A Tale of Two Spaces

Continuous phenotyping of cells and spatial pixels

Jiadong Mao[1], Jarny Choi[2] and Kim-Anh Lê Cao[1]

1 Melbourne Integrative Genomics & School of Mathematics and Statistics, University of Melbourne, Australia
2 Bioinformatics and Cellular Genomics, St Vincent's Institute, Australia

Email: jiadong.mao@unimelb.edu.au

Introduction

- We developed Φ-space (Phi-space), a computational framework for the continuous phenotyping of single-cell multiomics and spatial transcriptomics data.
- Cell type annotation, or the identification of typical cell identities, is a key step in single-cell and spatial omics analysis. Computational approaches typically utilise well-curated and annotated transcriptomics (eg bulk & single-cell RNA-seq) datasets as the reference to annotate query data.
- With the advancement of sequencing technologies, new query datasets are being generated using radically different platforms compared to the reference. Reference and query datasets can even belong to different omics types. See Table below for details.
- Φ-space bridges the gap between reference and query platforms to achieve robust and continuous annotation of query cells and spatial pixels. Continuous rather than discrete annotation is more useful for the discovery of new cell states and spatial patterns.
- We demonstrate the potentials of Φ-space in three case studies:

Reference	Query	Biological question
Bulk RNA-seq dendritic cell (DC) atlas	scRNA-seq from Rosa et al. (2018)	Validate that query fibroblasts have been reprogrammed into DCs after 9 days
scRNA-seq human bone marrow reference	Multiple batches of bone marrow scATAC-seq	Recover cell types defined by mRNA profiles using chromatin accessibility profiles
Azimuth integrated scRNA-seq human lung atlas	spatial transcriptomics of human non-small cell lung cancer tissues	Spatial distributions of cell identities and their interaction with tumour cells

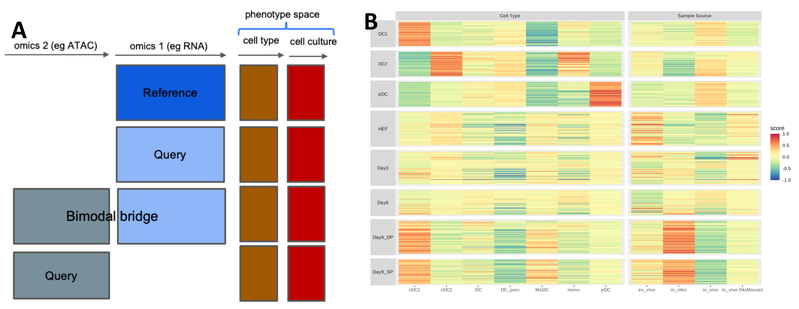
Method

To annotate a query dataset, Φ-space does the following:

1. Construct phenotype space. Choose an annotated reference dataset. Select reference sample phenotypic information to transfer to query data, eg cell types and culture methods. Train a multi-label soft/fuzzy classification model (eg partial least squares) on reference data.
2. Project query to phenotype space. Transfer trained model to annotate query dataset, resulting in a continuous phenotypic characterisation of query samples (eg **Fig B**).
3. Downstream visualisation and analysis based in phenotype space.

A Φ-space projects both reference and query to a constructed phenotype space for continuous phenotyping. If query has a different omics types, we require a bimodal dataset as the bridge.

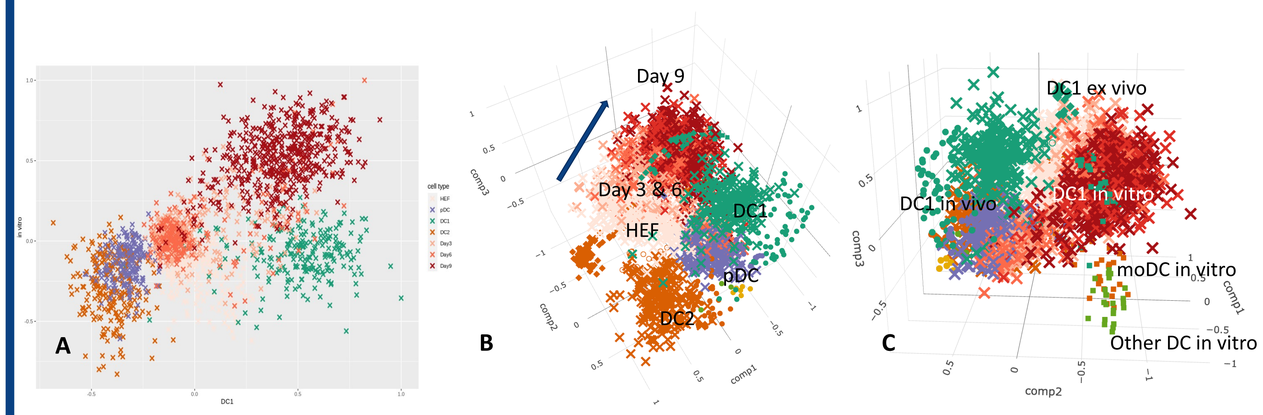
B Phenotype space representation of query cells (rows) from Case 1 below. Each column correspond to one phenotype, ie one axis in the phenotype space.



