Mathematical Frameworks for Integrative Analysis of Multi-omics Biological Data

This manuscript (permalink) was automatically generated from BIRSBiointegration/whitePaper@1d664a4 on July 1, 2020.

Authors

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

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

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

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

Abstract

Introduction

Multi-omics and multi-view single cell technologies

- Do multi-omics answer our biological questions?
    - Single cell community has naturally adopted a data-driven approach
    - Helpful to complement with hypothesis-driven and mechanistic-driven approaches
    - Technological advances will help refine our hypothesis (e.g. multi-modal studies focusing on a set of specific genes)
    - Can inform how different levels of regulation are influencing each other

- State-of-the-art

Brief overview of our three hackathon studies

- What do we hope to learn from multi-omics / multi view single cell studies

What bulk TCGA has taught us:

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

New single cell multi omics initiatives:

```
- Human Cell Atlas (HCA): assess variation in normal tissues
- Human Tumor Atlas Network (HTAN):
  Clinical, experimental, computational framework to generate three-
dimensional atlases of cancer transitions for diverse tumor types. Single-
cell, longitudinal, and clinical outcomes
```

- Are our technologies ready?

Do technologies capture the information we want?

Our hackathons featuring cutting edge technologies have showed that: - 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

- **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 learnings from some hackathons case studies that helped us obtain a broader picture and language

Outline of the paper:


Outline of the main themes following our brainstorming sessions from the hackathons (KA Le Cao)

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

Overview and biological question

Computational challenges

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

Methods for stats/math analyses and results summary

Spatial proteomics and cross-study analysis

Overview and biological question

Computational challenges

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

Methods for stats/math analyses and results summary

scNMT-seq as a case-study for epigenetic regulation

Overview and biological question

Computational challenges

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

Methods for stats/math analyses and results summary

Overview of common challenges and analytical methods spanning technologies and hackathon case studies

Short introduction explaining that we focus here on the common challenges across hackathons

Data integration

Selection of pre-processing method and/or variable selection for integration

- Why this is needed and how it was addressed across all hackathons
  - Normalisation / data transformation (seqFish), pre-processing, gene summaries (scNMT-seq) to variable selection (seqFish), these steps affect the downstream analyses, e.g. Alexis selected 19 genes whereas Zhu original paper based on 47 genes (difference in methods / processed data)
- Limitations and further developments required
  - No consensus reached as those are emerging data with no ground truth nor established biological results
  - Difficulty in reproducing original results (Alexis)

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

- Why this is needed
  - Common across all studies (seqFish: same features but not cells; scProt: same proteins, not cells but similar patients; scNMT-seq: same cells but not features)
  - We need some commonalities to anchor information across datasets
- How it was solved across all hackathons:
  - See table, in particular Partial feature overlap (but cells not matching) and No cell overlap but some / all feature overlap addressed by participants
- Limitations and further developments required
  - See table, Partial cell overlap (but no features matching) and No overlap were not addressed

Managing differences in scale and size

- Why this is needed
  - Datasets do 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)
- How it was solved across all hackathons
  - See table, projection based methods seem to work best across all studiesAlso note that pre-processing was performed (but still large datasets)
- Limitations and further developments required
  - Additional weighting is needed (e.g. Arora, Abadi).

Incorporation of existing knowledge (also in future directions)

- Why this is needed

  - ‘From discovery to detection’ (Meuleman + debrief), time is ripe to include more knowledge in our data driven approaches

- Further developments required

- Better gene ontology and hypothesis generated need to be validated (see debrief)

From generic to study specific approaches

Methods previously developed for bulk data that were applied to single cell across our hackathons

- Attempts to tweak existing methods and challenges associated
  - See table Generic approaches for supervised analysis / prediction, Study specific methods and Managing differences in scale: most methods were developed for bulk and applied with some tweaks (e.g dealing with missing values, size).
  - List methods that are either technology dependent (e.g. spatial) vs universal and how to choose them

What do we need for seamless methods extensions to single cell data

- Methods that can span technologies
- Benchmark data
- Better software infrastructure for methods benchmark

