CENTER FOR
COMPUTATIONAL BIOMEDICINE
HARVARD MEDICAL SCHOOL

scDiagnostics: diagnostic functions to assess the quality of cell type annotations in single-cell RNA-Seq data

scDiagnostics Package Authors

Authors and Contributors

Anthony Christidis [Creator & Author]
Computational Scientist,
Center for Computational Biomedicine,
Harvard Medical School

Andrew Ghazi [Author]
Statistical Geneticist,
Center for Computational Biomedicine,
Harvard Medical School

Smriti Chawla [Author]
Postdoctoral Fellow,
Center for Computational Biomedicine,
Harvard Medical School

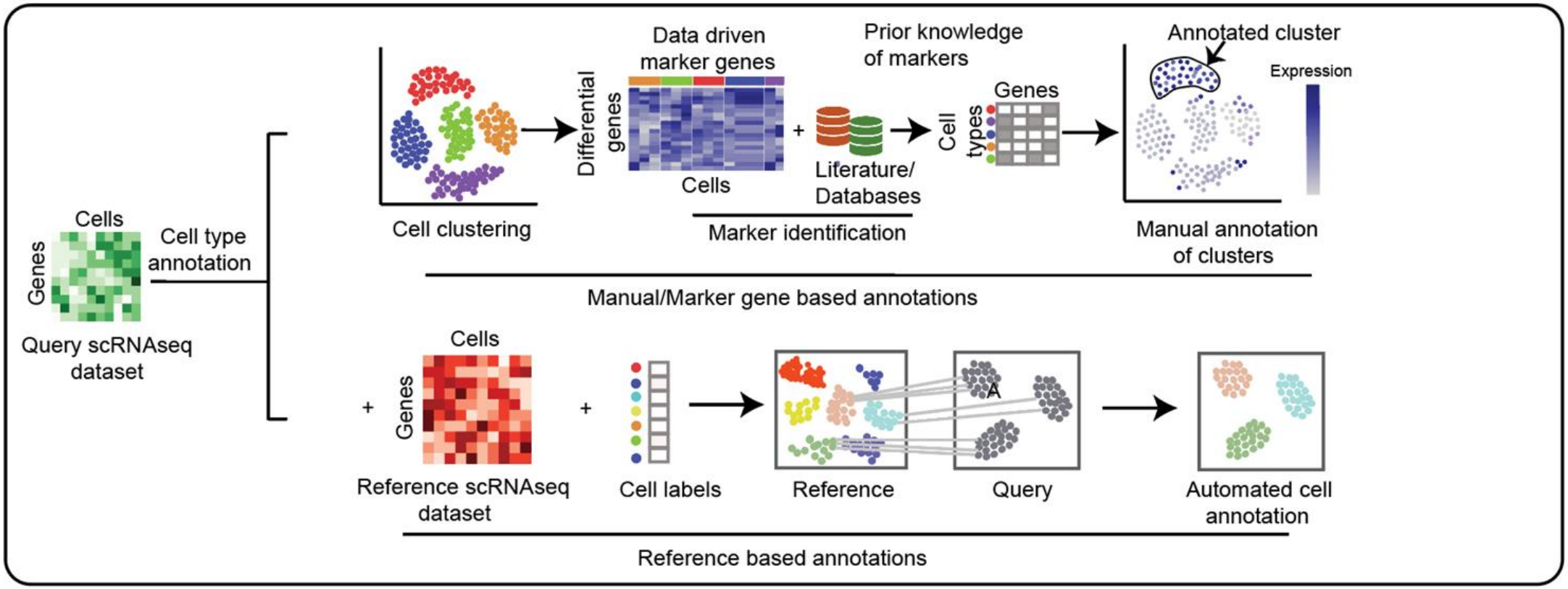
Nitesh Turaga [Contributor]
Consultant,
Center for Computational Biomedicine,
Harvard Medical School

Ludwig Geistlinger [Author]
Director of Computational Biology,
Center for Computational Biomedicine,
Harvard Medical School

Robert Gentleman [Author]
Founding Executive Director,
Center for Computational Biomedicine,
Harvard Medical School

