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

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

- **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

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

Draft outline

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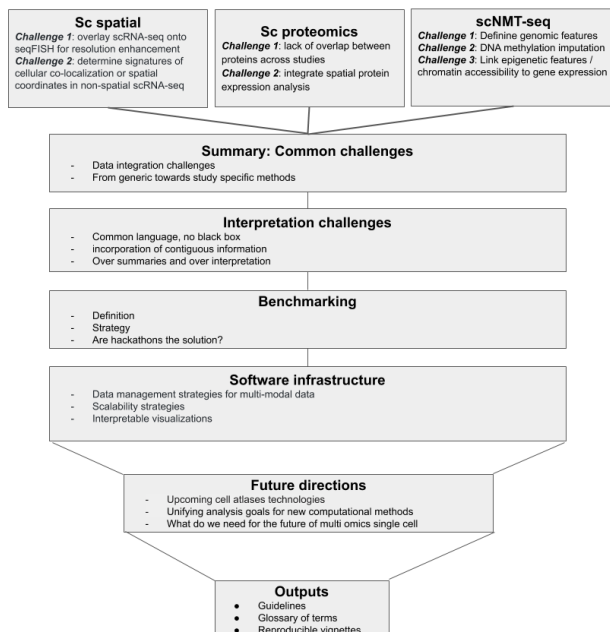

Figure caption: Main challenges discussed during our brainstorming sessions from the hackathons

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

Introduction

Comprehensive characterization of biological systems with multi-omics

- Single cell community has advanced technologies to enable concurrent processing of biological systems at multiple molecular resolutions
- The lack of prior knowledge and gold standard benchmark naturally leads to a data-driven approach

New single cell multi omics initiatives:

- Human Cell Atlas (HCA): assess variation in normal tissues
- Brain initiative and Allen Brain
- Human Tumor Atlas Network (HTAN): Single-cell, longitudinal, and clinical outcomes atlases of cancer transitions for diverse tumor types.
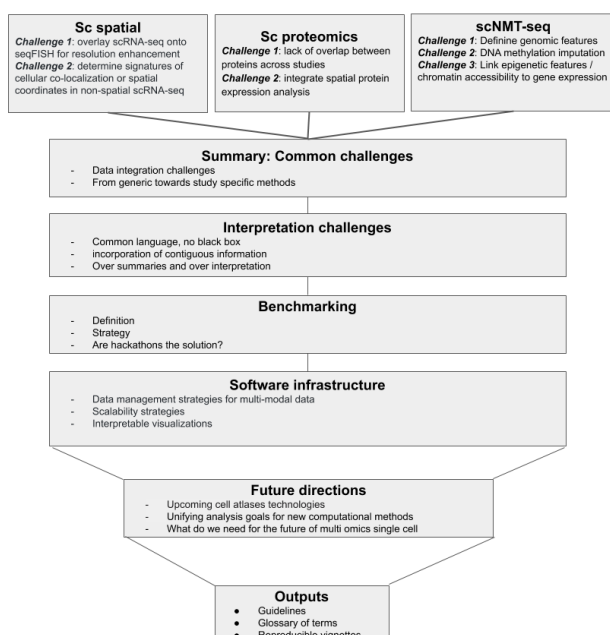
What bulk multi-omics (e.g. TCGA, ENCODE) have taught us:

- Type of omics that can answer a specific biological question
- The value of open resources for methodological developments
- New hypotheses

Using hackathons to illustrate analysis standards and challenges for capturing biological information from multi-omics technologies

- Brief overview of our three hackathon studies highlighting state of the art challenges (e.g., spatial transcriptomics, cross-study analysis, epigenetic regulation)

- Challenges include issues with noise and experimental design, Time lag between regulatory levels not addressed and many open questions remain (e.g methylation / gene expression), Direction of regulation not captured

- We present our findings from hackathon case studies that helped us obtain benchmarks and define a common language for multi-omics

- **Objectives of this paper**

  - Provide guidelines on tools / data / technologies / methods and needs to model the multi-scale regulatory processes in biological systems for a computational biologist audience

- **Outline and messages**

  - Cellular and molecular regulation is fundamentally multi-scale and captured by distinct data modalities
  - Traditional hypothesis-driven multi-omics/view studies only consider one facet of these technologies, but more can be learned through a holistic approach extending into atlases
  - We present our findings from hackathon case studies that helped us obtain a broader picture and language

  Outline of the paper:

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

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

Overview and biological question

Single cells are considered the smallest units and building blocks of each tissue, but 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 addition, most tissues are composed of multiple cell types whose identity and function can be inferred through their unique transcriptomic profile. In the last decade single-cell RNA-seq (scRNA-seq) played a key role to capture single cell gene expression profiles, which allowed us to map virtually 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.

More recently technologies have begun to emerge that can profile the transcriptome of single cells within their original environment. These technological advancements offer the possibility to examine how gene expression is influenced by cell-to-cell interactions or organized in a spatially coherent manner. One such approach is sequential single-molecule fluorescence in situ hybridization (seqFISH), which can identify single molecules at (sub)cellular resolution with high sensitivity. Nevertheless, and in contrast with scRNAseq, seqFISH and many other spatial transcriptomic technologies often pose significant technological challenges and hence the number of profiled genes per cell is usually restricted to a smaller number (10-100s).

To overcome the lack of spatial information with scRNAseq and the common limited coverage in spatial datasets, we sought to combine and integrate a matching scRNAseq and seqFISH dataset, since both were generated from the mouse visual cortex region. More specifically, in this hackathon 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. And in the opposite direction, we sought out how to transfer spatial information obtained from the seqfish dataset to that of the scRNAseq dataset.

Computational challenges

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

[suggestion 1]
Sequencing and imaging based single-cell transcriptomic profiling have complementary strengths. Whereas single-cell RNAseq generates transcriptome-wide information, it does not have spatial information. On the other hand, seqFISH (Lubeck 2014; Shah 2016) provides single-cell resolution spatial information, but typically profiles the expression level of only 100-300 genes. Although the newer generation of seqFISH technology (called seqFISH+) has greatly enhanced its capacity which can now be used to profile 10,000 genes (Eng 2019), the technology is significantly more complex and costly. As such, it is desirable to develop computational approaches to effectively integrate scRNAseq and seqFISH data analyses.

