REPUBLIC OF TÜRKİYE
YILDIZ TECHNICAL UNIVERSITY
GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

HERACLOMICS: A USER FRIENDLY LOCAL SERVER APP FOR SCRNA SEQ ANALYSIS

Doğukan UĞUN

MASTER OF SCIENCE THESIS

Department of Bioengineering

Program of Bioengineering

Supervisor

Assoc. Prof. Dr. Alper YILMAZ

April, 2025

REPUBLIC OF TÜRKİYE

YILDIZ TECHNICAL UNIVERSITY

GRADUATE SCHOOL OF SCIENCE AND ENGINEERING

HERACLOMICS: A USER FRIENDLY LOCAL SERVER APP FOR SCRNA SEQ ANALYSIS

A thesis submitted by Doğukan UĞUN in partial fulfillment of the requirements iort he degree of **MASTER OF SCIENCE** is approved by the committee on 09/04/2025 in Department of Bioengineering, Program of Bioengineering.

Assoc. Prof. Dr. Alper YILMAZ
Yildiz Technical University
Supervisor

Approved By the Examining Committee

Assoc. Prof. Dr. Alper YILMAZ, Supervisor
Yildiz Technical University _____

Prof. Dr. Nehir ÖZDEMİR ÖZGENTÜRK, Member
Yildiz Technical University _____

Assoc. Prof. Dr. Mehmet BAYSAN, Member
Istanbul Technical University _____

I hereby declare that I have obtained the required legal permissions during data collection and exploitation procedures, that I have made the in-text citations and cited the references properly, that I haven't falsified and/or fabricated research data and results of the study and that I have abided by the principles of the scientific research and ethics during my Thesis Study under the title of "HERACLOMICS: A USER FRIENDLY LOCAL SERVER APP FOR SCRNA SEQ ANALYSIS" supervised by my supervisor, Assoc. Prof. Dr. Alper YILMAZ. In the case of a discovery of false statement, I am to acknowledge any legal consequence.

Doğukan UĞUN

Signature

Dedicated to my family

and my best friends

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Assoc. Prof. Dr. Alper Yılmaz, for his invaluable guidance, encouragement, and expertise throughout this journey. His insights and unwavering support have played a pivotal role in shaping the direction of this research. Additionally, I extend my appreciation to my family, friends, and colleagues who have provided constant motivation and support during the completion of this work.

I hope that HERACLOMICS will contribute to the scientific community by facilitating efficient and insightful single-cell RNA sequencing analyses and that this work will serve as a foundation for future innovations in the field.

Doğukan Uğun

TABLE OF CONTENTS

LIST OF ABBREVIATIONS vii

LIST OF FIGURES viii

ABSTRACT x

ÖZET xii

1 INTRODUCTION 1

1.1 Background of Single Cell RNA Sequencing 1

1.2 Technological Advancements in scRNA-seq 2

1.3 Application of scRNA-seq 5

1.4 Key Analytical Step of scRNA-seq 6

1.5 Challenges in scRNA-seq Analysis 7

1.6 Computational Tools, Seurat, and R Shiny in scRNA-seq Analysis 8

1.7 Objectives of the Heraclomics 10

2 MATERIAL & METHODS 12

2.1 Overview of the platform 12

2.2 Platform Development 13

2.2.1 Architecture and Modularity 14

2.2.2 Use of R Shiny for Interactivity 14

2.2.3 Scalibility and Usability 15

2.3 Key Functionalities & Analysis Step 15

2.3.1 Data Loading 16

2.3.2 Quality Control 17

2.3.3 Doublet Removal 18

2.3.4 Harmony Integration and Dimensionality Reduction 19

2.3.5 Clustering 20

2.3.6 Data Correction 21

2.3.7 Labelling Clusters 22

2.3.8 Gene Co-expression Network 23

2.3.9 Gene Regulatory Network 24

2.3.10. Trajectory Analysis 25

2.3.11. Cell Communication 26

2.3.12. Differential Expression 26
3 RESULTS **28**
3.1 Starting App and Analyze Sample Data 28
3.1.1 Loading Data 28
3.1.2 Quality Control 31
3.1.3 Doublet Removal 34
3.1.4 Harmony Integration and Dimensionality Reduction 38
3.1.5 Clustering 44
3.1.6 Data Correction 48
3.1.7 Cluster Labelling 51
3.1.8 Gene Co-expression Network 54
3.1.9 Cell Communication 62
3.1.10. Differential Expression 67
3.2 Validation and Benchmarking 67
4 CONCLUSION **70**
4.1 Conclusion 70
REFERENCES **72**
A APPENDIX CHAPTER HEADING **76**
PUBLICATIONS FROM THE THESIS **77**

LIST OF ABBREVIATIONS

AUCell	Single-cell regulon activity scoring method
CellChat	Cell-cell communication inference tool
DESeq2	Differential Expression analysis package
GO	Gene Ontology
GRN	Gene Regulatory Networks
GSEA	Gene Set Enrichment Analysis
HPCAData	Human Primary Cell Atlas
ICARUS	Interactive web server for single-cell RNA-seq analysis
ImmGen	Immunological Genome Project
KEGG	Kyoto Encyclopedia of Genes and Genomes
MEGENA	Multiscale Embedded Gene Co-Expression Network Analysis
Monocle3	Trajectory inference tool for single-cell RNA-seq
MSigDB	Molecular Signatures Database
PCA	Principal Component Analysis
RTN	Regulatory Transcriptional Network
SCENIC	Single-cell regulatory network inference and clustering
scRNA-seq	Single-cell RNA sequencing
TF	Transcription Factor
t-SNE	t-Distributed Stochastic Neighbor Embedding
UMAP	Uniform Manifold Approximation and Projection

LIST OF FIGURES

Figure 1.1 Single-cell and spatial omic technologies overview **4**

Figure 1.2 Scaling of scRNA-seq experiments cell numbers reported in representative publications by publication date **5**

Figure 2.1 Flow Chart of Heraclomics **13**

Figure 3.1 Homepage and interface of Heraclomics **28**

Figure 3.2 Data upload and sample naming **29**

Figure 3.3 Data preview of uploaded data **30**

Figure 3.4 Quality control parameters **31**

Figure 3.5 Quality control violin plots **32**

Figure 3.6 Quality control scatter plots **33**

Figure 3.7 Doublet removal interface **35**

Figure 3.8 UMAP before doublet removal **36**

Figure 3.9 UMAP after doublet removal **37**

Figure 3.10 Harmony integration and dimensionality reduction interface **39**

Figure 3.11 UMAP and t-SNE plots of batch (if availables) and samples **40**

Figure 3.12 t-SNE plots of selected gene colored by expression **41**

Figure 3.13 Heatmaps of selected principal components (can be selected single or multiple) **42**

Figure 3.14 Elbow plot and PCA loading plots (top 10 differentiated gene labelled) **43**

Figure 3.15 Clustering interface and parameters **45**

Figure 3.16 3D Umap visualization of clusters and plot types **46**

Figure 3.17 2D Umap visualization of clusters and cell composition summary plot **47**

Figure 3.18 Data correction interface and parameters selection step **49**

Figure 3.19 Data correction UMAP, violin plot for RPLP1 and ridge plot output **50**

Figure 3.20 Cluster labelling interface and database selection **52**

Figure 3.21 Cluster labelling cluster label result table and UMAP plot **53**

Figure 3.22 Gene co-expression networks MEGENA parameters **55**

Figure 3.23 Gene co-expression networks network plot for selected parameters **56**

Figure 3.24 Gene co-expression networks module summary table **57**

Figure 3.25 Gene co-expression networks gene-module membership table first page **58**

Figure 3.26 Gene co-expression networks individual module network and interactive module selection **59**

Figure 3.27 Gene co-expression networks functional enrichment analysis selected module output table **60**

Figure 3.28 Gene co-expression networks bar plot and bubble plot for enriched GO terms of selected module **61**

Figure 3.29 Cell-cell communication cluster and signaling pathways selection with interface **63**

Figure 3.30 Cell-cell communication network visualization and pathway visualization according to cell clusters **64**

Figure 3.31 Cell-cell communication heatmap of signalling pattern between cell and interaction buble plot of communication probabilities **65**

Figure 3.32 Cell-cell communication significant ligand-receptor interaction table with evidence references **66**

ABSTRACT

HERACLOMICS: A User Friendly Local Server App For SCRNA Seq Analysis

Doğukan UĞUN

Department of Bioengineering

Program of Bioengineering

Master of Science Thesis

Supervisor: Assoc. Prof. Dr. Alper YILMAZ

HERACLOMICS is a local server app designed to empower users without R programming skills to perform comprehensive single-cell RNA-seq analysis. Featuring a tutorial-style interface, HERACLOMICS guides users through each stage of preprocessing, analysis, and visualization, with customizable quality control thresholds, adjustable dimensionality reduction, and clustering parameters supported by 2D and 3D UMAP and t-SNE visualizations. Flexible differential expression analysis between user-defined cell groups, coupled with gene set enrichment analysis, aids in identifying biologically impacted pathways. Multimodal data visualization is available for CITE-seq and 10X Genomics Multiome data, while a save-state function preserves progress to avoid repeating computational steps. HERACLOMICS also provides advanced tools for further analysis, such as cluster labeling based on marker genes, gene expression visualization across clusters, and identification of gene co-expression networks to reveal functional relationships. Gene regulatory networks are explored through transcription factors and target gene interactions, while trajectory analysis uncovers differentiation pathways using RNA velocity and pseudotime. Cell communication is examined through inferred ligand-receptor interactions, and differential

expression analysis identifies genes unique to selected cell groups. These features make HERACLOMICS a powerful and accessible tool for single-cell RNA-seq analysis, enabling users to complete a full workflow within 1–2 hours.

Keywords: Single-cell RNA-seq Analysis, user-friendly bioinformatics platform

ÖZET

HERACLOMICS: SCRNA Seq Analizi İçin Kullanıcı Dostu Bir Yerel Sunucu Uygulaması

Doğukan UĞUN

Biyomühendislik Anabilim Dalı
Biyomühendislik Programı
Yüksek Lisans Tezi

Danışman: Doç. Dr. Alper YILMAZ

HERACLOMICS, R programlama becerileri olmayan kullanıcıların kapsamlı tek hücreli RNA-seq analizi yapmalarını sağlamak için tasarlanmış yerel bir sunucu uygulamasıdır. Eğitim tarzı bir arayüze sahip olan HERACLOMICS, özelleştirilebilir kalite kontrol eşikleri, ayarlanabilir boyut azaltma ve 2D ve 3D UMAP ve t-SNE görselleştirmeleri tarafından desteklenen kümeleme parametreleriyle, kullanıcıları ön işleme, analiz ve görselleştirmenin her aşamasında yönlendirir. Kullanıcı tanımlı hücre grupları arasındaki esnek diferansiyel ifade analizi, gen seti zenginleştirme analiziyle birleştiğinde, biyolojik olarak etkilenen yolların belirlenmesine yardımcı olur. CITE-seq ve 10X Genomics Multiome verileri için çok modlu veri görselleştirmesi mevcuttur; bir durum kaydetme işlevi ise hesaplamalı adımların tekrarlanmasını önlemek için ilerlemeyi korur. HERACLOMICS ayrıca, işaretleyici genlere dayalı küme etiketleme, kümeler arasında gen ifadesi görselleştirme ve işlevsel ilişkileri ortaya çıkarmak için gen eş-ifade ağlarının belirlenmesi gibi daha ileri analizler için gelişmiş araçlar da sağlar. Gen düzenleyici ağlar transkripsiyon faktörleri ve hedef gen etkileşimleri

aracılığıyla araştırılırken, yörünge analizi RNA hızı ve psödozaman kullanarak farklılaşma yollarını ortaya çıkarır. Hücre iletişimi çıkarılan ligand-reseptör etkileşimleri aracılığıyla incelenir ve farklı ifade analizi seçili hücre gruplarına özgü genleri belirler. Bu özellikler HERACLOMICS'i tek hücreli RNA-seq analizi için güçlü ve erişilebilir bir araç haline getirir ve kullanıcıların 1-2 saat içinde tam bir iş akışını tamamlamasını sağlar.

Anahtar Kelimeler: Tek hücreli RNA-seq analizi, kullanıcı dostu biyoinformatik platformu

1
INTRODUCTION

1.1 Background of Single Cell RNA Sequencing

Single-cell RNA sequencing (scRNA-seq) represents a groundbreaking advancement in molecular biology, allowing researchers to examine gene expression at the resolution of individual cells. This contrasts sharply with traditional bulk RNA sequencing, which provides only averaged gene expression across entire cell populations, obscuring cellular heterogeneity. The advent of scRNA-seq has revolutionized our ability to identify rare cell types, study cellular differentiation, and investigate dynamic biological processes, making it an indispensable tool in modern biological and medical research (Lorenzi, 2021; Vafadarnejad, 2021; Sapub, 2024). This technology enables the dissection of complex tissues, revealing insights into developmental biology, cancer progression, and immune responses that were previously unattainable with conventional methods (Lorenzi, 2021). By providing a comprehensive view of cellular diversity, scRNA-seq has opened new avenues for personalized medicine, allowing for tailored therapeutic approaches based on the specific cellular makeup of individual patients (Vafadarnejad, 2021). RNA sequencing (RNA-seq), the precursor to scRNA-seq, emerged in the early 2000s as a revolutionary method for analyzing gene expression across entire populations of cells (Wang, Gerstein, & Snyder, 2009). Unlike microarrays, which require predefined probes, RNA-seq employs high-throughput sequencing to capture the transcriptome in an unbiased manner, enabling the detection of novel transcripts, splice variants, and gene fusions. Early applications of RNA-seq provided insights into transcriptional regulation, revealing complexities in eukaryotic gene expression that had been previously inaccessible. However, bulk RNA-seq's inability to resolve cell-to-cell variability posed a

significant limitation, particularly in heterogeneous tissues such as tumors or developing embryos (Tang et al., 2009). The first single-cell transcriptomic studies, conducted in the late 2000s, demonstrated the feasibility of isolating and sequencing RNA from individual cells (Tang et al., 2009). These pioneering efforts paved the way for technological advancements, such as microfluidics and droplet-based platforms, that have since propelled scRNA-seq into mainstream research. Over the past decade, scRNA-seq has undergone rapid evolution, transitioning from a niche technique to a widely accessible tool that is now integral to fields ranging from developmental biology to precision medicine (Kolodziejczyk et al., 2015). This progression has been driven by improvements in sensitivity, scalability, and computational methods, enabling researchers to analyze complex tissues at unprecedented resolution. The demand for comprehensive scRNA-seq analysis platforms arises from the complexity of the datasets generated, which are often vast, sparse, and multidimensional. Despite the existence of robust computational frameworks like Seurat and Scanpy, their reliance on programming skills limits accessibility for many biologists and clinicians. Bridging this gap requires the development of user-friendly tools that integrate advanced computational methods into intuitive interfaces, fostering broader adoption and application of scRNA-seq technology (Lorenzi, 2021; Sapub, 2024).

.

1.2 Technological Advancements in scRNA-seq

The rapid evolution of single-cell RNA sequencing (scRNA-seq) technologies has been instrumental in expanding the scope of cellular transcriptomic analysis. Among the pioneering methods, 10x Genomics Chromium and Drop-Seq leverage microfluidic systems to encapsulate single cells in droplets with barcoded beads, enabling high-throughput sequencing of thousands of cells in a single experiment (Macosko et al., 2015). These methods are particularly suited for large-scale studies, offering scalability and cost efficiency. In contrast, Smart-Seq2 focuses on full-length transcript coverage, making it ideal for detecting splice variants and lowly expressed genes (Picelli et al., 2013). Other innovations like Seq-Well and InDrop provide cost-effective solutions, particularly for laboratories with limited resources (Zilionis et al., 2017). Technologies like Fluidigm C1, which utilize

microfluidic chips for single-cell capture and library preparation, offer a high level of precision, albeit with lower throughput compared to droplet-based methods. These advancements have enabled researchers to address diverse biological questions, from characterizing rare cell populations to mapping differentiation trajectories. Each technology brings unique advantages, allowing researchers to tailor their approach based on the specific requirements of their studies. The development of barcoding techniques has also significantly enhanced scRNA-seq workflows. By assigning unique molecular identifiers (UMIs) to individual RNA molecules, these methods minimize amplification biases and improve quantification accuracy (Islam et al., 2014). Together, these innovations have propelled scRNA-seq into mainstream research, empowering scientists to explore cellular heterogeneity and uncover biological insights at an unprecedented scale. There are Imaging technologies other than Next generation sequencing technologies, NGS technologies in RNA sequencing categorizes technologies based on single-cell and spatial omics approaches, including plate-based, microwell-based and droplet-based systems where Imaging technologies includes Oligo-based probes and Epitope-based probes technologies as illustrated in Figure 1.1.

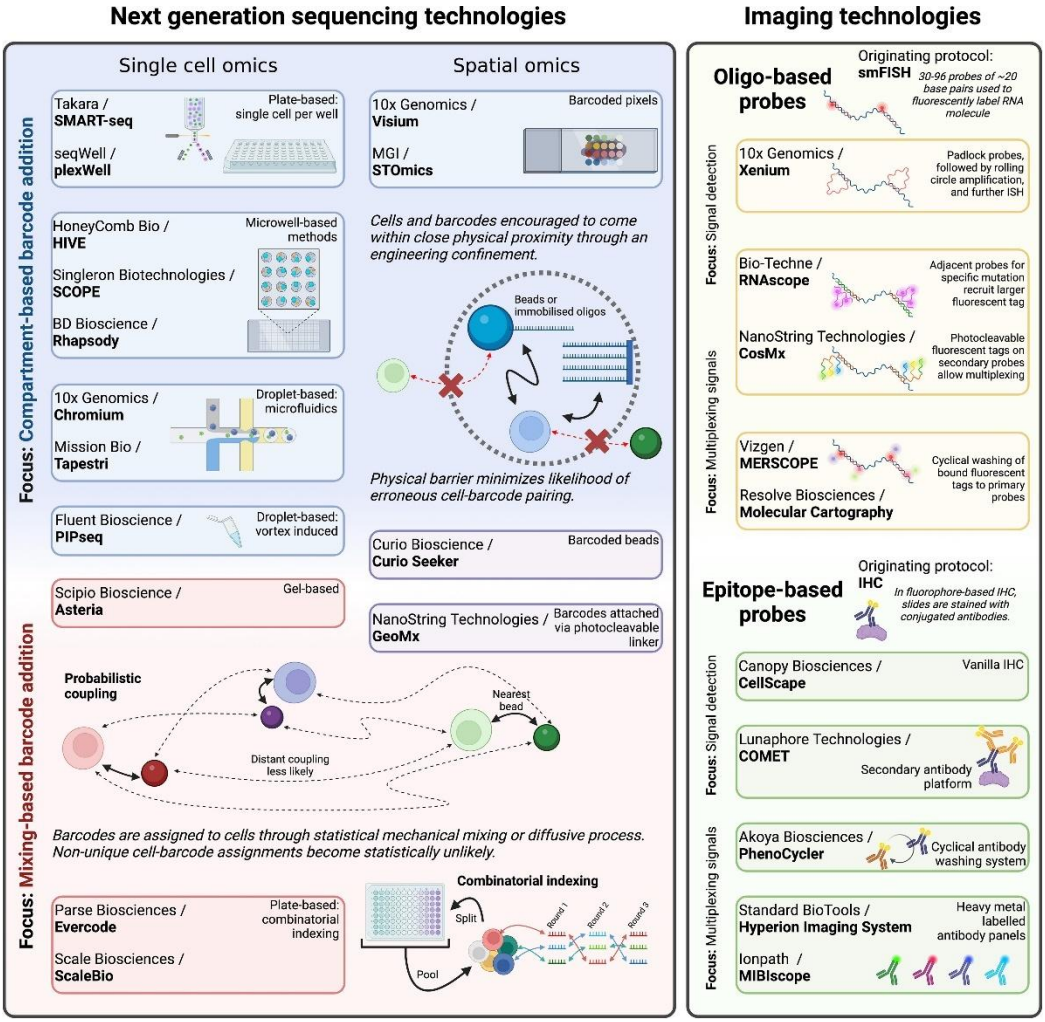


Figure 1.1 Single-cell and spatial omic technologies overview

Recent developments have significantly altered the number of cells profiled per research in addition to improving sensitivity and accuracy as shown in Figure1.2 (Svensson, Vento-Tormo, and Teichmann 2018). These days, it is possible to profile hundreds of thousands of individual cells simultaneously in a single study, which improves our capacity to identify various cell types and the signatures that correlate to them.