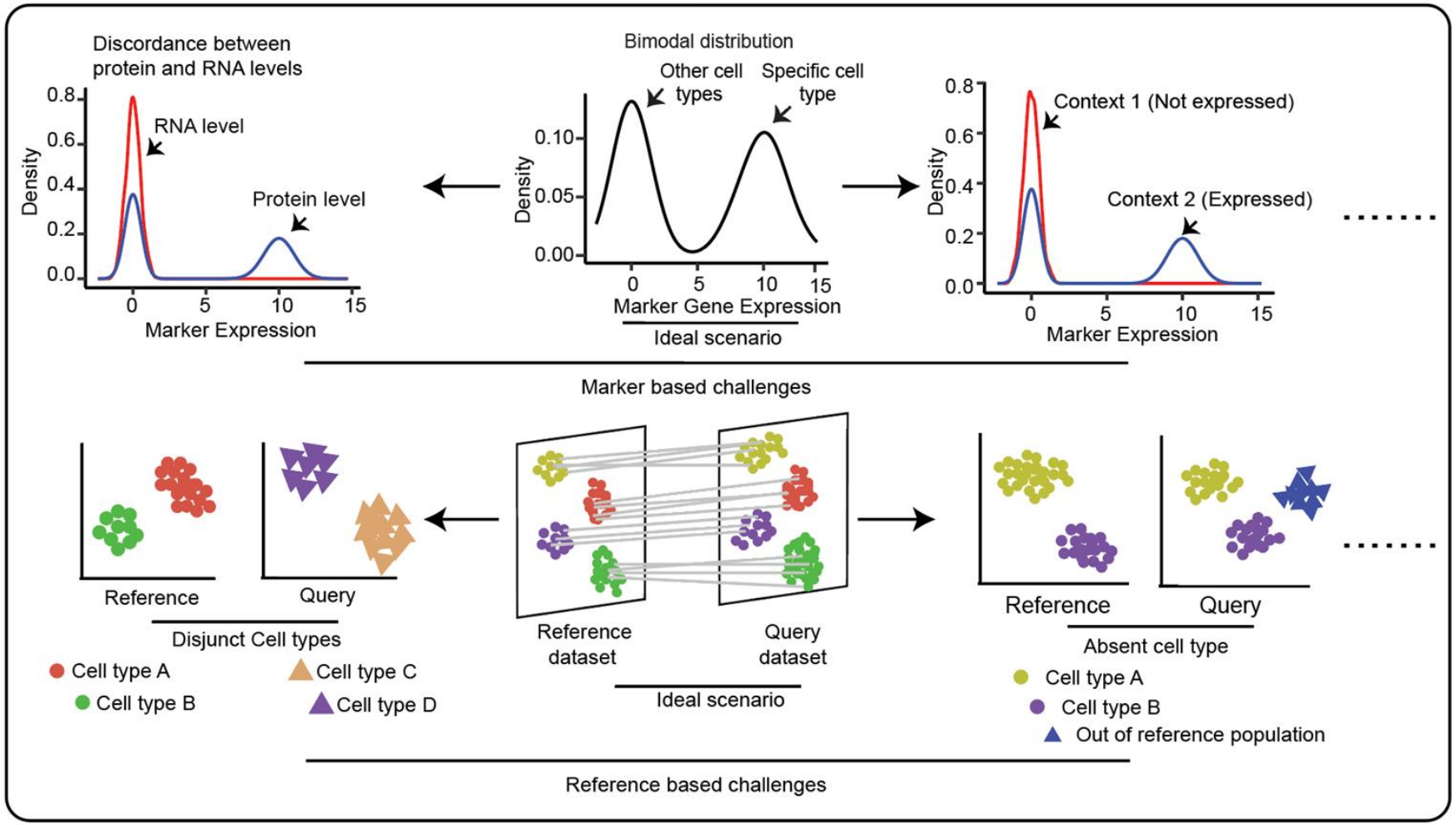
Background

Two Main Approaches to Cell Type Annotation



Background

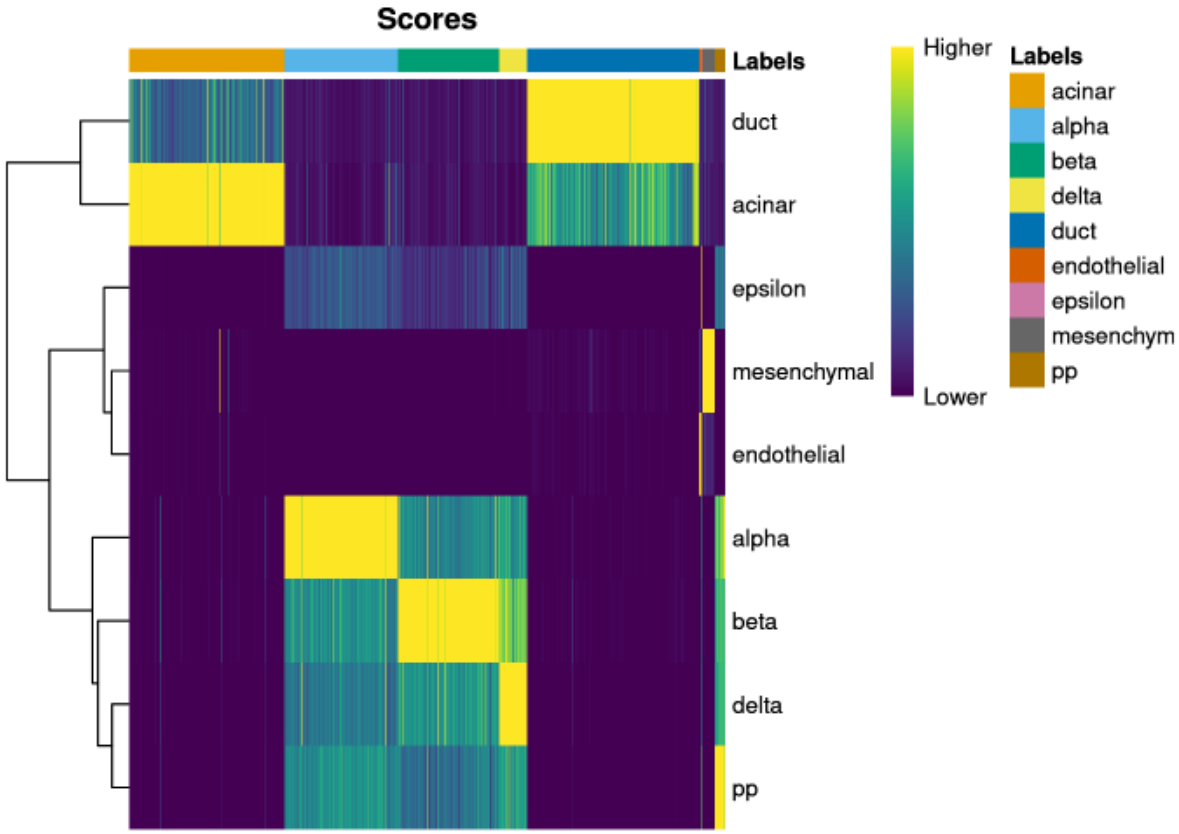
Undiagnosed Challenges in Cell Type Annotation



Previous Work

Annotation Diagnostics in SingleR

- **SingleR Annotation Diagnostics:**
  - Chapter 4 of SingleR book
- **Limited Scope:**
  - If some cell types are unique to query dataset, an annotation score is still given.
  - If a cell has an ambiguous cell type assignment, it is hard to see why.



Data

Reference and Query Datasets

- **Reference Data:**
  - Well-curated & expert-annotated
  - Cells have been labeled by domain experts using known marker genes and experimental validation (sometimes…)
  - Serves as ground truth for cell type annotation transfer
  - Used to train models, identify cell types, and serve as benchmark data for new annotation methods.
  - **Reference datasets in the `celldex` package:**
    - **Human Cell Atlas (HCA):** >100,000 cells from 15 human tissues
    - **Mouse Cell Atlas (MCA):** >100,000 cells from 14 mouse tissues.
    - **Human Peripheral Blood Mononuclear Cells (PBMCs):** >20,000 PBMCs from healthy donors, covering various immune cell types
    - And many more…

Data

Reference and Query Datasets

- **Query Data:**
  - new single-cell RNA-seq dataset that needs to be analyzed / annotated
  - cell types have been annotated by some method (e.g. annotation transfer), but their accuracy as not been assessed
  - **Goals:**
    - **Dataset alignment:** Ensure that query data aligns well with the reference data.
    - **Annotation Validation:** Assess to which extent the query cells have been well annotated.
    - **Anomalous Cell Detection:** Identify any potentially anomalous cells that may indicate issues with data quality or annotation.

