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

This manuscript (permalink) was automatically generated from BIRSBiointegration/whitePaper@3e0521c on August 25, 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 JalalAbadi**
  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

- **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

Draft outline

Full outline described in 01.outline.md as comment if you need to go back to the big picture.

Figures, Tables and online resources

Figures

- Figure 1: Outline of manuscript (Provided below)
- Figure 2: scRNA-seq + seqFISH main results
- Figure 3: scTargeted proteomics main results
- Figure 4: scNMT-seq main results
- Figure 5: Illustration of partial overlap between studies (at the cells and / or features level)
- Figure 6: Within study cross-validation
- Figure 7: Infrastructure of Giotto package as illustration

Tables

- Table 1: Summary of tasks and methods across all hackathons highlighting methods and common challenges (Provided below)
- Table 2: Glossary of terms, consensus terms will be used throughout the manuscript
- Table 3: Benchmarking single cell data sets
- Table 4: List of single cell analysis software

Online resources

- Online resource 1: Three hackathon datasets (github)
- Online resource 2: R packages with open source reproducible vignettes (12 vignettes)

Abstract

Multi-omics hackathon studies to illustrate standards and challenges

The rapid pace of technological development can be seen as only six years after single-cell sequencing was name method of the year, single-cell multimodal omics claimed the title [1]. 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. 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. Thus, a central question is what are the optimal analysis methods that can be carried forward from these data from which to learn novel biological processes.

The hackathon studies we selected for the Mathematical Frameworks for Integrative Analysis of Emerging Biological Data Workshop tailored independent challenges of data integration. The first challenge included spatial molecular profiling. While this technology is rapidly emerging, it often provides lower molecular resolution than 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 provide an unique opportunity to identify technology-specific challenges and unifying themes that are essential to effectively employ multi-omics datasets into new biological knowledge.

Figure 1: Main challenges discussed during our brainstorming sessions from the hackathons.

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

Overview and biological question

The first hackathon aimed to leverage the complemenatary 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 2: 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 SVM. 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 [3] (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) **Josh I need an update of that figure, could you change tasic to scRNA-seq in the caption?**. **D** Construction of a spatial network from cells' positions using Voronoi tessellation, where cell types were inferred from a Support Vector Machine (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 2A). 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 support vector machine (SVM) and supervised and semi-supervised generalized linear model regularized with elastic net penalty (enet and ssenet). The ssenet approach built a model iteratively: it combined both datasets and initially only retained the highest confidence labels, then gradually added more cell type labels until all cells were classified (Figure 2A). 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 combined 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 2A). 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.

Finally, 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 2D).

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

[suggestion 1]
How could one identify spatial patterns in the seqFISH data? This broad question can be divided into a number of specific tasks, such as detecting genes whose expression is spatially coherent, cell types whose spatial distribution is confined to distinct regions, recurrent multi-cell-type interaction clusters, etc, and the spatial scale may vary from subcellular all the way to tissue-wide organizations. While there have been abundant studies in the geo-spatial analysis domain, computational tools targeting specific spatial transcriptomic questions are still lacking. In this hackathon, one group tackled this challenge by aggregating gene expression data from neighboring cells followed by spatial clustering (Coullomb). Much more work in the future is need to further explore such information.

[suggestion 2]
Most analyses that were originally developed for scRNAseq data can be immediately applied to spatial transcriptomic datasets, however methods to extract sources of variation that originate from spatial factors are still sparse. To incorporate spatial information the cells from the seqFISH dataset were first connected through a spatial network based on Voronoi tessellation and then the expression of each individual cell was spatially smoothed by calculating the average gene expression levels over all the neighboring cells. This smoothened and aggregated data matrix was subsequently used to create a 2 dimensional UMAP from which clusters were identified through a density based clustering approach. The obtained cluster labels can then be mapped back to the original spatial locations for further visual inspection and analysis.

[suggestion 2 detailed]
Most analyses that were originally developed for scRNAseq data can be adapted for spatial transcriptomic datasets, however methods to extract sources of variation that originate from spatial factors are still sparse. The latter requires the development of 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. For this hackaton we aimed to incorporate the spatial information by connecting the cells from the seqFISH dataset through a spatial network based on Voronoi tessellation. Then, for each node, its RNA count data was aggregated with its first order neighbors' count data. The mean and standard deviation were computed for each gene in the gathered data in order to capture the global tendency as well as the variability in the area arround each node. Thus, each node has `nb_genes x nb_statistics (here 2)` variables. These "aggregation statistics" can be visualized on a 2D UMAP projection. These data were clustered in UMAP reduced spaces of dimensionality between 2 and 9, higher dimensions allowing to define more fine-grained clusters. The clustering was performed with HDBSCAN, a noise-aware density-based algorithm that can define arbitrary-shaped clusters. These clusters can then be visualized on the 2D UMAP projection and on the 2D spatial map of seqFISH data. The clusters are spatially coherent, some of them contain several cell types, and a given cell type is not necessarily limited to one specific cluster. During the exploratory phase consisting in varying the number of dimensions and the minium cluster size, a specific spot area was found clustered for several parameters combinations, suggesting it wasn't an artifact of the choice of parameters. "Differential expression" analysis was performed between this spot and the other areas, although we don't look at differences in gene data but in aggregation metrics. This area seems to correspond to a "regeneration hub", but this analysis has to be considered carefully and further analyzes is required to confirm this hypothesis. This "neighbors aggregation" method has been extended to aggregate RNA counts (or other node attributes) to higher orders of neighbors in order to define aggregation metrics on wider areas, which could be useful for analyzes of bigger tissues. One interesting extension would be to substract phenotypes contributions to RNA counts for each cell before performing the neighbors aggregation analysis in order to highlight genes that are modulated by spatial factors. But if we want to retrieve the mean expression of a cluster for cells belonging to it, we should first check cluster's convexity and be sure that no other cluster lies within it.

Spatial proteomics and cross-study analysis

Overview and biological question

The single-cell targeted proteomics hackathon investigated the tumor-immune microenvironment of primary breast cancer tissue. The challenge was the joint cross-study, cross-platform integrative analysis of single cell proteomics data that had low feature overlap, were generated on different antibody-based targeted proteomics technological platforms and in different laboratories [1, 2]. Mass cytometry (CyTOF) from Wagner et al. [4] measured 73 proteins in two panels (immune, tumor) in 194 tissue samples from 143 subjects, of which 6 patients had triple-negative negative breast cancer. 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]. 3 Whereas the first hackathon had matched seqFISH and and scRNA-seq data of samples from the same biological conditions, the data in this hackathon is obtained from different cohorts of patients posing a further challenge for analysis.

Whilst this is a formidable data integration challenge, it reflects the bioinformatics analysis of clinical teams who wish to compare and investigate data collected on different measurement platforms and cohorts of cancer patients. The first questions simply asked if partially-overlapping proteomic data collected on different patients with similar phenotypes could be integrated, and asked participants to test if measurements in one technology could be transferred and used to predict information in the second. The MIBI data provided spatial location and expression of proteins, therefore the challenge asked if the spatial expression patterns of proteins measured on mass-tag could be predicted. Secondly it raised questions unique to spatial 'omics technologies. Does spatial capture unique information, beyond cell compositions. Could information about the spatial location of immune cell populations in breast cancer be discovered in integrated analyses of these datasets. Finally the datasets had no overlap in patients, so how could heterogeneous phenotype information be used to integrate patient 'omics data with low feature and no tumor biological sample overlap.