Figure 1.2 Scaling of scRNA-seq experiments cell numbers reported in representative publications by publication date

1.3 Application of scRNA-seq

The versatility of single-cell RNA sequencing (scRNA-seq) has made it an indispensable tool across diverse research domains. In developmental biology, scRNA-seq enables researchers to map cellular differentiation trajectories, uncovering how stem cells develop into specialized cell types. This has provided critical insights into embryogenesis and organ development (Guo et al., 2010). In cancer research, this technology is instrumental in identifying tumor heterogeneity, revealing rare cancer subpopulations, and uncovering potential therapeutic targets (Tirosh et al., 2016). scRNA-seq also allows for the study of the tumor microenvironment, providing a deeper understanding of how cancer cells interact with their surroundings. In immunology, scRNA-seq is used to profile immune cell populations during infections, vaccinations, and autoimmune disorders. It has been pivotal in identifying novel immune cell subsets and understanding immune responses at the molecular level (Reyes et al., 2020). Neuroscientists utilize scRNA-seq to explore the diversity of neuronal populations, shedding light on brain function and diseases such as Alzheimer's and Parkinson's (Lake et al., 2016). This technology is equally valuable in regenerative medicine, where it aids in characterizing stem cell populations and evaluating their potential for tissue repair and therapeutic applications (Trapnell, 2015). Moreover, scRNA-seq has found applications in infectious disease research, enabling the study of host-pathogen interactions at the single-cell level. This has been particularly useful in understanding viral infections such as SARS-CoV-2 and their impact on specific

cell types. As scRNA-seq continues to evolve, its applications are expected to expand further, driving innovations across biological and medical sciences.

1.4 Key Analytical Step of scRNA-seq

Single-cell RNA sequencing (scRNA-seq) involves a series of analytical steps that ensure high-quality data and meaningful biological insights. These key steps include:

- Quality Control: Quality control is the first and most crucial step to filter out low-quality cells and genes. Common metrics used include the number of detected genes, total gene expression counts, and the percentage of mitochondrial gene expression. Cells with abnormal values for these metrics are excluded to ensure data accuracy.
- Normalization: Normalization adjusts for differences in sequencing depth and other technical biases, enabling fair comparisons across cells. Methods such as log-normalization or scaling by total counts are commonly employed.
- Dimensionality Reduction: To make high-dimensional data interpretable, dimensionality reduction techniques like Principal Component Analysis (PCA), t-SNE, or UMAP are applied. These techniques summarize data while preserving key relationships between cells.
- Clustering: Clustering groups cells into subpopulations based on their gene expression profiles. Algorithms such as Louvain or Leiden are widely used to identify distinct cell populations.
- Marker Gene Identification: Marker genes are identified for each cluster to characterize cell types. Differential gene expression analysis highlights genes that are uniquely expressed in each group.
- Integration: When multiple datasets are involved, batch effect correction and integration methods like Harmony or Seurat's CCA are applied to combine datasets into a unified framework.

- Trajectory Inference: In systems with dynamic processes, trajectory inference algorithms like Monocle or Slingshot map cells along developmental or temporal trajectories, revealing transitions between cell states.
- Visualization: Visualization tools, such as t-SNE and UMAP plots, are used to display the clustering and relationships among cells, providing intuitive insights into the data structure.

These analytical steps are nearly a must for scRNA-seq workflows, enabling researchers to derive insights into cellular heterogeneity, developmental pathways, and disease mechanisms.

1.5 Challenges in scRNA-seq Analysis

Despite its transformative potential, single-cell RNA sequencing (scRNA-seq) presents several analytical and technical challenges:

Data Sparsity: scRNA-seq data is inherently sparse, with many genes exhibiting zero expression in individual cells. This sparsity complicates downstream analysis and can lead to the loss of important biological signals (Svensson et al., 2017).

Batch Effects: Variability introduced by differences in experimental conditions, such as sample processing or sequencing platforms, can obscure true biological variation. Batch effect correction methods like Harmony and Seurat's integration pipelines are essential but add computational complexity (Haghverdi et al., 2018).

High Dimensionality: The large number of genes profiled per cell creates high-dimensional datasets that are challenging to visualize and interpret. Dimensionality reduction techniques, while effective, may oversimplify complex relationships (Becht et al., 2019).

Computational Demands: The size and complexity of scRNA-seq datasets require substantial computational resources, particularly for tasks like clustering, trajectory inference, and batch correction. This can be a barrier for smaller research groups with limited access to high-performance computing.

Cellular Annotation: Assigning cell types to clusters based on marker genes is often subjective and relies on prior knowledge, which may not always be available.

Automated tools like SingleR and CellTypist offer solutions but require careful validation.

Dropout Events: Dropout events, where genes are undetected due to low expression levels or technical noise, are prevalent in scRNA-seq data. This can bias downstream analyses and affect the reliability of inferred biological conclusions (Kharchenko et al., 2014).

Data Integration: Combining datasets from different studies or platforms introduces additional challenges, particularly when dealing with diverse cell populations or experimental conditions. Proper integration is crucial to ensure accurate biological interpretation (Stuart et al., 2019).

Ethical and Privacy Concerns: When analyzing patient-derived samples, ethical considerations and data privacy regulations must be addressed, particularly in the context of sharing and publishing data.

Addressing these challenges requires a combination of advanced computational methods, robust experimental designs, and collaborative efforts between biologists and bioinformaticians to maximize the utility of scRNA-seq data.

1.6 Computational Tools, Seurat, and R Shiny in scRNA-seq Analysis

The effective analysis of single-cell RNA sequencing (scRNA-seq) data relies on a suite of computational tools and frameworks, each designed to handle specific aspects of the workflow, from raw data preprocessing to visualization and interpretation.

Several computational frameworks have been developed to streamline scRNA-seq data analysis. Prominent among these are:

Seurat: One of the most widely used R-based frameworks, Seurat provides tools for normalization, dimensionality reduction, clustering, and integration of multiple datasets. Its intuitive functions and visualization capabilities make it a go-to choice for many researchers (Stuart et al., 2019).

Scanpy: A Python-based framework that offers similar functionalities to Seurat. It is particularly suited for large datasets and integrates well with other Python-based libraries.

Monocle: Focused on trajectory inference, Monocle is ideal for mapping cellular differentiation and dynamic processes.

Harmony: A batch correction tool that integrates with existing workflows to mitigate batch effects in scRNA-seq data.

SCENIC: Designed for inferring gene regulatory networks, SCENIC identifies transcription factors and their target genes within single cells.

Seurat has emerged as a cornerstone in the analysis of scRNA-seq data. Developed in R, Seurat provides a comprehensive suite of tools that cover the entire scRNA-seq pipeline. Key features include:

Data Integration: Using canonical correlation analysis (CCA) and mutual nearest neighbors (MNN), Seurat excels in integrating datasets from different experimental conditions.

Dimensionality Reduction: Seurat implements PCA, UMAP, and t-SNE for high-dimensional data visualization.

Clustering: The Louvain and Leiden algorithms are integrated into Seurat, enabling robust cell clustering.

Differential Expression Analysis: Seurat identifies marker genes that define cell clusters, aiding in cell type annotation.

Visualization: The framework supports a variety of plots, including feature plots, violin plots, and heatmaps, to facilitate data interpretation.

Seurat's extensibility and active community support have contributed to its popularity, making it an essential tool for scRNA-seq analysis.

While computational frameworks like Seurat provide robust functionalities, their reliance on programming expertise can be a barrier for many biologists. This is where R Shiny plays a transformative role.

R Shiny is an R package that allows developers to build interactive web applications for data analysis and visualization. Its relevance in scRNA-seq analysis lies in its ability to:

Streamline Workflows: By integrating tools like Seurat into a user-friendly interface, R Shiny democratizes access to advanced analytical methods.

Enhance Visualization: Interactive plots generated using R Shiny enable researchers to explore data dynamically, identifying patterns and relationships more effectively.

Facilitate Collaboration: Web-based applications allow teams to collaborate remotely, sharing insights and annotations in real time.

Enable Customization: Applications can be tailored to specific research needs, incorporating modules for quality control, clustering, trajectory analysis, and more.

In the context of this project, R Shiny serves as the backbone for Heraclomics, providing an intuitive interface for biologists and clinicians to interact with complex scRNA-seq data. By bridging the gap between computational power and usability, R Shiny ensures that the insights derived from scRNA-seq data are accessible to a broader audience.

1.7 Objectives of the Heraclomics

The primary objective of this thesis is to address the challenges faced in single-cell RNA sequencing (scRNA-seq) analysis by developing Heraclomics, an interactive web-based platform. This project aims to bridge the gap between computational complexity and accessibility, enabling researchers to conduct advanced analyses without requiring extensive programming expertise. Heraclomics integrates established computational frameworks, such as Seurat, into an intuitive R Shiny interface, ensuring usability for a wide range of users. The thesis begins by evaluating the current state-of-the-art technologies and methodologies in scRNA-seq analysis, identifying limitations in existing frameworks. This evaluation informs the development of Heraclomics, which incorporates critical functionalities

like quality control, clustering, trajectory inference, and differential expression analysis. A key goal is to provide a seamless and user-friendly interface that simplifies scRNA-seq workflows and empowers biologists to derive meaningful insights efficiently. Validation of Heraclomics is conducted through case studies using real-world datasets. These studies benchmark the platform's performance, comparing it to existing tools in terms of accuracy, usability, and efficiency. Overall, the thesis aspires to democratize scRNA-seq analysis, making it accessible to researchers from diverse backgrounds and enabling breakthroughs in biological discovery through innovative computational solutions.

2
MATERIAL & METHODS

2.1 Overview of the platform

Heraclomics is an interactive web-based platform specifically designed to streamline single-cell RNA sequencing (scRNA-seq) analysis. Built on the R Shiny framework, Heraclomics combines the power of established computational tools, such as Seurat, with an intuitive user interface to provide researchers with a comprehensive solution for analyzing scRNA-seq data. By integrating robust analytical frameworks with dynamic visualizations, Heraclomics bridges the gap between computational complexity and accessibility, catering to both computational biologists and experimental scientists. The platform is built with modularity at its core, allowing users to progress sequentially through key steps in the scRNA-seq workflow. Each module in Heraclomics is designed to address a specific stage of the analysis pipeline as can seen in Figure2.1, including quality control, dimensionality reduction, clustering, differential expression analysis, and trajectory inference. The modular architecture ensures that users can perform complete analyses or focus on specific components as needed, offering flexibility and scalability for datasets of varying sizes and complexities. A distinguishing feature of Heraclomics is its ability to integrate datasets from different experimental conditions or sequencing platforms. Through advanced methods such as Harmony and Seurat's integration tools, the platform effectively corrects for batch effects, enabling seamless comparison and integration of datasets. Additionally, Heraclomics supports multiple input file formats, including 10Xdata, .rds, and .txt count matrix, ensuring compatibility with diverse data sources. Interactive visualizations play a pivotal role in Heraclomics. By leveraging libraries such as Plotly and ggplot2, the platform provides dynamic plots for quality control,

clustering, and trajectory analysis. These visualizations enable users to intuitively explore their data, identify patterns, and derive meaningful biological insights. Moreover, Heraclomics incorporates customizable parameters at every step, empowering users to tailor the analysis to their specific research needs. Designed with usability in mind, Heraclomics eliminates the steep learning curve often associated with computational tools. Its user-friendly interface ensures that researchers without programming expertise can navigate and utilize advanced scRNA-seq analytical methods.

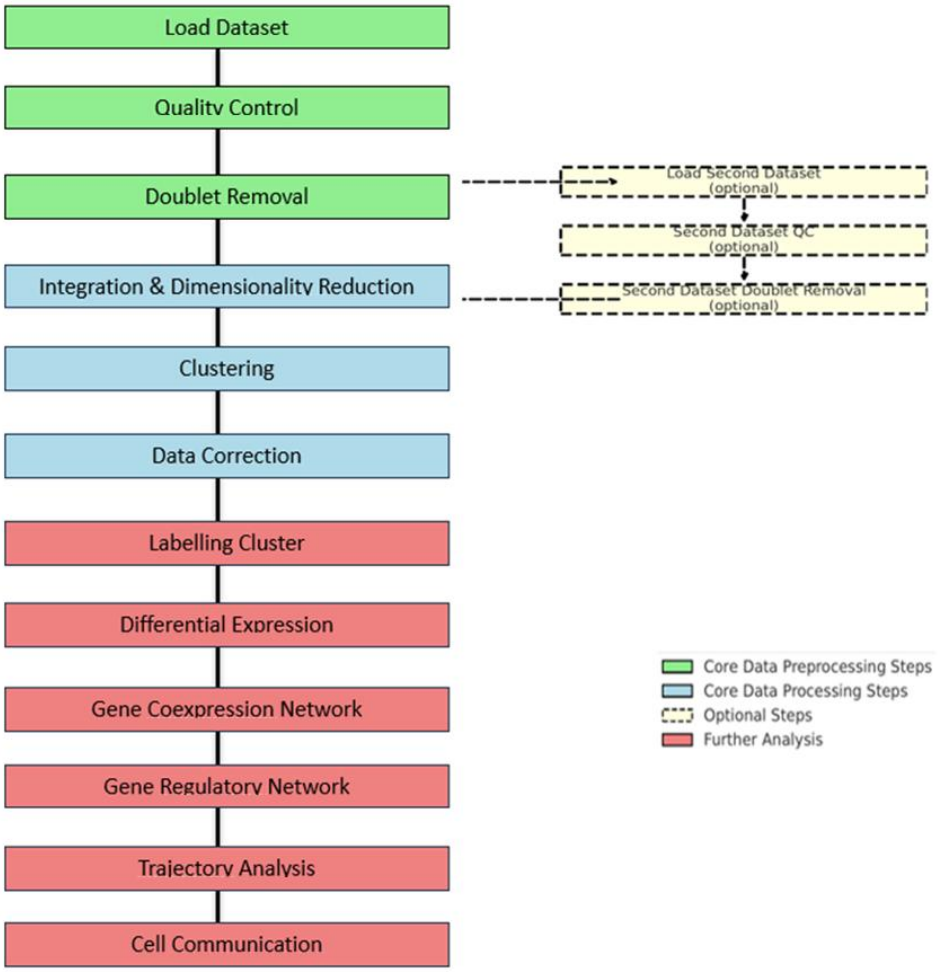


Figure 2.1 Flow Chart of Heraclomics

2.2 Platform Development

The development of Heraclomics is anchored in a modular architecture powered by R Shiny, a framework that enables the creation of interactive web applications directly from R. The platform's modular design is not only crucial for scalability

but also ensures that users can engage with each step of the scRNA-seq analysis pipeline independently or as part of a cohesive workflow. Each module corresponds to a specific stage in the pipeline, such as data loading, quality control, clustering, or trajectory inference, and is interconnected to allow seamless transitions between analytical tasks.

2.2.1 Architecture and Modularity

The architecture of Heraclomics follows a layered approach to ensure robustness and flexibility. The backend is built using Seurat and other R libraries for data handling, computation, and integration (Stuart et al., 2019; Korsunsky et al., 2019). The frontend, developed in R Shiny, focuses on creating a user-friendly interface that abstracts the complexity of these computations. By adopting a modular structure, Heraclomics allows for the integration of additional functionalities without disrupting existing workflows. Each module is self-contained, featuring its own inputs, outputs, and dependencies, which enhances maintainability and adaptability to new analytical methods or user requirements. For instance, the quality control module offers customizable thresholds for filtering cells and genes, while the clustering module enables dynamic adjustments to resolution parameters. This modularity ensures that Heraclomics can cater to a wide spectrum of users, from novice researchers requiring default workflows to advanced users seeking custom analyses (Korsunsky et al., 2019).

2.2.2 Use of R Shiny for Interactivity

R Shiny is pivotal in providing Heraclomics with its interactive capabilities. By leveraging Shiny's reactive programming model, the platform automatically updates outputs and visualizations in response to user inputs. This interactivity enhances the user experience by offering real-time feedback and dynamic adjustments (Chang et al., 2021). For example, users can modify clustering parameters and instantly visualize the impact on t-SNE or UMAP plots. Similarly, quality control filters can be fine-tuned interactively, with immediate updates to the displayed metrics and plots (Dean & Mahoney, 2018). Shiny's seamless integration with Plotly and ggplot2 enables the generation of interactive plots that allow users to zoom, pan, and select data points for deeper exploration (Sievert, 2020). This interactive fosters a more engaging analysis process, empowering users to identify

trends and anomalies efficiently. Additionally, the use of Shiny modules facilitates the compartmentalization of functionality, ensuring that each analytical step remains independent yet interoperable with the overall platform.

2.2.3 Scalibility and Usability

The design of Heraclomics prioritizes scalability and usability. The modular framework ensures that the platform can handle datasets of varying sizes, from small pilot studies to large-scale multi-sample experiments. Its user-friendly interface minimizes the learning curve, enabling researchers with limited computational expertise to perform advanced analyses. Furthermore, the platform's architecture allows for the addition of new tools and methods, ensuring that it remains adaptable to the evolving needs of the scRNA-seq community. By combining modularity, interactivity, and scalability, Heraclomics provides a powerful yet accessible solution for single-cell RNA sequencing analysis. Its development reflects a commitment to making advanced computational tools more inclusive, enabling researchers across disciplines to unlock the full potential of their data.

2.3 Key Functionalities & Analysis Step

Heraclomics implements a comprehensive and modular analysis workflow designed to handle the complexities of single-cell RNA sequencing (scRNA-seq) data. The workflow begins with an introduction, which provides users with an overview of the platform's capabilities and navigation guidelines. Following this, the data loading step enables users to upload datasets in various formats, preview data structures, and validate file integrity to ensure compatibility. Quality control measures are then applied to filter datasets based on metrics like the number of detected genes and mitochondrial content, ensuring high-quality data for downstream analysis. Once the primary dataset is preprocessed, the platform offers a doublet removal step to identify and exclude artificial cell doublets, enhancing the accuracy of subsequent analyses. For studies requiring comparative or integrative analysis, users can load a second dataset and perform similar quality control and doublet removal steps. Integration and dimensionality reduction are then applied, allowing for batch effect correction and visualization of high-dimensional data using techniques like PCA, t-SNE, or UMAP. Clustering modules group cells into

subpopulations based on gene expression profiles using algorithms like SLM, Louvain or Leiden. Data correction modules address confounding factors such as cell cycle effects, further improving cluster resolution. Advanced analyses follow, including trajectory inference to map dynamic processes, differential expression to identify key genes, and cluster labeling to assign biologically meaningful names to identified subpopulations. Further steps include constructing gene co-expression networks to explore gene-gene interactions, inferring gene regulatory networks to understand transcription factor activity, and analyzing intercellular communication through ligand-receptor interactions. This modular workflow ensures that users can tailor the analysis to their specific research questions while benefiting from a robust and flexible platform.

