CENTER FOR
COMPUTATIONAL BIOMEDICINE
HARVARD MEDICAL SCHOOL

scDiagnostics: Diagnostic Tools to Assess the Cell Type Assignment Quality in Single-Cell RNA-Seq

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

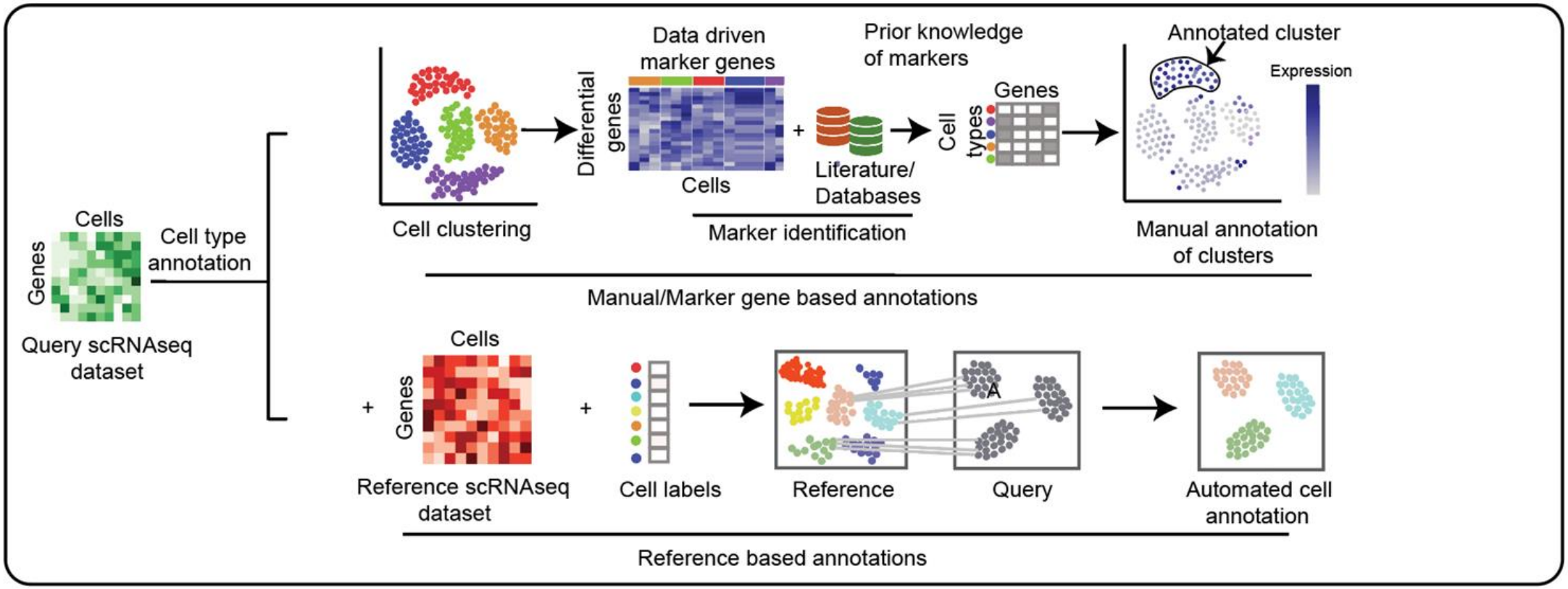
Nitesh Turaga [Contributor]
Consultant,
Center for Computational Biomedicine,
Harvard Medical School