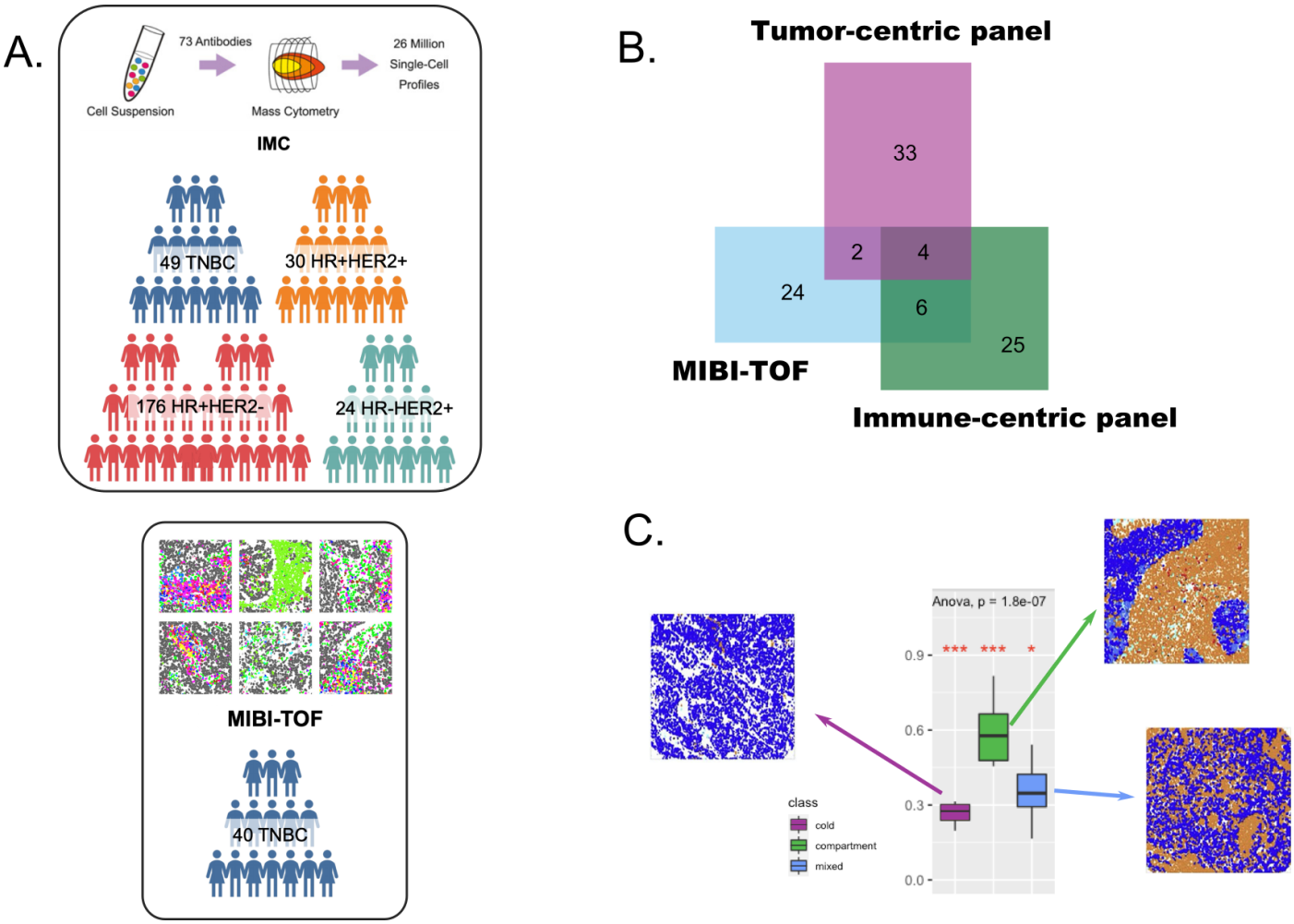


Figure 3: **A** Overview of sc_targeted proteins hackathon challenge. **B** Challenge 1: Partial to no overlap between protein features across studies. **C** Challenge 2: Spatial analysis with Moran's index, a computed on Gabriel graph (using both dummy variables and protein expression measurements) that differs significantly between groups. Boxplot of Moran's index values on tumor/immune status with examples from Keren et al. [5] corresponding to each tumor category. Red asterisks indicate significance of an ANOVA of each group with all others, and the p-value from an overall ANOVA across the three groups is reported above.

Computational challenges

Challenge 1: Lack of overlap between protein features across studies

The low number of features precluded integration of features at the level of gene set or pathways. There were only 20 proteins that were assayed in both studies (Figure 3A-B). The majority of features were cell-type markers or biomarkers targets of breast cancer therapeutic intervention. The limited overlap in these studies necessitated the use of surrogate measures of cross- study association.

Although the overlap in proteins was low, many proteins were cell-type markers, providing the opportunity to perform cross-study integration of cell type proportions in tumor tissue samples. Hackathon participants applied several semi-supervised and supervised algorithms to transfer cell labels and cell compositions from one dataset to the second. To capture the hierarchical structure of cell lineage, Lauren Hsu (Harvard) applied a simple random forest (RF) approach to perform feature transfer learning of cell type labels. An adaptation of the prediction strength approach described in Tibshirani [10] demonstrated model robustness: first, a model was trained on the labeled dataset and 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, she assessed the ability of the second model to recover the correct original labels when making predictions on the labeled dataset.

Yingxin Lin (Univ of Sydney) mapped the cells from CyTOF to imaging with spatial information by solving an entropic regularization optimal transport problem [4, 11], utilising 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. By clustering on the imputed expression matrix, she was able to identify a sub tumour cell type that is not revealed in the original matrix.

However the scales of protein expression was a possible limiting factors When integrating cell compositions using the correlation expression of protein markers, some cell markers were expected on a range of cell types (e.g., CD45), whereas others are more specialized and represent 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, the sensitivity and specificity of protein markers for cell types, tissue and disease heterogeneity

The assignment of cell type relied on manually curated protein annotation, and was dependent on domain-specific knowledge; for example that 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. Participants expressed a 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. There is a need for protein expression atlases to support cell type classification and potentially if this could be developed from large scale consortiums IHC of proteins (Human Protein Atlas [5, 6]), although the antibodies used and their performances might vary between labs.

Standards/QC/Normalization

Challenge 2: spatial protein expression analysis

Whereas the CyToF mass spectrometry provided protein expression and counts/composition of cells in breast tumor-immune environment, the MIBI-TOF data provided spatial information that quantified cell attributes (shape/size/spatial coordinates) in addition to expression levels, thus providing 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 (eg tumor-immune boundary) or a polygon, which is a closed plane defined by a number of lines. A polygon 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. The neighborhood of polygons can be defined with a Euclidean distance or sphere, by a number of bounded cells or other measures, many of which were developed in geographical information science or ecology and assess if a spatially measured variable has a random, dispersed or clustered pattern [7].