2.3.1 Data Loading

The data loading step in Heraclomics provides a comprehensive, user-friendly solution for uploading and preparing single-cell RNA sequencing (scRNA-seq) datasets, streamlining the workflow for users. This module supports multiple file formats, including 10X matrix, .txt (count matrix), and .rds Seurat objects, ensuring flexibility and compatibility across various data sources. By leveraging advanced tools like Seurat for data preprocessing (Stuart et al., 2019) and the DT (Data Tables) library for efficient file manipulation (Dowle & Srinivasan, 2021), Heraclomics delivers robust performance even with large datasets. The R Shiny framework enables an interactive interface, allowing users to easily upload datasets, preview their structure, and validate file integrity through automated checks (Chang et al., 2021). Key features include a summary view of the uploaded dataset, which provides essential metrics such as cell counts, gene counts, facilitating immediate insights into the data's structure and content. The module also incorporates automated format conversion, ensuring that data is prepared in formats compatible with downstream modules as Seurat objects. In addition, the system is equipped with robust error-handling mechanisms that detect common issues, such as missing metadata, duplicated variables, or incompatible file formats, and provide users with clear guidance on how to address them. Supporting tools and libraries include Seurat for handling raw data, enabling smooth integration into the analysis pipeline (Stuart et al., 2019), and DT, which ensures efficient reading and manipulation of large files (Dowle & Srinivasan, 2021). The interactive features powered by Shiny

make it possible for users to interactively validate datasets, navigate through the module with ease, and resolve errors before proceeding to the next steps in the analysis pipeline (Chang et al., 2021).

2.3.2 Quality Control

The quality control step in Heraclomics is critical for ensuring the integrity and reliability of single-cell RNA sequencing (scRNA-seq) data. This module utilizes a suite of libraries to identify and filter out low-quality cells and genes, minimizing noise and maximizing the biological relevance of the dataset. Quality control metrics include the number of detected genes per cell, total gene expression counts, and the percentage of mitochondrial and ribosomal RNA content. Cells that fall outside user-defined thresholds for these metrics are flagged and excluded, thereby preventing biases in downstream analysis. Heraclomics employs the Seurat library (Stuart et al., 2019) to implement quality control measures, including filtering and statistical evaluation of cells and genes. Visualization libraries such as ggplot2 (Wickham, 2016) and Plotly (Sievert, 2020) are used to generate violin plots, scatter plots, and other interactive visualizations, allowing users to inspect quality metrics and adjust thresholds dynamically. The Shiny framework (Chang et al., 2021), in conjunction with additional libraries like shinyjs, shinyWidgets, and shinyalert, provides real-time feedback and interactivity, enabling users to fine-tune filtering parameters and immediately observe the impact on the dataset. The cowplot (Wilke, 2024) library is integrated to produce composite figures, offering users an overview of multiple quality metrics in a single visualization. HTMLwidgets are employed to enhance interactivity, enabling zooming, panning, and selection of specific data points for detailed inspection.

This comprehensive suite of tools makes the quality control module a robust and user-friendly solution. By integrating statistical analysis with interactive visualization, Heraclomics empowers researchers to refine their data, ensuring that only high-quality cells and genes are used in subsequent analyses. The module leverages the Seurat framework for implementing these filters, while visualization libraries like ggplot2 and Plotly are used to generate interactive plots. These plots, including violin plots and scatter plots, provide an intuitive way for users to inspect the distribution of quality metrics and adjust thresholds interactively. The R Shiny interface ensures that users can dynamically refine filtering criteria and

immediately see the impact of their adjustments on the dataset. By integrating visualization with real-time analysis, Heraclomics empowers researchers to perform high-quality preprocessing with confidence.

2.3.3 Doublet Removal

The doublet removal step in Heraclomics is designed to identify and exclude artificial cell doublets, which occur when two cells are encapsulated together during droplet-based sequencing. Doublets can introduce significant noise and confound downstream analyses, making their detection an important, albeit optional, preprocessing step. Users have the flexibility to skip this step if doublet identification is not deemed necessary for their analysis. Heraclomics employs the DoubletFinder package (McGinnis et al., 2019), integrated with the Seurat framework, to predict and flag potential doublets. This is achieved by simulating artificial doublets and comparing them against real cell profiles. Users can fine-tune key parameters such as the expected doublet rate and neighborhood size to suit the specifics of their dataset and experimental design. To aid visualization, the module leverages t-SNE and UMAP plots to display doublets and singlets interactively. Cells flagged as doublets are color-coded, allowing users to visually inspect their distribution and decide whether to exclude them. ggplot2 (Wickham, 2016) and Plotly (Sievert, 2020) are used to generate these visualizations, providing a clear and intuitive interface for exploring potential doublets. The R Shiny framework (Chang et al., 2021) enhances interactivity by enabling users to dynamically adjust parameters and immediately observe the effect on flagged doublets. This real-time feedback fosters a more informed decision-making process. While doublet removal is not mandatory, it is highly recommended for datasets prone to this artifact, as it ensures a cleaner dataset for subsequent steps like clustering and differential expression analysis. By integrating statistical rigor, advanced visualization, and user-friendly interactivity, the doublet removal step in Heraclomics provides researchers with a robust and customizable approach to data preprocessing. This ensures that downstream analyses are based on high-quality, singlet cell data, improving overall accuracy and reliability. This step incorporates user-adjustable parameters, including the expected doublet rate and neighborhood size, allowing flexibility to adapt to varying experimental conditions. Visualizations generated using ggplot2 and Plotly enable users to inspect predicted doublets interactively,

examining distributions and confidence scores. Once flagged, doublets can be excluded from the dataset with a single click, ensuring a clean and reliable input for downstream analysis. The integration of Shiny provides real-time interactivity, enabling users to adjust thresholds dynamically and observe their impact immediately. This combination of statistical rigor, computational efficiency, and user-friendly design makes the doublet removal step in Heraclomics an essential component of the preprocessing workflow. By ensuring that only singlet cells are included, Heraclomics enhances the accuracy and reliability of all subsequent analyses.

2.3.4 Harmony Integration and Dimensionality Reduction

The Harmony Integration and Dimensionality Reduction step in Heraclomics addresses batch effects, normalization, and high-dimensional data challenges inherent in single-cell RNA sequencing (scRNA-seq) datasets. Batch effects, which arise from differences in experimental conditions or sequencing platforms, can obscure true biological variation. This step ensures seamless integration of datasets while preserving meaningful biological signals. Normalization methods are a key component of this step. Heraclomics provides two primary normalization options: SCTransform and log-normalization. SCTransform (Hafemeister & Satija, 2019) performs variance-stabilizing transformations, enabling accurate integration and downstream analysis, while log-normalization scales and transforms data to make expression values comparable across cells. The outputs from these normalization methods feed into dimensionality reduction and visualization steps, providing users with a variety of exploratory tools. Heraclomics utilizes the Harmony algorithm (Korsunsky et al., 2019) to correct batch effects and integrate datasets effectively. Harmony aligns shared biological signals across datasets by identifying mutual nearest neighbors and adjusting for batch-specific variability. This integration facilitates downstream analyses by enabling the comparison of data from different conditions. Dimensionality reduction techniques, such as Principal Component Analysis (PCA), t-Distributed Stochastic Neighbor Embedding (t-SNE), and Uniform Manifold Approximation and Projection (UMAP), are applied to represent high-dimensional data in two or three dimensions. PCA is used as a preprocessing step to identify significant components of variation. UMAP and t-SNE provide low-dimensional representations that are well-suited for visualizing cellular

relationships and subpopulations. Outputs of this step include UMAP gene expression plots, t-SNE gene expression plots, dimensionality reduction heatmaps, elbow plots for identifying optimal components, and PCA loadings plots. Interactive visualizations are generated using Plotly (Sievert, 2020) and ggplot2 (Wickham, 2016), allowing users to explore integrated datasets dynamically. UMAP and t-SNE plots highlight cluster structures and enable researchers to inspect relationships between cell populations. Additional tools like cowplot and ggrepel enhance the clarity and layout of plots. The integration of R Shiny (Chang et al., 2021) enhances user interactivity by enabling dynamic parameter adjustments, such as the number of neighbors for UMAP or the perplexity value in t-SNE. Supporting libraries, including shinyjs, shinyalert, shinyWidgets, dplyr (Wickham et al., 2023), and harmony, ensure efficient processing and interactive feedback throughout the step. By combining Harmony integration, advanced normalization techniques, and dimensionality reduction, Heraclomics ensures that scRNA-seq datasets are harmonized and visually interpretable. This step plays a crucial role in preparing data for clustering and other downstream analyses, enabling researchers to uncover meaningful biological insights from their integrated datasets.

2.3.5 Clustering

The clustering step in Heraclomics is designed to identify distinct cell populations based on gene expression profiles, a critical component for understanding cellular heterogeneity in single-cell RNA sequencing (scRNA-seq) datasets. This module employs robust clustering algorithms, such as Louvain, SLM and Leiden, integrated with the Seurat framework (Stuart et al., 2019), to group cells into meaningful clusters. These algorithms operate on the nearest-neighbor graph, which is constructed from the principal component analysis (PCA) results. The nearest-neighbor graph is a fundamental structure that captures local relationships between cells in high-dimensional space, enabling accurate community detection. Users can adjust key clustering parameters dynamically, including the resolution for the Louvain and Leiden algorithms. Higher resolutions yield finer clusters, while lower resolutions merge similar groups into broader categories. The platform also allows the customization of UMAP and t-SNE parameters, such as the number of neighbors and perplexity values, to optimize the visualization of clustering results.

These adjustments provide users with the flexibility to tailor their analysis to specific datasets and biological questions. Visualization is a core feature of the clustering module, offering both 2D and 3D visualizations through UMAP and t-SNE plots. These visualizations generated using ggplot2 (Wickham, 2016), Plotly (Sievert, 2020), and cowplot, enable users to explore the spatial distribution of clusters interactively. The interactive interface, powered by Shiny (Chang et al., 2021), enhances user experience by providing real-time feedback when adjusting parameters. For instance, users can immediately observe changes in cluster assignments and their spatial distribution as they modify resolution or visualization settings. The platform supports both 2D and 3D visualizations, offering greater flexibility for exploring complex datasets. To ensure computational efficiency, the clustering module utilizes dplyr (Dowle & Srinivasan, 2021) for rapid data manipulation. By integrating advanced algorithms, customizable parameters, and interactive visualizations, Heraclomics empowers researchers to identify and analyze novel cell populations effectively. This step serves as a cornerstone for downstream analyses, including differential expression and trajectory inference, facilitating the discovery of meaningful biological insights. By combining powerful algorithms, advanced visualizations, and user interactivity, the clustering module provides a comprehensive and accessible solution for scRNA-seq data analysis, facilitating the discovery of novel cell populations and insights into cellular function.

2.3.6 Data Correction

The data correction step in Heraclomics is designed to address confounding effects in single-cell RNA sequencing (scRNA-seq) datasets, such as cell cycle variation, batch-specific biases, or other unwanted sources of variation. By correcting these effects, this step ensures that downstream analyses accurately reflect true biological differences rather than technical artifacts. Heraclomics employs Seurat's cell cycle scoring and regression methods (Stuart et al., 2019) to mitigate the influence of cell cycle effects. Users can visualize the distribution of cells across cell cycle phases (G1, S, G2/M) and apply corrections interactively. Heraclomics also provides the ability to correct and analyze gene-specific effects or pathway-related variations. Users can input specific genes or select pathways of interest using the msigdbr library, which integrates curated gene sets from databases like MSigDB (Liberzon

et al., 2015). This functionality allows users to account for potential biases or influences exerted by specific genes or gene sets on clustering and downstream analyses. The platform generates multiple outputs to help users assess and interpret the effects of data correction. These include UMAP and t-SNE plots updated after correction, violin plots to display distributions of key metrics, and ridge plots to examine variability across specific groups. These visualizations are generated using ggplot2 (Wickham, 2016), Plotly (Sievert, 2020), and viridis for enhanced color schemes. Interactive features, powered by R Shiny (Chang et al., 2021), enable users to dynamically adjust parameters for correction and instantly view the effects on visualizations. The inclusion of shinyWidgets and shinyalert enhances interactivity and user feedback throughout the process. Additional libraries, such as msigdbr and htmlwidgets, provide extended functionality for pathway analysis and interactive visualization, respectively. To ensure computational efficiency, the data correction module uses dplyr (Dowle & Srinivasan, 2021) for rapid data manipulation, while Matrix supports sparse matrix operations. By integrating robust statistical methods, advanced visualizations, and dynamic interactivity, the data correction step in Heraclomics empowers researchers to refine their datasets, ensuring biological relevance and readiness for downstream analyses.

2.3.7 Labelling Clusters

The cluster labeling step in Heraclomics allows users to assign meaningful biological identities to the clusters identified during the clustering process. This step is crucial for interpreting cell types, states, or functions within single-cell RNA sequencing (scRNA-seq) datasets. Heraclomics integrates robust tools and reference datasets to streamline the annotation process, including SingleR (Aran et al., 2019) and the scType framework (Ianevski et al., 2019), enabling automated and manual cluster labeling. SingleR performs reference-based annotation by comparing gene expression profiles of clusters to well-curated reference datasets. These include "Human Primary Cell Atlas" (HPCAData) (Mabbott et al., 2013), "Blueprint Encode" (BlueprintEncodeData) (Stunnenberg et al., 2016), "Database Immune Cell Expression" (DatabaseImmuneCellExpressionData) (Monaco et al., 2019), "ImmGen" (ImmGenData) (Heng et al., 2019), "Monaco Immune Data" (MonacoImmuneData) (Monaco et al., 2019), "Novershtern Hematopoietic Data" (NovershternHematopoieticData) (Novershtern et al., 2011), "Mouse RNAseq"

(MouseRNAseqData) (Yizhak et al., 2013), and "ENCODE and Blueprint Epigenomics" (ENCODEBlueprintData) (The ENCODE Project Consortium, 2012). This automated annotation approach provides users with suggested labels for each cluster, enhancing reproducibility and consistency in cluster characterization. UMAP and t-SNE plots offer spatial representations of clusters, providing clear visual feedback for annotation decisions. By integrating annotation tools with visualization capabilities, this step provides researchers with a flexible and efficient approach to cluster labeling. The cluster labeling module also includes statistical outputs, such as marker gene enrichment scores and correlation values, to validate annotation accuracy. This ensures that the labels assigned to clusters are robust and biologically meaningful, facilitating downstream analyses like differential expression, trajectory inference, and cell communication studies.

2.3.8 Gene Co-expression Network

The Gene Co-Expression Networks module in Heraclomics identifies and visualizes networks of genes that exhibit correlated expression patterns across cells. This step employs MEGENA (Song & Zhang, 2015) to construct co-expression networks using Planar Filtered Networks (PFNs) and multiscale clustering to identify gene modules. These networks help researchers explore relationships between genes and uncover potential regulatory roles in cellular processes. Heraclomics integrates a range of libraries to enhance functionality, including Seurat (Stuart et al., 2019) for data preprocessing, MEGENA for network analysis, and igraph (Csardi & Nepusz, 2006) for advanced graph-based operations. Functional enrichment analysis is performed using clusterProfiler (Yu et al., 2012) and org.Hs.eg.db (Carlson, 2023), allowing researchers to link gene modules to biological pathways. Visualization libraries such as ggplot2 (Wickham, 2016), enrichplot, and Plotly (Sievert, 2020) provide interactive and publication-quality graphical outputs. Key outputs of this module include a network plot displaying gene connections, a module summary table highlighting key characteristics of each identified module, and hub gene identification to pinpoint highly connected genes within the network. Additional outputs include gene module membership lists, visualizations of individual modules, and enrichment analysis results that reveal pathways and biological processes associated with each module. Interactive features, built using Shiny (Chang et al., 2021) and shinyWidgets, allow users to

customize parameters for network construction, such as correlation thresholds and clustering resolution. Users can also explore modules interactively through network visualizations, selecting specific nodes or edges to examine relationships and underlying biological implications. This module provides a comprehensive approach for understanding the functional organization of the transcriptome and identifying potential regulatory mechanisms within single-cell datasets.

2.3.9 Gene Regulatory Network

The Gene Regulatory Networks (GRN) module in Heraclomics enables the inference and visualization of transcriptional regulatory relationships in single-cell RNA sequencing (scRNA-seq) datasets. By leveraging advanced network analysis and visualization tools, this module identifies transcription factors (TFs) and their target genes, providing insights into the regulatory architecture that drives cellular processes and states. The module integrates multiple libraries to perform comprehensive GRN construction and analysis. GENIE3 (Huynh-Thu et al., 2010) is employed to infer TF-target relationships based on gene expression data, while RTN (Castro et al., 2016) reconstructs transcriptional regulatory networks. AUCell (Aibar et al., 2017) is used to score TF activity across cells, enabling the identification of active regulons. For network visualization, igraph (Csardi & Nepusz, 2006) and visNetwork provide interactive graph-based representations, allowing users to explore complex regulatory interactions. Preprocessing and cluster-level analysis are handled by Seurat (Stuart et al., 2019), while pheatmap generates heatmaps to display regulon activity. Plotly (Sievert, 2020) is used to create dynamic UMAP visualizations, and msigdbr supports functional enrichment analysis of regulatory networks. Additionally, Heraclomics leverages shiny, shinydashboard, shinyjs, and shinyalert to create an intuitive and user-friendly interface, enabling researchers to dynamically adjust parameters and explore GRNs in real time. The use of tidyverse and tidyr facilitates efficient data manipulation and transformation. Key outputs of this module include UMAP visualizations showing regulon activity across cells and clusters, heatmaps that highlight the activity of transcriptional regulators in specific cell populations, and interactive network plots that allow detailed exploration of TF-target interactions. A regulon table provides a summary of identified regulons, including associated genes and activity scores, ensuring comprehensive insights into regulatory networks. These

features empower researchers to study transcriptional regulation, identify key TFs, and uncover lineage-specific regulatory programs in scRNA-seq datasets.

2.3.10 Trajectory Analysis

The Trajectory Analysis module in Heraclomics enables the reconstruction and visualization of dynamic cellular processes by inferring pseudotime trajectories from single-cell RNA sequencing (scRNA-seq) datasets. By analyzing the progression of cells through developmental or transitional states, this module provides insights into lineage relationships, key regulators, and gene expression dynamics. The module integrates multiple libraries to ensure robust trajectory inference and interactive exploration. monocle3 (Cao et al., 2019) is employed for pseudotime estimation, dimensionality reduction, and trajectory reconstruction, while SeuratWrappers facilitates seamless integration with clustering results from Seurat. Visualization is powered by plotly (Sievert, 2020) and ggplot2 (Wickham, 2016), generating interactive and publication-ready outputs. Libraries such as igraph (Csardi & Nepusz, 2006) enable network-based trajectory modeling, and shiny, shinyWidgets, shinyjs, and shinyalert create an intuitive and user-friendly interface for real-time exploration. Data manipulation and presentation are supported by dplyr and DT for efficient computation and tabular outputs. Key outputs of the module include trajectory graph visualizations that depict pseudotime paths overlaid on dimensionality reduction spaces, cluster visualizations that annotate cell groupings along trajectories, and partition visualizations that highlight distinct lineages or branches. Gene expression dynamics along pseudotime are represented through interactive plots, allowing users to explore the expression patterns of key genes across transitions and branches. Interactive features enable real-time adjustment of pseudotime parameters, cluster resolutions, and branch detection thresholds, providing flexibility and detailed insights into lineage-specific transitions and regulatory mechanisms. This module is pivotal for studying dynamic biological processes such as differentiation, immune responses, and cancer progression. By integrating advanced methodologies with interactive exploration, the trajectory analysis module provides researchers with a comprehensive platform to uncover pseudotime relationships and their biological significance.