In this hackathon, the participants were provided with seqFISH (Zhu 2018) and scRNAseq (Tasic 2016) data corresponding to the mouse visual cortex and challenged to accurately identify cell-types by integrating both datasets. Cell type labels, derived from scRNAseq analysis (Tasic 2016) and previous seqFISH/scRNAseq integration (Zhu 2018) were also provided as reference. A variety of computational approaches were applied to achieve this goal, including: supervised classification with support vector machines (Coulumb, Xu), semi-supervised self-training (Singh), and unsupervised matrix factorization methods (Sodicoff). While the methodologies are different, a number of themes recur, such as the importance of gene selection and batch effect correction. As expected, the ability to identify refined cell-type structure relies on the selection of cell-type specific marker genes in seqFISH data, suggesting a potential benefit of using single-cell RNAseq data to guide seqFISH experimental design. Batch effect is another important factor affecting the accuracy of data integration. While a number of batch effect correction methods have been developed (COMBAT, Seurat, Scanorama, etc), it remains challenging to distinguish technical from biological variations if the biological samples do not match exactly.

[suggestion 2]
The mouse visual cortex consists of multiple complex cell types, however the number of profiled genes in the seqFISH dataset is limited to 125 genes. Moreover, these genes were not prioritized based on their ability to discriminate between cell types and thus assigning the correct cell identity is challenging. In contrast the scRNAseq dataset is transcriptome wide, thus including the 125 aforementioned genes. As such we first used all genes to identify the cell type labels for each cell in the scRNAseq data with high certainty. Next, we leveraged that information to build a classifier based on (a subset of) the 125 common genes only. This classifier could subsequently be applied to the seqFISH dataset in order to assign cell types to each cell with high probability.

During the hackathon we tested various machine learning or data integration models, but also noted that initial dataset normalization strategies might have a significant impact on the final results (see common challenges [**???**]). Although unique molecular identifier (UMI) based scRNAseq and seqFISH can both be considered as count data, we observed dataset specific biases that could be attributed to either platform (imaging vs sequencing) or sample specific sources of variation. Here we opted to apply a quantile normalization approach which forces a similar expression distribution for each shared gene. The hackathon participants selected two machine learning classifiers, a supervised support vector machine (SVM) and a semi-supervised lasso and elastic-net regularized generalized linear model (glmnet). To further improve the SVM model the participants tried multiple kernels and searched for optimal hyperparameters using a combined randomized and zoomed search. In addition, different flavors of recursive feature elimination was used to find the optimal or minimum number of genes needed to correctly classify the majority of the cells. Importantly, participants tested different classification accuracy metrics to alleviate the major class imbalance in the dataset, since more than 90% of cells were excitatory or inhibitory neurons. The glmnet approach used a iterative model building approach, which combines both datasets and initially only retains the highest confidence labels and then gradually adds more cell type labels until all cells have been classified. This type of self training approach might in fact be more generalizable to other datasets. Finally, a data integration approach based on integrative non-negative matrix factorization (NMF) was used by applying the previously published LIGER approach. By integrating both datasets in a similar subspace based on shared factors, cell type labels can be transferred using a nearest neighbor approach.

Challenge 2: determine signatures of cellular co-localization or spatial coordinates in non-spatial scRNA-seq

[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 (Coulumb). 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.

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. (2019) 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 (Keren et al. 2018). (Figure 3a)

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 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.

[Fig1a] (Create pictogram of sc_targarted proteins hackathon challenge)

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 3b). 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.

Figure 3B
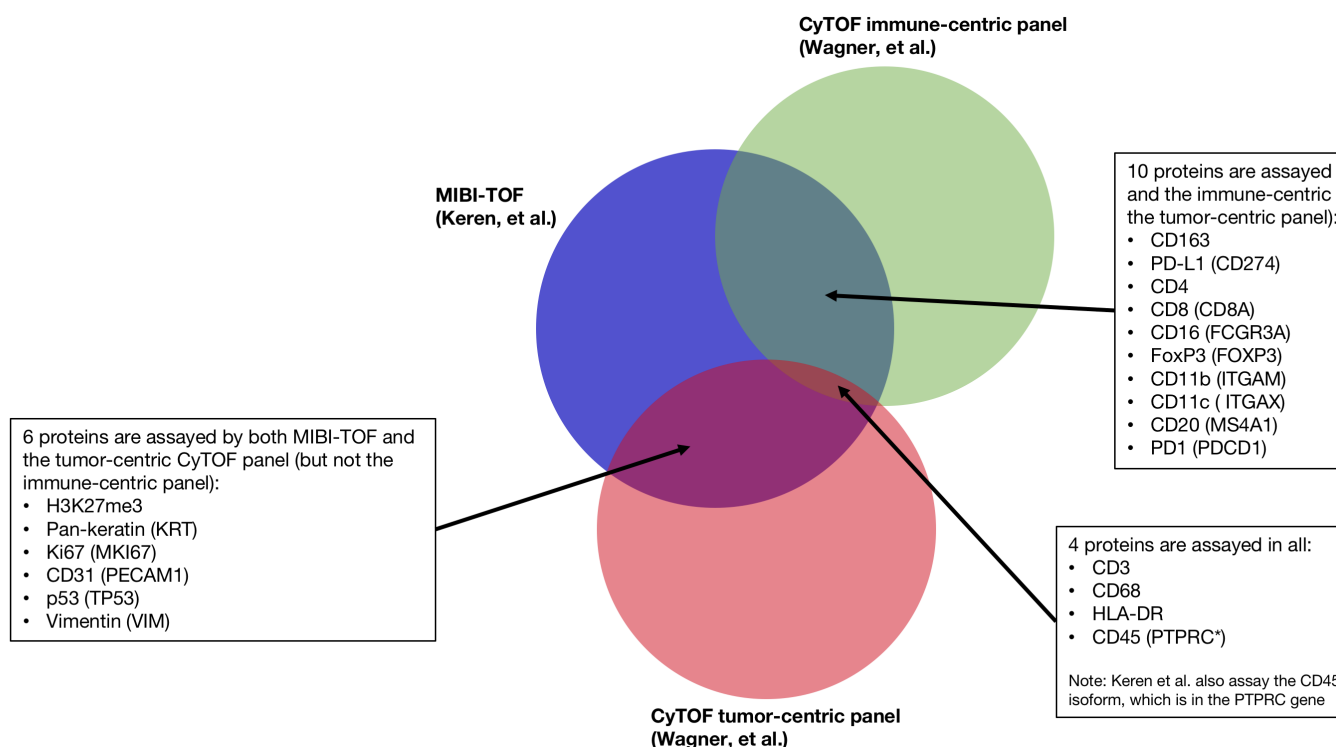
Although the overlap in proteins was low, many proteins were cell-type markers, providing the opportunity to perform cross-study integration of cell type proportions in tumor tissue samples. Hackathon participants applied several semi-supervised and supervised algorithms to transfer cell labels and cell 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 (2012 https://doi.org/10.1198/106186005X59243) 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 implemented using RStan and topic modelling [Needs input from Pratheepa]