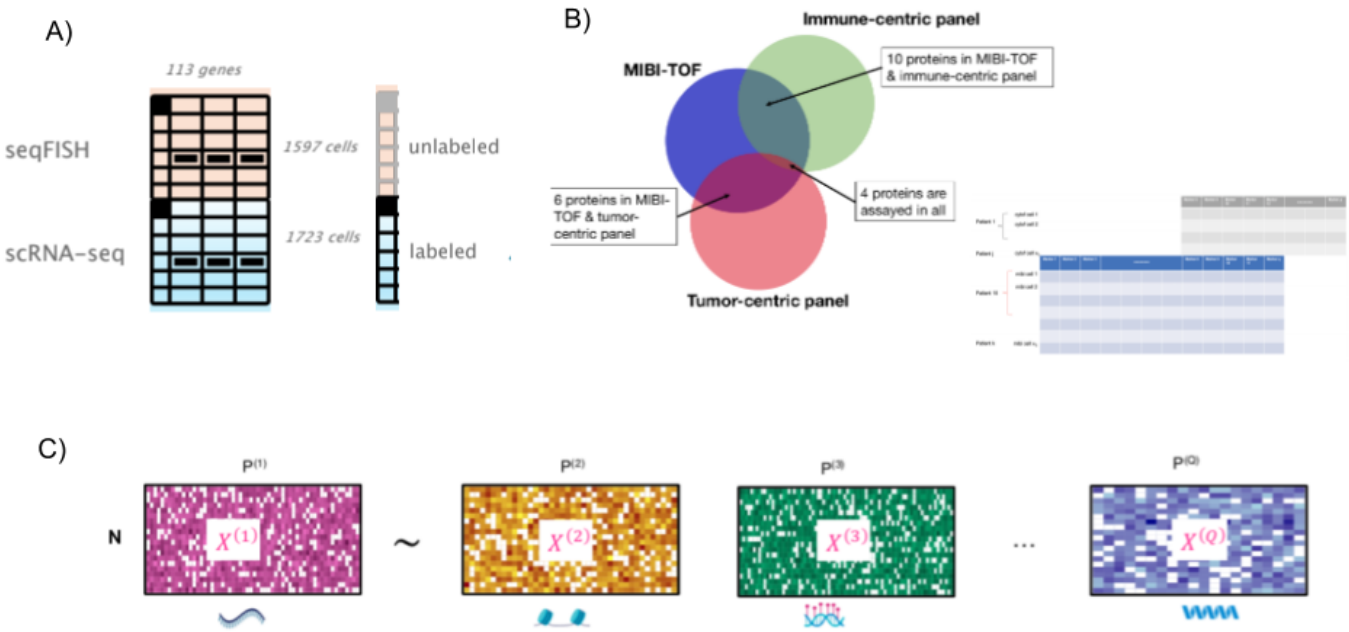


Figure caption:

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]
B. scProteomics: a small number of features overlap (proteins) but patients and cells do not overlap. Data imputation (?), spatial co-localisation or cell type prediction was performed [credit: Lauren Hsu and Pratheepan Jenagan]
C. scNMT-seq: cells are matching across assays but features do not overlap. Data integration was performed [credit: Al Abadi]

Challenges for interpretation

Interpretation requires a good understanding of the methods (no black boxes, a common language)

Communicating within the field: what approaches are we talking about?

Ref to Glossary of terms

Interpretation of the output from the analyses of the data is facilitated by the incorporation of contiguous information.

Redundant biological knowledge and incorporation of information from databases are important in the process. Biological Interpretations: bridges to data bases such as KEGG, Gene Ontology, HCA. Validation through complementary data.

Visualization tools for interpretation and communication

Examples: Brushing UMAP.

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

Issues of over-discretization, over-simplification

Example 1: The notion of cell-type is insufficient (Communication challenge with biologists about tradeoffs between focusing on rare cell types vs. more “continuous” view on cell types).

Problem with loss of information in the desire to simplify.

Over interpretation of graphical outputs

Counterexamples

Techniques and challenges for benchmarking methods

We must first define what we are 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

- Simulation is useful for having known truth, but it is difficult to simulate realistic covariance structure across features and across data modalities.

- 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.