2.3.11 Cell Communication

The Cell Communication module in Heraclomics enables the analysis and visualization of intercellular signaling networks within single-cell RNA sequencing (scRNA-seq) datasets. This module uses CellChat (Jin et al., 2021) to infer ligand-receptor interactions between cell populations, quantifying their strength and significance to uncover patterns of cellular communication. Preprocessing and clustering integration are facilitated by Seuratwrapper (Stuart et al., 2019), while network analysis and visualization are supported by igraph (Csardi & Nepusz, 2006) and visNetwork. Visualization tools such as ggplot2 (Wickham, 2016) and ggalluvial enhance the clarity and interpretability of signaling pathway representations. Additional libraries, including NMF for identifying signaling patterns, shinyWidgets and shinyalert for interactive feedback, and DT for tabular data summaries, provide a user-friendly and dynamic experience. Key outputs of this module include communication network visualizations depicting ligand-receptor interactions between cell populations, specific signaling pathway visualizations highlighting the role of distinct pathways, communication patterns heatmaps that display interaction intensities, bubble plots summarizing interaction significance, and tabular summaries of significant ligand-receptor pairs. The interactive R Shiny interface allows users to dynamically filter and explore communication networks, adjust parameters, and inspect specific interactions in detail. For example, network visualizations are interactive, enabling users to select nodes and edges to examine relationships, while heatmaps and bubble plots provide high-level overviews of communication patterns. This module is essential for exploring intercellular communication mechanisms in various biological contexts, including development, immune responses, and tissue homeostasis. By identifying critical signaling pathways and interactions, researchers can uncover novel regulatory mechanisms, functional roles of cell populations, and potential therapeutic targets.

2.3.12 Differential Expression

The Differential Expression (DE) module in Heraclomics identifies genes that are differentially expressed between groups of cells, such as clusters or conditions, within single-cell RNA sequencing (scRNA-seq) datasets. This module uses DESeq2 (Love et al., 2014) for robust statistical testing of differential expression,

providing accurate and reproducible results. Preprocessing and integration with cluster-level annotations are handled by dplyr and DT for efficient data manipulation and tabular output. Visualization tools, including ggplot2 (Wickham, 2016) and plotly (Sievert, 2020), enable interactive exploration of DE results, while pheatmap facilitates the creation of heatmaps to display the expression of top-ranked genes across groups. Key outputs of this module include a comprehensive DE results table listing significant genes, their fold changes, adjusted p-values, and expression metrics. An interactive volcano plot highlights significant genes based on their fold changes and statistical significance, allowing users to dynamically filter and inspect results. The MA plot provides an overview of gene expression changes relative to their mean expression levels, while a heatmap of top genes visualizes the expression patterns of the most significant genes across cell groups or conditions. The interactive interface, powered by R Shiny and shinyalert, allows users to adjust parameters for differential testing, such as significance thresholds and fold change cutoffs, and dynamically update visualizations. The use of plotly enables interactive features such as zooming, panning, and point selection for detailed inspection of specific genes. Tabular outputs generated with DT provide an organized and sortable view of DE results, facilitating downstream analysis. This module is critical for identifying biomarkers, understanding gene expression differences between conditions, and exploring the molecular drivers of cellular heterogeneity. By integrating robust statistical methods with interactive visualizations, the DE module provides a comprehensive and user-friendly platform for analyzing differential expression in scRNA-seq datasets.

3
RESULTS

3.1 Starting App and Analyze Sample Data

This chapter demonstrates how Heraclomics works, showcasing its interface shown in Figure3.1, functionality, and workflow steps. The figures included provide evidence of the platform's performance and usability. It also presents the analytical outputs generated by Heraclomics using the Human Breast Cancer Block A, Section 1 dataset retrieved from 10X Genomics (10X Genomics, n.d.).

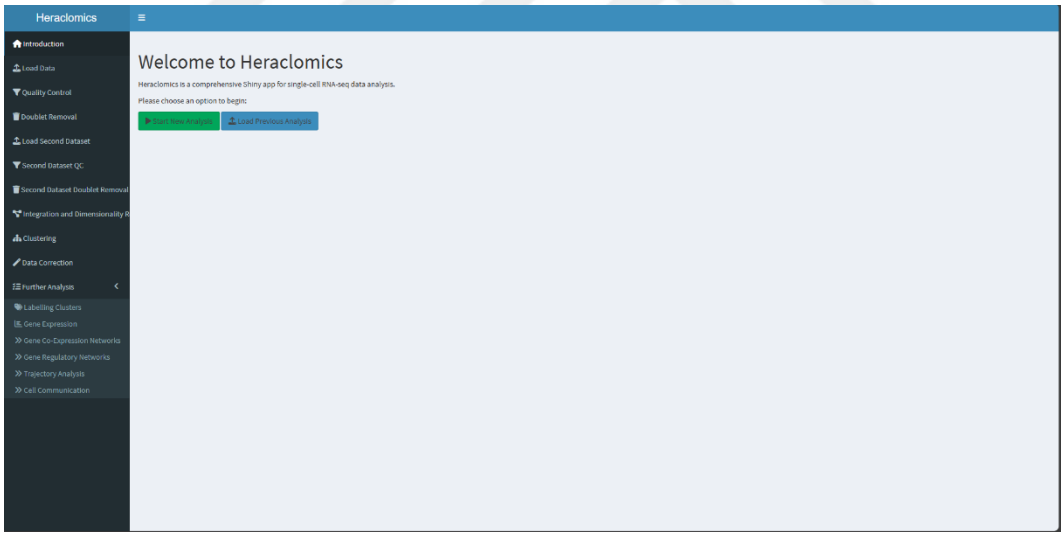


Figure 3.1 Homepage and interface of Heraclomics

3.1.1 Loading Data

Loading data process and turning it to Seurat objects works very simple and fast according to system. In Figure 3.2 simplicity of loading variety of data types and naming data can be seen. In the Figure 3.3 shows first overview of loaded data which given as a table

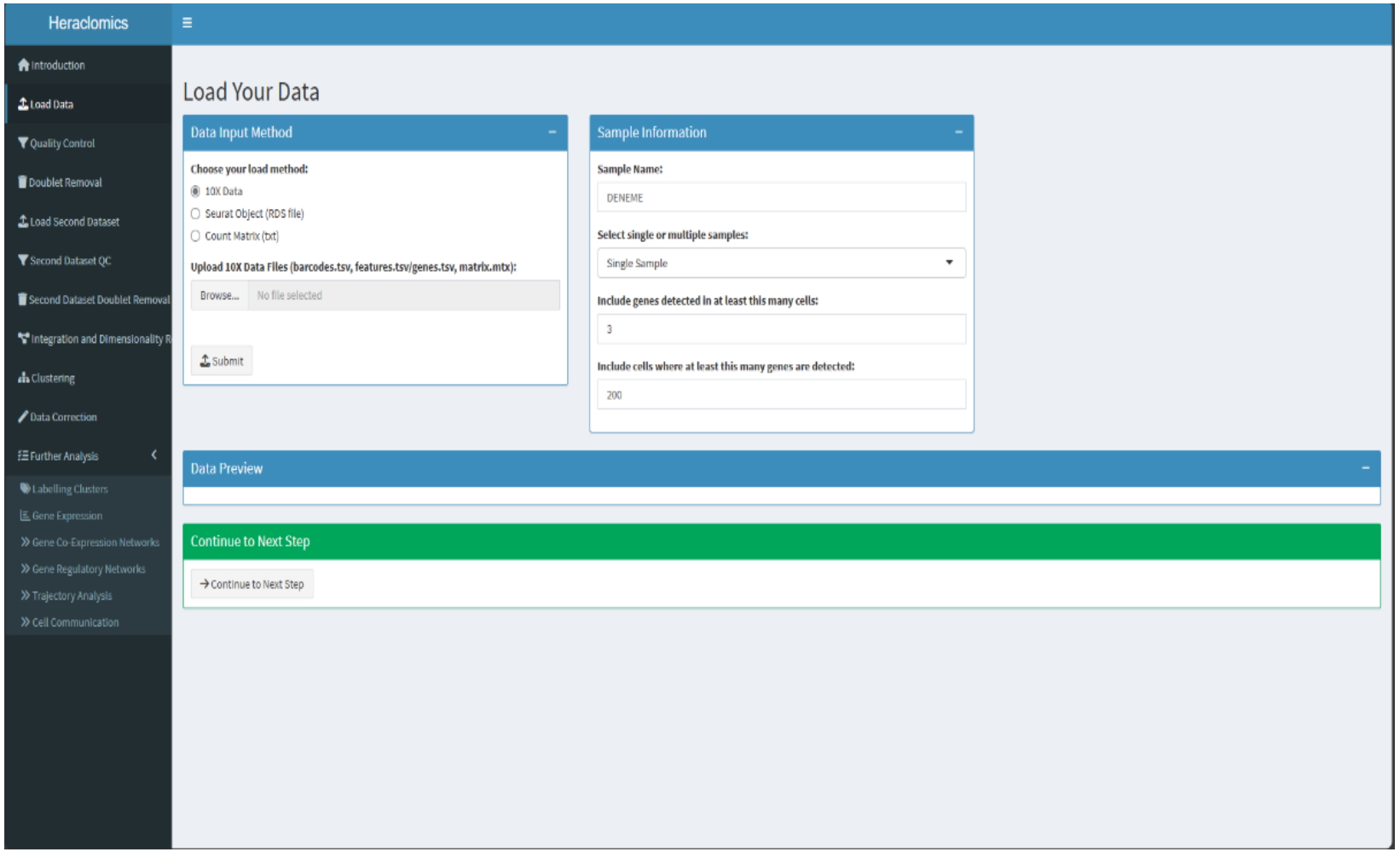


Figure 3.2 Data upload and sample naming

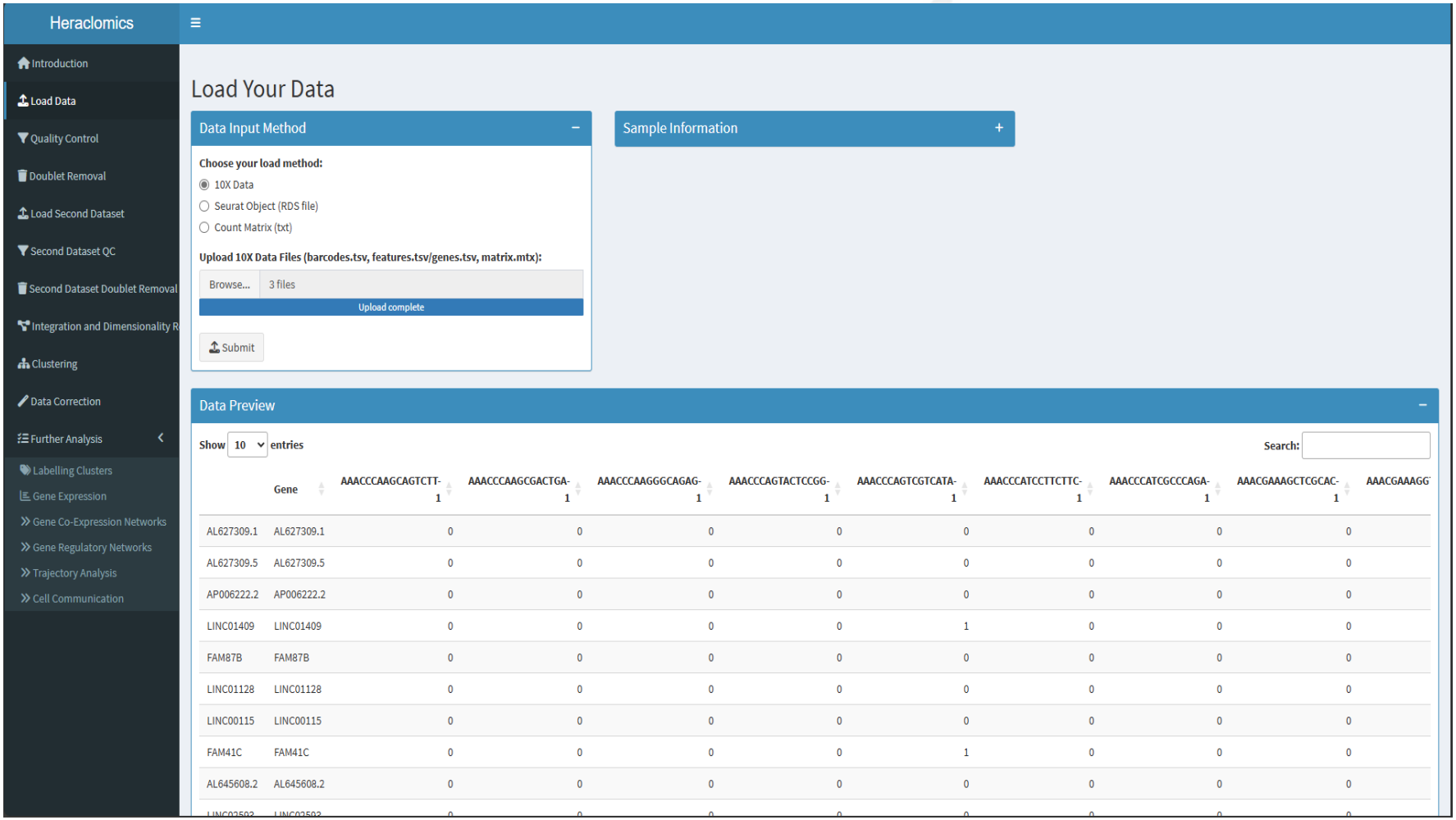


Figure 3.3 Data preview of uploaded data

3.1.2 Quality Control

The quality control process in Heraclomics ensures that only high-quality cells are retained for downstream analysis. This step involves filtering cells based on metrics such as the number of detected genes, total RNA counts, ribosomal and mitochondrial content, which are visualized interactively to allow users to refine thresholds dynamically can be seen in Figure 3.4. Violin plots (Figure 3.5) show the distributions of key quality metrics across cells, including mitochondrial gene percentage and detected gene counts. These visualizations help users identify potential outliers or low-quality cells. In Figure 3.6, a scatter plot illustrates the relationship between mitochondrial content and total RNA counts, with cells failing the user-defined thresholds highlighted and flagged for removal. After applying the quality control filters, the dataset is significantly refined, retaining only high-quality cells while removing those likely to introduce noise. This process ensures that the remaining data is biologically meaningful and ready for downstream analysis.