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.

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)

scNMT-seq as a case-study for epigenetic regulation

Overview and biological question

The maturation of single-cell sequencing technologies has enabled the identification of transcriptional profiles associated with lineage diversification and cell fate commitment[1]. Yet, the accompanying epigenetic changes and the role of epigenetic layers in driving cell fate decisions still remains poorly understood[2].

scNMT-seq is one of the first experimental protocols that enable simultaneous quantification of RNA expression and epigenetic information from individual cells[3]. 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 readouts 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 readouts from the same cell.

Data set description

Gastrulation marks a major lineage specification event in mammalian embryos, accompanied by profound transcriptional rewiring and epigenetic remodelling [4]. scRNA-seq studies have identified major transcriptional changes associated with cell fate commitment, but the accompanying epigenetic reprogramming and the coordination between different epigenetic layers remains poorly understood.

In this hackaton we used a data set where scNMT-seq was applied to mouse gastrulation [5]. A total of XXX cells across four developmental stages (E4.5 to E7.5) passed quality control for all three data modalities. 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 (D=XXX), CpG islands (D=XXX), p300 binding sites (D=XXX), CTCF binding sites (D=XXX) and DHS (D=XXX) open sites.

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 evaluate whether the same stages can also be discerned on the basis of DNA methylation.

Computational integration

We considered three computational strategies:

- **PLS (as implemented in DIABLO[6]):** projection to least squares model. A sparse generalisation of canonical correlation analysis that maximises covariation between prespecified pairs of data modalities.
- **LIGER** [7]: unsupervised non-negative matrix factorisation 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** [XXX]: a multi-omics supervised clustering algorithm inspired from `survClust`.

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 readouts, and (3) the challenging interpretability of individual dinucleotides. To address these problems, DNA methylation measurements are typically aggregated over predefined 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 succesful to distinguish heterogeneous cell types, but it does not improve interpretabiliy and it leads to a massively large feature set. The alternative strategy is to adopt a supervised approach where 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. Notably we observe (...)

[DESCRIBE RESULTS]

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 readout 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 [Figure 5 + Table 1]

We summarize the common main challenges faces across all hackathons, and the common approaches adopted to analyse multi-omics single cell data.

Summary of hackathon study-specific methods

- **Table** describes method, foundation in the context of previous bulk and single cell literature, and technology dependence
  - Attempts to tweak existing methods and challenges associated in hackathons
  - List methods that are either technology dependent (e.g. spatial) vs universal and how to choose them

Dependence on pre-processing method and/or variable selection

- These steps are key and affect downstream analyses
  - Normalization / data transformation (seqFish), pre-processing, gene summaries (scNMT-seq) to variable selection (seqFish)
  - reproducibility difficult / no consensus. e.g. Alexis selected 19 genes whereas Zhu original paper based on 47 genes (difference in methods / processed data)
- Hackathon data pre-processed enable better comparisons across methods
- No consensus reached as those are emerging data with no ground truth nor established biological results

Approaches for partial overlap of information (cells / features) and how to predict (cell type, dataset) using another data set

- Overlap in each study
  - seqFish: same features but not cells; scProt: same proteins, not cells but similar patients; scNMT-seq: same cells but not features
  - How it was solved (Table)
- Anchoring information across datasets or studies is needed (Figure)
- Incorporation of existing biological knowledge
  - 'From discovery to detection' (Meuleman + debrief), time is ripe to include more knowledge in our data driven approaches
- Challenge: Partial cell overlap (but no features matching) and No overlap were not addressed

Managing differences in scale and size for datasets that do not match cells or features

- Hackathons datasets did not match cells or features.
  - scNMTQ-seq: MOFA limitation when # features vary (and size of datasets).
  - seqFish: greedy approach to select the best gene subset (Alexis, size); consider batch effect removal method (Amrit, scale)
- Consensus on projection based methods, even if pre-processing was applied (Table)
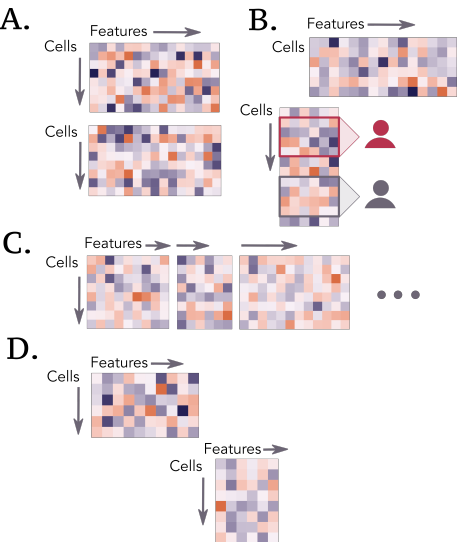- Additional weighting is needed (e.g. Arora, Abadi).