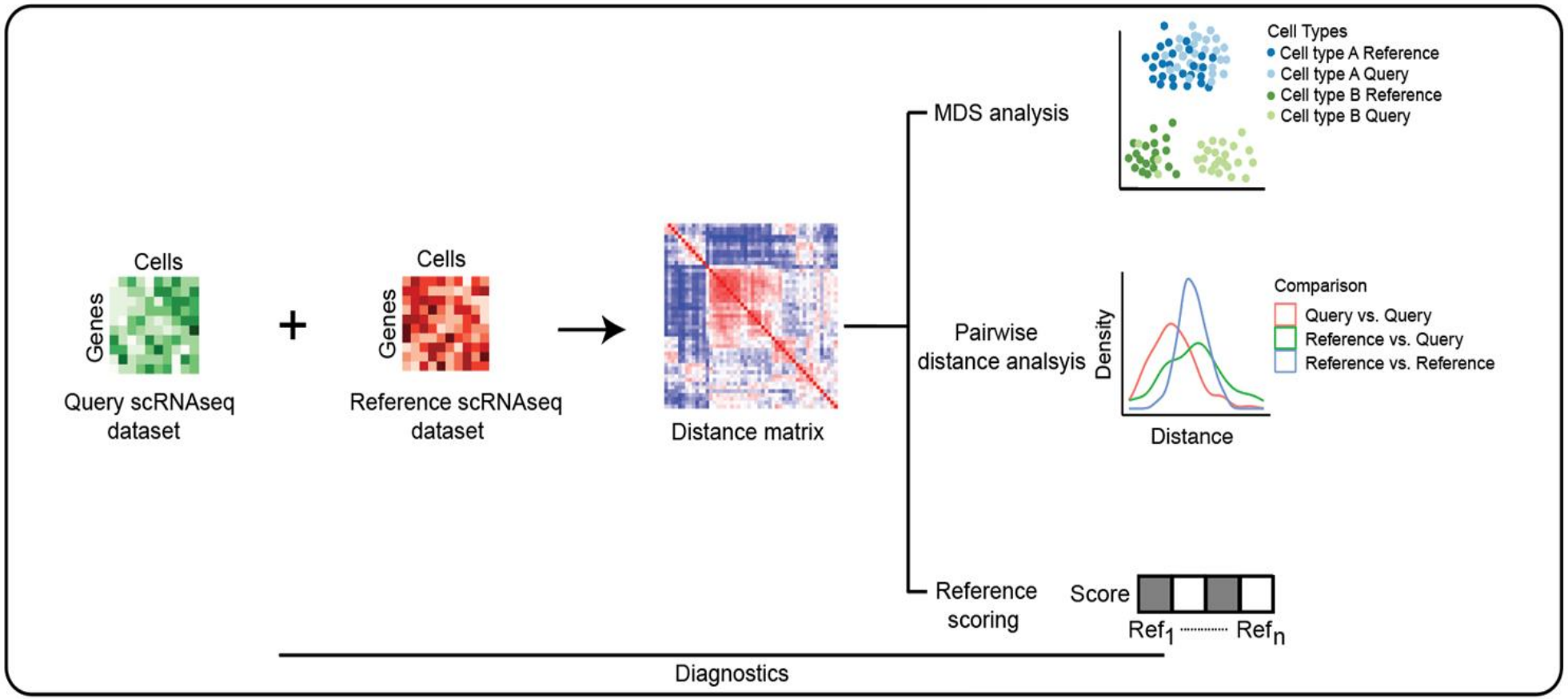
Introducing scDiagnostics

Enhancing the Standard scRNA-seq Cell Annotation Pipeline

- **Core Functionality:**
  - **Data Alignment Checking:** Ensures that the query dataset is well aligned with the reference dataset, verifying the consistency of gene expression patterns.
    - E.g. PCA subspace comparison, canonical correlation analysis, Wasserstein Distances.
  - **Annotation Validation:** Confirms the accuracy of cell type annotations in the query dataset, using statistical and computational methods.
    - E.g. discriminant space projections, PCA projections, distance and correlation-based analyses.
  - **Anomalous Cell Detection:** Identifies potentially anomalous cells, ensuring robust and accurate results.
    - E.g. isolation forests, direction of outlyingness (future work).
- **Impact:**
  - Enhances the reliability and accuracy of scRNA-seq data analysis.
  - Provides confidence in the results obtained from downstream analysis.