Case 1: Have those fibroblasts become DC-like after 9 days?

Rosa et al. (2018) proposed a method to generate dendritic cells (DCs) by reprogramming human esophageal fibroblasts (HEF). We use the human DC atlas of Elahi et al. (2022) to construct a phenotype space for characterising DC identities of reprogrammed HEFs.

- Query: scRNA-seq containing induced DCs and some control cell types.
- Reference: bulk RNA-seq atlas of DCs from different culture methods (in vivo, in vitro etc).
- Challenge:
  - Platform effects: bulk vs single-cell
  - Unknown cell identities: After reprogramming for 3, 6 and 9 days, HEF is expected to experience transition in cell identity to become more and more DC like.

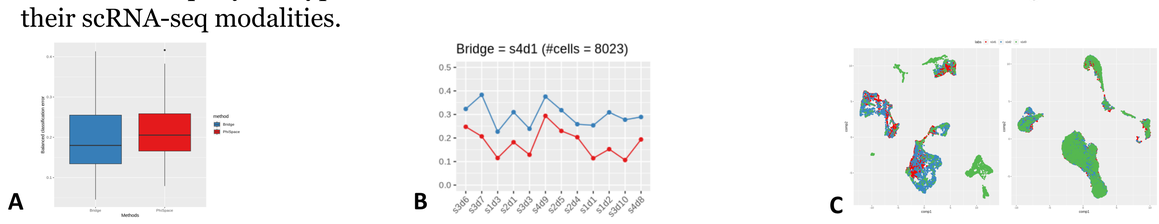


Phenotype space visualisation of query cells (crosses) and reference samples (other shapes, present in **B** & **C** only). **A** Query cells (crosses) along the classical DC type 1 (DC1) and in vitro dimensions in phenotype space. A cell identity transition can be observed: after 9 days of reprogramming, HEF becomes more DC1 like with strong in vitro identity. **B** & **C** First 3 PCs of phenotype space representations of reference and query. Annotation in **B** (**C**) are query (reference) cell types. Day 9 cells are close to DC1 in vitro bulk samples in the reference (dark green squares), whereas the control DC1 cells in the query data are closer to reference DC1 in vivo samples (dark green circles).

Case 2: Can RNA reference annotate ATAC query?

Φ-space is designed for continuous phenotyping, but it is also good at discrete phenotyping, ie cell type label transfer. We use 13 batches of bimodal scRNA+ATAC-seq of bone marrow mononuclear cells (BMMC) generated by Luecken et al. (2021) for a benchmark study. We compare the classification accuracy of Φ-space to Seurat bridge integration (Hao et al., 2023).

- Reference: integrated human BMMC scRNA-seq from Stuart, Butler, et al. (2019).
- Cross-validation (CV) setting: at each CV iteration, we use one of the 13 batches of bimodal scRNA+ATAC-seq as the bridge, and query the scATAC-seq modality of the remaining 12 batches.
- Platform effects: scRNA-seq vs scATAC-seq
- Ground truth query cell types: we use Azimuth BMMC reference to annotate cells in all 13 batches using their scRNA-seq modalities.

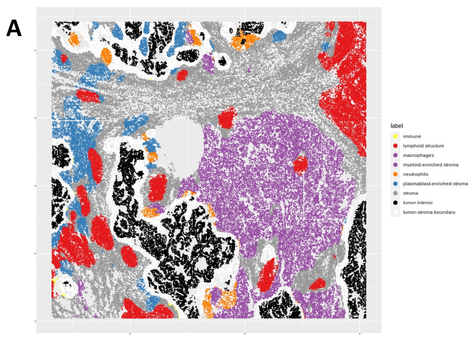


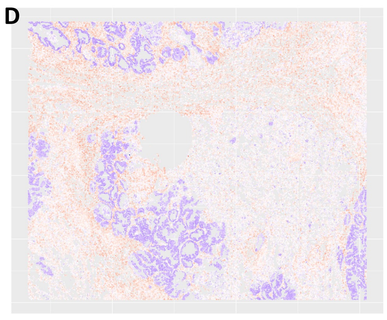
A Overall the performances of Bridge integration and Φ-space (using PLS-based classification) are very close, although Φ-space was less variant. **B** When using batch s4d1 as bridge, Φ-space outperformed Bridge integration for all query sets. **C** UMAPs of ATAC features (left) and their phenotype space representations (right) of cells from 3 batches. The phenotype space representations showed little batch effects.

