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

Anthony Christidis,
Computational Scientist,
Core for Computational Biomedicine,
Harvard Medical School

Andrew Ghazi,
Statistical Geneticist,
Core for Computational Biomedicine,
Harvard Medical School

Ludwig Geistlinger,
Director of Computational Biology,
Core for Computational Biomedicine,
Harvard Medical School

Background and Motivation

Understanding the Importance of Accurate scRNA-seq Analysis

- **Introduction to scRNA-seq:**
  - Single-cell RNA sequencing (scRNA-seq) enables the study of gene expression at the individual cell level.
  - It provides high-resolution insights into cellular diversity, states, and functions.
- **Importance of Data Quality:**
  - Proper alignment and accurate cell annotation are critical for meaningful biological interpretations.
  - Errors in these steps can lead to incorrect conclusions and wasted resources.
- **Common Challenges:**
  - **Batch effects:** Variability introduced during sample processing and sequencing.
  - **Misannotation:** Incorrect identification of cell types due to overlapping gene expression profiles or manual errors.

Integration into Current Workflow

Enhancing the Standard scRNA-seq Analysis Pipeline

- **Current Workflow Steps:**
  - **Quality Control:** Identifies and removes low-quality cells and potential technical artifacts.
  - **Normalization:** Adjusts for differences in sequencing depth and other technical variations.
  - **Feature Selection:** Selects the most informative genes (features) for downstream analysis.
  - **Dimensionality Reduction:** Transforms into lower-dimensional space (e.g., PCA, t-SNE, UMAP).
  - **Clustering:** Groups cells into clusters to identify distinct cell populations.
  - **Marker Gene Detection:** Identifies genes that are uniquely or highly expressed in specific clusters.
  - **Cell Type Annotation:** Assigns labels to cell clusters based on known cell type signatures.
  - **(New Step) Annotation Diagnostics:** Ensure cell type annotation is accurate.

Context and Data Overview

Understanding the Reference and Query Datasets

- **Reference Data:**
  - **Description:** The reference dataset is 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 rigorous validation techniques.
  - **High-Quality Data:** This dataset serves as a gold standard for comparison due to its high quality and reliability.
  - **Usage:** Used to train models, identify cell types, and serve as a benchmark for new data.

```
# Load library
library(scDiagnostics)
# Load reference data (processed HeOrganAtlasData(tissue = c("Marrow")) dataset)
data("reference_data")
```

Context and Data Overview

Understanding the Reference and Query Datasets

- **Query Data:**
  - **Description:** The query dataset consists of new single-cell RNA-seq data that needs to be analyzed.
  - **Unknown Annotations:** Initial annotations are provided, but their accuracy has not been confirmed.
  - **Analysis Goals:**
    - **Alignment Check:** Ensure that the new data aligns well with the reference data.
    - **Annotation Validation:** Verify that the cell type annotations in the query data are accurate.
    - **Anomalous Cell Detection:** Identify any potentially anomalous cells that may indicate issues with data quality or annotation.

```
# Load query data
data(“query_data")
```

Context and Data Overview

Understanding the Reference and Query Datasets

```
# Compare expert and SingleR annotation

table(Expert_Annotation = query_data$expert_annotation,
      SingleR = query_data$SingleR_annotation)

                       SingleR

Expert_Annotation  B_and_plasma  CD4  CD8  Myeloid

   B_and_plasma             136    0    0       10

   CD4                        0  190    1       13

   CD8                        0  134  196       29

   Myeloid                    0    0    0       40
```

Integration into Current Workflow

Enhancing the Standard scRNA-seq Analysis Pipeline

- **New Step: Introducing scDiagnostics**
- **Fits into the workflow after cell type annotation.**
  - **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.

