714 · PMID: 32391016 · PMCID: PMC7194113

37. **Machine learning: a probabilistic perspective**
Kevin P. Murphy
MIT Press (2012)
ISBN: 9780262018029

38. **Dimensionality reduction for visualizing single-cell data using UMAP**
Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel WH Kwok, Lai Guan Ng, Florent Ginhoux, Evan W Newell
Nature Biotechnology (2018-12-03) https://doi.org/gfkwzq
DOI: 10.1038/nbt.4314 · PMID: 30531897

39. **DIABLO: an integrative approach for identifying key molecular drivers from multi-omics assays**
Amrit Singh, Casey P Shannon, Benoît Gautier, Florian Rohart, Michaël Vacher, Scott J Tebbutt, Kim-Anh Lê Cao
Bioinformatics (2019-09-01) https://doi.org/ggpt9c
DOI: 10.1093/bioinformatics/bty1054 · PMID: 30657866 · PMCID: PMC6735831

40. **Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments**
Luyi Tian, Xueyi Dong, Saskia Freytag, Kim-Anh Lê Cao, Shian Su, Abolfazl JalalAbadi, Daniela Amann-Zalcenstein, Tom S. Weber, Azadeh Seidi, Jafar S. Jabbari, … Matthew E. Ritchie
Nature Methods (2019-05-27) https://doi.org/gf3jhp
DOI: 10.1038/s41592-019-0425-8 · PMID: 31133762

41. **Benchmarking single-cell RNA-sequencing protocols for cell atlas projects**
Elisabetta Mereu, Atefeh Lafzi, Catia Moutinho, Christoph Ziegenhain, Davis J. McCarthy, Adrián Álvarez-Varela, Eduard Batlle, Sagar, Dominic Grün, Julia K. Lau, … Holger Heyn
Nature Biotechnology (2020-04-06) https://doi.org/ggrbbh
DOI: 10.1038/s41587-020-0469-4 · PMID: 32518403

42. **Systematic comparison of single-cell and single-nucleus RNA-sequencing methods**
Jiarui Ding, Xian Adiconis, Sean K. Simmons, Monika S. Kowalczyk, Cynthia C. Hession, Nemanja D. Marjanovic, Travis K. Hughes, Marc H. Wadsworth, Tyler Burks, Lan T. Nguyen, … Joshua Z. Levin
Nature Biotechnology (2020-04-06) https://doi.org/ggrksw
DOI: 10.1038/s41587-020-0465-8 · PMID: 32341560 · PMCID: PMC7289686

43. **Accounting for technical noise in single-cell RNA-seq experiments**
Philip Brennecke, Simon Anders, Jong Kyoung Kim, Aleksandra A Kołodziejczyk, Xiuwei Zhang, Valentina Proserpio, Bianka Baying, Vladimir Benes, Sarah A Teichmann, John C Marioni, Marcus G Heisler
Nature Methods (2013-09-22) https://doi.org/gbd3mc
DOI: 10.1038/nmeth.2645 · PMID: 24056876

44. **Splatter: simulation of single-cell RNA sequencing data**
Luke Zappia, Belinda Phipson, Alicia Oshlack
Genome Biology (2017-09-12) https://doi.org/gc3h3g
DOI: 10.1186/s13059-017-1305-0 · PMID: 28899397 · PMCID: PMC5596896

45. **A Sparse PLS for Variable Selection when Integrating Omics Data**
Kim-Anh Lê Cao, Debra Rossouw, Christèle Robert-Granié, Philippe Besse

Statistical Applications in Genetics and Molecular Biology (2008-01-18) https://doi.org/cw7zft
DOI: 10.2202/1544-6115.1390 · PMID: 19049491

46. **Sparse principal component analysis via regularized low rank matrix approximation**
Haipeng Shen, Jianhua Z. Huang
Journal of Multivariate Analysis (2008-07) https://doi.org/b7x3cc
DOI: 10.1016/j.jmva.2007.06.007

47. **Quantifying the Association between Gene Expressions and DNA-Markers by Penalized Canonical Correlation Analysis**
Sandra Waaijenborg, Philip C. Verselewel de Witt Hamer, Aeilko H Zwinderman
Statistical Applications in Genetics and Molecular Biology (2008-01-23) https://doi.org/bpzb68
DOI: 10.2202/1544-6115.1329 · PMID: 18241193

48. CCA **: An *R* Package to Extend Canonical Correlation Analysis**
Ignacio Gonzalez, Sébastien Déjean, Pascal Martin, Alain Baccini
Journal of Statistical Software (2008) https://doi.org/gf4f5m
DOI: 10.18637/jss.v023.i12