Kris Sankaran examined the extent to which expression data could be used to predict spatial properties of tissue samples. To build predictors, cells were first clustered (K = 20) on the basis of protein expression. Sample-level expression summaries were defined as the proportions of cells belonging to each cluster. To build the spatial response variables, a K-nearest neighbor graph was obtained from cell centroids (K = 5). For each cell, the average distance to its 5 nearest neighbors was computed, reflecting its local density. Further, the entropy of the cluster memberships across nearest neighbors was found, reflecting local heterogeneity. To summarize samples' cell ecosystems, cell-level statistics were averaged across each sample's cells. A random forest model trained from expression to spatial predictors achieved an average cross validation RMSEs of (tk) for neighborhood size and entropy, respectively, relative to baselines of () obtainable by predicting the mean.

Dr. Pratheepa Jeganathan applied topic modelling and defined five topic trained on protein expression and cell compositions in the CyToF data were sufficient to predict cell co-locations, in 10% MIBI-ToF Test data. Pratheepa Jeganathan (Stanford) applied a Bayesian modelling approach based on latent Dirichlet allocation (Blei, Latent dirichlet allocation Journal of machine Learning research 3.Jan (2003): 993-1022). Topic modeling was used to identify the dominated topics and assign spatial location of MIBI-TOF cells to the CyTOF data or vice-versa, based on the topic distribution in each cell (**Ref topic modelling?**). 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. Cells from MIBI-TOF were depicted in five clusters (link to vignette) and were consistently based on the observed and predicted marker expression, but these clusters were not identified with only observed marker expressions. [further details from Pratheepa available in pdf file in debrief folder]

Yingxin Lin (Sydney) examined the prognostic performance of different higher level spatial metrics. She measured protein autocorrelation using Moran's Index (I) with a sphere distance, cell type localisation using nearest neighbour correlation, or cell type interaction composition, Ripley's L-function. High-dimensional Cox models with fused lasso penalty and random forest survival models were fitted utilising different features, including clinical features such as tumour stage, tumour grade, age and tumour size, as well as features like cell type composition. Evaluating by the c-index via cross-validation, the spatial metrics are found to be predictive, especially in triple negative breast cancer where clinical features such as grade are poorly prognostic.

Lauren Hsu (Harvard) also considered Moran's I but used a graph-based neighborhood measure (Gabriel graph, based on Delaunay triangulation) instead of a sphere euclidean distance, and found that Moran's I differed significantly between the three prognostic tumor scores described by Keren, et al. [5](Figure 3C).

Need for development of spatial measure - different in dimensions of RNA v proteins

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

Another question often faced in cross-study integration is integration of biological samples that are non-overlapping but have similar phenotypes. The aim is to identify biomarkers from the different omics data to predict the same phenotype, and, more importantly, to explore how the markers

selected from multiple datasets are in agreement with or distinct from each other. The integration of the markers from each dataset should enable to extend biological knowledge that is not available by single omics data. To solve the challenge, phenotypical data (such as the cell attributes) are the critical factors that should be used to link the two datasets (Figure 3D).

The participants were successful at data integration using patient phenotype measures such as grade, stage and overall survival. Breast cancer is highly heterogeneous, and multiple breast cancer molecular subtypes have been described [8, 9]. Both MIBI and Jackson data used different approaches to cell type annotation and had 13 proteins in common.

Borrowing from ecology and french school of ordination, Chen Meng (Munich) described this problem as a case of the fourth corner problem (or RLQ). Briefly, given two omics data where both rows (features) and columns (samples) are non overlapping, and phenotypical data available for each omics data, multiplying the two phenotypical factors will derive a bridging matrix that links the features of two omics data. We should note that the two phenotypical matrices need to be multipliable, i.e. the phenotypical data should describe the same phenotypical factors over the samples in the corresponding dataset. The Chessel fourth corner RLQ is a matrix decomposition method to solve the problem [11; doi:10.1111/ecog.02302]. It decomposes the bridging matrix (phenotypical matrix) into components, each of which often represents a specific phenotypical pattern in the data. The loading matrix of each of the omics data indicate how a feature is correlated with phenotypical factors.

scNMT-seq as a case-study for epigenetic regulation

Overview and biological question

The first two hackathons leveraged paired datasets from complementary technologies that measure the same molecular modality at distinct resolutions. The aim was to develop integration strategies to gain a holistic view into and enable high molecular and spatial resolution of biological systems. Datasets spanning disparate molecular scales, such as DNA and RNA level measurements, are critical to further resolve the regulatory networks that mediate cell fate decisions. For example, the maturation of single-cell sequencing technologies has enabled the identification of transcriptional profiles associated with lineage diversification and cell fate commitment[12]. Yet, the accompanying epigenetic changes and the role of epigenetic layers in driving cell fate decisions still remains poorly understood[13].

scNMT-seq is one of the first experimental protocols that enable simultaneous quantification of RNA expression and epigenetic information from individual cells[14]. Briefly, in scNMT-seq cells are incubated with a GpC methyltransferase enzyme that labels accessible GpC sites via DNA methylation. Thus, after bisulfite sequencing, GpC methylation marks can be interpreted as direct read-outs for chromatin accessibility. This stands in contrast to CpG methylation marks, which can be interpreted as endogenous DNA methylation. In addition, 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 hackathon leverages scNMT-seq data to assess the performance of data integration strategies to infer complex gene regulatory mechansisms driving cell fate decisions.

Data set description

Gastrulation marks a major lineage specification event in mammalian embryos, accompanied by profound transcriptional rewiring and epigenetic remodelling [6].

In this hackaton we used a data set where scNMT-seq was applied to study mouse gastrulation [15] by profiling cells across four developmental stages (E4.5 to E7.5). For simplicity, in the workshop 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 for the two modalities. Overview of data is shown in Figure ??A.

Applying standard dimensionality reduction algorithms confirms that all three embryonic stages can be separated on the basis of RNA expression. The task of the workshop was to leverage the multi-faceted nature of measurements to better resolve the single-cell subpopulations from distinct embyonic stages.

Computational integration

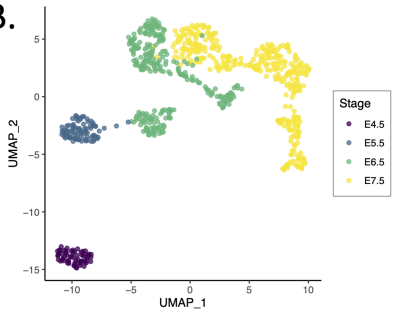
We considered three computational strategies:

- **block sPLS** [16]:** block sparse Projection to Latent Structures model implemented in mixOmics[17]. A sparse generalization of canonical correlation analysis that maximizes covariation between multiple sets of data modalities.
- **LIGER** [18]: unsupervised non-negative matrix factorization model for manifold alignment. It 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.
- **MOSAIC**: a multi-omics supervised integrative clustering algorithm inspired by `survClust` [19].