Context and Data Overview

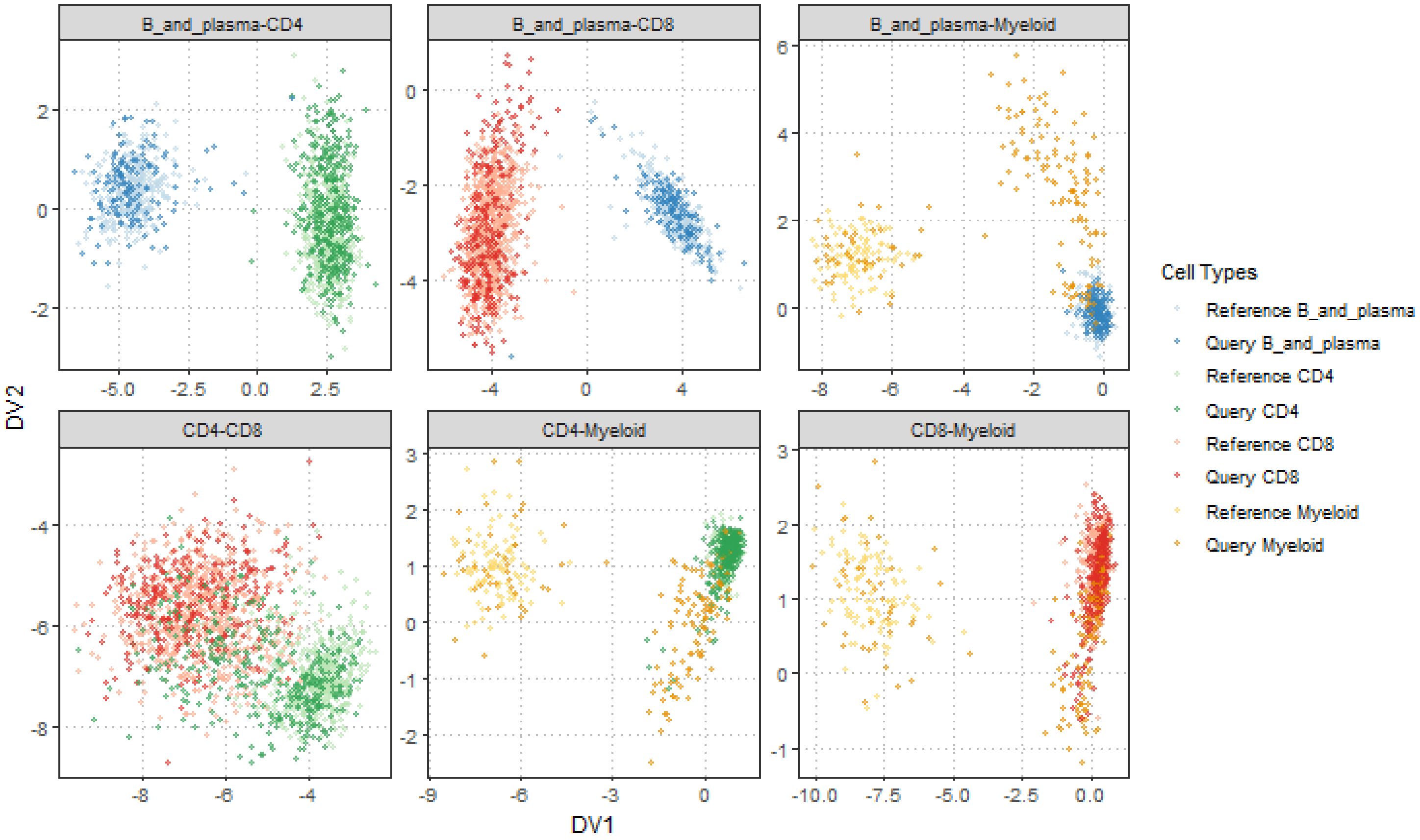
Understanding the Reference and Query Datasets