Figure caption: A) scSpatial: required overlap of features (genes), but cells do not overlap. Cell type prediction for seqFISH data was performed based on scRNA-seq (known) [credit: Amrit Singh / Kris S] B) scProteomics: a small number of features overlap (proteins) but patients and cells do not overlap. Data imputation (?), spatial co-localization or cell type prediction was performed [credit: Lauren Hsu and Pratheepan Jenagan / Kris S] C) scNMT-seq: cells are matching across assays but features do not overlap. Data integration was performed [credit: Al Abadi / Kris S] D) Nor cells nor features are matching [credit: Chen Meng / Kris S]

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 {#tbl:example-id}

Common challenges	Tasks	seqFISH case study	Sc targeted proteomics	scNMT-seq
	Normalisation, data transformation, pre-processing	Data distribution checks (Coullomb, Singh) HVG (Xu)	VSN (Meng) Arcsinh transformation (Jeganathan). Inverse transformation (Jenagan) Selection of patients (Jenagan)	Summaries of DNA measurements (data provided in hackathon)

Common challenges	Tasks	seqFISH case study	Sc targeted proteomics	scNMT-seq
Overlap	**Partial feature overlap** (cells not matching)		**Imputation**: Direct inversion with latent variables (Sankaran) Optimal transport to predict protein expression (Lin) KNN averaging (Jenathan*)* **No imputation:** Biological Network Interaction (Foster**)**	
	Partial cell overlap (no features matching)		MBPCA (Meng*)	
	No cell overlap (partial or complete feature overlap)	Averaging nearest neighbors in latent space to impute unmeasured expression values (Coullomb?)	Topic modelling to predict cell spatial co-location or spatial expression (Jenathan, partial feature overlap) Transfer cell type label with Random Forest (Hsu, complete feature overlap?)	LIGER - NMF (Welch, complete feature overlap)
	No overlap		RLQ (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)		Supervised clustering MOSAIC (Arora) (bulk) Lasso in regression-type models (bulk)
	Cell type prediction	Projections and clustering (Sodicoff*) SVM (Coullomb, Xu) ssEnet (Singh) (all bulk)		
Study specific approaches	**Spatial analysis**	HMRF Voronoi tesselation (Coullomb) (bulk)	Spatial autocorrelation (Moran index, Hsu, Lin) Select spatial discriminative features: Moran Index, NN correlation, Cell type interaction composition, L function (Lin) (all bulk?)	
Managing differences in scale	**Data integration**	LIGER - NMF (Sodicoff*) *(sc)* ComBat (Singh) Projection methods MFA, sGCCA (Singh*)* UMAP / tSNE (Sodicoff) (sc)	Multi-block PCA Weighting matrices based on their similarities: Data wise (STATIS, MFA) (Chen*) (bulk) Scale CyTOF and MIBI-TOF to the range of CyTOF values (Jenagan)	LIGER - NMF (Welch) (bulk) Projection to Latent Structures (Abadi) (bulk) MOSAIC as weighted approach (Arora) (bulk)
Inclusion of additional information			Survival prediction: Cox regression based on spatial features (Lin)	Include annotated DHS Index to anchor new/unseen data from DNase-seq, (sc)ATAC-seq, scNMT-seq, etc. for *De novo* peak calling(Meuleman*) (bulk)

Challenges for interpretation

There are many difficulties involved of the understanding and communication of results from complex data sets and analyses as we have seen in [**???**] (sec:scNMTQ-seq), [**???**](sec:scProteomics), [**???**](sec:scSpatial).

We'll separate out the different challenges into different levels.

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

The first is the communication 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.

Supervised versus unsupervised

One simple delineation between methods 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 [8].

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 [9].

A schematic summary of the different stages in interpretation is provided here:
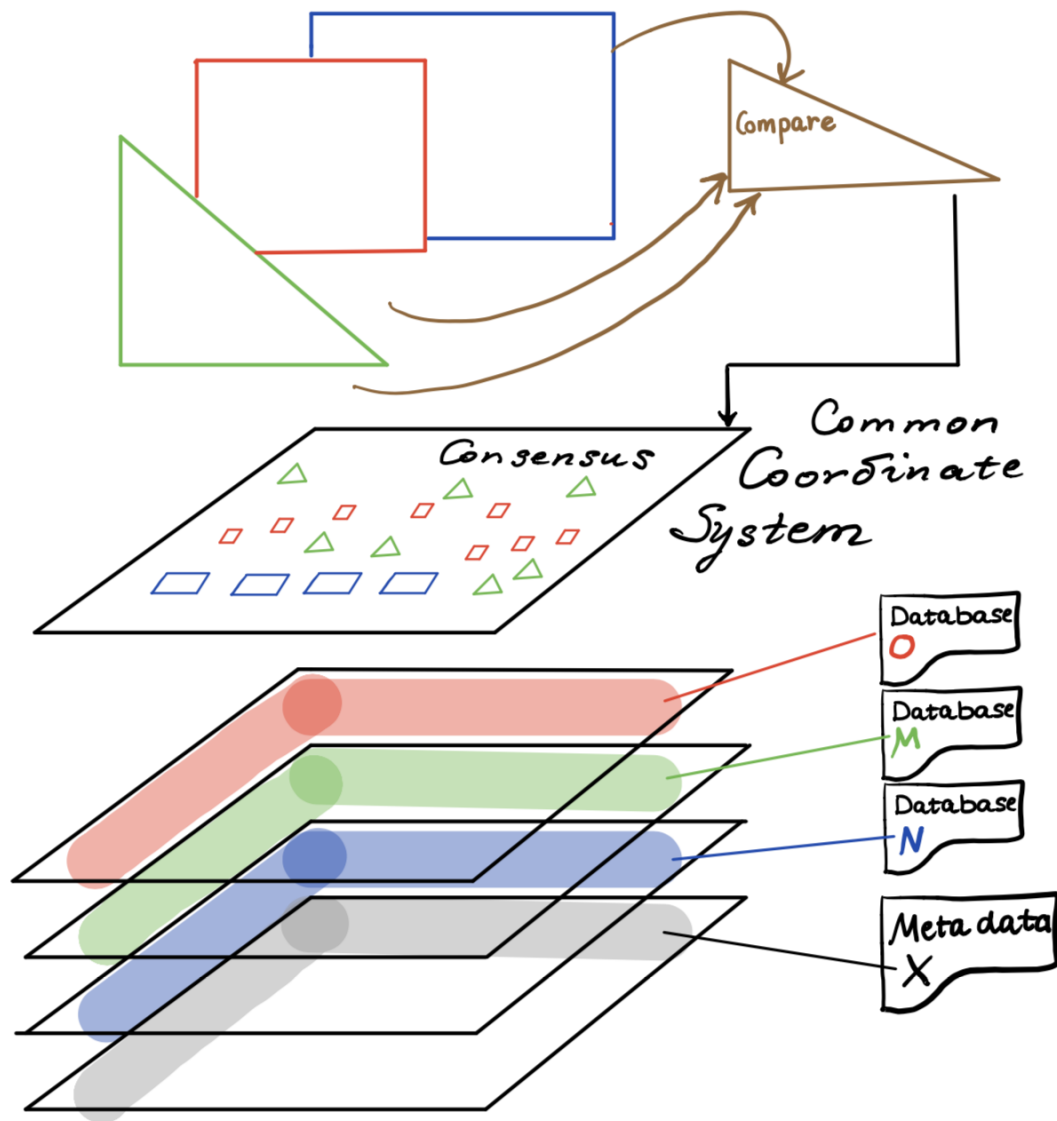
Figure 1: Schematic diagram of stages of interpretaion and integration of data sources