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

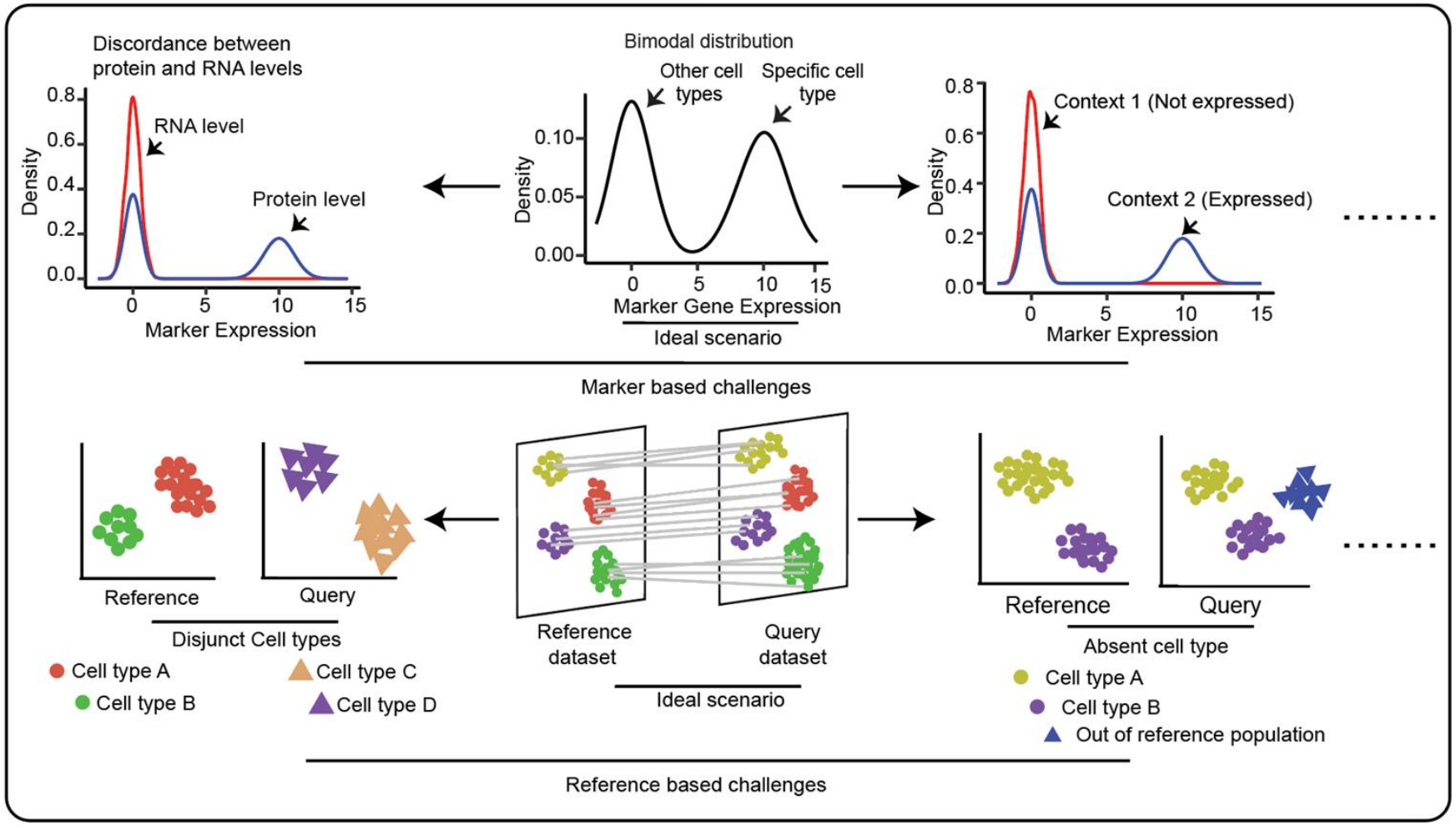
Background and Motivation

Two Main Approaches to Cell Type Annotation



Background and Motivation

Statistical Challenges in Cell Type Annotation



Previous Work in Annotation Diagnostics

Annotation Diagnostics in SingleR

- **SingleR Annotation Diagnostics:**
  - Chapter 4 of SingleR book
- **Methodological Limitations:**
  - **Uncertain Accuracy:** The diagnostics methods may not always accurately reflect true cell types, leading to potential misassignments.
  - **Limited Scope:** The available methods often fail to address the full complexity of biological variability and technical artifacts.
- **Visualization Limitations:**
  - **Inadequate Tools:** SingleR lacks comprehensive visualization options to effectively interpret annotation results.
  - **Poor Clarity:** Current visualization methods do not provide clear insights into annotation quality or highlight potential issues effectively.

Context and Data Overview

The Reference and Query Datasets

- **Reference Data:**
  - **Description:** The reference dataset should be a well-curated, expertly annotated collection of single-cell RNA-seq data.
  - **Expert Annotations:** Cells in this dataset have been accurately identified and labeled by domain experts using known marker genes and experimental validation (sometimes…).
  - **High-Quality Data:** This dataset serves should serve as ground truth for cell annotation transfer.
  - **Usage:** Used to train models, identify cell types, and serve as a benchmark data for new annotation methods.
  - **Example References Datasets in Bioconductor package celldex:**
    - **Human Cell Atlas (HCA):** This dataset contains scRNA-seq data from over 100,000 cells from 15 human tissues, including blood, bone marrow, brain, kidney, liver, lung, pancreas, and more.
    - **Mouse Cell Atlas (MCA):** Similar to HCA, this dataset contains scRNA-seq data from over 100,000 cells from 14 mouse tissues.
    - **Human Peripheral Blood Mononuclear Cells (PBMCs):** This dataset contains scRNA-seq data from over 20,000 PBMCs from healthy donors, covering various immune cell types.
    - And many more…