49. **HIGHLIGHTING RELATIONSHIPS BETWEEN HETEROGENEOUS BIOLOGICAL DATA THROUGH GRAPHICAL DISPLAYS BASED ON REGULARIZED CANONICAL CORRELATION ANALYSIS**
I. GONZÁLEZ, S. DÉJEAN, P. G. P. MARTIN, O. GONÇALVES, P. BESSE, A. BACCINI
Journal of Biological Systems (2011-11-21) https://doi.org/bmbjf5
DOI: 10.1142/s0218339009002831

50. **A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis**
D. M. Witten, R. Tibshirani, T. Hastie
Biostatistics (2009-04-17) https://doi.org/fd4g54
DOI: 10.1093/biostatistics/kxp008 · PMID: 19377034 · PMCID: PMC2697346

51. **Sparse Canonical Correlation Analysis with Application to Genomic Data Integration**
Elena Parkhomenko, David Tritchler, Joseph Beyene
Statistical Applications in Genetics and Molecular Biology (2009-01-06) https://doi.org/b7x4jb
DOI: 10.2202/1544-6115.1406 · PMID: 19222376

52. **Integrative analysis of gene expression and copy number alterations using canonical correlation analysis**
Charlotte Soneson, Henrik Lilljebjörn, Thoas Fioretos, Magnus Fontes
BMC Bioinformatics (2010-04-15) https://doi.org/dtxhsx
DOI: 10.1186/1471-2105-11-191 · PMID: 20398334 · PMCID: PMC2873536

53. **Gene expression signatures modulated by epidermal growth factor receptor activation and their relationship to cetuximab resistance in head and neck squamous cell carcinoma**
Elana J Fertig, Qing Ren, Haixia Cheng, Hiromitsu Hatakeyama, Adam P Dicker, Ulrich Rodeck, Michael Considine, Michael F Ochs, Christine H Chung
BMC Genomics (2012) https://doi.org/gb3fgp
DOI: 10.1186/1471-2164-13-160 · PMID: 22549044 · PMCID: PMC3460736

54. **Identifying multi-layer gene regulatory modules from multi-dimensional genomic data**
W. Li, S. Zhang, C.-C. Liu, X. J. Zhou
Bioinformatics (2012-08-03) https://doi.org/f4d488
DOI: 10.1093/bioinformatics/bts476 · PMID: 22863767 · PMCID: PMC3463121

55.**:{unav)**
Aedín C Culhane, Guy Perrière, Desmond G Higgins
BMC Bioinformatics (2003) https://doi.org/fckzmd
DOI: 10.1186/1471-2105-4-59 · PMID: 14633289 · PMCID: PMC317282

56. **Extensions of Sparse Canonical Correlation Analysis with Applications to Genomic Data**
Daniela M Witten, Robert J. Tibshirani
Statistical Applications in Genetics and Molecular Biology (2009-01-09) https://doi.org/b45jtg
DOI: 10.2202/1544-6115.1470 · PMID: 19572827 · PMCID: PMC2861323

57. **MOGSA: Integrative Single Sample Gene-set Analysis of Multiple Omics Data**
Chen Meng, Azfar Basunia, Bjoern Peters, Amin Moghaddas Gholami, Bernhard Kuster, Aedín C. Culhane
Molecular & Cellular Proteomics (2019-08-09) https://doi.org/ggf3j3
DOI: 10.1074/mcp.tir118.001251 · PMID: 31243065 · PMCID: PMC6692785

58. **Consistency and overfitting of multi-omics methods on experimental data**
Sean D McCabe, Dan-Yu Lin, Michael I Love
Briefings in Bioinformatics (2020-07) https://doi.org/gghpmf
DOI: 10.1093/bib/bbz070 · PMID: 31281919 · PMCID: PMC7373174

59. **Bootstrapping cluster analysis: Assessing the reliability of conclusions from microarray experiments**
M. K. Kerr, G. A. Churchill
Proceedings of the National Academy of Sciences (2001-07-24) https://doi.org/cgpp6p
DOI: 10.1073/pnas.161273698 · PMID: 11470909 · PMCID: PMC55356

60.**:{unav)**
Sandrine Dudoit, Jane Fridlyand
Genome Biology (2002) https://doi.org/drffcb
DOI: 10.1186/gb-2002-3-7-research0036 · PMID: 12184810 · PMCID: PMC126241

61. **A Three-Gene Model to Robustly Identify Breast Cancer Molecular Subtypes**
Benjamin Haibe-Kains, Christine Desmedt, Sherene Loi, Aedin C. Culhane, Gianluca Bontempi, John Quackenbush, Christos Sotiriou
JNCI: Journal of the National Cancer Institute (2012-02-22) https://doi.org/fzb27r
DOI: 10.1093/jnci/djr545 · PMID: 22262870 · PMCID: PMC3283537