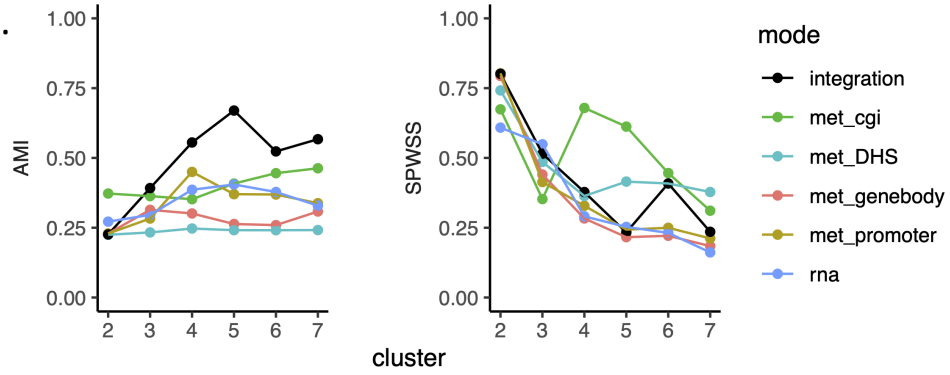
A.

Data view	Name	N (# of cells)	P (# of features)	Missing values
Gene expression	rna	799	18345	0%
Promoter methylation	met_promoter	799	17179	34%
Genebody methylation	met_genebody	799	17559	18%
CGI methylation	met_cgi	799	14080	45%
DHS methylation	met_DHS	799	6673	59%

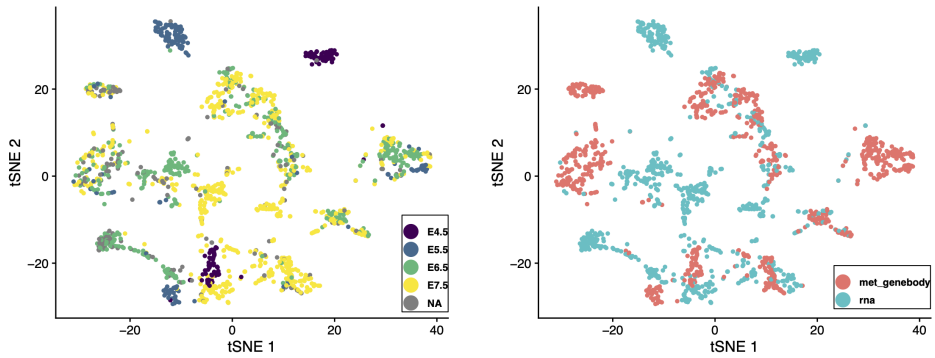
B.



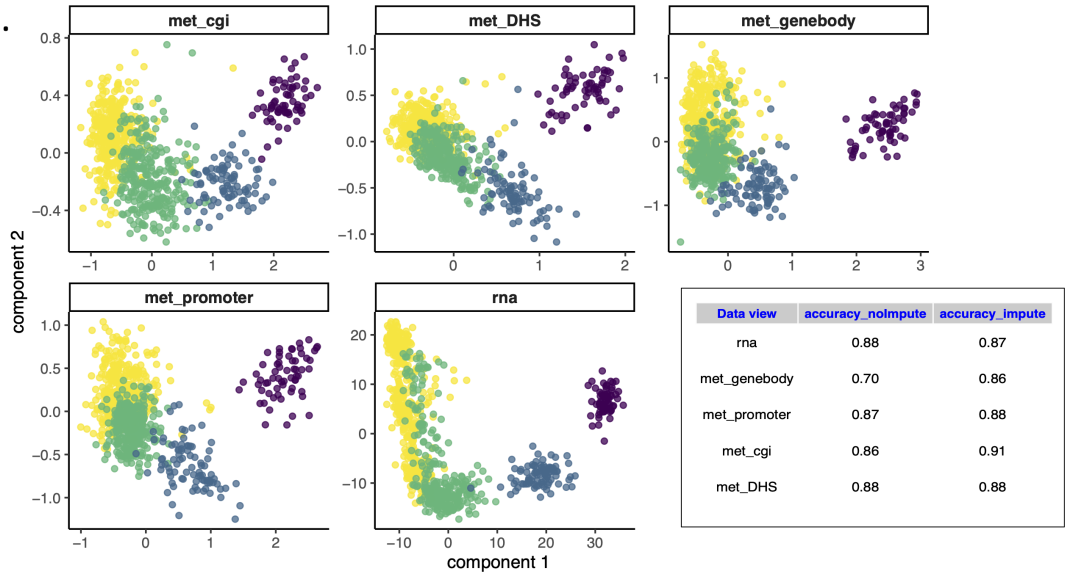
C.



D.



E.



Data view	accuracy_noImpute	accuracy_impute
rna	0.88	0.87
met_genebody	0.70	0.86
met_promoter	0.87	0.88
met_cgi	0.86	0.91
met_DHS	0.88	0.88

Overview of hackathon analyses. **A** A summary of a subset of data types used. Challenges 1 and 2: defining genomic contexts, varying dimensions and different amount of missing values across various modalities. **B** Uniform Manifold Approximation and Projection of RNA measurements using 671 highly variable genes resolves the various embryonic stages shown by different colors. **C** Supervised analysis using view-specific and integrative distance measures (MOSAIC). The integration finds 5 clusters of cell populations based on Adjusted Mutual Information (AMI) and Standardized Pooled Within Sum of Squares (SPWSS) and outperforms individual analyses. **D** Joint clustering of cells finds common and

view-specifc patterns of variation. LIGER uses non-negative matrix factorization to integrate transcriptome and genebody methylome and align cells in a common subspace. The poor alignment suggests decoupling of the gene expression and genebody methylation during gastrulation and highlights the importance of linking relevant epigenetic features to transcriptome (challenge 3). **E** The shared variation among views is largely driven by embryonic stage. Unsupervised integration with variable selection using Projection to Latent Structures finds covarying components from RNA and methylation data which separate the cells from each stage in all views (mean balanced accuracy of kmeans using components 1 and 2 = 88%). Each point represents a cell and the axes represent the derived components for each view. The method either omits the missing values (challenge 2) while integration or incorporates the inferred values in the integration (shown). Balanced accuracy measures for a k-means clustering analysis using the extracted components from the two scenarios shows that genebody methylation components have considerably improved in predicting the embryonic stage after imputation (bottom right).

Computational challenges

Challenge 1: defining genomic features

The first challenge 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 (1) the sparsity levels, (2) the binary nature of the read-outs, and (3) 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 input features.

There are two common strategies to define genomic elements. The first one is to use a running window approach across the entire genome. This strategy has been successful to distinguish heterogeneous cell types, but it does not improve interpretabiliy and it typically leads to an enormously large feature set. The alternative strategy is to adopt a biologically-informed approach where genic annotations (e.g. genebody, exon, or promoter regions) as well as ChIP-seq data or chromatin accessibility information is employed to restrict the feature space to genomic regions of regulatory potential.

[DESCRIBE RESULTS]

Our results confirm that the appropriate selection of the feature space is critical for a successful integration with RNA expression.

Challenge 2: Missing values in DNA methylation

Because of the low amounts of starting material, single-cell bisulfite sequencing protocols are limited by incomplete CpG coverage. Nonetheless, in contrast to scRNA-seq, missing data can be discriminated from dropouts.

Two strategies were put forward in the workshop to handle missing values. The first is to define an inference framework that omits missing values, as done in PLS and MOSAIC. The second approach, for methods that do not handle missing information, including LIGER, is to perform *a priori* imputation of DNA methylation values.