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 [10] 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 [11] 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, [12] 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 [13]. 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.
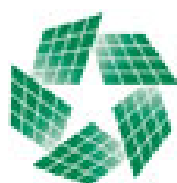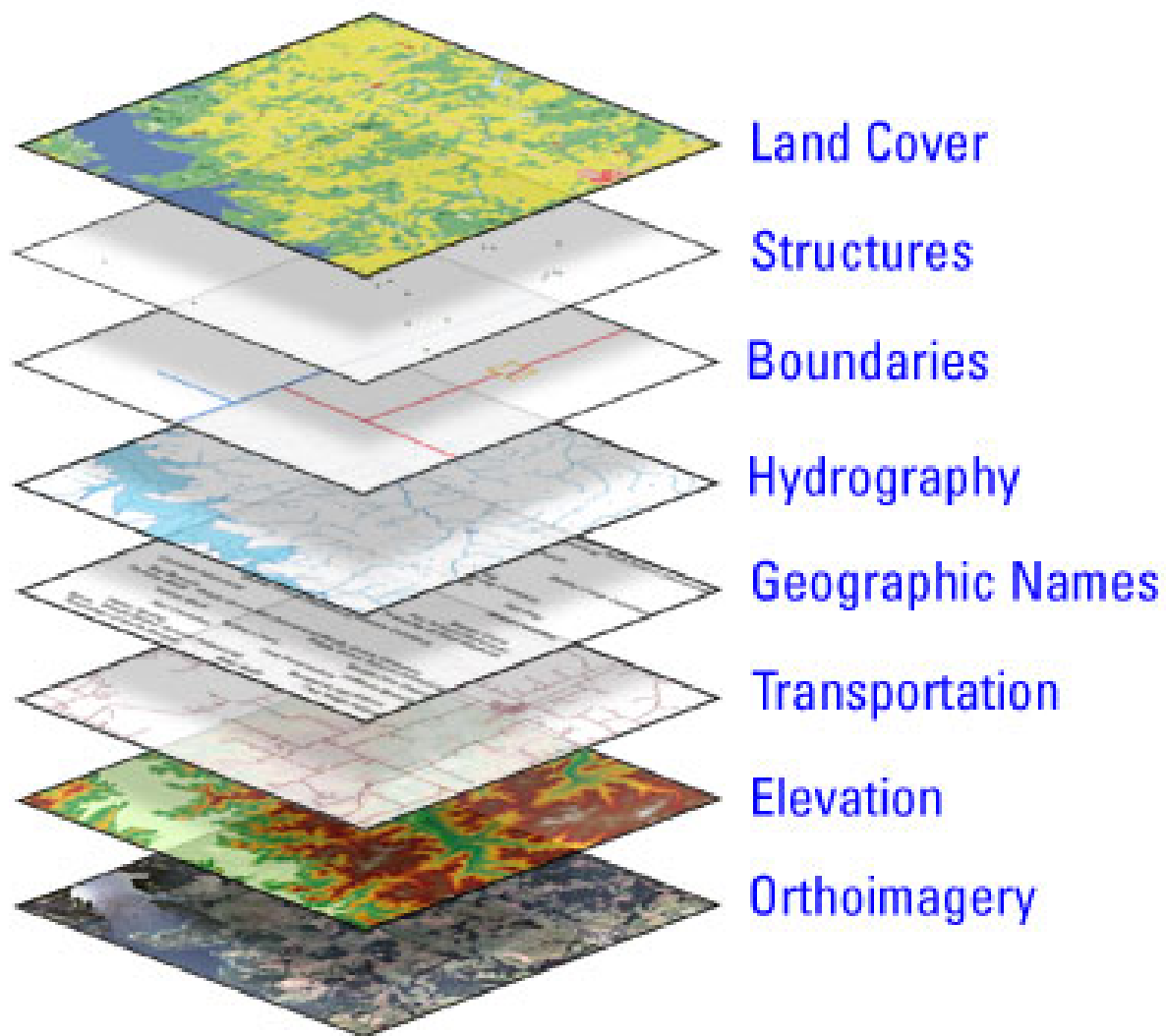
Figure 2: Standards in Geographic Information Systems enable the integration of multiple layers of data

