#215 A flexible protein-RNA multiomic framework for hierarchical cell typing using marker-derived metagene scores

Dan McGuire[1], Claire Williams[1], Lidan Wu[1], Evelyn Metzger[1], Yi Cui[1], Patrick Danaher[1], Bhattacharjee, Sayani[1], Prajan Divakar[1], Mirko Corselli[1], Joseph Beechem[1]

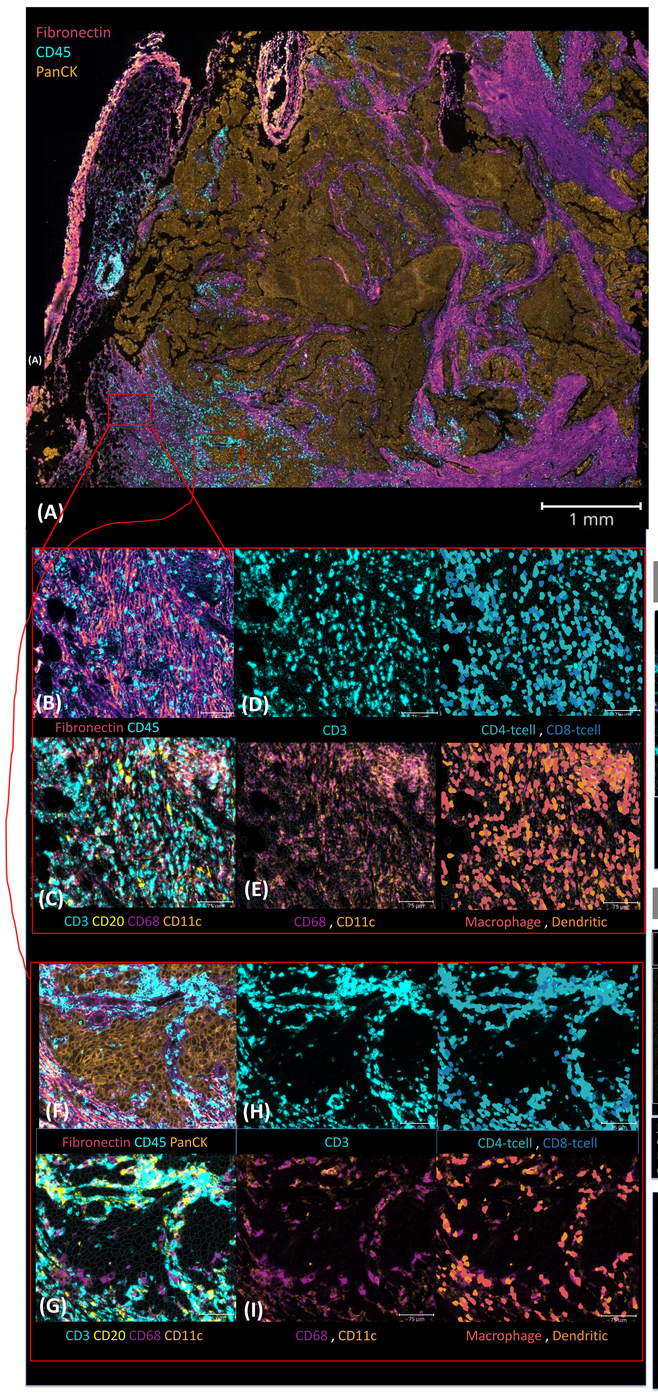
1. Bruker Spatial Biology, Seattle WA 98109

Abstract

Single-cell spatial multiomic assays, which jointly measure RNA transcripts and protein markers, offer unprecedented opportunities for understanding the complex interplay between transcriptional state and cellular phenotype in complex tissues. However, cell typing in these datasets remains challenging: protein signals can bleed across segmented cell boundaries, creating mixed or incompatible marker profiles (e.g., simultaneous detection of CD3 and CD20), while transcriptomic measurements are often sparse and stochastic, limiting the reliability of individual marker genes. To address these challenges, we developed HieraType, a flexible hierarchical cell typing framework that integrates prior biological knowledge with data-driven modeling.