Here we compared the integration performance for PLS ~and MOSAIC~ with and without imputation. We observe that after inferring the missing values in methylation data using a nearest neighbors approach, the axes of concordnat variation with gene expression better resolve the population structure in some epigentic regions. To quantify any improvement in biological relevance of the extracted components of shared variation, we used a supervised analysis which used the number of embryonic stages as the number of clusters. Notably, the balanced accuracy of prediction based on genebody methylation components improved from 70% for the original data to 86% after inferring the missing values.

Challenge 3: Linking epigenetic features to gene expression

One of the main advantages of single-cell multi-modal assays 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 (in promoters and enhancers), H3K4me3 (in promoters) and H3K36me3 (in gene bodies), 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 be (partialy) 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 and it is mandatory to restrict the feature space to avoid a complex multiple testing problem.

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. Still, 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 to 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 across all studies, whilst cell prediction required machine or statistical learning methods (SVM, Enet) and spatial analysis Hidden Markov random field or Moran's Index. As the computational methodologies span technologies, so do the

central challenges highlighted in each hackathon. For example the accuracy of the analysis depends critically on data pre-processing, (normalization, upstream feature selection), differences in scale across data sets and overlap (or lack thereof) features and cells (Figure 4). In many cases, this preprocessing can yield data mapping to common molecular features, such as genes, that can be the focus of the integration. However, as we highlighted particularly in the example of the single cell proteomics challenge, many multi-omics questions may have limited shared genes of features between studies. In this case, analysis methods investigating hierarchical structure among the 'omics, cell and phenotype layers and applied a number of measures of higher order concordance are crticial for successful cross-study intergation. Even in cases with matching molecular features, analyses focusing on hierarchical structure 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 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 2Letter), 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, Coullomb selected 19 genes (*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 ([**???**]{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 6673 to 18345, Figure ??). Some projection-based methods, such as MOFA [20], require a similar number of features in each data set, whilst others such as PLS / sGCCA [16] are not limited by such data setting and enable flexible analysis (Abadi). 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), batch effect removal approaches, such as Combat [21] (Singh), or weighting specific data sets (Arora, Abadi) were considered and all offer 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 4) 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 [16]) 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). 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 ([**???**]{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].

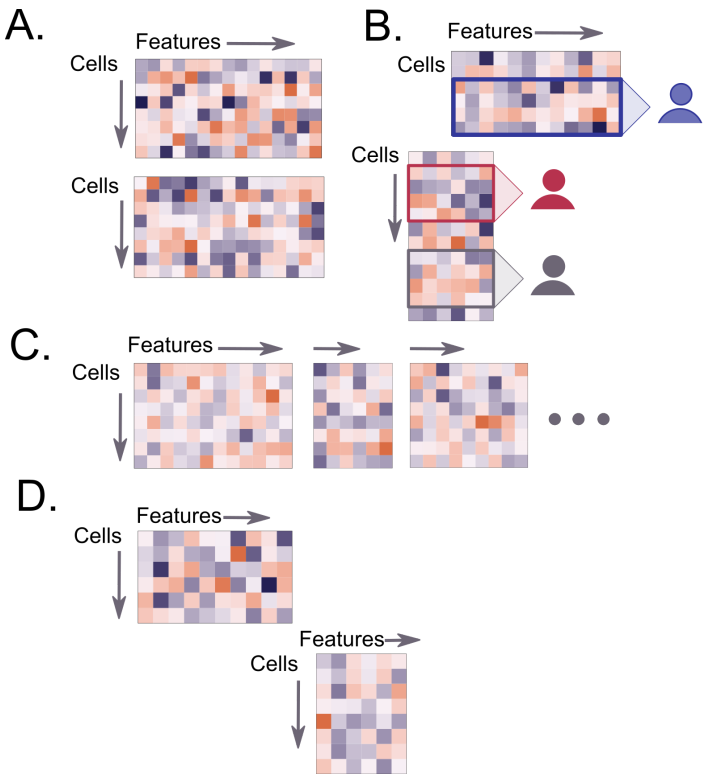


Figure 4: **A.** Overlap of features (genes) but not cells (e.g. spatial transcriptomics where cell type prediction for seqFISH data was performed based on scRNAseq where cell types are known. **B.** Partial overlap of features (proteins) but no overlap of cells (e.g. spatial proteomics that required 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** {#tbl:common}

Common challenges	**Tasks**	**sc Spatial**	**sc targeted proteomics**	**sc NMT-seq**

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 [22] (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 [23] (Sodicoff) (sc) ComBat (Singh) Projection methods MFA, sGCCA [16] (Singh*) (bulk) UMAP/tSNE (Sodicoff) (sc)	Multi-block PCA [24] Weighting matrices based on their similarities: STATIS, MFA (Chen*)(bulk) Scale MIBI-TOF to the range of CyTOF values (Jenagan)	LIGER [23] (Welch) (sc) Projection method sGCCA [16] (Abadi) (bulk) Multi Omics Supervised Integrative Clustering with weights (Arora) (bulk)
Overlap	**Cell overlap** (features not matching)			**Dimension reduction and projection methods:** LIGER [23] (Welch) (sc) sGCCA [16] (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 [24] (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 [23] (Welch)
	No cell overlap (partial feature overlap)		Topic modeling to predict cell spatial co-location or spatial expression (Jenathan, partial feature overlap)	

Common challenges	Tasks	sc Spatial	sc targeted proteomics	sc NMT-seq
	No overlap		RLQ [25] (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 [23] (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 [26].

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 [27].

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

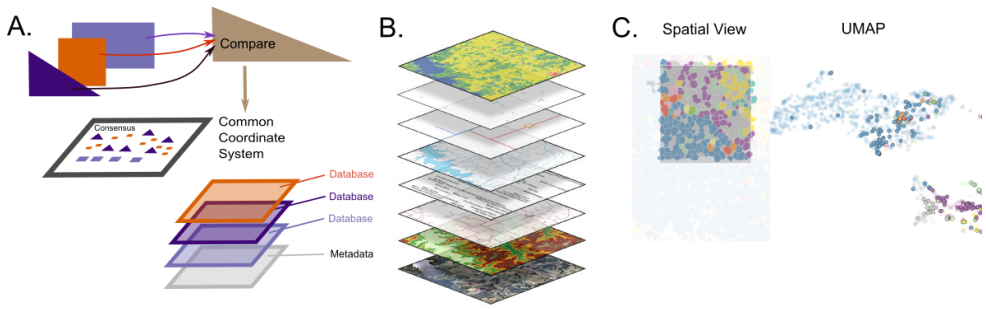


Figure 5: **A** Schematic diagram of stages of interpretaion and integration of data sources. **B** Standards in Geographic Information Systems enable the integration of multiple layers of data. **C** Brushing an UMAP with a covariate can illustrate the dynamics of cell changes (**Kris: I have this feeling 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 [28] 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 [29] 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, [30] 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 [31]. 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 5B).

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 5A 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 [32], 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 [33] (Figure 5C).