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 the Figure [**???**](fig:InterpWorkflow) 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 [14], 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 [15]:
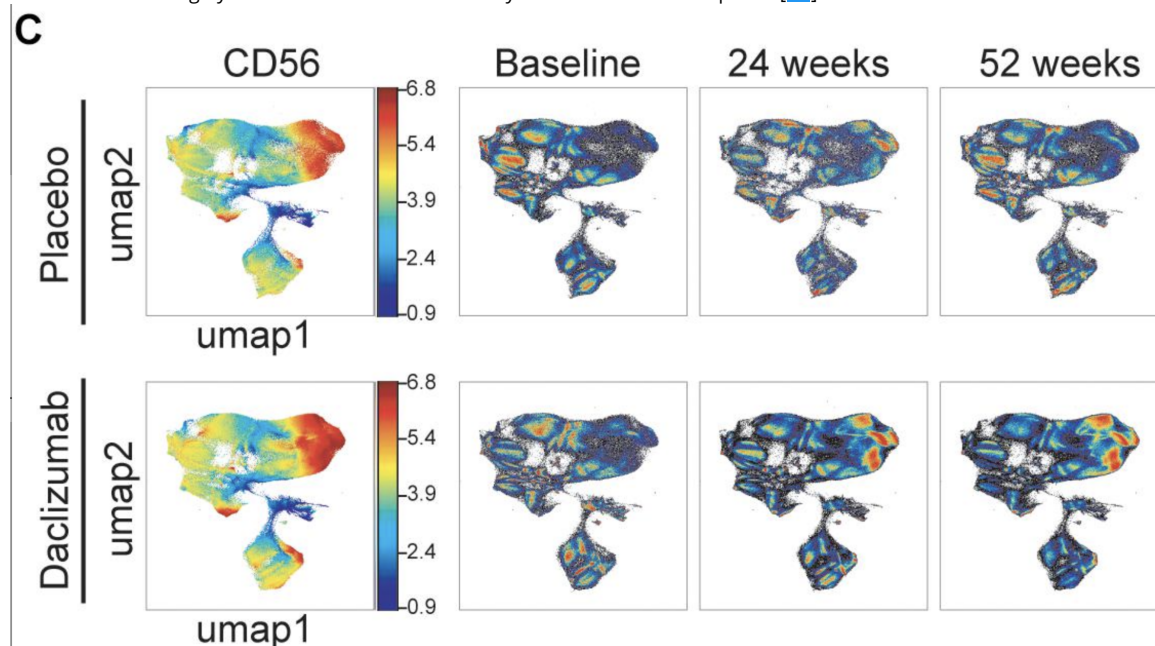
Visualization tools for interpretation and communication to biologists

An example of effective visual interpretation tools is interactive brushing of UMAP plot, see [**???**](fig:BrushUMap) by Kris Sankaran.

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: [8]

Factor Analysis: [9]

Statis, conjoint analysis: [10]

The French way: [11]

Overview and connections of methods: KS [12]

Kevin Murphy: Probabilistic Machine Learning, MIT Press [**???**] [16] [16]

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: [14]

UMAP: [17] https://arxiv.org/abs/1802.03426v2

Spatial tumor and immune cells: [13]

CD56 Immune cell coloring, paper with C. Blish: [15]

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

Often, the goal in benchmarking is recovery of known cell types with processing of raw data, quantification, and clustering. The Adjusted Rand Index (ARI) or other metrics for partitions are used.

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.

Strategies for benchmarking

Benchmarking multi-modal methods on typical multi-modal datasets is inherently difficult, as we rarely know the ground truth [18]. Simulation is useful for having known truth, but it is difficult to simulate realistic covariance structure across features and across data modalities (**Figure**).

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.

Cross-validation within study

For example the cross-validation analysis of the scNMT-seq dataset was performed using MOFA+ (**Figure**). Such an 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 [19,20,21]. Cross-validation using folds or leave-one-out has likewise been used in many multi-modal method development papers [6,22,23,24,25,26,27,28,29,30,31,32].

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 [33].

Finally, we note that for assessing clustering, a number of papers have suggested to resampling or data-splitting strategies to determine prediction strength [34,35,36,37]. 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.

Using hackathon studies as community benchmarking

to fill here, could mention the dream challenge and link back to our own experience / learnings

Figure 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] **B)** scNMT-seq study: correlations with linear projections (MOFA+) evaluated with cross-validation [credit: Mike Love]. **DOI missing**
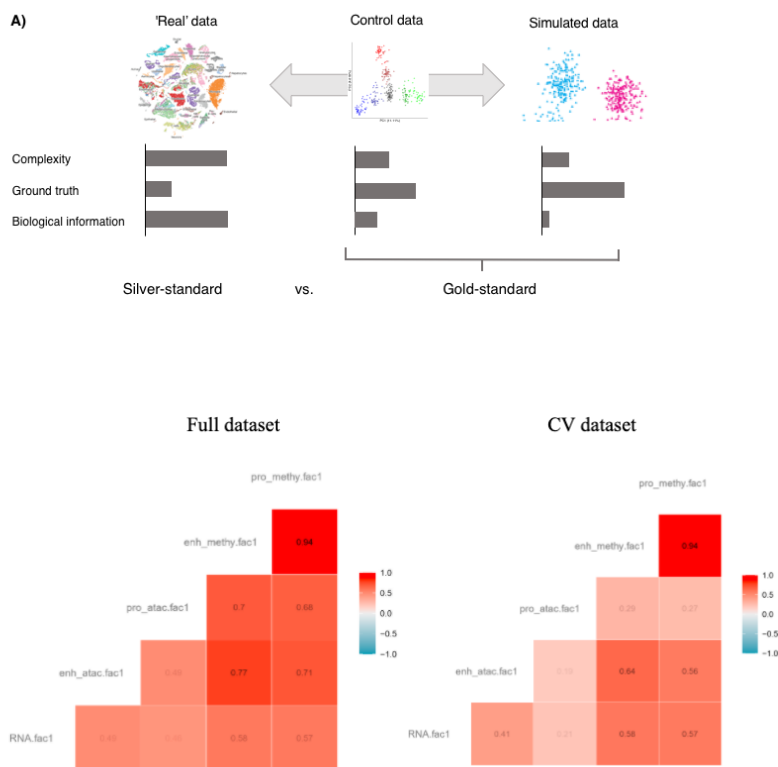
Software strategies to enable analyses of multimodal single cell experiments

In this chapter we review the situation of scientists who create and use analytic software for visualization and inference in multimodal single-cell experiments. Our discussion is necessarily limited in scope, but we provide pointers to concrete details when relevant.