62. **A simple, scalable approach to building a cross-platform transcriptome atlas**
Paul W Angel, Nadia Rajab, Yidi Deng, Chris M Pacheco, Tyrone Chen, Kim-Anh Lê Cao, Jarny Choi, Christine A Wells
bioRxiv (2020-03-11) https://doi.org/gg898g
DOI: 10.1101/2020.03.09.984468

63. **A federated ecosystem for sharing genomic, clinical data**
The Global Alliance for Genomics and Health
Science (2016-06-09) https://doi.org/ggctm3
DOI: 10.1126/science.aaf6162 · PMID: 27284183

64. **GrimoireLab - Software Development and Community Analytics platform** https://chaoss.github.io/grimoirelab/

65. http://ceur-ws.org/Vol-987/3.pdf

66. **Bioconductor - Home** https://bioconductor.org/

67. **Bioconductor build/check results** https://bioconductor.org/checkResults/

68. https://bioconductor.org/support

69. **MultiAssayExperiment**
Marcel Ramos [Aut, Cre], Levi Waldron [Aut], MultiAssay SIG[Ctb]
Bioconductor (2017) https://doi.org/gg6p3d
DOI: 10.18129/b9.bioc.multiassayexperiment

70. **Software for the Integration of Multiomics Experiments in Bioconductor**
Marcel Ramos, Lucas Schiffer, Angela Re, Rimsha Azhar, Azfar Basunia, Carmen Rodriguez, Tiffany Chan, Phil Chapman, Sean R. Davis, David Gomez-Cabrero, … Levi Waldron
Cancer Research (2017-10-31) https://doi.org/gcj278
DOI: 10.1158/0008-5472.can-17-0344 · PMID: 29092936 · PMCID: PMC5679241

71. **ExperimentHub**
Bioconductor Package Maintainer
Bioconductor (2017) https://doi.org/gg6p3c
DOI: 10.18129/b9.bioc.experimenthub

72. **DelayedArray**
Hervé Pagès
Bioconductor (2017) https://doi.org/gg5tw4
DOI: 10.18129/b9.bioc.delayedarray

73. **rhdf5**
Bernd Fischer [Aut], Gregoire Pau [Aut], Mike Smith [Aut, Cre]
Bioconductor (2017) https://doi.org/gg5tw6
DOI: 10.18129/b9.bioc.rhdf5

74. **mbkmeans: fast clustering for single cell data using mini-batch *k*-means**
Stephanie C. Hicks, Ruoxi Liu, Yuwei Ni, Elizabeth Purdom, Davide Risso
bioRxiv (2020-05-27) https://doi.org/gg5tw3
DOI: 10.1101/2020.05.27.119438

75. **mbkmeans**
Yuwei Ni, Davide Risso, Stephanie Hicks, Elizabeth Purdom
Bioconductor https://doi.org/gg5tw5
DOI: 10.18129/b9.bioc.mbkmeans

76. **SCANPY: large-scale single-cell gene expression data analysis**
F. Alexander Wolf, Philipp Angerer, Fabian J. Theis
Genome Biology (2018-02-06) https://doi.org/gc22s9
DOI: 10.1186/s13059-017-1382-0 · PMID: 29409532 · PMCID: PMC5802054

77. **SingleCellExperiment**
Aaron Lun [Aut, Cph], Davide Risso [Aut, Cre, Cph]
Bioconductor (2017) https://doi.org/gg5wfr
DOI: 10.18129/b9.bioc.singlecellexperiment

78. **GenomicRanges**
H. Pagès P. Aboyoun
Bioconductor (2017) https://doi.org/gg6rfz
DOI: 10.18129/b9.bioc.genomicranges

79. **SPOTlight: Seeded NMF regression to Deconvolute Spatial Transcriptomics Spots with Single-Cell Transcriptomes**
Marc Elosua, Paula Nieto, Elisabetta Mereu, Ivo Gut, Holger Heyn
bioRxiv (2020-06-04) https://doi.org/gg6rfx
DOI: 10.1101/2020.06.03.131334

80. **Points of view: Color blindness**
Bang Wong
Nature Methods (2011-06-01) https://www.nature.com/articles/nmeth.1618
DOI: 10.1038/nmeth.1618

81. **Color coding**
Bang Wong
Nature Methods (2010-08) https://doi.org/dhm3mz
DOI: 10.1038/nmeth0810-573 · PMID: 20704014

82. **The viridis color palettes** https://cran.r-project.org/web/packages/viridis/vignettes/intro-to-viridis.html

83. **dtm2451/dittoSeq**
Daniel Bunis
(2020-08-26) https://github.com/dtm2451/dittoSeq