Visualization tools for interpretation and communication to biologists

An example of effective visual interpretation tools is interactive brushing of UMAP plot, see Figure 5C 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: [26]

Factor Analysis: [27]

Statis, conjoint analysis: [28]

The French way: [29]

Overview and connections of methods: KS [30]

Kevin Murphy: Probabilistic Machine Learning, MIT Press [**???**] [34] [34]

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: [32]

UMAP: [35] https://arxiv.org/abs/1802.03426v2

Spatial tumor and immune cells: [31]

CD56 Immune cell coloring, paper with C. Blish: [33]

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

Definition of benchmarking

Visualizations and biological assessment of marker gene lists resulting from multi-omics analyses provide critical interpretation of integrative analysis of high-throughput data. Additional quantitative metrics are essential to establish to delineate biologically-relevant features from features resulting from either comptuational and techical artifacts. Beyond interpretation, such quantitative benchmarks are also essential to enable unbiased comparison between the numerous analysis methods used for analysis. For example, the goal of multi-platform single cell data analysis is often the recovery of known cell types through computational methods. In this case, The adjusted Rand Index (ARI) or other metrics for partitions can directly compare known cell types with the results from processing of raw data, quantification, and clustering. In cases where cell types or biological features are unknown a priori, one may also attempt to benchmark methods for their ability to discover known relationships between data modalities, e.g. gene regulatory relationships observed between chromatin accessibility and gene expression. However, this is made difficult by the fact that these

relationships are not fully known at the single cell level which provides the very motivation for the data generation and analysis. Cross-validation within a study and cross-study validation can be used to assess whether solutions found by multi-modal methods generalize to held-out observations or held-out studies.

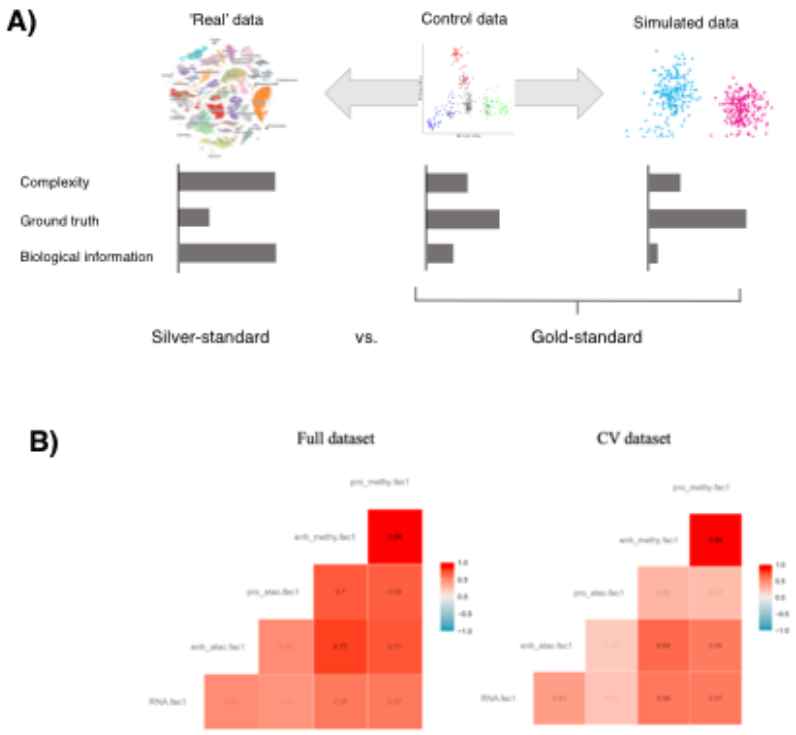


Figure 6: **A** Silver standard: Svensson et al. (2017) Nat Methods; Wang et al. (2019) bioRxiv; Cole et al. (2019) Cell Systems; Zhang et al. (2017) bioRxiv; Soneson et al. (2018) Nat Methods; Saelens et al. (2019) Nat Biotechnol; Gold standard control data: Tian et al. (2019) Nat Meth; Freytag et al. (2018) F1000Res; Gold standard simulated data with the splatter R package appia et al. (2017) Genome Biol [credit: Matt Ritchie] (**Refs to be added proper**). **B** scNMT-seq study: correlations with linear projections (MOFA+) evaluated with cross-validation [credit: Wancen Mu and Michael Love].

Strategies for benchmarking

Benchmarking multi-modal methods on typical multi-modal datasets is inherently difficult, as we rarely know the ground truth [36]. Simulation methods that generate high-throughput data in silico are useful for having known truth, but it is difficult to simulate realistic covariance structure across features and across data modalities (Figure 6A). Moreover, these simulations can inadvertently embed the same underlying assumptions as the computational methods employed for analysis, introducing further biases into benchmark studies. Therefore, high-throughput datasets with a known ground truth are also critical for multi-omics studies and robust testgrounds for future hackathon studies that were widely discussed throughout the workshop.

Cross-validation within study

One strategy for quantitative assessment of multi-model methods is cross-validation (CV). While CV cannot determine the accuracy of a method, it can be used to assess a method's self-consistency. For example a cross-validation analysis of the scNMT-seq dataset was performed as part of the hackathon for this meeting using MOFA+ (Figure 6B). Such 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.

Evaluation of methods using permutation or cross-validation has been performed previously, typically to optimize a tuning parameter or other aspects of model selection. Permutation has been used to create null datasets, either as demonstration that a method is not overfitting, or for tuning parameter selection, where the optimal parameter setting should produce an objective that is far from the null distribution [37,38,39]. Cross-validation using folds or leave-one-out has likewise been used in many multi-modal method development papers [17,40,41,42,43,44,45,46,47,48,49,50].

A challenge with within study cross-validation is how to match dimensions of latent space across folds. Previous evaluations of multi-modal methods using cross-validation have focused on the top factor, swapping the sign of the project as needed to align the top latent factor across folds [51].

Finally, we note that for assessing clustering, a number of papers have suggested to resampling or data-splitting strategies to determine prediction strength [10,52,53,54]. For clustering of cells into putative cell types or cell states, such previously developed techniques could be applied in a multi-modal setting.

Cross-study validation would assess if relationships discovered in one dataset present in other datasets, potentially looking across single cell and bulk.

Creating benchmarking datasets

(**add examples from Google Doc**)

Benchmark datasets for single cell studies have largely centered around measuring sequencing depth and diversity of cell types derived from a single assay of interest (e.g. scRNAseq). A benchmark dataset serves a few purposes:

- Provides ground truth for the intended effect of exposure in a proposed study design.
- Provides validation for a data integration task for which a new computational method may be proposed.

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 to a certain exposure.

Both the intended effects and data integration task rely on study design that takes into account:

- Biological and technical variability via replicates, block design, and randomization.

- Power analysis for the intended effect or data integration task.

- Dependencies between modalities, for e.g. gene expression depending on gene regulatory element activity, 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, and benchmark datasets may be established for commonly used combinations of modalities or technologies, towards specific data integration tasks.

Software strategies to enable analyses of multimodal single cell experiments

Benchmark datasets, analysis pipelines, and development of multimodal genome-scale experiments are all enabled through community-developed, open source software and data sharing platforms. Our hackathons were designed to leverage and build on these open frameworks to enable the use and evaluation of robust benchmark strategies, and to support the enhancement and extension of the components themselves. Specifically, we provided easy to use data packages to distribute the multi-omics data, and we developed vignettes that reproducibly demonstrate each of the analysis approaches employed for these datasets (https://github.com/BIRSBiointegration/Hackathon).