Introducing scDiagnostics

Enhancing the Standard scRNA-seq Cell Annotation Pipeline



Example

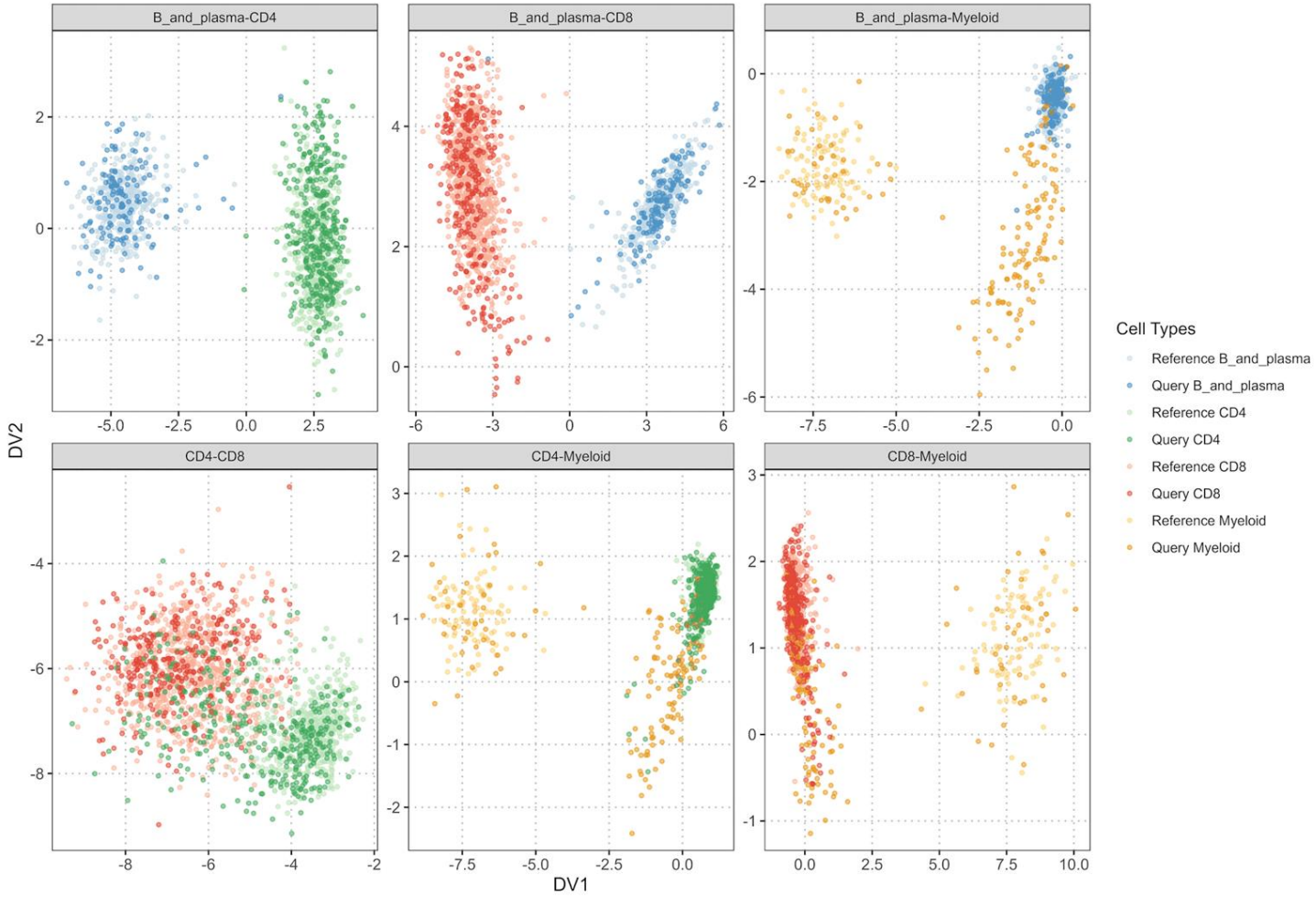
Cell Type Annotation With SingleR

- `HeOrganAtlasData(tissue = c("Marrow"))` from the `scRNAseq` package
- Standard OSCA-based preprocessing (QC, normalization, etc.)
- 70% of data as “reference” and remaining 30% as “query” data
- Annotation transfer from reference to query via `SingleR`

	SingleR Annotation			
Expert Annotation	B and Plasma	CD4	CD8	Myeloid
B and Plasma	139	0	0	10
CD4	0	190	1	13
CD8	0	134	196	29
Myeloid	0	0	0	40