Basic aims

We take it for granted that **openness** is a *sine qua non* for computational tooling in this area. All components need to be accessible for full vetting by the community, so licensing in Creative Commons, Artistic, or GNU frameworks is expected. We also aim for a **coordinated approach**, so that duplication of effort between groups working on similar problems can be avoided. Finally, we seek solutions that are **efficient** and avoid **lock-in**. Real-time improvements to the tool-set should be feasible, respecting needs for stability, reliability, and continuity of access to evolving components.

These objectives are fluid and open to interpretation. Community engagement and communication are important to achieving desired goals in this domain.

Key questions

- 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 ensure that visualization methods that are central to multimodal single cell analysis are usable by researchers with visual impairments?

These questions will ultimately be answered through the creation of a **software ecosystem**[38,39,40]. As an example of an ecosystem of broad scope, we cite bioconductor.org [41]. This project produces code in R for **data representation and data services** for many data modalities used in genome-scale experimentation. Bioconductor's resources for achieving **scalability**[42] include tools for analyzing massive data resources with tunable RAM footprints[43], and tooling for supporting fault-tolerant parallel distributed computing in various cluster and cloud contexts [44]. Finally, Bioconductor supports *developers* who seek to build broad user bases by providing multiplatform/multistream **continuous integration/continuous delivery** of contributed packages [45], and *users* with different skill sets by articulating standards for documentation, and testing, and by hosting community forums and workshops[46].

Data management strategies

A ready-to-use integrative data class. The Waldron group at City University of New York has defined a Bioconductor class implementing an abstract data type called "multiassay experiment"[47]. This is relevant for multimodal single-cell experiments as each mode will be characterized by a different collection of features on possibly non-overlapping collections of samples. Metadata on features is bound directly into the class instance. For example, genes and transcripts can be enumerated using Ensembl catalog identifiers, represented as GRanges instances[48]; regions of accessibility from, e.g., ATAC-seq experiments, may be defined using genomic coordinates in a clearly specified reference build. Metadata on samples includes all relevant information on experimental conditions such as treatment, protocol, and date of technical processing. Figure 3 shows how this class can be used to amalgamate and annotate results of a seqFISH experiment.
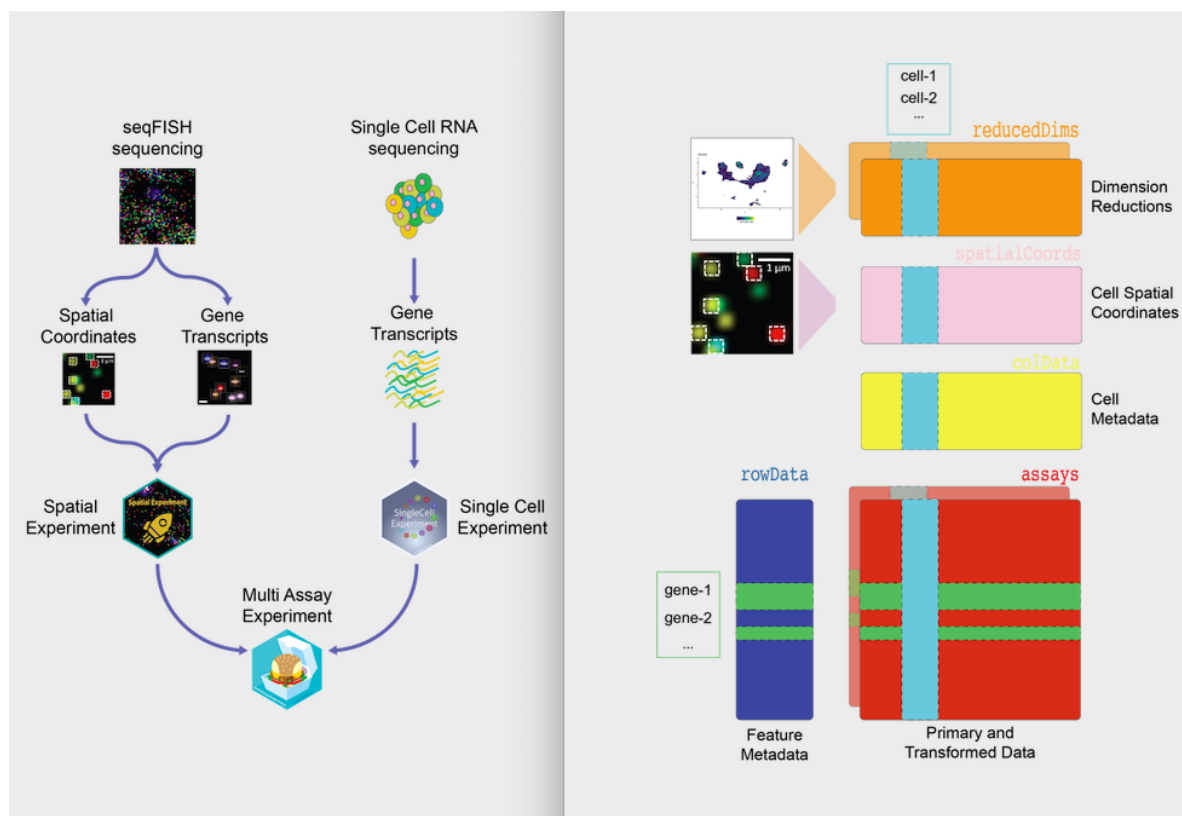
Figure 3: Left: Combination of seqFISH-based SpatialExperiment and SingleCellExperiment instances into a MultiAssayExperiment. Right: details of the SpatialExperiment class design.