Case 3: How to characterise cell state variations in situ?

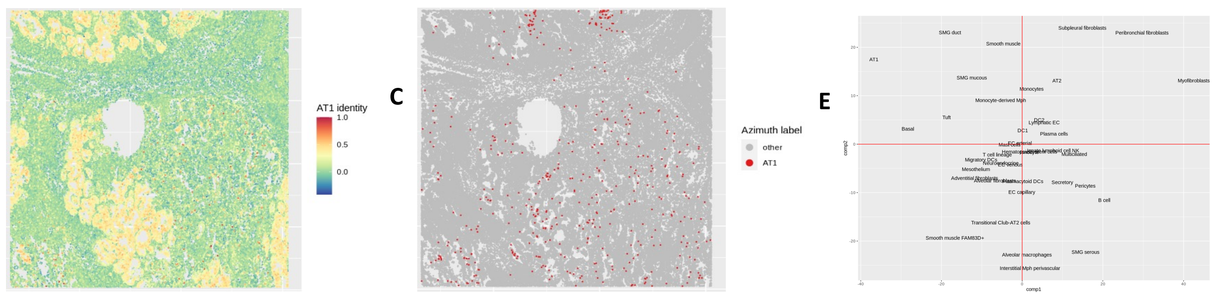
Spatially transcriptomics (ST) allows in situ measurement of transcripts. Different platforms generate transcriptional reads at different resolution. We applied Φ-space to both super- and sub-cellular resolutions. Here we only show results for the sub-celluar resolution.

- Query: ST (Nanostring CosMx) of a non-small cell lung cancer tissue sample (~20mm² area).
- Reference: Azimuth integrated human lung atlas (Sikkema et al., 2023).
- Challenges:
  - Platform effects: reference is whole transcriptome but query only contains ~960 transcripts
  - New cell types: use healthy cell types in reference to characterise cancer cell types





A Partitioning of tissue area into spatial niches done by He et al. (2022). Each point represents centroid of a cell. **B** Φ-space prediction of alveolar type 1 (AT1) cell identity in situ, showing that tumour areas have stronger AT1 identity. **C** Discrete phenotyping using Seurat (Stuart et al., 2019) does not show an association between AT1 and tumour niche. **D** & **E** PCA of cell identity maps for all 39 cell types. PC1 loading in **D** shows a strong negative correlation with tumour niche, showing that the dominating mode of variation in spatial cell identity distributions correspond to cancerous tissue areas. **E** shows PC1 and PC2 scores for all cell types, where PC2 is correlated with muscle and fibroblasts identities.



Conclusions and future work

- Continuous phenotyping based on soft classification showed great potentials in bridging the gaps between reference and query datasets, often generated using radically different platforms. This is useful for maximising the utilisation of existing bulk and single-cell datasets.
- Future work will focus on downstream analysis within the phenotype spaces. For example, in Case 3 we only analysed 1 tissue sample, but a more systematic approach is needed for a large number of tissue samples. Spatial cell identity maps (eg Fig B in Case 3) for a single sample could be organised as a three-way tensor, which can be viewed as a 'tissue-specific cell identity atlas'. Then tensor analysis methods can be used to extract shared and differential spatial patterns across different tissue samples. This creates new and exciting avenues for further method development.

References

Elahi et al. (2022). The human dendritic cell atlas: An integrated transcriptional tool to study human dendritic cell biology. J Immunol
Hao et al. (2023). Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat Biotech
He et al. (2022). High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. Nat biotech
Luecken et al. (2021). A sandbox for prediction and integration of DNA, RNA, and protein data in single cells. NeurIPS 2021
Monaco et al. (2019). RNA-Seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types. Cell Rep
Rajab et al. (2021). An integrated analysis of human myeloid cells identifies gaps in in vitro models of in vivo biology. Stem Cell Rep
Rosa et al. (2018). Direct reprogramming of fibroblasts into antigen-presenting dendritic cells. Sci Immunol
Sikkema et al. (2023). An integrated cell atlas of the lung in health and disease. Nat Med
Stuart, Butler, et al. (2019) Comprehensive integration of single-cell data. Cell

Acknowledgement: ARC DP DP200102903

St Vincent's Institute MEDICAL RESEARCH