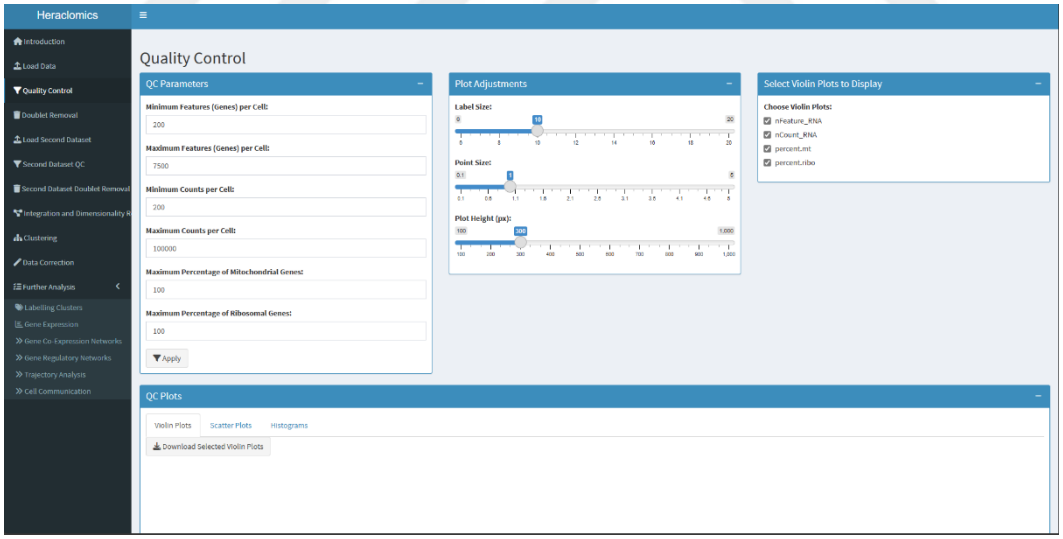


Figure 3.4 Quality control parameters

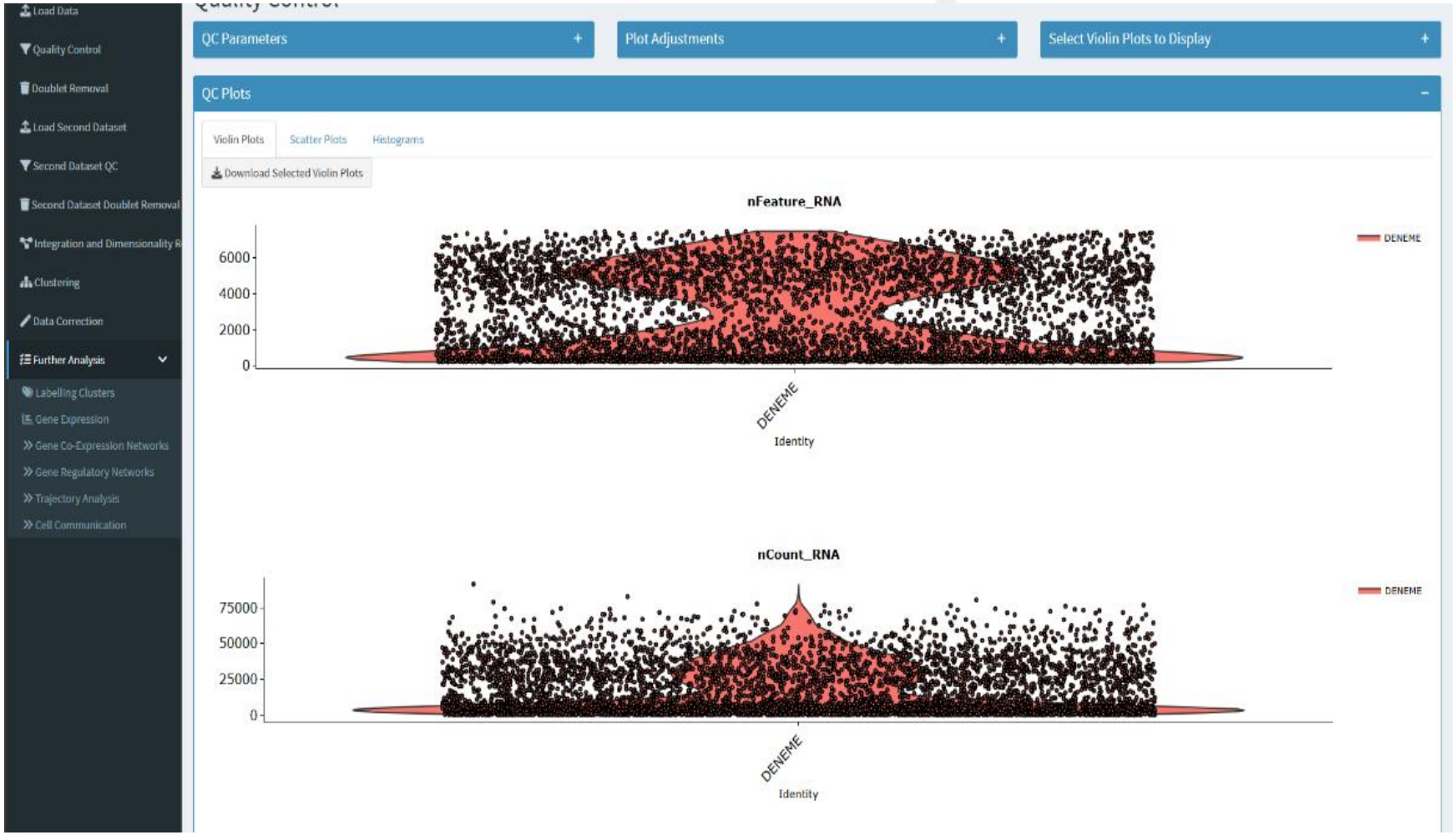


Figure 3.5 Quality control violin plots

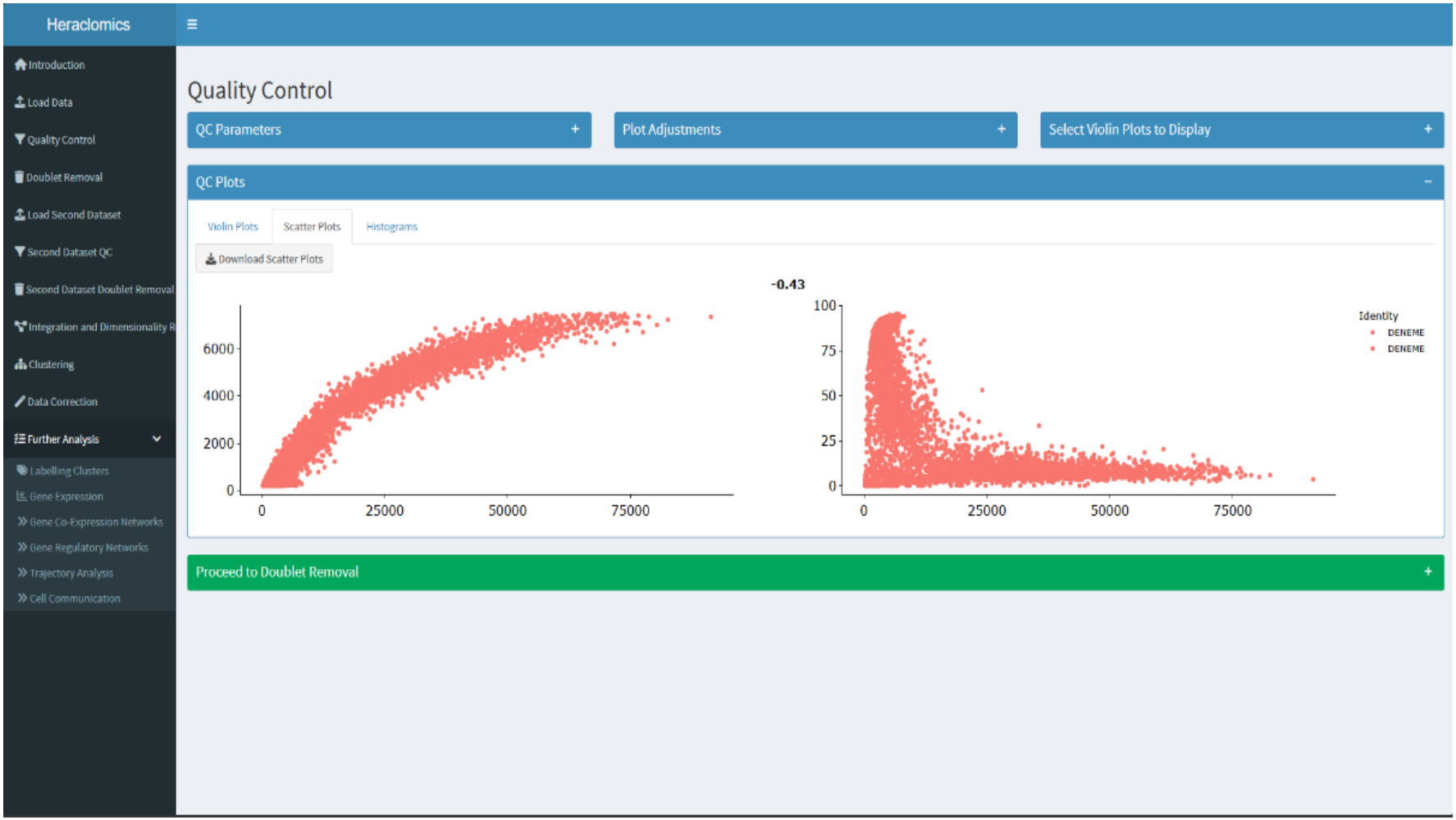


Figure 3.6 Quality control scatter plots

3.1.3 Doublet Removal

The doublet removal step in Heraclomics plays a critical role in ensuring the quality of single-cell RNA sequencing data by identifying and removing doublets—instances where two cells are captured in the same droplet as define in material and method section, leading to mixed gene expression profiles. This step provides an interactive and user-friendly interface, as shown in Figure 3.7, where users can start or skip the doublet removal process, adjust visualization parameters such as label size, point size, and plot dimensions, and select either UMAP or t-SNE for visualization. In Figure 3.8, a UMAP plot illustrates the distinction between singlets (in blue) and doublets (in red) before removal, enabling users to visually assess the distribution of doublets across the dataset. After the removal process, Figure 3.9 demonstrates the updated UMAP plot with only singlets remaining, reflecting the success of the doublet removal step. Heraclomics also allows users to simulate doublets, providing a deeper understanding of how doublets appear in their dataset, which can guide informed decisions on removal. Importantly, this step is optional, allowing users flexibility to proceed without removal if no significant doublets are detected or if the analysis does not require it. The interactive and visual features of this module ensure that users can confidently refine their data, improving the reliability of downstream analyses. This module underscores Heraclomics' emphasis on providing robust data processing tools while maintaining ease of use and flexibility for researchers.

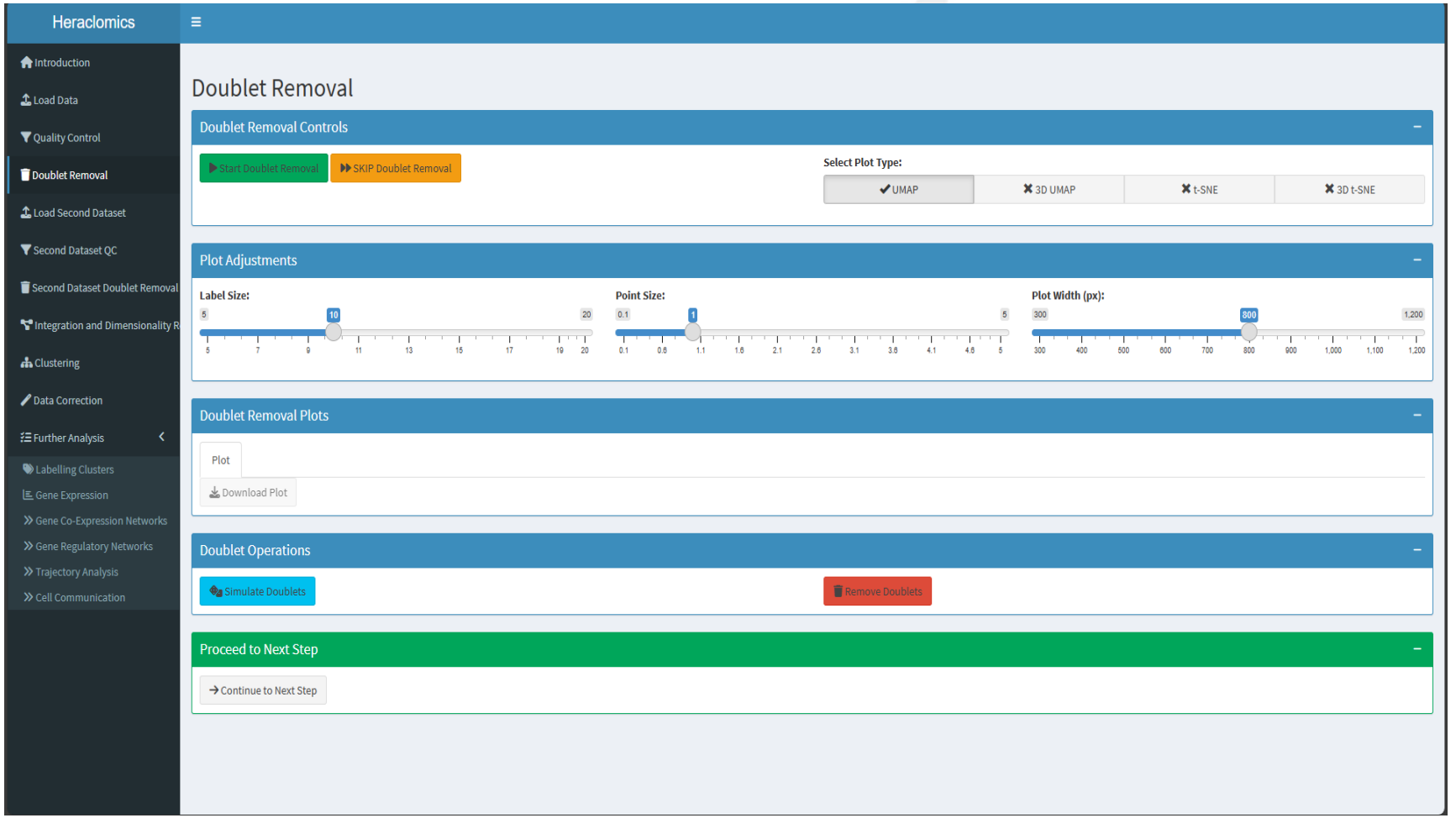


Figure 3.7 Doublet removal interface

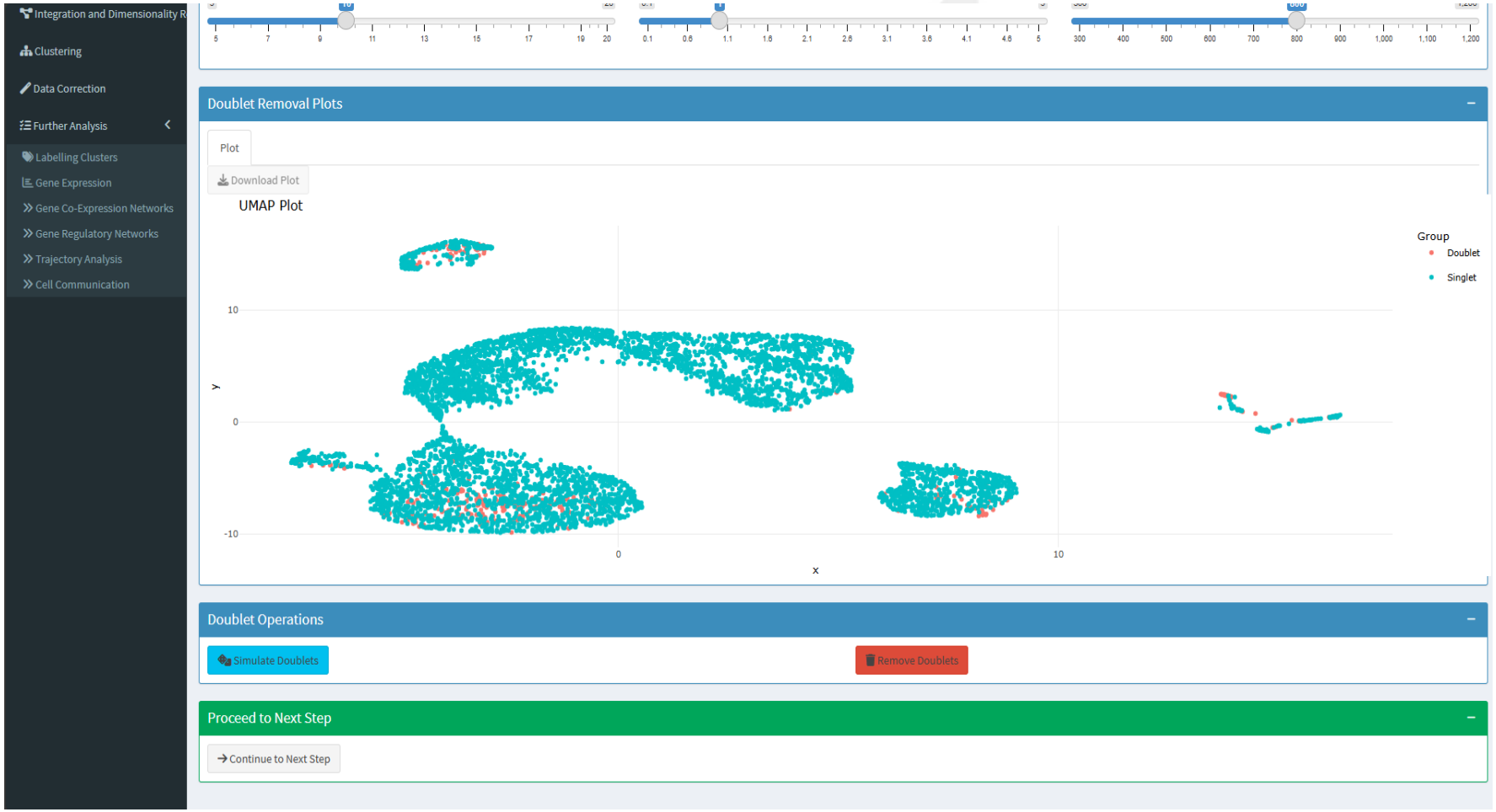


Figure 3.8 Umap before doublet removal

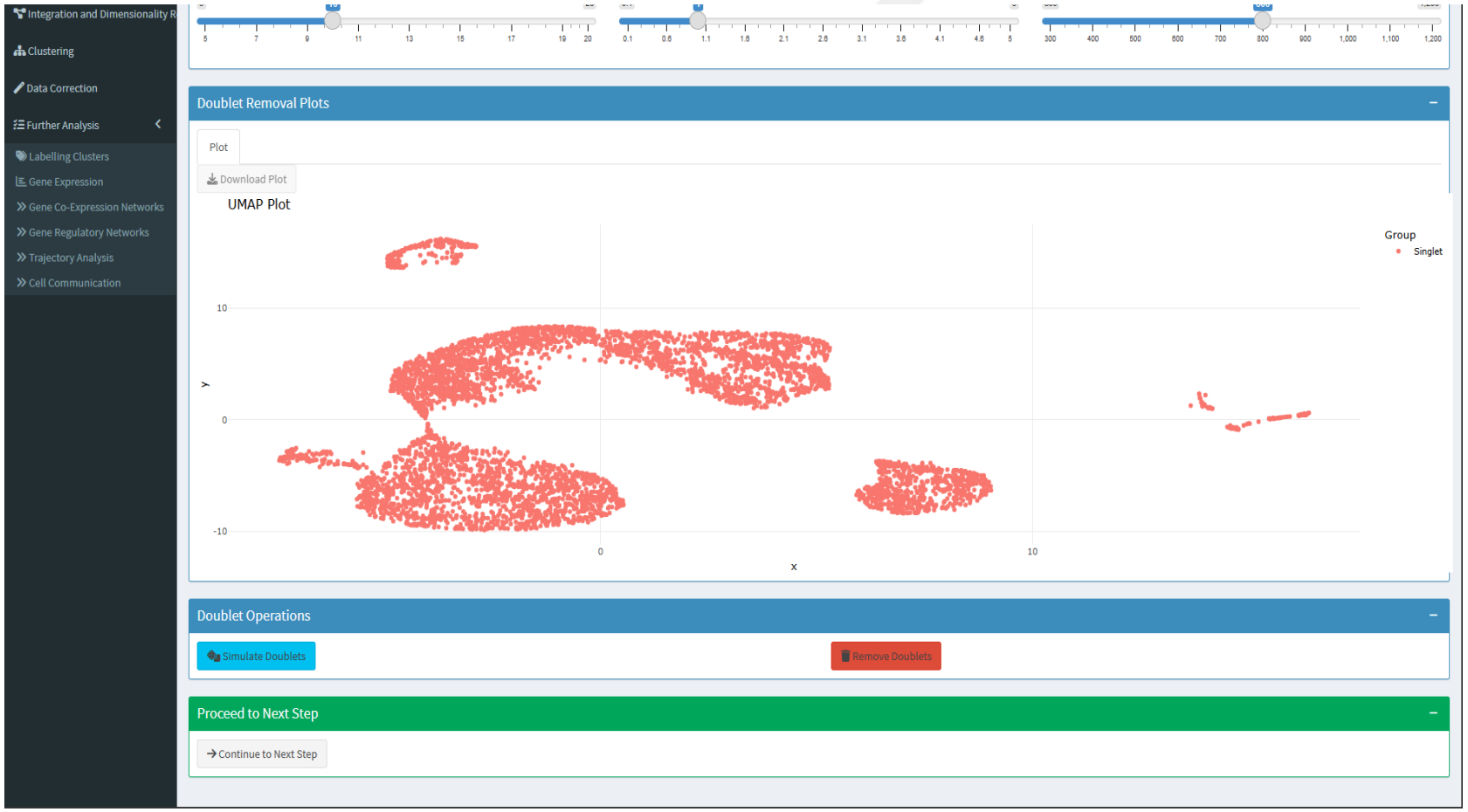


Figure 3.9 UMAP after doublet removal

3.1.4 Harmony Integration and Dimensionality Reduction

The Harmony Integration and Dimensionality Reduction step in Heraclomics is designed to address batch effects, normalize datasets, and reduce high-dimensional data for visualization and downstream analysis. This step integrates multiple datasets while ensuring that meaningful biological signals are preserved, allowing researchers to work with harmonized data if available. As shown in Figure 3.10, the user interface allows researchers to specify the datasets to integrate and select a normalization method, such as SCTransform or LogNormalization. Once the preferred method is chosen, the Harmony algorithm is applied, effectively correcting batch effects while maintaining biological variability.

The outputs of this step provide a comprehensive overview of the integrated data. Figure 3.11 illustrates the integrated dataset visualized using UMAP, with samples color-coded by batch, showcasing the seamless merging of datasets. For gene-specific insights, Figure 3.12 highlights the expression levels of a selected gene across the dataset using t-SNE plots, aiding in identifying distinct expression patterns. The Dimensionality Reduction Heatmap, shown in Figure 3.13, emphasizes the contribution of principal components to data variability, while the Elbow Plot in Figure 3.14 helps determine the optimal number of components for further analysis by visualizing the variance explained by each component and the PCA Loadings Plot identifies the top principal components, highlighting the genes contributing most to data variability.

This step leverages advanced algorithms like Harmony for integration and utilizes dimensionality reduction techniques such as UMAP, t-SNE, and PCA to ensure data consistency and clarity. The interactive interface enhances user experience, allowing selecting genes and principal componants for visualization and allows users to download output figures. By refining datasets and reducing technical artifacts, this step lays the foundation for accurate clustering and downstream biological interpretations.

