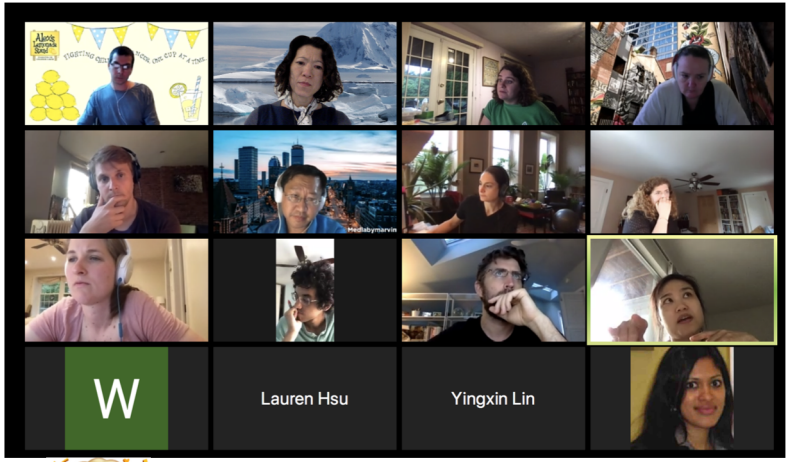

Brainstorming is hard work!

Debrief: summary of analyses and common challenges

Kim-Anh Lê Cao (University of Melbourne)
Casey Greene (University of Pennsylvania)

What’s the fad with multi-omics?

- Do they answer our biological questions?
- Are we technology-ready?
- Are technologies capturing the information we want?
- How do we work out which modality / technology is best
- What are the (many) missing pieces of the puzzle?

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 tell us how different levels of regulation are influencing each other

Are we technology-ready? Are technologies capturing the information we want?

- Our keynotes presented cutting-edge technologies, but ...
  - 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

How do we work out which modality / technology is best to answer a biological question?

The Atlas strategy

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

What are the (many) missing pieces of the puzzle?

- Functional annotations: “what do these genes do within the cell?”
  - Incorporation of prior knowledge (e.g. GO), how to best incorporate prior knowledge (post hoc interpretation, model fitting, etc)

- Experimental design for multi-omics differ from single modalities designs
  - We’ll work it out once we know which modality is useful!

- Multi omics/modalities atlases
  - Balance between consortium level, discovery-driven multi-omics profiling vs. small-scale discovery driven vs. well planned, hypothesis-driven multi-omics research

Common challenges across the 3 hackathons studies

- Partial to no overlap of information (features, cells)
- Inclusion of uncertainty / unknown in our methods (missing value, methyl rates)
- Experimental designs for multi-omic studies
- Data-driven approaches may obscure the biological questions
- Use of a given omic as surrogate for prediction (cells, features, temporal measurements ...)
- Focus on the low hanging fruit (?)
    - Balance computational costs with simpler, faster methods and visualisation
    - Variation / noise too high
- Is our methodological hypothesis matching the biological hypothesis?
    - E.g. statistical correlation == biological association?

Generic vs context specific approaches

- ‘Sometimes it is just a tweak’ (it is not)

- How to facilitate generic towards context specific developments
  - Domain knowledge
  - Toolkits / mega packages shared across the community for easier benchmarking and application of methods (#benchmark_theme and #software_theme)

Methods used across all hackathons (in progress)

Please fill / amend in shared google doc & #summary_theme

Tasks	**seqFISH**	**Sc targeted proteomics**	**scNMT-seq**
Normalisation, transformations, pre-processing	Check data distribution HVG	Row and column wise, VSN Data wise (STATIS, MFA)	
Managing differences in scale (see also: data integration)		Inverse transformation	
Partially overlapping features (imputation)		Optimal transport Topic modelling Direct inversion to predict spatial embeddings Graph based convolution	
No overlap between cells			LIGER (based on NMF)
No overlap between cells or features		RLQ Transfer cell type label with RF	
Classification	SVM self training ENet Balanced error rate		Supervised clustering (MOSAIC)
Feature selection	Recursive Feature Elimination	Spatial discriminative features (spatial autocorrelation / NN correlation)	Lasso in regression-type models
Cell type prediction	projections / clustering SVM ssEnet		
Spatial analysis	HMRF Voronoi tesselation	Moran index, NN correlation / cell type interaction composition, L function Delaunay / Gabriel neighbourhood	
DE analysis	Based on summary statistics		
Data integration	LIGER (NMF) UMAP / tSNE	Multi-block PCA Weighting matrices based on their similarities Phenotype overlapping Correspondence analysis	Projection to Latent Structures LIGER
Include clinical features		Cox regression based on spatial features	

Data integration: what is coming next? (#future_theme)

- Integration across studies (different cells, e.g. atlases)
- Integration with partial feature overlap (e.g. sc Targeted proteomics)
- Integration across studies + multi omics studies