84. **Highly multiplexed molecular and cellular mapping of breast cancer tissue in three dimensions using mass tomography**
Raúl Catena, Alaz Özcan, Laura Kütt, Alex Plüss, Peter Schraml, Holger Moch, Bernd Bodenmiller, IMAXT Consortium
Cold Spring Harbor Laboratory (2020-05-25) https://doi.org/gg87jf
DOI: 10.1101/2020.05.24.113571

85. **ZipSeq: barcoding for real-time mapping of single cell transcriptomes**
Kenneth H. Hu, John P. Eichorst, Chris S. McGinnis, David M. Patterson, Eric D. Chow, Kelly Kersten, Stephen C. Jameson, Zev J. Gartner, Arjun A. Rao, Matthew F. Krummel
Nature Methods (2020-07-06) https://doi.org/gg87jd
DOI: 10.1038/s41592-020-0880-2 · PMID: 32632238

86. **Orchestrating single-cell analysis with Bioconductor**
Robert A. Amezquita, Aaron T. L. Lun, Etienne Becht, Vince J. Carey, Lindsay N. Carpp, Ludwig Geistlinger, Federico Marini, Kevin Rue-Albrecht, Davide Risso, Charlotte Soneson, … Stephanie C. Hicks
Nature Methods (2019-12-02) https://doi.org/ggdxgx
DOI: 10.1038/s41592-019-0654-x · PMID: 31792435 · PMCID: PMC7358058

87. **Integrating single-cell transcriptomic data across different conditions, technologies, and species**
Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, Rahul Satija
Nature Biotechnology (2018-04-02) https://doi.org/gc87v6
DOI: 10.1038/nbt.4096 · PMID: 29608179 · PMCID: PMC6700744

88. **The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution**
Orit Rozenblatt-Rosen, Aviv Regev, Philipp Oberdoerffer, Tal Nawy, Anna Hupalowska, Jennifer E. Rood, Orr Ashenberg, Ethan Cerami, Robert J. Coffey, Emek Demir, … Xiaowei Zhuang
Cell (2020-04) https://doi.org/ggtkzd
DOI: 10.1016/j.cell.2020.03.053 · PMID: 32302568 · PMCID: PMC7376497

89. **The Human Cell Atlas**
Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, … Human Cell Atlas Meeting Participants
eLife (2017-12-05) https://doi.org/gcnzcv
DOI: 10.7554/elife.27041 · PMID: 29206104 · PMCID: PMC5762154

90. **The Human Cell Atlas: from vision to reality**
Orit Rozenblatt-Rosen, Michael J. T. Stubbington, Aviv Regev, Sarah A. Teichmann
Nature (2017-10-26) https://doi.org/gfgkr8
DOI: 10.1038/550451a · PMID: 29072289

91. **Multivariate analysis of multiblock and multigroup data**
A. Eslami, E. M. Qannari, A. Kohler, S. Bougeard
Chemometrics and Intelligent Laboratory Systems (2014-04) https://doi.org/f52wrr
DOI: 10.1016/j.chemolab.2014.01.016

92. **Dimension reduction techniques for the integrative analysis of multi-omics data**
Chen Meng, Oana A. Zeleznik, Gerhard G. Thallinger, Bernhard Kuster, Amin M. Gholami, Aedín C. Culhane
Briefings in Bioinformatics (2016-07) https://doi.org/f83qvd
DOI: 10.1093/bib/bbv108 · PMID: 26969681 · PMCID: PMC4945831

93. **Robust meta-analysis of gene expression using the elastic net**
Jacob J. Hughey, Atul J. Butte
Nucleic Acids Research (2015-07-13) https://doi.org/f7nnbm
DOI: 10.1093/nar/gkv229 · PMID: 25829177 · PMCID: PMC4499117

94. **Biomarker definitions and their applications**
Robert M Califf
Experimental Biology and Medicine (2018-02-06) https://doi.org/gcxh8n
DOI: 10.1177/1535370217750088 · PMID: 29405771 · PMCID: PMC5813875

95. **Biomarker signatures of aging**
Paola Sebastiani, Bharat Thyagarajan, Fangui Sun, Nicole Schupf, Anne B. Newman, Monty Montano, Thomas T. Perls
Aging Cell (2017-04) https://doi.org/d2cm
DOI: 10.1111/acel.12557 · PMID: 28058805 · PMCID: PMC5334528

96. **Biomarker Panels in Critical Care**
Susan R. Conway, Hector R. Wong
Critical Care Clinics (2020-01) https://doi.org/d2cn
DOI: 10.1016/j.ccc.2019.08.007 · PMID: 31733684

---

1. Supposes a thesis (e.g. the guilt of an accused man) is supported by a great deal of circumstantial evidence of different forms, but in agreement with each other; then even if each piece of evidence is in itself insufficient to produce any strong belief, the thesis is decisively strengthened by their joint effect.↩