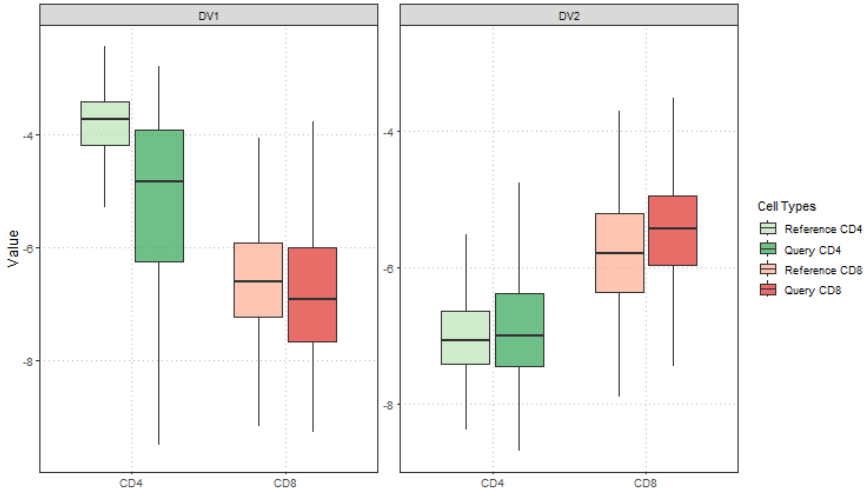
```
# Visualize cell types in (binary) discriminant spaces
disc_output <- calculateDiscriminantSpace(reference_data = reference_data,
                                          query_data = query_data,
                                          ref_cell_type_col = "expert_annotation“,
                                          query_cell_type_col = " SingleR_annotation ")


# Visualize output via scatterplot
plot(disc_output, plot_type = "scatterplot")
plot(disc_output, cell_types = "CD4-CD8", plot_type = "boxplot")
```

Context and Data Overview

Understanding the Reference and Query Datasets





Context and Data Overview

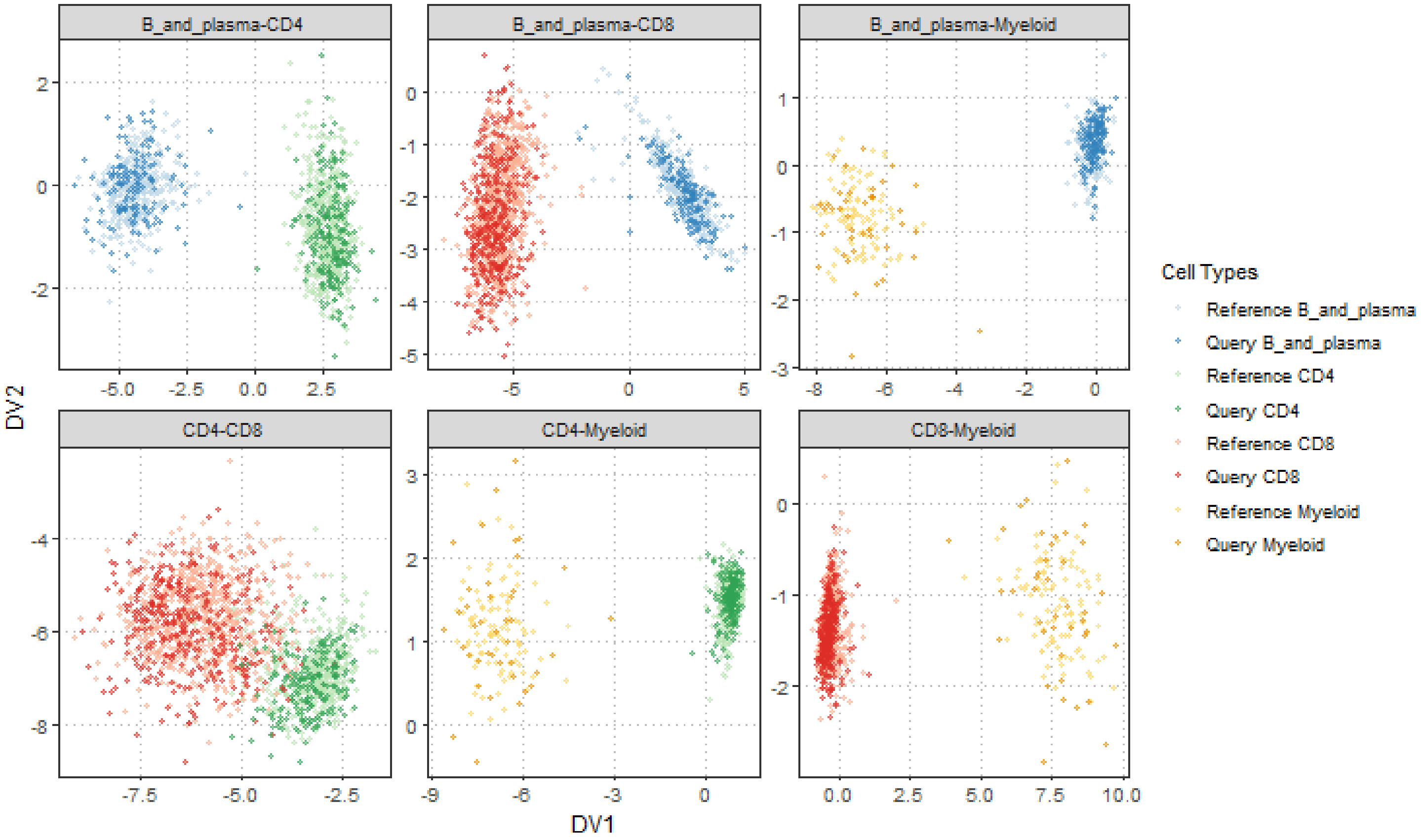
Understanding the Reference and Query Datasets