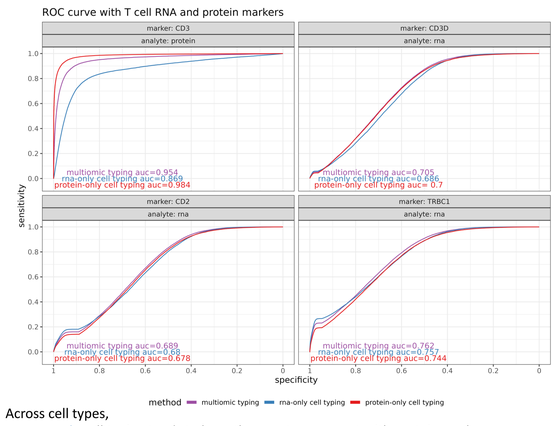
SMI is for research use only and not for use in diagnostic procedure.

A Case Study with Spatial Multiomics Breast Cancer Tissue



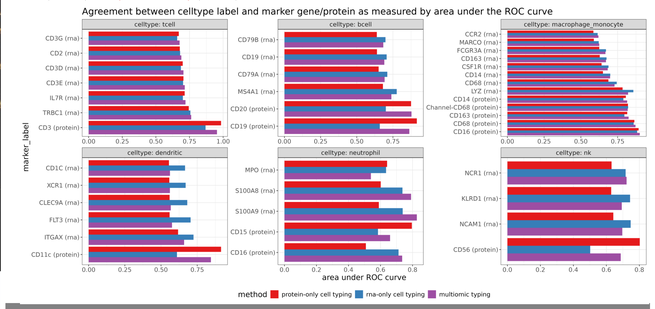
Performance gain from using RNA *and* Protein

The same HieraType pipeline was processed using rna-only, protein-only, or both modalities (multiomics). Performance was measured by calculating the concordance between the continuous protein / RNA marker expression and the binary call for the corresponding cell type. Below, we see multiomics for all three RNA markers, with vast improvement for CD3 protein expression compared to rna-only .

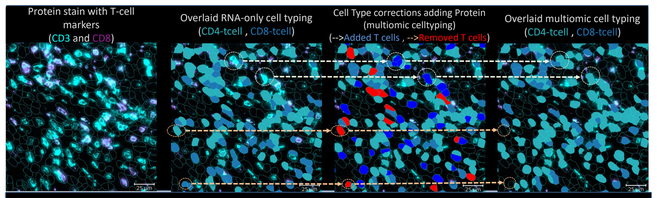


Across cell types,

- rna-only cell typing tends to have the worst agreement with protein markers
- protein-only tends to have the worst agreement with RNA markers
- multiomics **tends to be best / or nearly best across both modalities**
- **Across all markers and celltypes, multiomics has highest average AUC (0.723)** compared to rna-only (0.712) and protein-only (0.697)

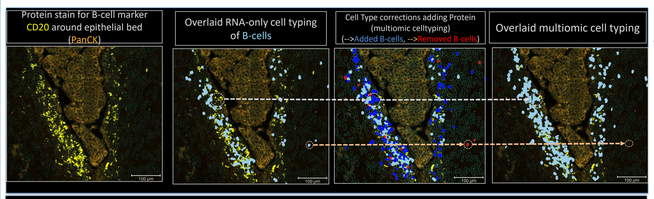


Example: T cell type refinement with Multiomics



Left: Small tissue section with CD3 and CD8 protein signals. Overlaying RNA-only based T cell calls covers much of these signals. Right: T cell call modifications from combined multiomic RNA+ Protein cell typing. Red cells are changed to a non-T cell label, while Blue colored cells are changed to T-cell with additional protein context. These changes largely reflect the underlying protein stain.