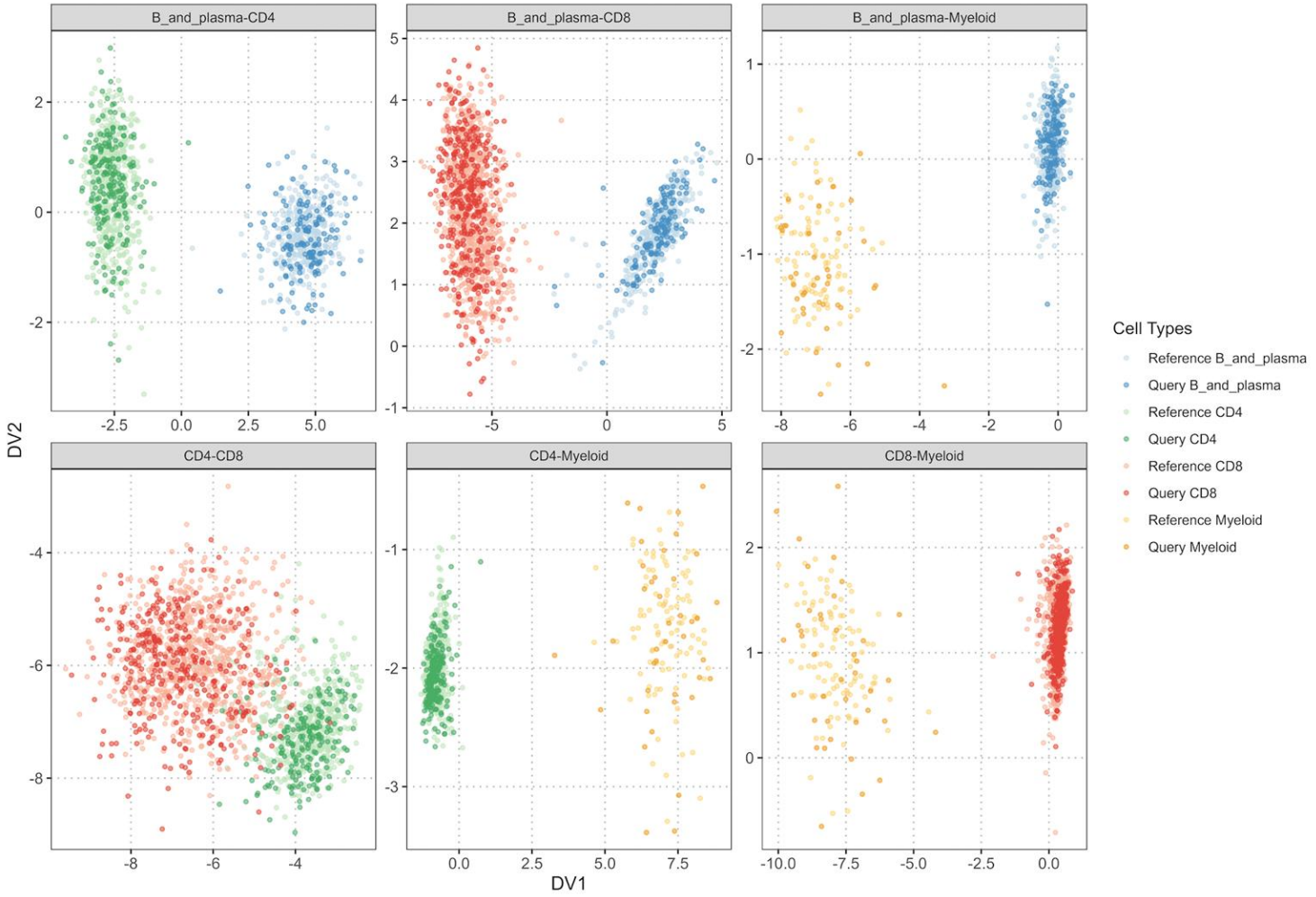
Visualization of SingleR Annotation

“Poor” Annotation in Discriminant Space



Visualization of Expert Annotation

“Good” Annotation in Discriminant Space



Workshop Materials

Package Information and Links

- **Package (development) GitHub repository**
- **Package website**
- **Workshop Materials:**
  - Repository
  - Slides
  - Vignette
  - Docker image
  - Galaxy workshop
- **Package will soon be available on Bioconductor.**

CENTER FOR
COMPUTATIONAL BIOMEDICINE
HARVARD MEDICAL SCHOOL

Live Workshop Session