Context and Data Overview

The Reference and Query Datasets

- **Query Data:**
  - **Description:** The query dataset consists of new single-cell RNA-seq data that needs to be analyzed.
  - **Annotations:** The query cells have been annotation by some method (e.g. annotation transfer), but their accuracy as not been assessed.
  - **Analysis Goals:**
    - **Alignment Check:** Ensure that the new 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.
  - **Annotation Validation:** Confirms the accuracy of cell type annotations in the query dataset, using statistical and computational methods.
  - **Anomalous Cell Detection:** Identifies potentially anomalous cells, ensuring robust and accurate results.
- **Impact:**
  - Enhances the reliability and accuracy of scRNA-seq data analysis.
  - Provides confidence in the results obtained from downstream analysis.

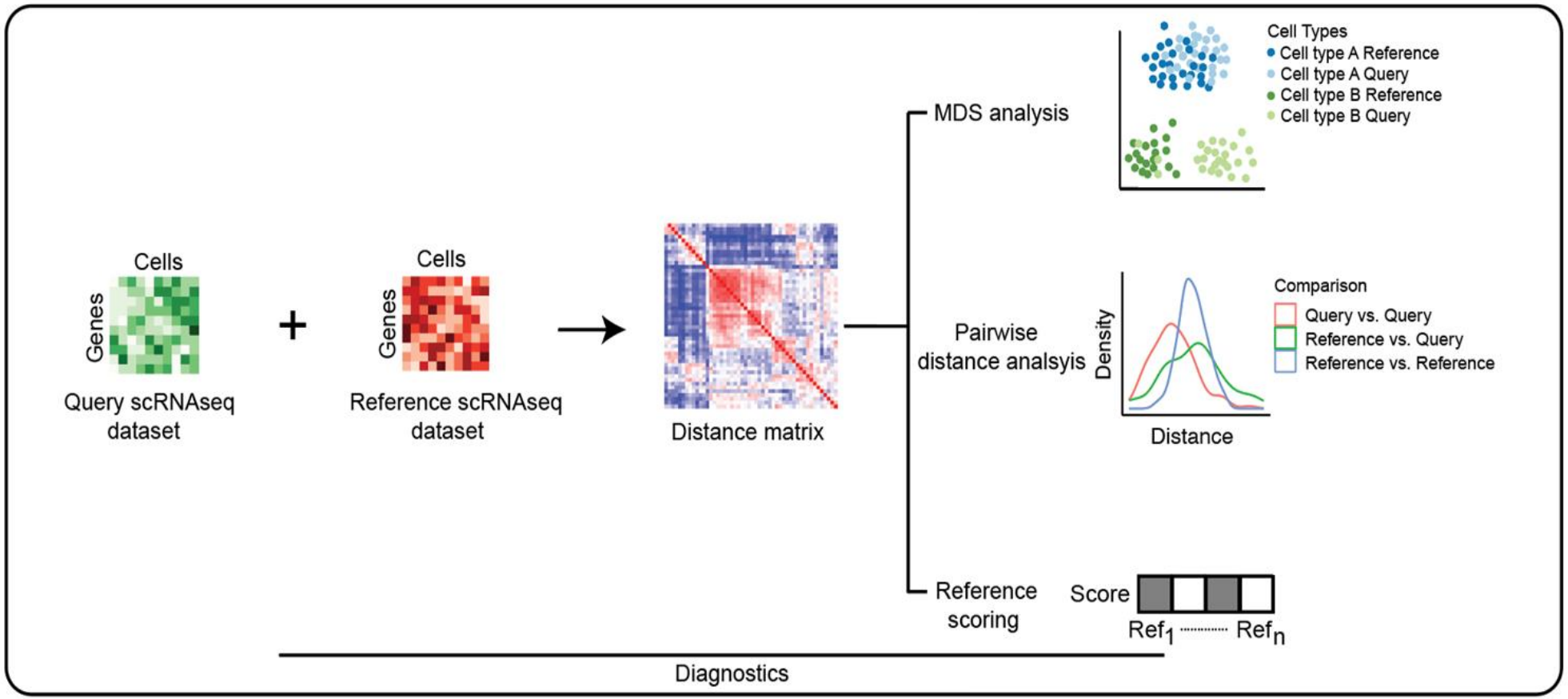
Introducing scDiagnostics

Enhancing the Standard scRNA-seq Cell Annotation Pipeline

- **Core Functionality:**
  - **Data Alignment Checking:** Ensures that the query dataset is well aligned with the reference dataset, verifying the consistency of gene expression patterns.
  - **Annotation Validation:** Confirms the accuracy of cell type annotations in the query dataset, using statistical and computational methods.
  - **Anomalous Cell Detection:** Identifies potentially anomalous cells, ensuring robust and accurate results.
- **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 Application

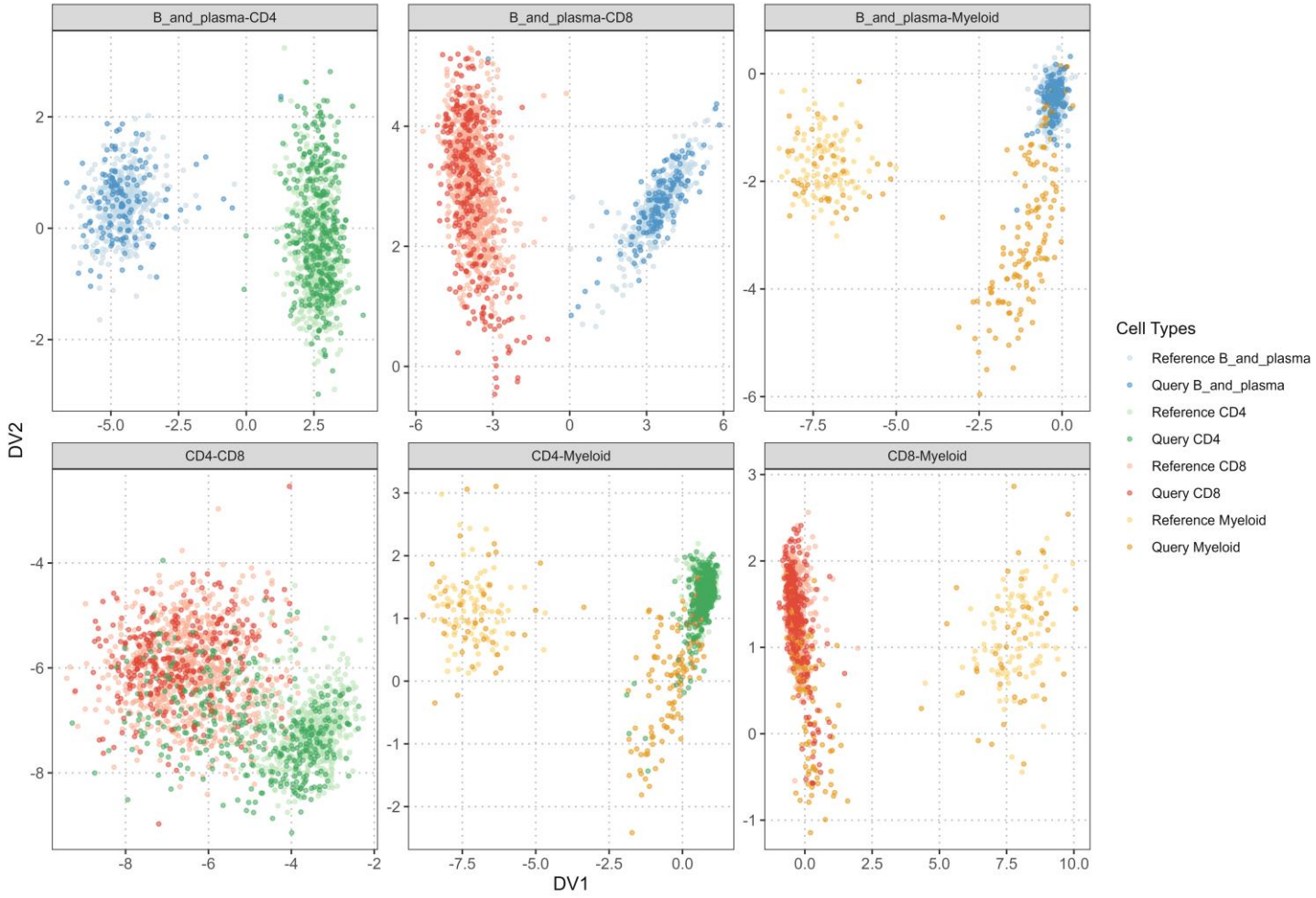
SingleR Annotation Diagnostics

- HeOrganAtlasData(tissue = c("Marrow")) Data in scRNAseq.
- Curation/processing of data (QC, normalization, etc.).
- Take 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