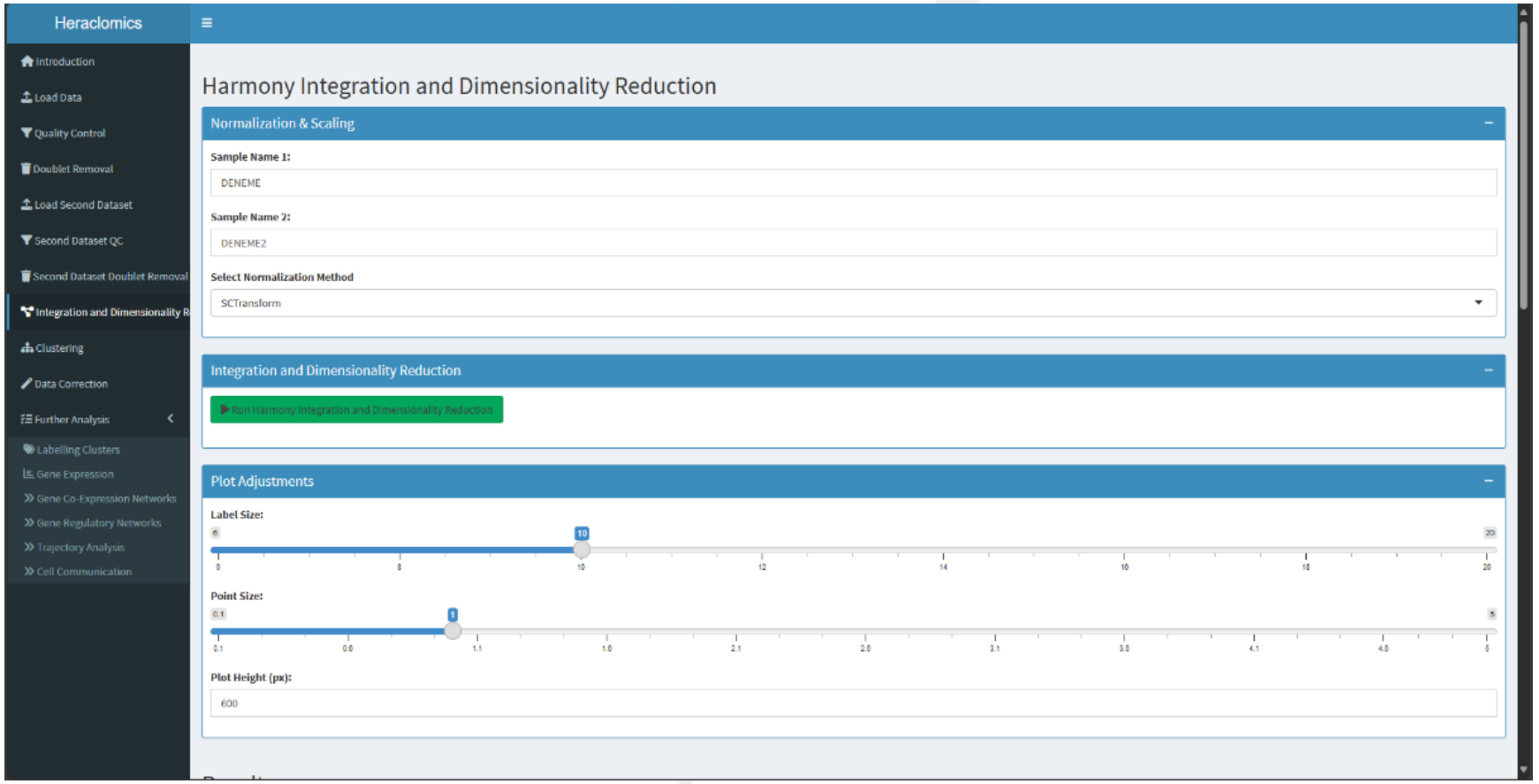


Figure 3.10 Harmony integration and dimensionality reduction interface

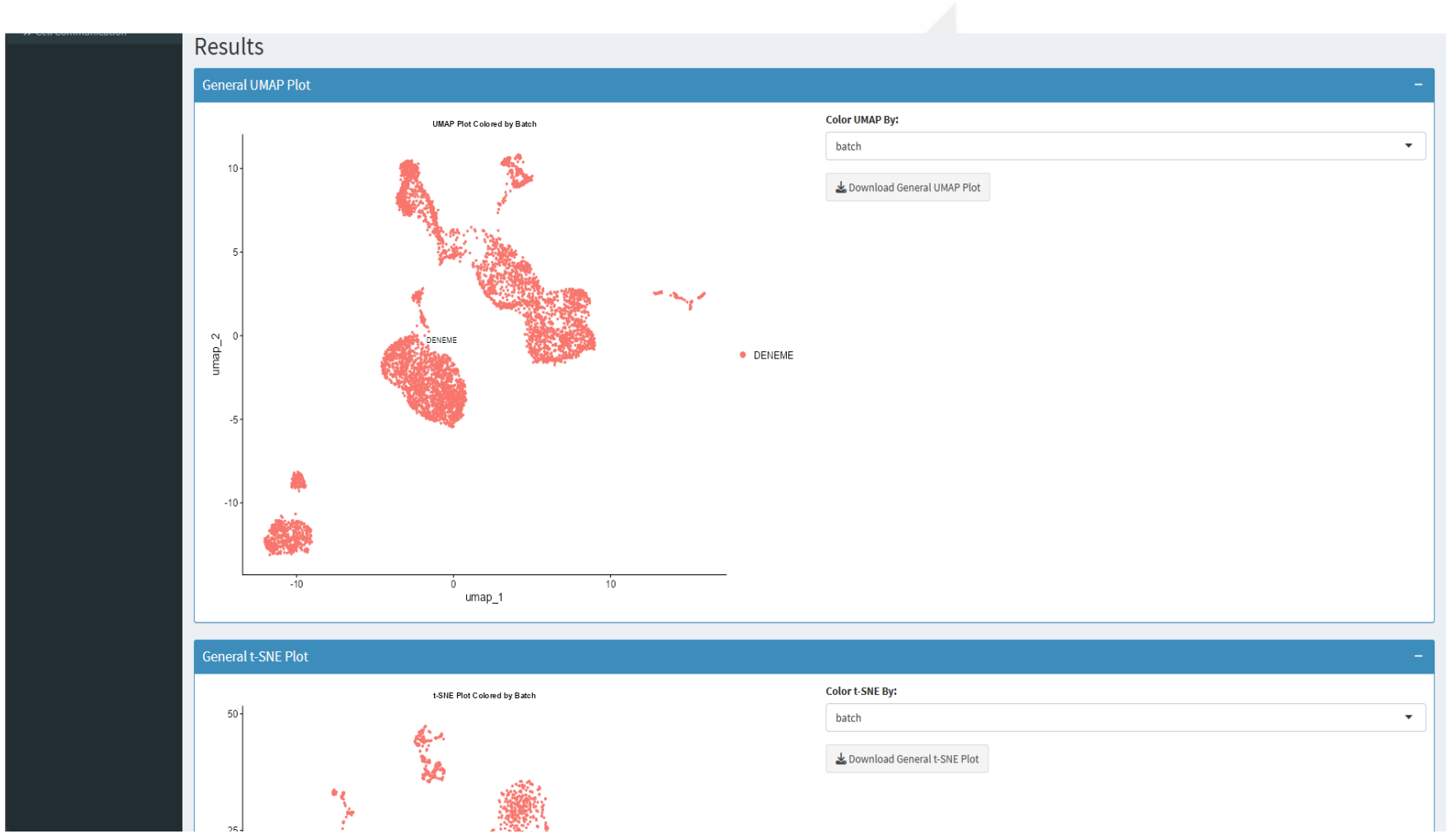


Figure 3.11 UMAP and t-SNE plots of batch (if availables) and samples

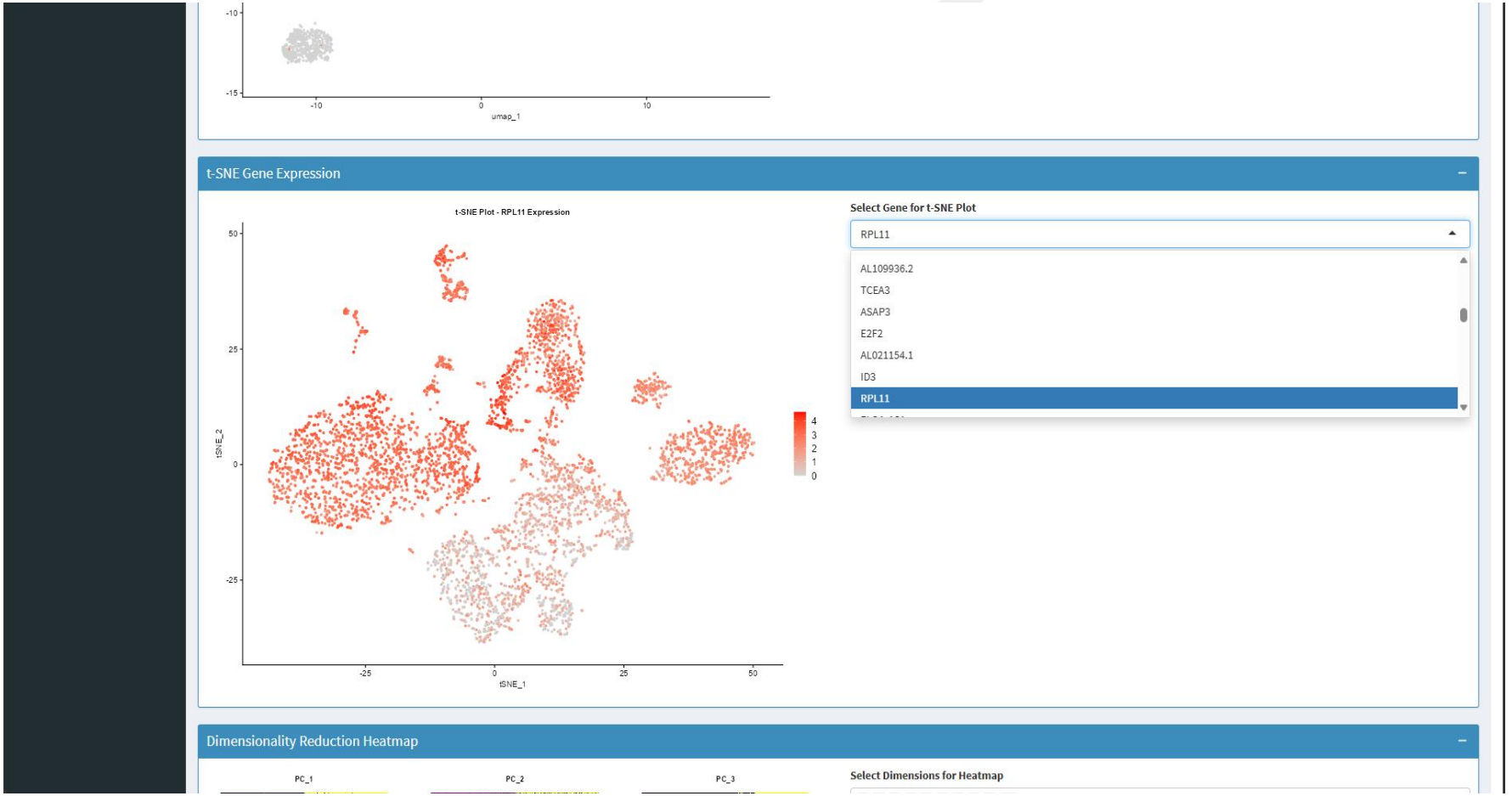


Figure 3.12 t-SNE plots of selected gene colored by expression

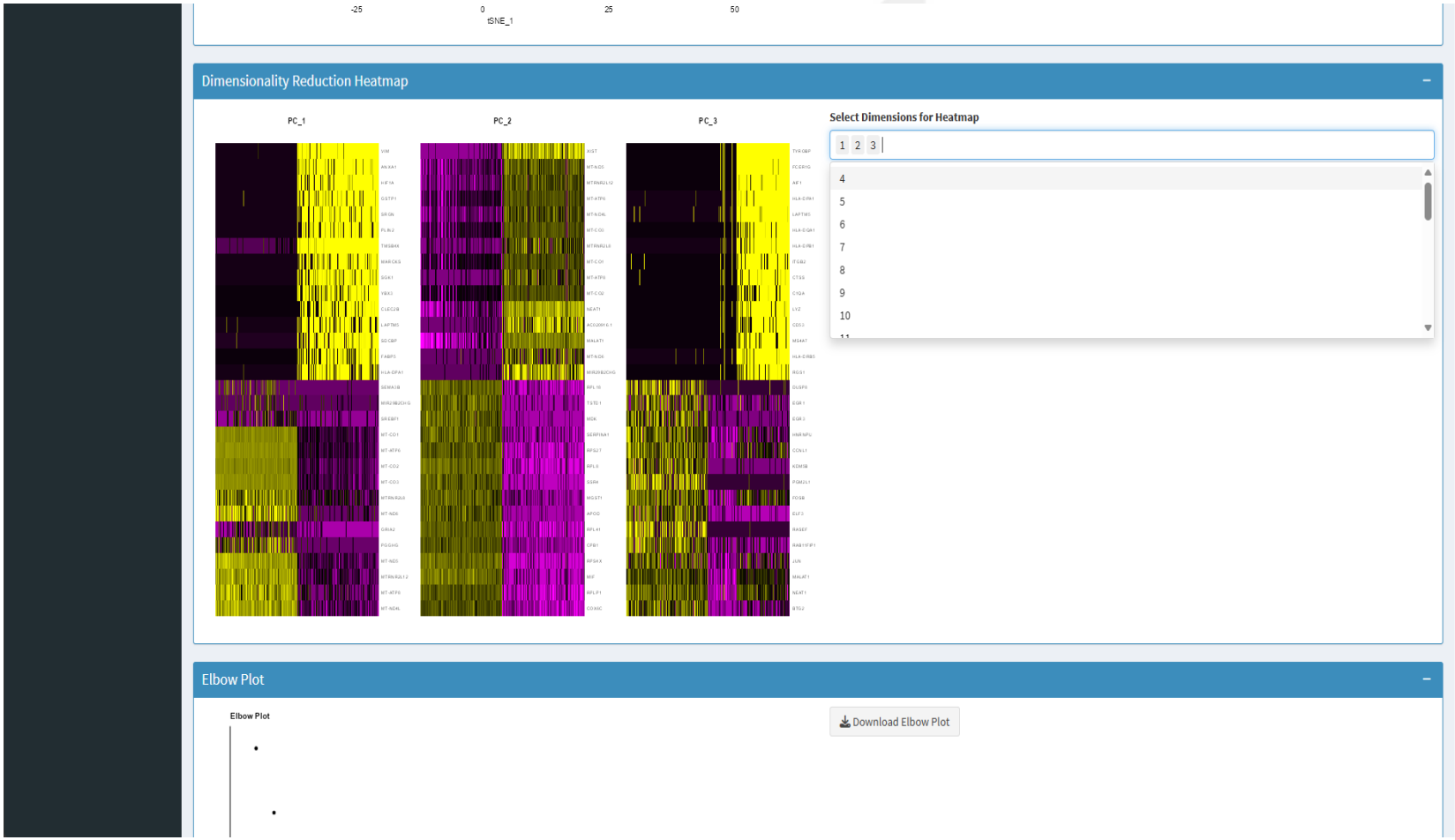


Figure 3.13 Heatmaps of selected principal components (can be selected single or multiple)



Figure 3.14 Elbow plot and PCA loading plots (top 10 differentiated gene labelled)

3.1.5 Clustering

The clustering step in Heraclomics enables the identification of distinct cell populations based on their gene expression profiles, a critical component in understanding cellular heterogeneity in single-cell RNA sequencing datasets. The platform provides a highly interactive and customizable interface, as shown in Figure 3.15, where users can adjust parameters for clustering algorithms, such as the nearest-neighbor graph construction, resolution, and dimensionality. Key parameters include the number of dimensions used for clustering, k.param (the number of nearest neighbors considered), and the number of trees (n.trees) used in graph construction. By optimizing these inputs, users can refine clustering accuracy and tailor the analysis to specific datasets. Heraclomics uses robust clustering algorithms such as SLM, Louvain and Leiden via the Seurat framework, enabling precise identification of cellular subpopulations with the algorithm of users' choice. The platform supports both UMAP and t-SNE plots for cluster visualization, offering users the flexibility to switch between 2D and 3D representations. Figure 3.16 displays a 3D UMAP plot, while Figure 3.17 shows the 2D UMAP plot, highlighting the spatial relationships between clusters.

Additionally, the clustering module provides a Cell Composition Summary (Figure 3.17), which visualizes the proportion of cells in each cluster, enabling an overview of cellular composition across the dataset. This information is critical for understanding the biological diversity within the sample.

Interactive features, powered by R Shiny, allow users to dynamically adjust clustering and visualization parameters while receiving immediate feedback on their analyses. This step ensures a comprehensive exploration of cell population structures, setting the stage for downstream analyses such as differential expression and pathway enrichment.