Example: B cell type refinement with Multiomics



Left: Small tissue section with lots of CD20 B cell protein signals. Overlaying RNA-only based B-cell calls misses a large proportion of these, likely due to relatively low RNA signal for B cell markers. Right: B cell call modifications from combined multiomic RNA+ Protein cell typing. Majority of these changes are Blue colored cells which are updated to a B-cell label with additional protein context.

Legend for figure to left (A-I)

High-level spatial multiomic view of breast cancer tissue **(A)** and zoomed in view **(B)** of a dense cell region filled with immune and stromal activity. Here, Fibronectin expression appears to dominate over CD45, yet individual immune markers highlight large proportion of immune cell activity including T cells **(D)** and myeloid cells **(E)**.

(F) Another zoomed in view of densely packed region of cells with rich mixture of immune, stromal, and epithelial protein signals. Overlapping and mixed evidence across cell boundaries challenges standard dissociative approaches for assigning cell types **(F-G)**. HieraType algorithm largely overcomes these challenges **(H-I)** by aggregating correlated expression signals of both RNA and protein across biologically relevant genes and similar cells; stabilizing noisy features and preserving interpretability.

Conceptual Overview for HieraType Marker based cell typing

Flexible construction of Protein and RNA markers for a cell type

For each cell type to be called, a 'marker list' is defined. These may include both protein and RNA markers

- **'Index markers'** are definitional and fairly specific to the cell type.
- **'Predictors'** are secondary markers with supporting signal, which may be expressed in only a subset of cells or in other cell types
- Manual marker specification allows **direct control** of what signals influence cell type calls

```
print(HieraType::markerslist_multiomic_tcellmajor$cd8t)

$index_marker
[1] "CD8A"       "CD8B"       "CD8_protein"

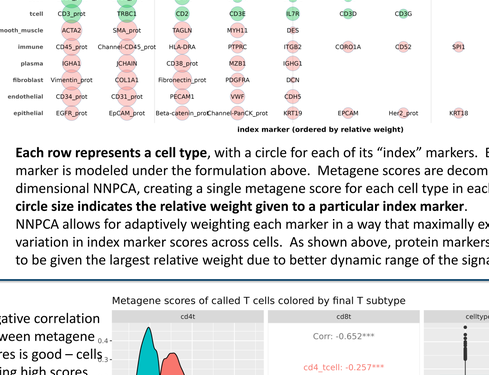
$predictors
 [1] "CD8_protein"   "CD3_protein"   "CCR7_protein"   "CD45RA_protein" "CD27_protein"   "GZMA_protein"
 [7] "GZMB_protein"  "LAG3_protein"  "TCF7_protein"   "CD3D"           "CD3E"           "CD3G"
[13] "CD8A"          "CD8B"          "IL7R"           "CD27"           "SELL"           "CCR7"
[19] "TCF7"          "LEF1"          "FOXP1"          "BACH2"          "KLF2"           "CD28"
[25] "ICOS"          "STAT3"         "STAT5"          "RUNX3"          "BATF"           "EOMES"
[31] "TBX21"         "PRDM1"         "PDCD1"          "LAG3"           "TIGIT"          "HAVCR2"
[37] "GZMK"          "GZMA"          "GZMB"           "PRF1"           "GNLY"           "KLRG1"
[43] "KLRD1"         "IFNG"          "CCL5"           "CX3CR1"         "NKG7"           "FAS"
[49] "FASLG"         "IL2"           "CD244"          "CXCR3"          "KLRB1"          "TOX"
[55] "NR4A2"         "BCL6"          "CXCL13"
```

Hierarchical Framework

Model estimated Prob(Immune) → carried over as prior/weight for immune subtyping

Prob(Tcell) → carried over to T subtyping

… →



Major 'Level 1' Cell type classes

Major Immune cell types

T cell subtypes

CD8 T cell subtypes

Arbitrary hierarchy supported; can extend subtypes into deeper levels or redefine hierarchy.