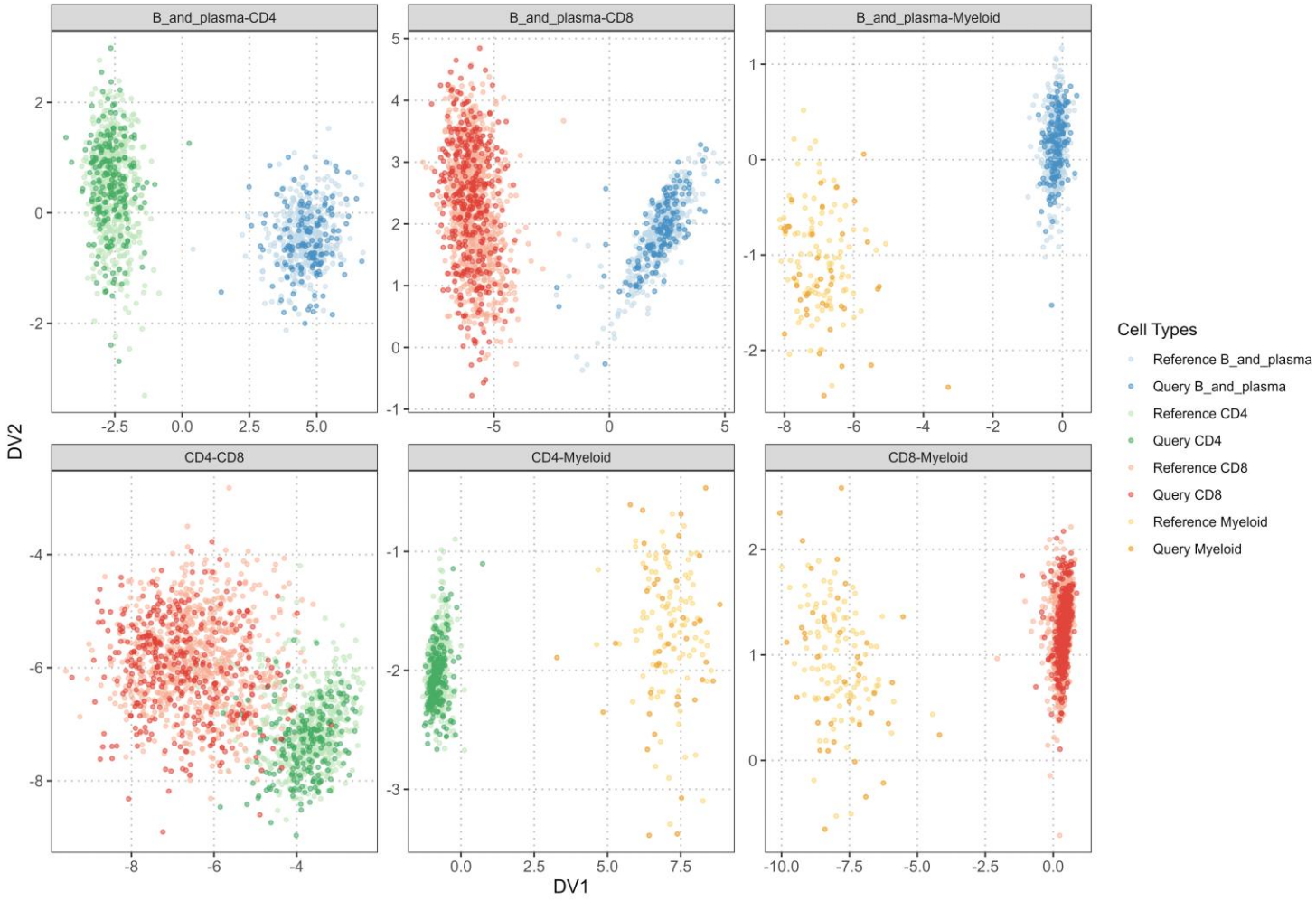
Visualization of SingleR Annotation

“Poor” Annotation Diagnostics



Visualization of Expert Annotation

“Good” Annotation Diagnostics



Workshop Materials

Material 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