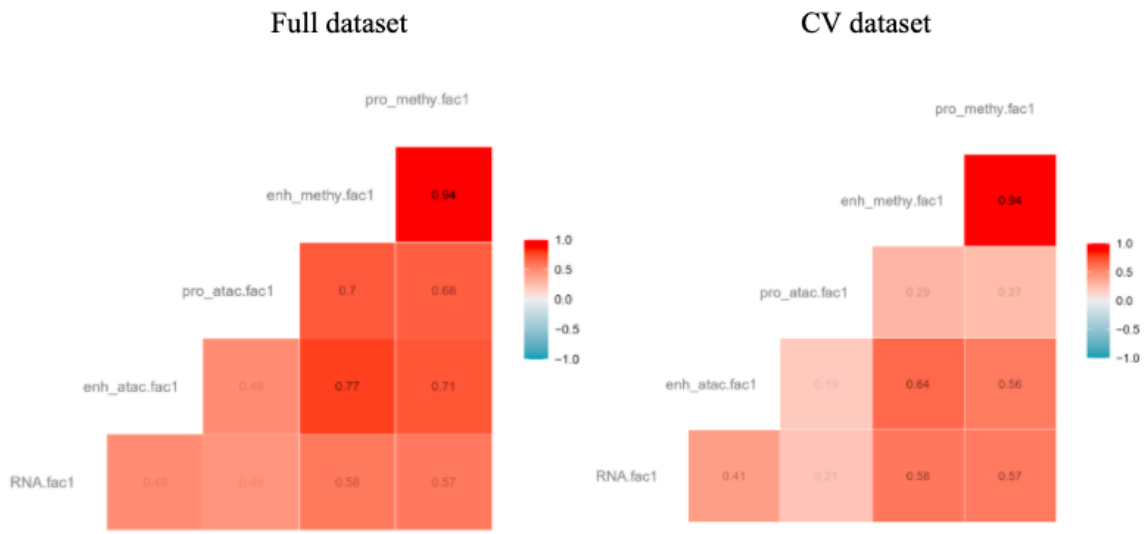
For multi-modal assays, while the intended effects can vary based on the leading biological questions, one may abstract out common data integration tasks such as co-embedding, mapping or correlation, and inferring causal relationships. We distinguish data integration from further downstream analyses that may occur on integrated samples such as differential analysis of both assays wrt 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 can be performed. For example the following cross-validation analysis of the scNMT-seq dataset was performed using MOFA+



A challenge with within study cross-validation is how to match dimensions of latent space across folds. (add examples from Google Doc of papers that have performed either permutation or cross-validation to assess model performance)

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

Software strategies to enable analyses of multimodal single cell experiments

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?

Data management strategies

- Abstract data type: “multiassay experiment”. This reflects the idea that each mode will be characterized by a different collection of features on possibly non-overlapping collections of samples. The metadata on features should be clearly and conventionally defined. For example, genes and transcripts are enumerated using Ensembl catalog identifiers; regions of accessibility are defined using genomic coordinates in a clearly specified reference build. Metadata on samples must include all relevant information on experimental conditions such as treatment, protocol, and date of technical processing.

(More fodder-AS)

Key points: 1) What do we want to store and share? Data object vs analysis object. How would the design change based on what is stored? 2) Do we need a flexible, universal framework (e.g. MAE) or an experiment class for every possible combination of modalities or technologies? 3) Do we have adequate data representation for all “assays”?

- Multi-modal single cell data may consist of multi-assay measurements from the same cell (e.g. citeSeq, 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). A data container for a multi assay analysis must hold

1. Assay slots containing variables or features from multiple modalities (e.g. gene expression units from sc-RNAseq and protein units from sc-proteomics). In some cases, the feature may be multidimensional (e.g. spatial coordinates, locations of eQTLs).
2. Observations or cell identities
3. Metadata for sample of origin for the individual cells, e.g. study, center, phenotype, perturbation.
4. A map between the different assays to enable analysis

- The MAE is such a Bioconductor container for overlapping observations, and may serve as a starting point for further expansion. Besides the primary data elements for storing “data objects”, the summarizedExperiment class offers attributes and Methods for storing results of analysis, as an “analysis object”.
  - While common assays such as RNAseq and ATACseq have well-defined data representations (e.g. transcript names), data representation need to be defined newer assays, which may need multiple dimensions for adequate definition (e.g. x, y, z coordinates for images).
  - 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.

(Note, standard BioC container/class terms may not be correctly used. End of fodder - AS)

- Serializations and data access methods for
  - spatial transcriptomics
  - scNMT-seq ...

Scalability strategies

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. Therefore, it is desirable to aim towards presenting scientific information in a manner that is as accessible as possible for all readers. Color vision 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 [1] as a default setting in our visualizations. Several colorblind-friendly palettes exist (e.g., see R packages viridis and dittoSeq) 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.