Multi-modal single cell data may consist of multi-assay measurements from the same cell (e.g. CITE-seq, sci-CAR) or integration of multi-assay measurements from distinct cells from the same or distinct starting samples. A sample here refers to the biological specimen of origin (tissue A from individual X). The MultiAssayExperiment class includes 1) Assay slots containing variables or features from multiple modalities (e.g. gene expression units from scRNA-seq and protein units from sc-proteomics). 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. 3) A map between the different assays to enable analysis

Of note: - The observations of different modalities may not be directly comparable (e.g. RNA may be measured from individual cells but spatial transcriptomics may cover a few cells in the matched area).
- In the absence of universal standards, the metadata may vary from analysis to analysis. - It is crucial that data containers use consistent assay access methods (possibly through methods inheritance. e.g. from `SummarizedExperiment`). This will ensure less redundancy in development process and allow powerful implementation strategies.

- Serializations and data access methods for
  - spatial transcriptomics
  - scNMT-seq
  - other combinations

Reducing barriers to interpretable visualizations

Color is a powerful data visualization tool that helps representing the different dimensions of our increasingly complex and rich scientific data. Color vision deficiencies affect a substantial portion of the population[49]. Therefore, it is desirable to aim towards presenting scientific information in a manner that is as accessible as possible for all readers. Color deficiency leads to difficulties in perceiving patterns (the basis for the Ishihara's color vision tests) in multi-colored figures. In rare cases, the perceived patterns; e.g. in heatmaps and reduced dimension plots, can differ between individuals with normal and color deficient vision.

One strategy to address these issues is to include colorblind friendly visualizations [50,51] as a default setting in our visualizations. Several colorblind-friendly palettes exist (e.g., see R packages viridis[52] and dittoSeq[53]) and can be integrated into data presentation as the default option. Even with these palettes in place, it is desirable to limit the number (about 8-10 at a maximum) of colors in visualizations. To reduce the dependence on colors, one solution would be to include additional visual cues to differentiate regions (hatched areas) or cells (point shapes). Overall, a broader discussion regarding the accessibility of our figures that is not just limited to color vision deficiencies would be greatly beneficial towards improving data accessibility. Perhaps one tool to address broader accessibility could be the inclusion an "accessibility caption" accompanying figures which "guide" the reader's perception of the images.

US Government tools for accessibility

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

Type	Brief name (link)	Description
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
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

Scalability strategies

Connecting the hackathon activities to the longer-term software strategy for multimodal single-cell experiments

The contributed challenge analyses as well as the used datasets may be collated into software packages which include their specific dependencies. In such a setting, the analysis packages can simply use the corresponding dataset package(s) as a dependency. Additionally, if the challenge datasets have undergone preprocessing it is essential that all the preprocessing steps are outlined and made available in the dataset packages. This will facilitate transparency and analysis reproducibility, as well as allow Continuous Integration (CI) of the analyses and preprocessing changes and Continuous Delivery (CD) of the analysis reports. The CI/CD workflow may also be automated on a hosted server and containerized reports can be generated for enhanced efficiency and portability, respectively.

Discussion

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

Consensus Term	Related Terms	Description	Citation
network	graph, adjacency matrix	A set of *nodes*, representing objects of interest, linked by *edges*, representing specific relationships between nodes.	?
node	vertex	Element of interest in a network and linked to other nodes. For example: people, cells, proteins or genes. Nodes can have several properties called *attributes* like cell type or position.	?
edge	link	The relationship between 2 nodes in a network. For example: friendship in social networks, cells in contact in a spatial network, or gene-gene interactions in a gene regulatory network.	?
concordant	concordant, coherent, consistent	?	33
contributions	variable weights, loadings, eigenvector, axis, direction, dimension, coefficients, slopes	Contributions of the original variables in constructing the components.	32, 54
latent factors	variates, scores, projections, components, latent/hidden/unobserved variables/factors	Weighted linear combinations of the original variables.	32, 54
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.	32, 12, 55
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.	32, 54, 56
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.	57
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.	58, 59
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. **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

2. **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

3. **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

4. **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

5. **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

6. **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

7. **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

8. **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

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

10. **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

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

12. **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

13. **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

14. **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

15. **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

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

17. **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

18. **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

19. **:{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

20. **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

21. **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

22. **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

23. **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

24. **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

25. 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

26. **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

27. **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

28. **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

29. **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

30. **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

31. **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

32. **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

33. **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

34. **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

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

36. **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

37. **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

38. **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

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

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

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

42. **On System Scalability**
Charles B. Weinstock, John Goodenough
Figshare (2018) https://doi.org/gg47cw
DOI: 10.1184/r1/6575879.v1

43. **DelayedArray: A unified framework for working transparently with on-disk and in-memory array-like datasets**
Hervé Pagès at fredhutch.org>, with contributions from Peter Hickey hickey at gmail.com>, Aaron Lun monkeys with keyboards at gmail.com>
Bioconductor version: Release (3.11) (2020) https://bioconductor.org/packages/DelayedArray/

44. **BiocParallel: Bioconductor facilities for parallel evaluation**
Bioconductor Package Maintainer, Martin Morgan, Valerie Obenchain, Michel Lang, Ryan Thompson, Nitesh Turaga
Bioconductor version: Release (3.11) (2020) https://bioconductor.org/packages/BiocParallel/

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

46. https://bioconductor.org/support

47. **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

48. **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

49. https://tinyurl.com/y4emdyvr

50. **Points of view: Color blindness**
Bang Wong

Nature Methods (2011-06-01) https://www.nature.com/articles/nmeth.1618
DOI: 10.1038/nmeth.1618

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

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

53. **dtm2451/dittoSeq**
Daniel Bunis
(2020-06-20) https://github.com/dtm2451/dittoSeq

54. **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

55. **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

56. **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

57. **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

58. **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

59. **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.↩