Subtypes can be defined for any cell type class

Constructing Cell Type Metagene Scores

$$Y = WY\eta + WX\beta + X\gamma + \epsilon$$

Pearson residuals of the **index marker** gene/protein

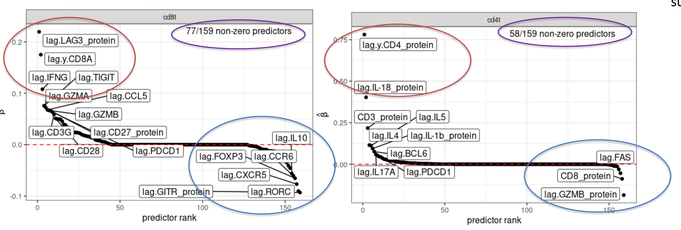
Smooth average of Pearson residuals of the **index marker** across similar cells in PCA space

Smooth average of potentially relevant **predictor gene/protein expression** across similar cells in PCA space

Predictor gene/protein expression

For each index gene in a cell type, construct a metagene score leveraging information across similar cells in PCA space and biologically relevant predictor genes

Cell type metagene scores from Non-linear least squares regression

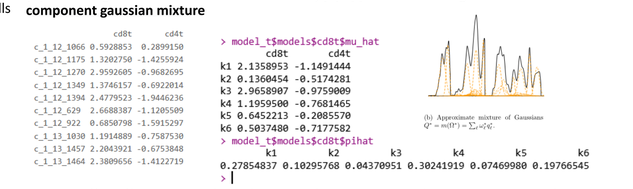


- Metagene weights are restricted to be **positive** for markers associated with the cell type, **negative** for markers only associated with other cell type classes in the same level of hierarchy.
- Non-useful genes shrink to zero, negligible impact of specifying the wrong markers in markerlist.
- Sign restriction enforces the appropriate direction of covariance among relevant gene expression within a cell type.
- Learning weights directly from data means no requirement of a full panel reference profile, flexible across platforms and modalities, i.e., learning cell types from dissociative scRNAseq reference

Probabilistic Clustering of continuous cell type metagene scores

Let cells with positive metagene scores for a celltype be "anchor cells" (A). Metagene scores for a cell type may not be normally distributed or follow any known parametric distribution. Following **Slawski Et. Al (2),** we can approximate the distribution of metagene scores across anchor cells for a celltype using a *mixture* of normals

(A) Ex: Metagene scores for "cd8t" anchor cells, (B) approximated using an 'overfitted' 6-component gaussian mixture



Overall Cell Type probability can be computed using Bayes rule

Joint probabilities for either CD8/CD4 T cell

```
cell_ID best_class best_score      cd8t      cd4t
  <char>     <char>      <char>     <char>    <char>
  c_1_100_1      cd4t      0.1813    0.1135    0.1813
c_1_100_1000     cd4t      0.4077    0.2529    0.4077
c_1_100_1016     cd8t      0.9854    0.9854 0.003229
c_1_100_1018     cd4t      0.9905 0.0003658    0.9905
c_1_100_1025     cd4t      0.4179 4.149e-06    0.4179

  c_1_99_989     cd4t      0.8318 3.422e-05    0.8318
  c_1_99_990     cd8t      0.9561    0.9561   0.03587
  c_1_99_996     cd4t      0.5882    0.4114    0.5882
  c_1_99_997     cd4t      0.6296    0.3704    0.6296
  c_1_99_998     cd4t      0.8868 6.186e-06    0.8868
```

Let y_c be the cell type category for cell c, $scores_c$ the corresponding metagene scores, and t denoting a possible cell 'type' or category (for $t = 1, \ldots, T$).