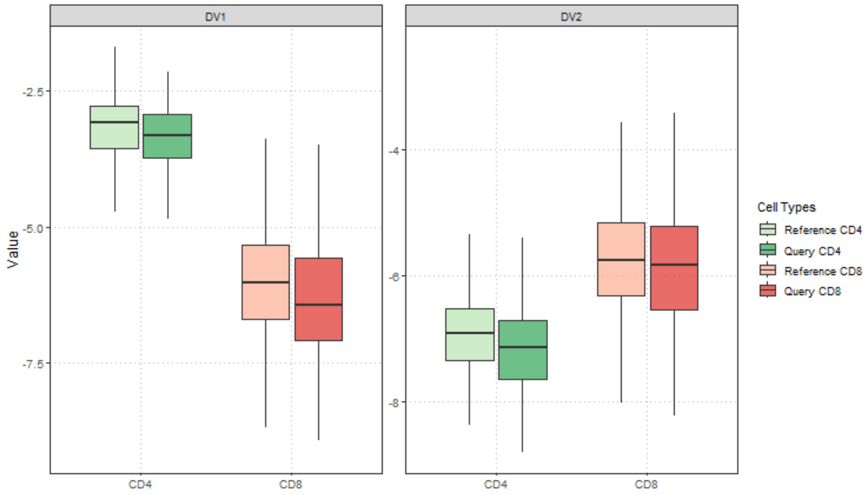
```
# Visualize cell types in (binary) discriminant spaces
disc_output <- calculateDiscriminantSpace(reference_data = reference_data,
                                          query_data = query_data,
                                          ref_cell_type_col = "expert_annotation“,
                                          query_cell_type_col = " expert_annotation ")


# Visualize output via scatterplot
plot(disc_output, plot_type = "scatterplot")
plot(disc_output, cell_types = "CD4-CD8", plot_type = "boxplot")
```

Context and Data Overview

Understanding the Reference and Query Datasets





Key Features of scDiagnostics

Enhancing Data Quality and Annotation Accuracy

- **Core Functionalities:**
  - **Alignment Checking:** Assess how well the query dataset matches the reference dataset in terms of gene expression patterns.
  - **Annotation Validation:** Use statistical and computational methods to confirm the accuracy of cell type labels.
  - **Anomalous Cell Detection:** Identify potentially anomalous cells at the global level (e.g., outliers) and cell-specific level (e.g., misclassified cells).
  - **Quality Control Measures:** Additional checks to ensure data integrity and reliability.
- **Innovative Aspects:**
  - Unique algorithms for alignment assessment.
  - Advanced validation techniques that go beyond simple cross-checks.
- **Relevance:**
  - Helps researchers maintain high standards in data analysis.
  - Facilitates accurate biological discoveries and insights.

Workshop Materials

Material Information and Links

- **Link to development repository:** https://github.com/ccb-hms/scDiagnostics
- **Link to pkgdown development website:** https://ccb-hms.github.io/scDiagnostics/index.html
- **Workshop Materials:** https://github.com/AnthonyChristidis/scDiagnosticsBioc2024Demo
  - Slides available here.
  - Vignette available here.
- **Galaxy workshop (with Docker container):** https://workshop.bioconductor.org/
- **Package will soon be available on Bioconductor.**