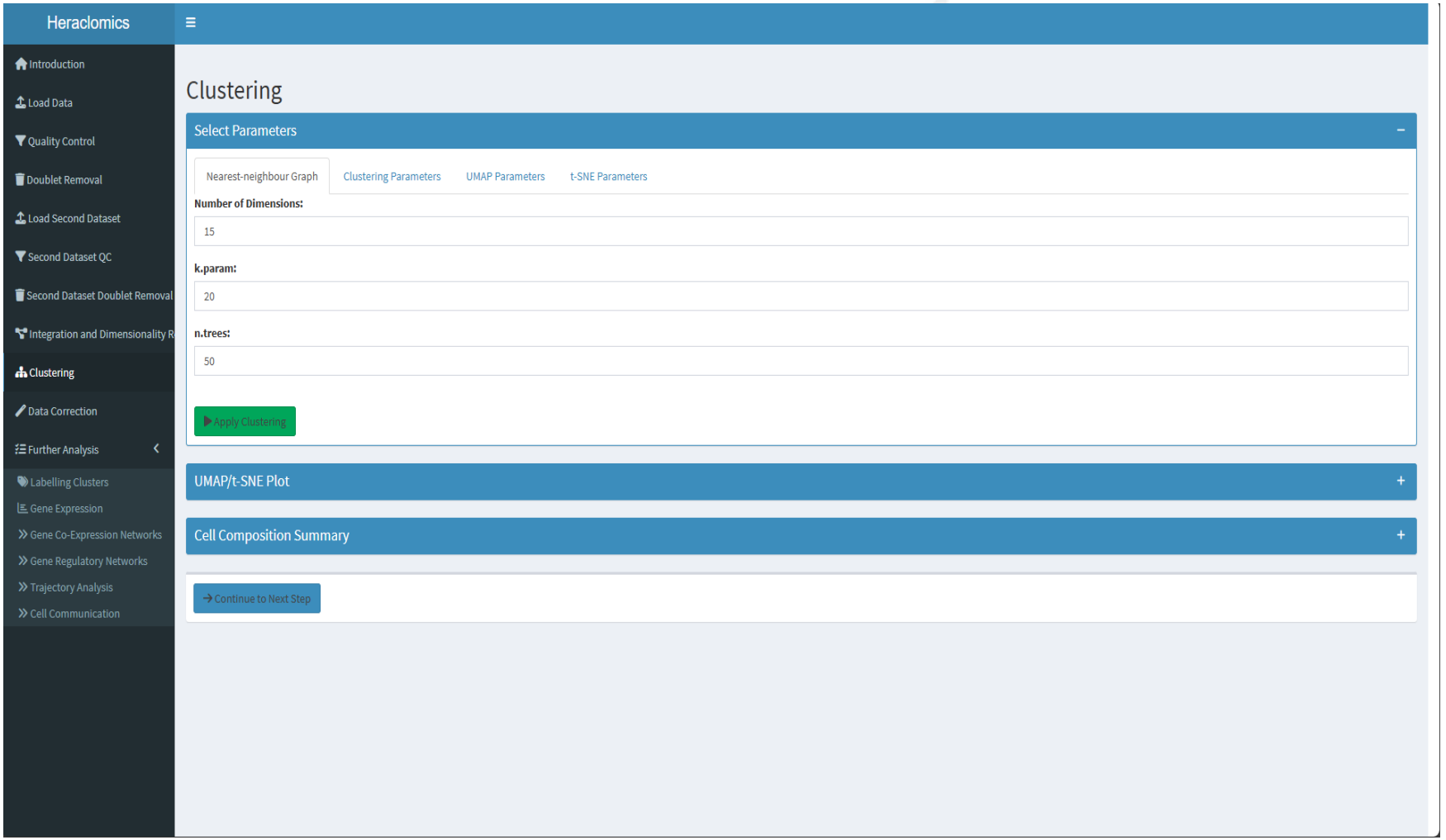


Figure 3.15 Clusterin interface and parameters

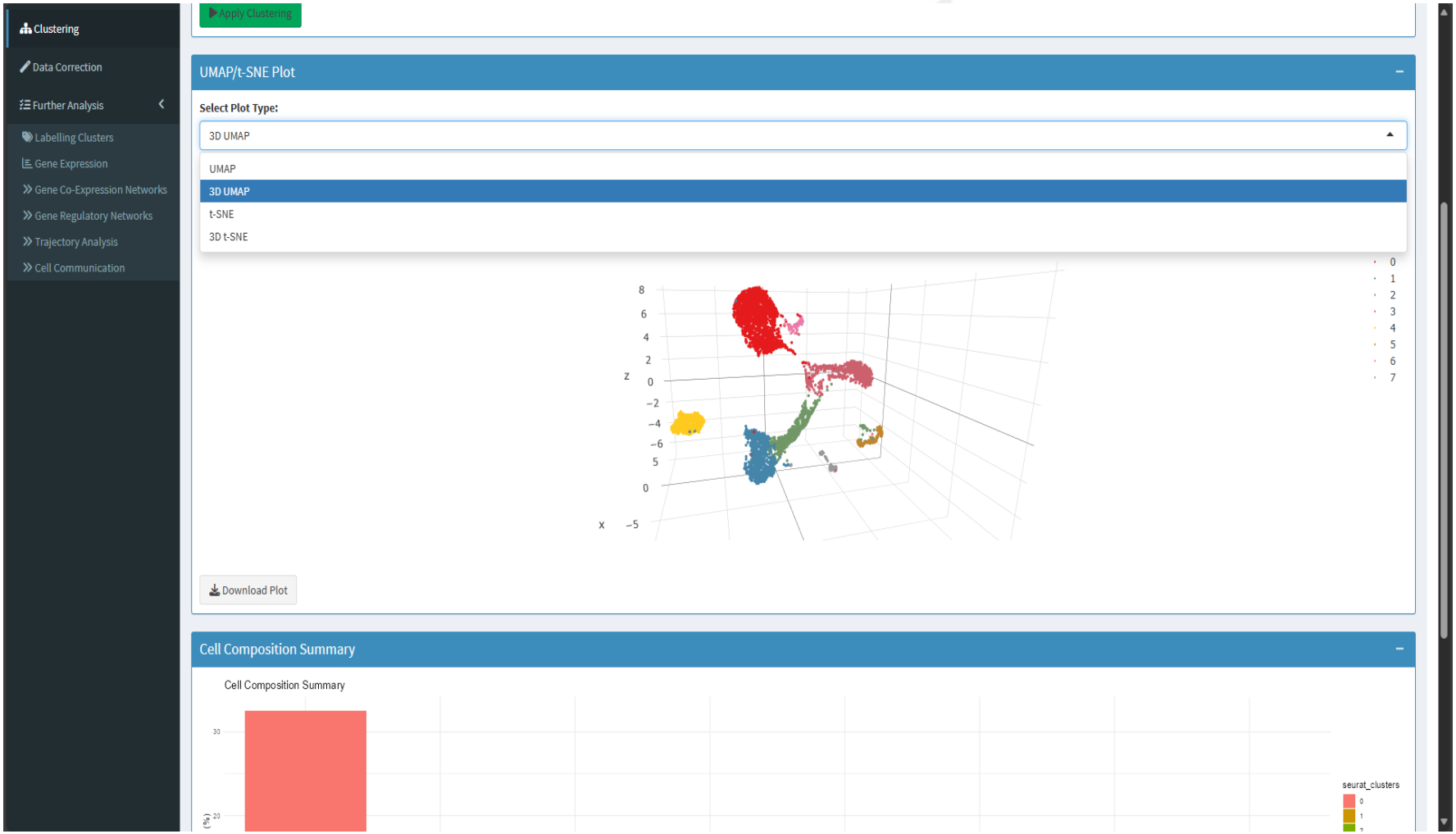


Figure 3.16 3D Umap visualization of clusters and plot types

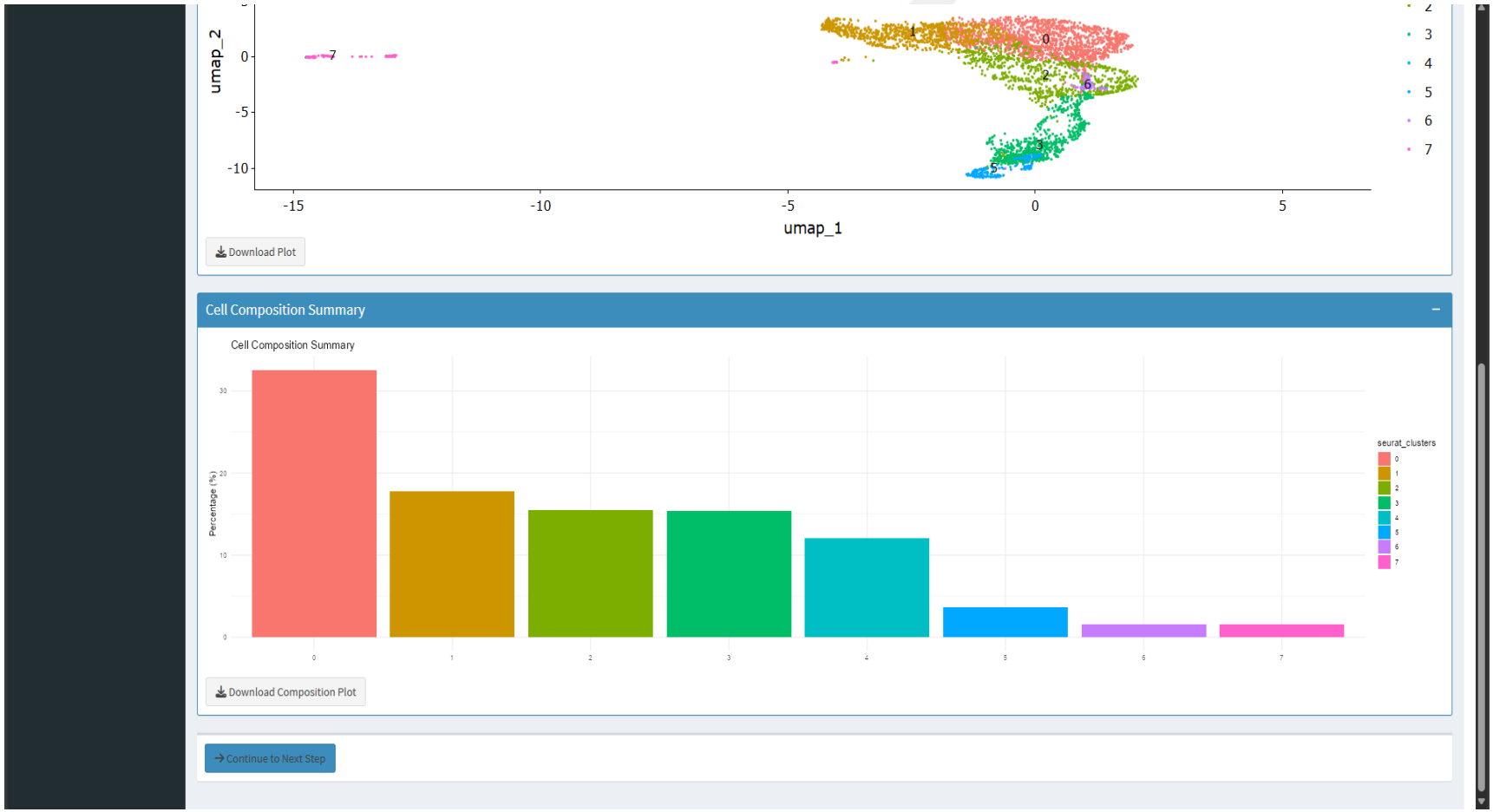


Figure 3.17 2D Umap visualization of clusters and cell composition summary plot

3.1.6 Data Correction

The data correction step in Heraclomics is designed to refine single-cell RNA sequencing (scRNA-seq) datasets by addressing unwanted technical and biological biases, such as cell cycle effects, batch-specific artifacts, and other confounding factors. This step ensures that subsequent analyses focus solely on true biological variation. The interactive interface, as depicted in Figure 3.18, provides users with multiple parameters to customize the correction process. Users can select specific effects to regress, such as cell cycle phases (S and G2/M scores), mitochondrial gene percentages, ribosomal content, or nFeature_RNA. Users can also refine datasets by specifying genes of interest or entire pathways for correction. For instance, Figure 3.18 demonstrates how a user-selected pathway, "Cytoplasmic Ribosomal Proteins," is corrected alongside gene-specific corrections for "RPLP1." Such targeted corrections are invaluable for removing specific sources of variability that may obscure meaningful insights.

The outputs of this step are diverse and visually informative. The UMAP Plot After Correction (Figure 3.19) displays the updated clustering structure post-correction, emphasizing the impact of regression on dataset harmonization. Additional visualizations, such as Violin Plots for gene expression distribution (Figure 3.19) and Ridge Plots for cell cycle phase distribution (Figure 3.19), provide detailed summaries of how the data has been adjusted. These outputs enable users to assess the effectiveness of the correction step and ensure data quality before proceeding.

Heraclomics integrates state-of-the-art tools such as Seurat's regression methods for cell cycle and pathway correction and employs advanced visualization libraries like ggplot2 and Plotly to create interactive, user-friendly outputs. The platform's dynamic features, powered by R Shiny, allow real-time updates and feedback on parameter adjustments, ensuring that researchers can iteratively refine their datasets for optimal results. By addressing confounding factors systematically, the data correction module ensures that downstream analyses are robust, reliable, and biologically meaningful.

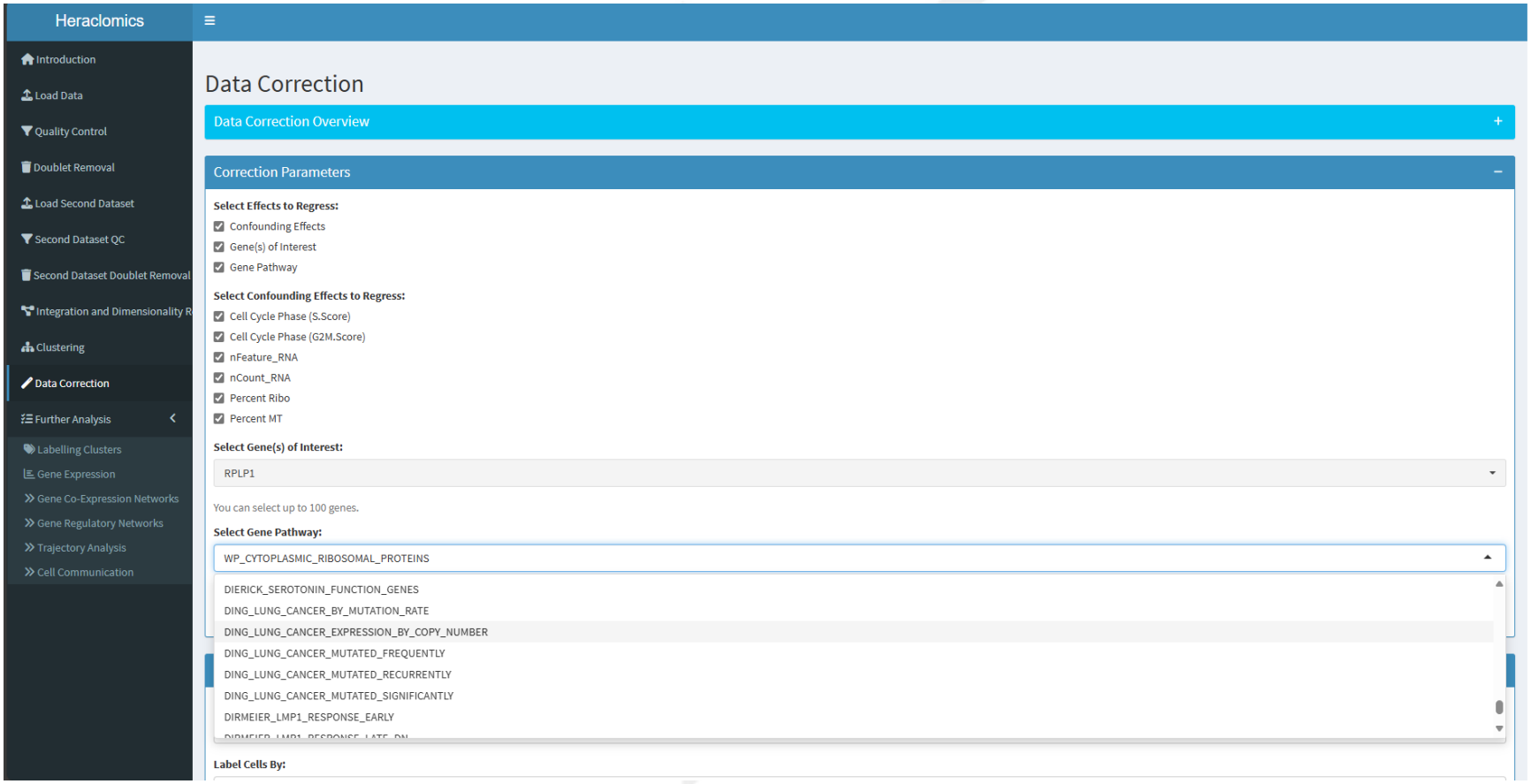


Figure 3.18 Data correction interface and parameters selection step

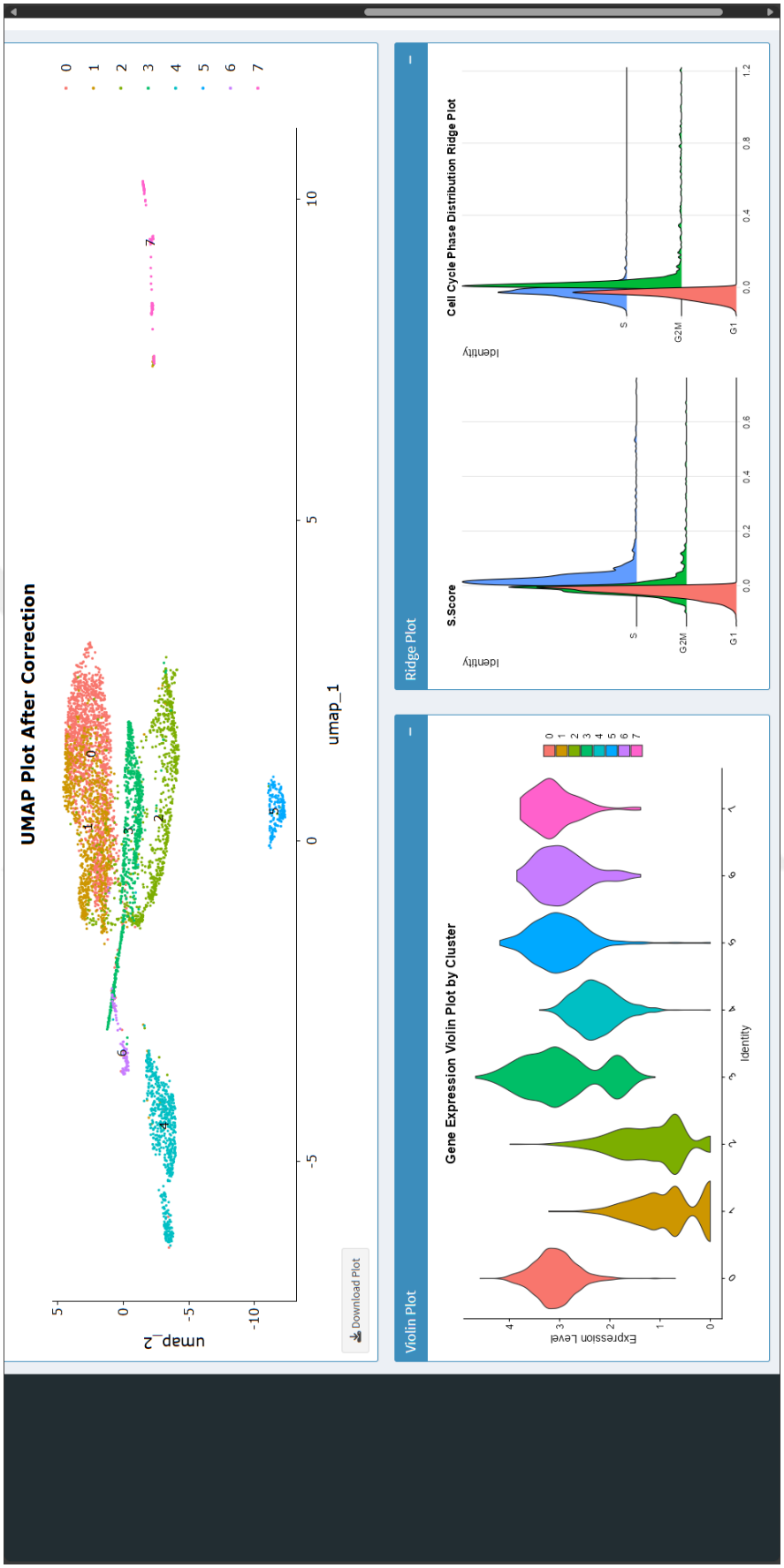


Figure 3.19 Data correction UMAP, violin plot for RPLP1 and ridge plot output

3.1.7 Cluster Labelling

The cluster labeling step in Heraclomics assigns meaningful biological identities to the clusters identified during the analysis process. Users can label clusters interactively using reference datasets such as the Human Primary Cell Atlas, Blueprint Encode, Database Immune Cell Expression, ImmGen, Monaco Immune Data, Novershtern Hematopoietic Data, Mouse RNAseq, and ENCODE, as shown in Figure 3.20. This feature enables researchers to characterize cell types accurately within their dataset.

The results include a cluster table that displays each cluster's label, providing a clear and concise summary of cluster identities. Additionally, the UMAP plot in Figure 3.21 visualizes the labeled clusters, with each cluster colored and labeled according to its assigned cell type. This visualization facilitates the identification of distinct cell populations and their spatial relationships within the dataset.

Users can refine cluster labels by incorporating their marker genes or selecting different reference datasets to optimize the annotation. Heraclomics offers an easy-to-use interface for this task, enhancing the efficiency and accuracy of the labeling process. The ability to download cluster labels further supports reproducibility and downstream analysis.