Reference 1: Color Coding

Reference 2: Points of View: Color Blindness

Viridis Color Palettes

An overview of the issues with impaired color perception

US Government tools for accessibility

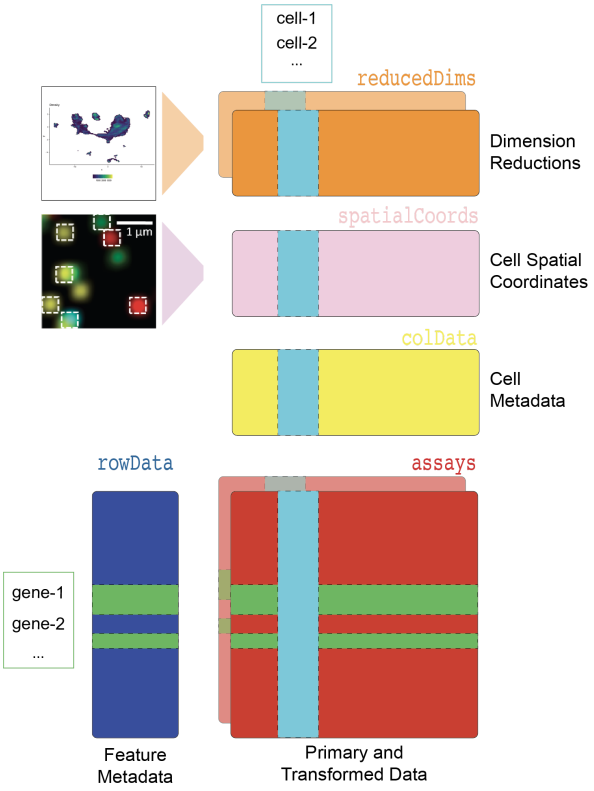
Details of working components – trimmed

you can interact with underlying data at google sheet

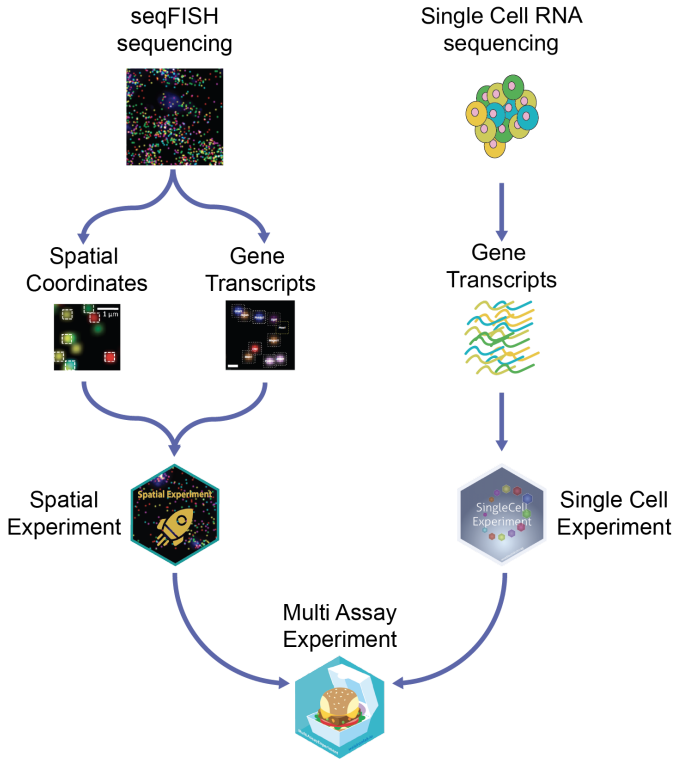
Type	Brief name (link)	Description
Matlab package	CytoMAP	CytoMAP: A Spatial Analysis Toolbox Reveals Features of Myeloid Cell Organization in Lymphoid Tissues
Matlab package	histoCAT	histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data
Python library	PyTorch	General 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

Type	Brief name (link)	Description
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

Here is the schematic of SpatialExperiment class from Dario Righelli.



Here is the schematic of how seqFISH data are stored inot the SingleCellMultiModal package from Dario Righelli.



Discussion

Emerging analytical methods and technologies

Community needs for data structures, analysis methods, etc

Glossary

Consensus term	Synonyms	Description
Network	Graph	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.

References

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