A wide array of open-source genomics frameworks for multi-platform single cell data are also developing in R, Python, and in other software frameworks that enable graphical user interfaces for point-and-click exploration. We take it for granted that **openness** is a *sine qua non* for computational tooling in bioinformatics. This widespread availability combined with standardized licensing in Creative Commons, Artistic, or GNU frameworks enables all components need to be accessible for full vetting by the community, thereby facilitating widescale adoption, extension, and collaboration. These open source software efforts also 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 like GitHub. Our hackathons directly leveraged such code sharing through GitHub, with contributed analyses reports [https://github.com/BIRSBiointegration/Hackathon/blob/master/analysis-vignettes.md] compiled into reproducible workflows [using GitHub Actions] enabling continuous integration of changes to source codes and continuous delivery of vignettes as well as containerized snapshots of the analyses environments. For challenge studies conducted in R, the source code and vignettes were assembled into R packages which took the loaded libraries as dependencies. For those carried out in Python, the requierments were included in the source files for automatic installation and deployment[using https://github.com/fastai/fastpages].

While GitHub enables code distribution and accessibility, it does not address issues of useability or adoption. The open source and community needs for multi-omics go beyond code sharing and ultimately addressed through robust software ecosystems [55,56,57]. As an example of an ecosystem of broad scope, we cite bioconductor.org [58]. Bioconductor supports developers who seek to build broad user bases by providing multiplatform/multistream continuous integration/continuous delivery of contributed packages [59], and users with different skill sets by articulating standards for documentation, and testing, and by hosting community forums and workshops [60]. 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. The further community engagement and communication from the workshop were critical to identify additional key software goals needed to advance the methods and interpretation of multi-omics. In this section, we reflect on the software infrastructure that enabled our analytic frameworks for inference and visualization of the multimodal single-cell experiments. Notably, the hackathons highlighted the following central challenges that require further community-level software development.

- How should multimodal single cell data be managed for interactive and batch analyses?
- What methods will help software developers create scalable solutions for multimodal single cell analysis?
- How can we implement robust visualization methods for multimodal single cell analysis and ensure their accessiblity to researchers with visual impairments?

First, reproducibility and transparency are crucial aspect in hackathons, starting with data availability. Providing data to the community is a long standing issue. A particular challenge of data distribution for multimodal single-cell experiments is that each mode of data is characterized by a different collection of features on possibly non-overlapping collections of samples from similar biological conditions. Common data structures to store and operate on these data collections can support data dissemination with robust metadata and implementation of the common analysis frameworks for multi-omics data interpretation that can be employed across diverse measurement technologies. As an example, Bioconductor has built a ready-to-use integrative data class `multiAssayExperiment` on the S4 object oriented structure in R [61]. The `multiAssayExperiment` class includes the following features to support multi-platform genomics data analysis 1) Storage for variables or features from multiple data modalities (e.g. gene expression units from scRNA-seq and protein units in sc-proteomics), either from the same cells or distinct cells from the same or distinct starting samples or biological specimen of origin. In some cases, the feature may be multidimensional (e.g. spatial coordinates, locations of eQTLs).
2) Metadata for sample of origin for the individual cells, e.g. study, center, phenotype, perturbation, in all the variables or features stored. 3) A map between the datasets from the different assays to enable analysis.

The `multiAssayExperiment` class is designed so that the variables or features stored for each assay in `multiAssayExperiment` each independently use Bioconductor data structures for the stored data modality and bind metadata annotating the molecular features for each assay into this class instance. For example, genes and transcripts can be enumerated using gene labels from Ensembl [62] as catalog identifiers, represented as genomic regions through `GRanges` instances [63], etc.

In our context, input data across all hackathons were pre-processed with steps documented, and the data were stored in `MultiAssayExperiment` [64] objects which allows collating of standard data formats as well as powerful data access and data processing methods. The datasets were centrally managed and hosted on `ExperimentHub` [65] and the `SingleCellMultiModal` package [doi:10.18129/B9.bioc.SingleCellMultiModal] was used to query the relevant datasets for each analysis. Figure 7 shows how this class was used to amalgamate and annotate the multimodal dataset consisting of seqFISH and scRNA-seq experimental data employed for this first hackathon. To combine these two different experiments, the seqFISH data were stored into the SpatialExperiment S4 class object, while the scRNA-seq data were stored into a SingleCellExperiment Bioconductor class object [66]. Then, these objects were easily stored into a MultiAssayExperiment class object and released with the SingleCellMultiModal Bioconductor package [**???**]. <!–Is this an unique storage feature of Bioconductor or is there something comparable in python or other bioinformatics ecosystems that we should discuss as well to be comprehensive?–!> We note that this infrastructure was readily suitable and employed for the scNMT-seq and the seqFISH+scRNA-seq (Fig. 7) datasets. However, the lack of overlap between samples in the scProteomics hackathon revealed an important area of future work to develop software infrastructure that can link biologically related datasets without direct feature or sample mappings for multi-omics analysis.

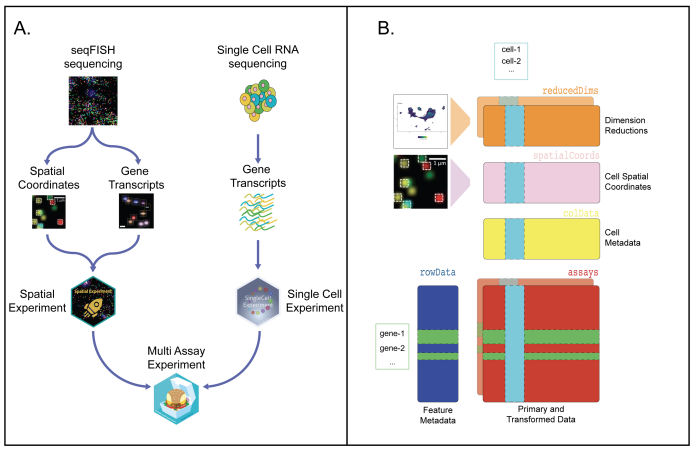


Figure 7: Panel A: Combination of seqFISH-based SpatialExperiment and SingleCellExperiment instances into a MultiAssayExperiment. Panel B: details of the SpatialExperiment class design.

In addition to data storage, several hackathon contributors (Al Abadi, Patheepa Jeganathan) used the `multiAssayExperiment` class to implement software for the analysis methods they employed to integrate the multi-modal single cell data (scNMT-seq, sc Proteomics, **pointers to vignettes**), allowing for easier preprocessing, transformation, extraction of spatial information from raster objects, addition of cell information and visualization **Al, details please if needed**.

These hackathons highlighted the further challenges of summarizing assay-specific genomic features for multi-platform data integration. While RNA-seq has well-defined units and IDs (e.g., transcript names), other assays may be summarized at different genomic scales, e.g., gene promoters, exons, introns, or gene bodies. The `GenomicRanges` package [67] may be used to compute summaries at different scales and overlap 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 [68] may be used to deconvolute multi-cellular spots signal. Finally, in the absence of universal standards, the metadata available may vary from analysis to analysis. A potential solution is to define the minimum set of metadata variables necessary for each assay, and for pairs of assays to be comparable for common analyses. <!–Is this problem solved through the multiAssayExperiment and/or broader Bioconductor ecosystem? If so how and if not what are areas of future work needed as a field.–!> In addition to standardizing data processing, the substantial size of single cell data also require new strategies that are emerging to allow for scalability of *storing* and *access* of large datasets [69,70] and for emerging scalable *algorithms* that allow for data to be stored in memory or on disk such as unsupervised clustering of cell types in either R [71,72] or Python [73].