We compute the overall posterior probability that y_c is a particular type t, as

$$P(y_c = t|scores_c) = \frac{P(scores_c|y_c = t)P(y_c = t)}{\sum_{l=1}^{T} P(scores_c|y_c = l)P(y_c = l)}$$

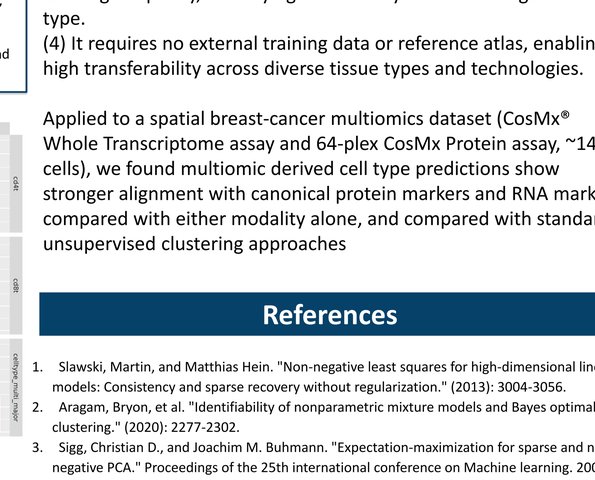
Using the individual mixture models for a particular cell type with K mixture components, we have

$$P(scores_c|y_c = t) = \sum_{j=1}^{K} P(scores_c|k = j, y_c = t)P(k = j|y_c = t)$$

The prior probability for a particular cell type class t can be estimated by counting the number of cells used to estimate the mixture model in each cell type, and computing a proportion amongst all possible cell type classes, i.e.,

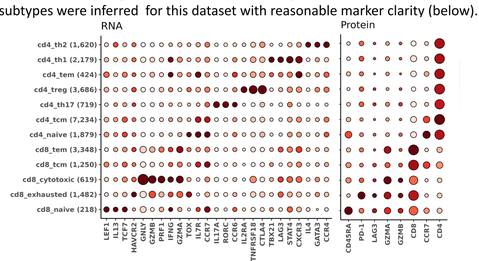
$$P(y_c = t) = \frac{\sum_{c=1}^{C} I(scores_{c,t} > 0)}{\sum_{l=1}^{T}\sum_{c=1}^{C} I(scores_{c,l} > 0)}$$

Cell type confidence scores increase/decrease directly with marker expression



Granular T subtyping demonstration

Granular T subtyping is omitted from previous figures for sake of clarity, but CD4 and CD8 T subtypes were inferred for this dataset with reasonable marker clarity (below).



Software and blogpost links

- https://github.com/Nanostring-Biostats/CosMx-Analysis-Scratch-Space/tree/Main/_code/HieraType
- https://nanostring-biostats.github.io/CosMx-Analysis-Scratch-Space/posts/multiomics-hieratype
- https://nanostring-biostats.github.io/CosMx-Analysis-Scratch-Space/posts/immune-cell-typing-smooth

Conclusions

To our knowledge, HieraType is the first dedicated method for hierarchical cell typing in spatial multiomics and incorporates several key advances:

(1) It aggregates correlated expression signals across biologically related genes and proteins to construct cell type metagene scores; stabilizing noisy features and preserving interpretability.

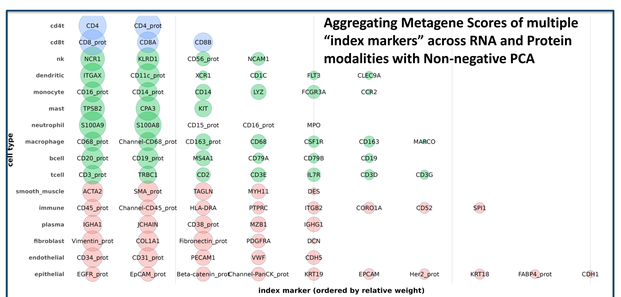
(2) It formalizes marker-based cell typing through probabilistic modeling, transforming an ad hoc process into a reproducible, quantitative framework.

(3) It adaptively weights RNA and protein contributions based on local signal quality, identifying which analytes best distinguish cell type.

(4) It requires no external training data or reference atlas, enabling high transferability across diverse tissue types and technologies.