In addition to robust comptuational metrics, several hackathons developed novel data visualization strategies as essential components for interpreting multi-modal data. Often these visualization strategies rely on heatmaps or reduced dimension plots, and utilize color to represent the different dimensions. For many, these colors and low dimensional plots facilate pattern detection and interpretation of these increasingly complex and rich high-throughput, multi-platform single cell data in standard. However, relying on color for interpretation leads to difficulties in perceiving patterns (the basis for the Ishihara's color vision tests) in multi-colored figures for a substantial proportion of the population with color vision deficiencies. Moreover, relying on perceived patterns through color can result in different data interpretations between individuals, particularly reflecting their individual color detection capacities. One strategy to present standardized scientific information accessible information to all readers is to include colorblind friendly visualizations [74,75] as a default setting, using palettes such as form the R packages viridis [76] and dittoSeq [77], while limiting to a number of 10 colors. Additional visual cues to differentiate regions (hatched areas) or cells (point shapes) can also reduce the dependence on colors. The inclusion an "accessibility caption" accompanying figures

which “guide” the reader’s perception of the images would 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.

Details of working components – trimmed

you can interact with underlying data at google sheet

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
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

Type	Brief name (link)	Description
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 analyses and reference datasets from the Mathematical Frameworks for Integrative Analysis of Emerging Biological Data Workshop demonstrated the power of hackathons to have the potential both to 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 designed specifically to span state-of-the-art multi-omics challenges to map the epigenetic, molecular, and cellular interaction across both time and sample populations. While on the one hand coupling single cell measurements can inherently simplfy the challenge of linking disparate biological scales, layering new sets of molecular measurements intrinsically increases complexity of the corresponding analyses to interpret these data. The computational needs hinge on the underlying biological question being asked as well as the data themselves. For example, different considerations must be made for cross-modal integration higlighted in the seqFISH and scNMT-seq challenges than modeling with unmatched measurements in the scProteomics challenge. Regardless, through these hackathons we identified several common analysis themes spanning algorithmic advances, interpretation, benchmarking, and software infrastructure necessary for biological interpretation. Notably, all cases require methods for dealing with data quality, data loss by summarization, timing variances both between and within omics layers, and batch effects will be necessary challenges to overcome. Conversely, the ability to infer regulatory relationships between different omics and interpret these relationship with data visualization strategies present new research opportunities.

Technologies to profile biological systems at a single cell resolution and across molecular scales are advancing at an unprecidented pace. Analytically, these advances require the computational analysis commuity to pursue research that can first enable robust, data set specific analyses that are tailored to a specific biology or measurement technology and secondly scale and adapt to these rapid advances. For example, the hackathons highlighted current technologies for spatial molecular profiling, which currently require profiling samples with multiple technologies to balance high resolution spatial and molecular profiling across data modalities. While the data are not spatially resolve, the scNMT-seq challenge highlights the potential of further inference of gene regulation through concurrent profiling of RNA, methylation, and chromatin state. While posed as disparate analysis challenges, integration of matched datasets between these spatial and epigenetic profiling techniques could further resolve the dependence of cell-type and cellular-interactions on regulatory networks. Likewise, the workshop keynote by Bernd Bodenmiller higlighted new extensions to these technologies that further enable three dimensional spatial molecular profiling [78] and still other technologies are exanding to allow for temporally resolved profiling [79]. Integration strategies that are aware of these future directions and the mathematical analysis challenges that span technologies will be most adept to advance biological interpretation from these emerging datasets. Implementing these analysis tools requires further robust software ecosystems, including as examples Bioconductor [80], Biopython, Scanpy [73], Seurat [81], and Giotto [82], in which users create these analysis approaches and anticipate stable yet adaptive data structures that are robust for these emerging technologies. The size of these emerging datasets, particularly in the context of their application to atlas projects including notably the Human Tumor Atlas Network [83], Human Cell Atlas [84], Allen Brain Initiative, and ENCODE <!–are there others or citations I should be using for this?–!> as key examples, comptuational efficiency and scalability of these implementations are becoming ever more critical.

Emerging analytical methods and technologies

Community needs for data structures, analysis methods, etc

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.	?

Consensus Term	Related Terms	Description	Citation
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	?	51
contributions	variable weights, loadings, eigenvector, axis, direction, dimension, coefficients, slopes	Contributions of the original variables in constructing the components.	17, 86
latent factors	variates, scores, projections, components, latent/hidden/unobserved variables/factors	Weighted linear combinations of the original variables.	17, 86
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.	17, 30, 87
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.	17, 86, 88
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.	89

Consensus Term	Related Terms	Description	Citation
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.	90, 91
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. **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

11. **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

12. **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

13. **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

14. **scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells**
Stephen J. Clark, Ricard Argelaguet, Chantriolnt-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

15. **Multi-omics profiling of mouse gastrulation at single-cell resolution**
Ricard Argelaguet, Stephen J. Clark, Hisham Mohammed, L. Carine Stapel, Christel Krueger, Chantriolnt-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

16. **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

17. **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

18. **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

19. **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

20. **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/https://doi.org/10.1186/s13059-020-02015-1
DOI: https://doi.org/10.1186/s13059-020-02015-1

21. **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

22. **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

23. **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

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

25. **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

26. **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

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

28. **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

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

30. **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

31. **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

32. **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

33. **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

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

35. **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

36. **Sparse canonical methods for biological data integration: application to a cross-platform study**
Kim-Anh Lê Cao, Pascal GP Martin, Christèle Robert-Granié, Philippe Besse
BMC Bioinformatics (2009-01-26) https://doi.org/frn49f
DOI: 10.1186/1471-2105-10-34 · PMID: 19171069 · PMCID: PMC2640358

37.**:{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

38. **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

39. **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

40. **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

41. **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

42. **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

43. 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

44. **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

45. **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

46. **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

47. **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

48. **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

49. **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

50. **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

51. **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

52. **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

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

54. **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

55. **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

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

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

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

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

60. https://bioconductor.org/support

61. **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

62. **Ensembl 2020**
Andrew D Yates, Premanand Achuthan, Wasiu Akanni, James Allen, Jamie Allen, Jorge Alvarez-Jarreta, M Ridwan Amode, Irina M Armean, Andrey G Azov, Ruth Bennett, … Paul Flicek
Nucleic Acids Research (2019-11-06) https://doi.org/ggqp72
DOI: 10.1093/nar/gkz966 · PMID: 31691826 · PMCID: PMC7145704

63. **Software for Computing and Annotating Genomic Ranges**
Michael Lawrence, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, Vincent J. Carey
PLoS Computational Biology (2013-08-08) https://doi.org/f5cmfg
DOI: 10.1371/journal.pcbi.1003118 · PMID: 23950696 · PMCID: PMC3738458

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

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

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

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

68. **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

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

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

71. **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

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

73. **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

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

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

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

77. **dtm2451/dittoSeq**
Daniel Bunis
(2020-08-23) https://github.com/dtm2451/dittoSeq

78. **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

79. **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

80. **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

81. **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

82. **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

83. **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

84. **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

85. **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

86. **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

87. **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

88. **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

89. **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

90. **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

91. **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.↩