Applied to a spatial breast-cancer multiomics dataset (CosMx® Whole Transcriptome assay and 64-plex CosMx Protein assay, ~147k cells), we found multiomic derived cell type predictions show stronger alignment with canonical protein markers and RNA markers compared with either modality alone, and compared with standard unsupervised clustering approaches

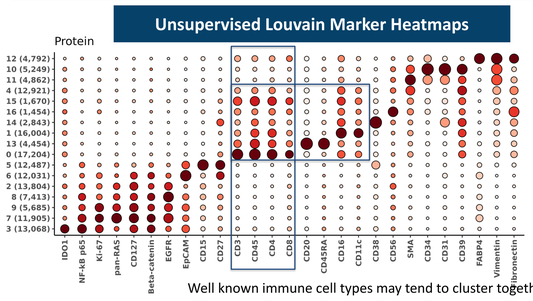
Aggregating Metagene Scores of multiple "index markers" across RNA and Protein modalities with Non-negative PCA



Each row represents a cell type, with a circle for each of its "index" markers. Each index marker is modeled under the formulation above. Metagene scores are decomposed via 1-dimensional NNPCA, creating a single metagene score for each cell type in each cell. Above, **circle size indicates the relative weight given to a particular index marker**.
NNPCA allows for adaptively weighting each marker in a way that maximally explains the variation in index marker scores across cells. As shown above, protein markers typically tend to be given the largest relative weight due to better dynamic range of the signals.

Performance Comparison and Challenges with Unsupervised Clustering

Unsupervised Louvain Marker Heatmaps

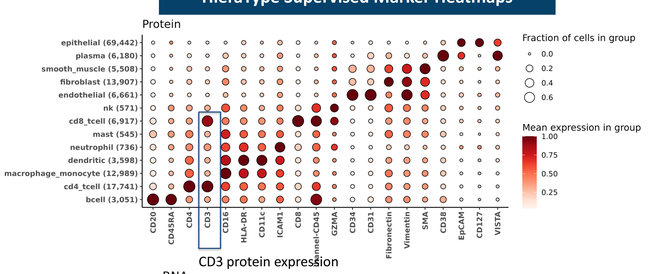


Well known immune cell types may tend to cluster together, or fail to differentiate between known subtypes (i.e., CD8+ vs. CD4+ T cells)

Clusters may be defined by lack of expression, or cells with low signal / few counts.

Unsupervised clusters easily defined by "cell state" rather than "cell type". i.e., variability in markers of proliferation that cluster across multiple cell types

HieraType Supervised Marker Heatmaps

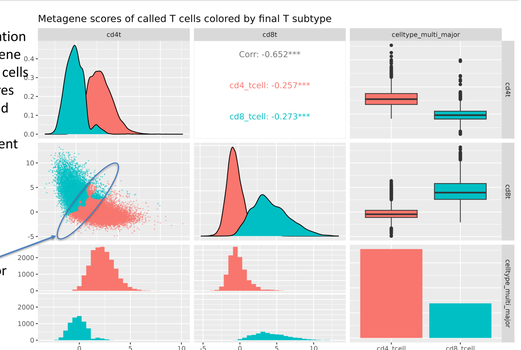


CD3 protein expression concentrated within T cell classes

Easy in practice to check if markers reflect user-specified cell type definitions

Negative correlation between metagene scores is good – cells having high scores for one class and low for another make for confident cell type calls

Lower confidence for more ambiguous cells reflected in probabilistic clustering



References

1. Slawski, Martin, and Matthias Hein. "Non-negative least squares for high-dimensional linear models: Consistency and sparse recovery without regularization." (2013): 3004-3056.
2. Aragam, Bryon, et al. "Identifiability of nonparametric mixture models and Bayes optimal clustering." (2020): 2277-2302.
3. Sigg, Christian D., and Joachim M. Buhmann. "Expectation-maximization for sparse and non-negative PCA." Proceedings of the 25th international conference on Machine learning. 2008.

-Bruker Confidential-