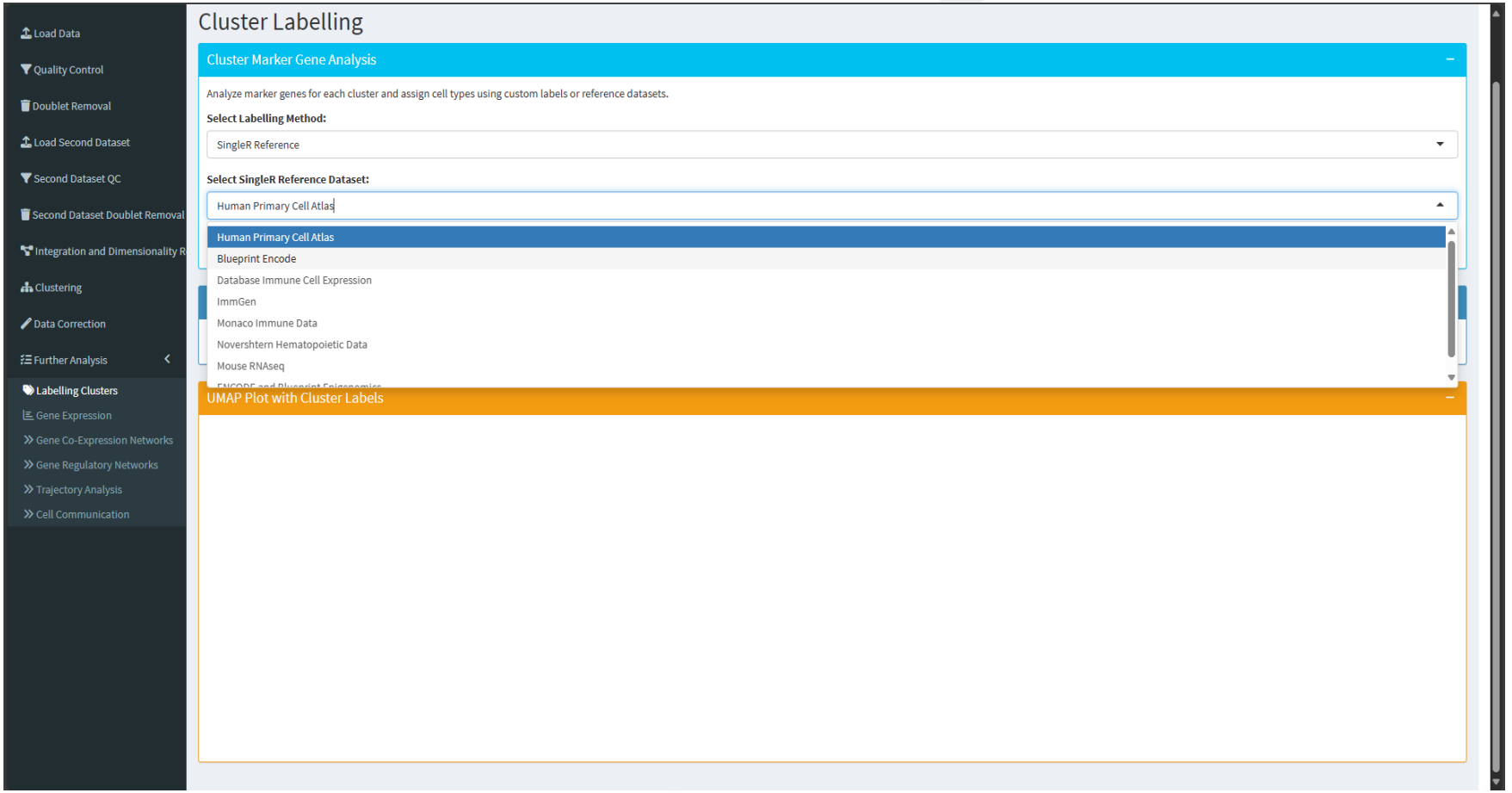


Figure 3.20 Cluster labelling interface and database selection

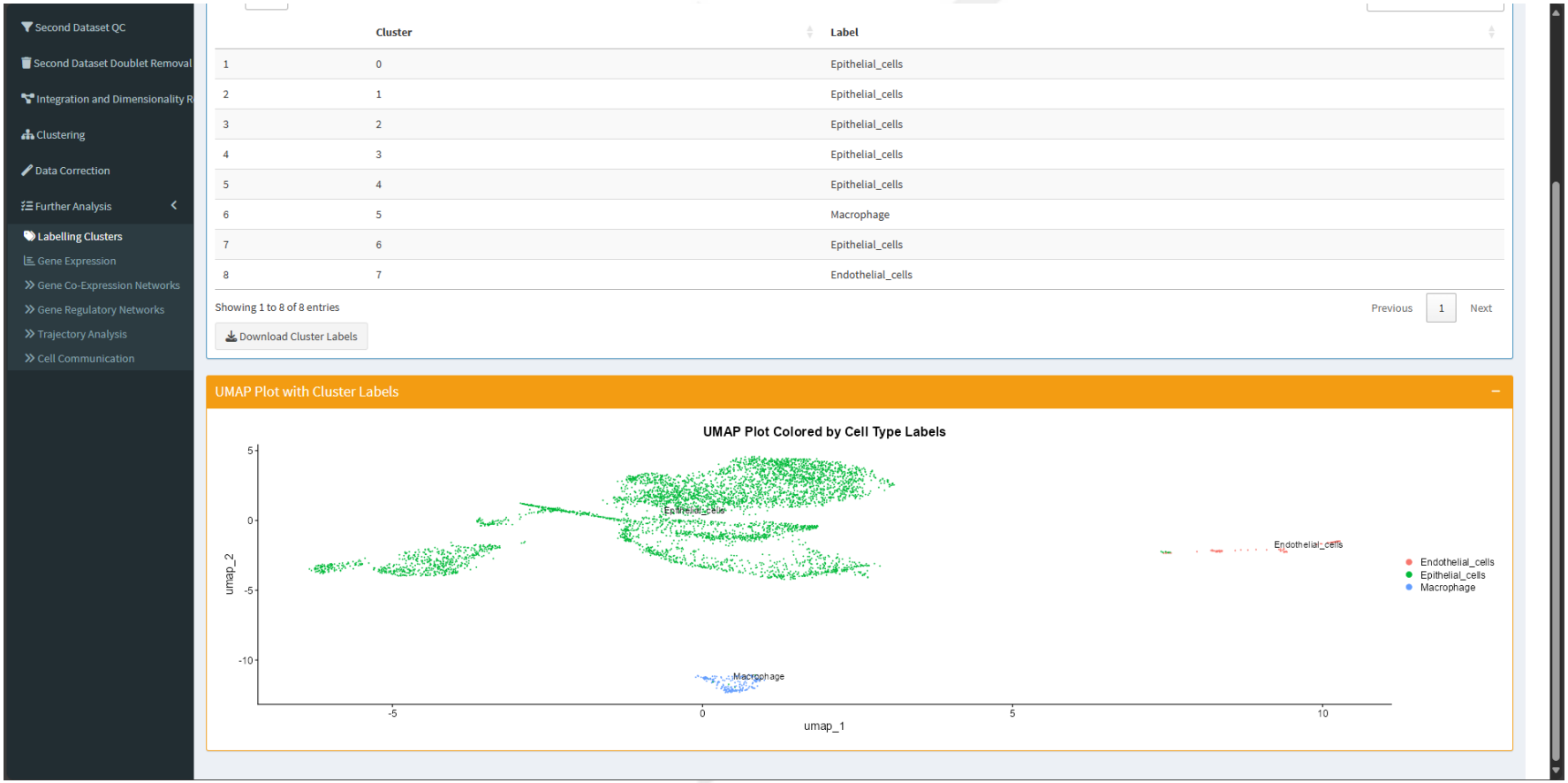


Figure 3.21 Cluster labelling cluster label result table and UMAP plot

3.1.8 Gene Co-expression Network

The Gene Co-Expression Networks step in Heraclomics allows users to explore gene co-expression patterns by constructing and analyzing multiscale co-expression networks. Using MEGENA (Multiscale Embedded Gene Co-Expression Network Analysis), the platform provides detailed insights into gene co-regulation and functional interactions. The user interface, as depicted in Figure 3.22, provides customizable options for network construction, such as selecting the input method (e.g., variable genes or differentially expressed genes), defining the number of genes, and adjusting parameters like p-value thresholds and module size. Figure 3.23 showcases the Network Plot, where nodes represent genes, and edges indicate co-expression relationships. This visualization offers a high-level view of the gene co-expression structure, enabling users to detect densely connected regions that may correspond to biologically significant modules. The Module Summary tab, shown in Figure 3.24, presents a detailed table summarizing the detected gene modules. This table includes information on module size, parent-child relationships, hub genes, and p-values, allowing researchers to focus on modules with significant biological implications. The Gene-Module Membership tab (Figure 3.25) provides a comprehensive list of individual genes and their associated modules. This searchable table allows users to quickly locate specific genes and understand their roles within the network. In Figure 3.26, the Module Visualization tab highlights the structure of individual modules, showing the connections between genes within a selected module. This targeted view helps researchers focus on specific modules of interest and investigate their internal connectivity. The Enrichment Analysis tab (Figure 3.27) offers insights into the functional relevance of gene modules. Enriched Gene Ontology (GO) terms are listed in a table, including details such as fold enrichment and adjusted p-values. Accompanying visualizations, such as the bar plot and bubble plot in Figure 3.28, highlight the top enriched terms for selected modules, enabling users to interpret the biological significance of their findings. Heraclomics integrates advanced tools for network construction and analysis, leveraging MEGENA for module detection and R-based visualization libraries like ggplot2 and Plotly for interactive outputs.

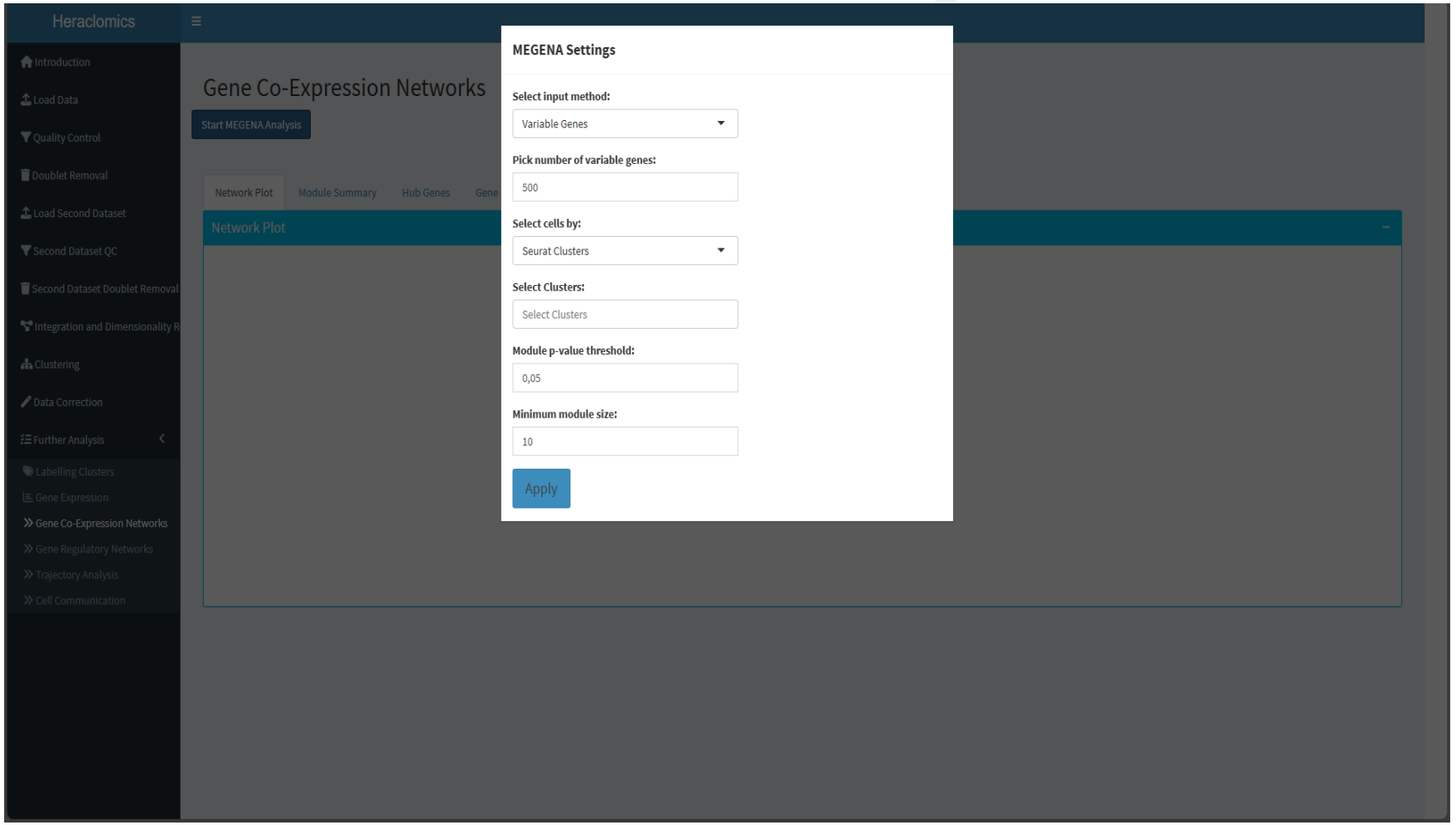


Figure 3.22 Gene co-expression networks MEGENA parameters

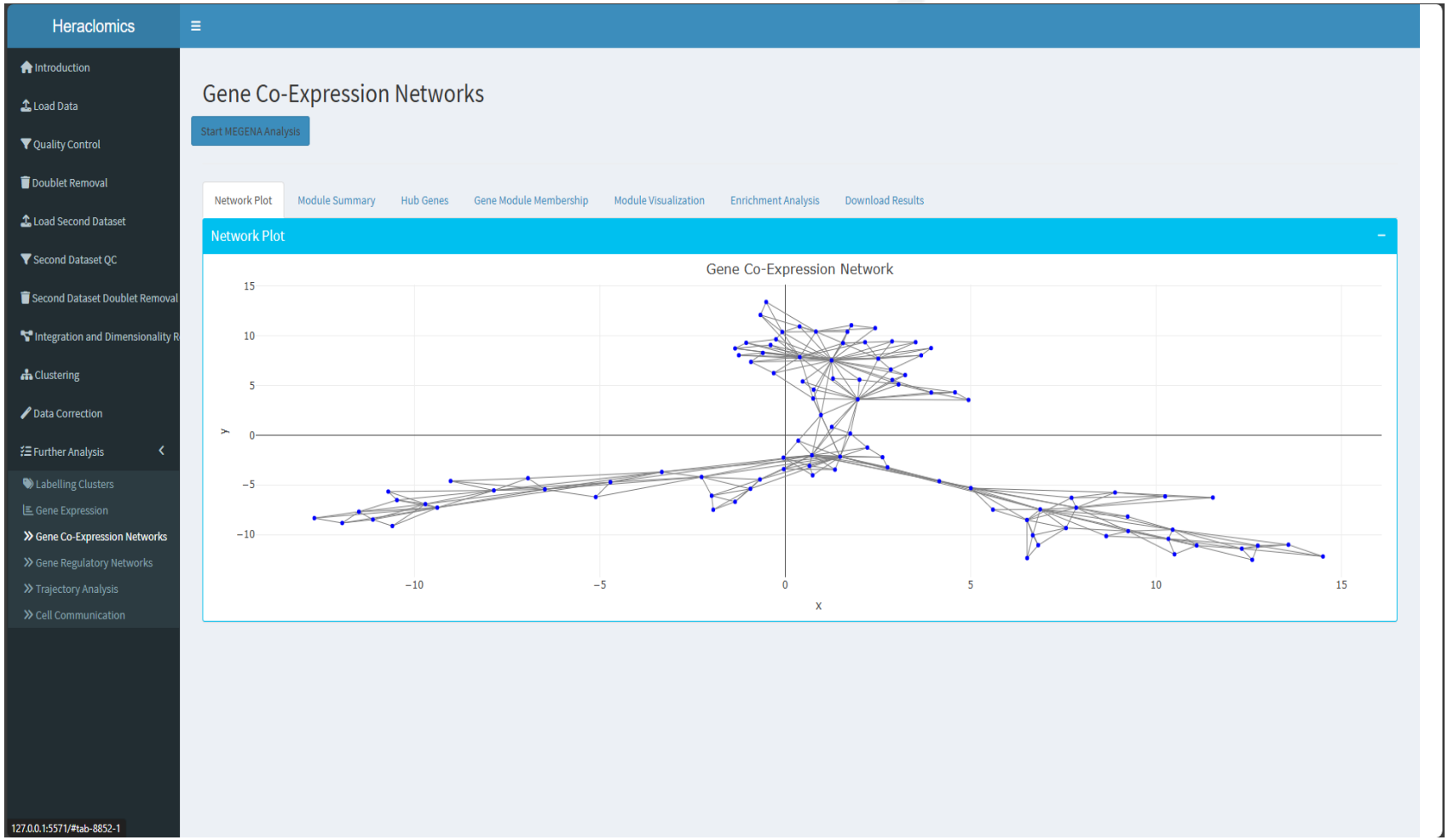


Figure 3.23 Gene co-expression networks network plot for selected parameters

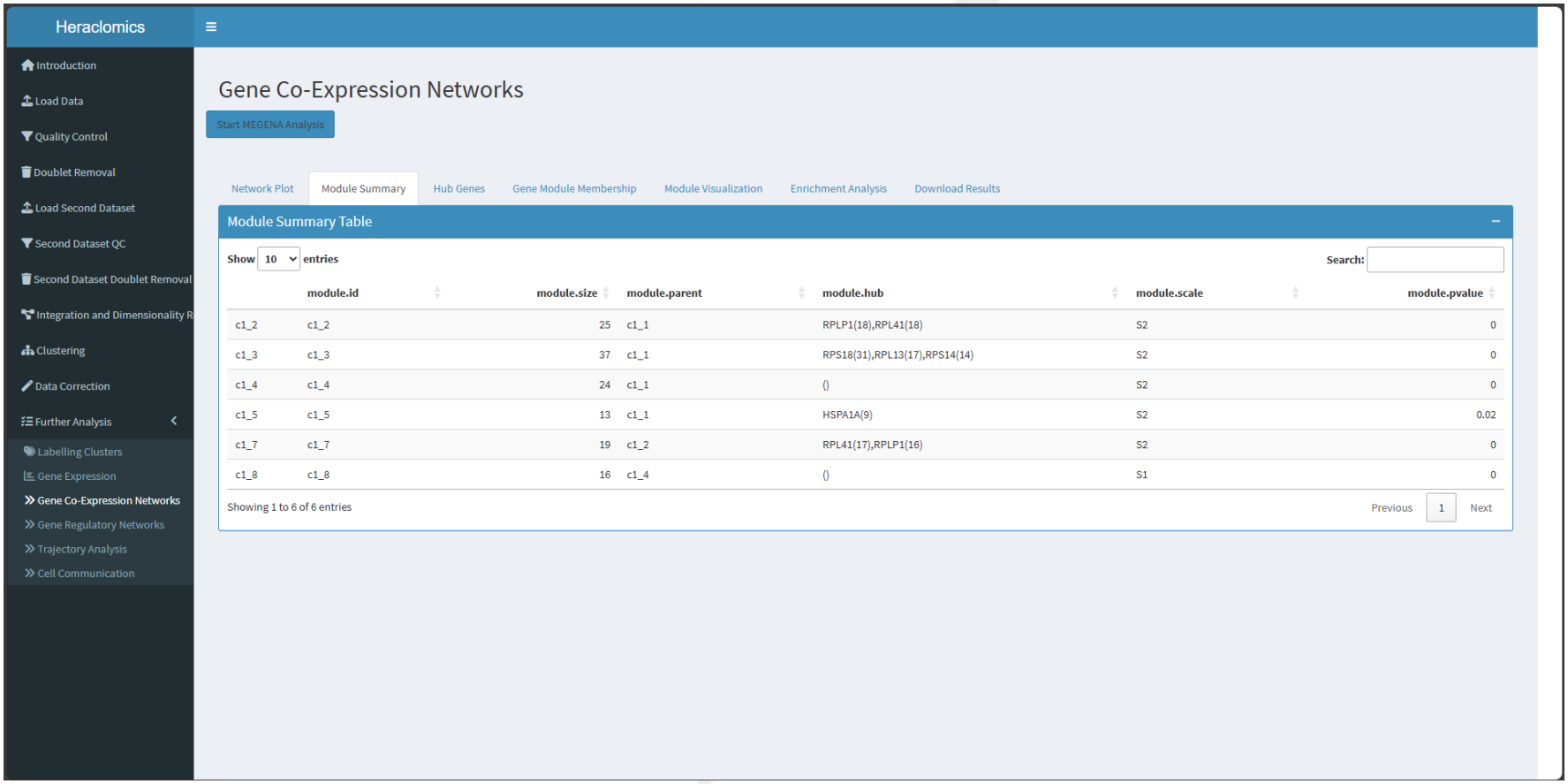


Figure 3.24 Gene co-expression networks module summary table

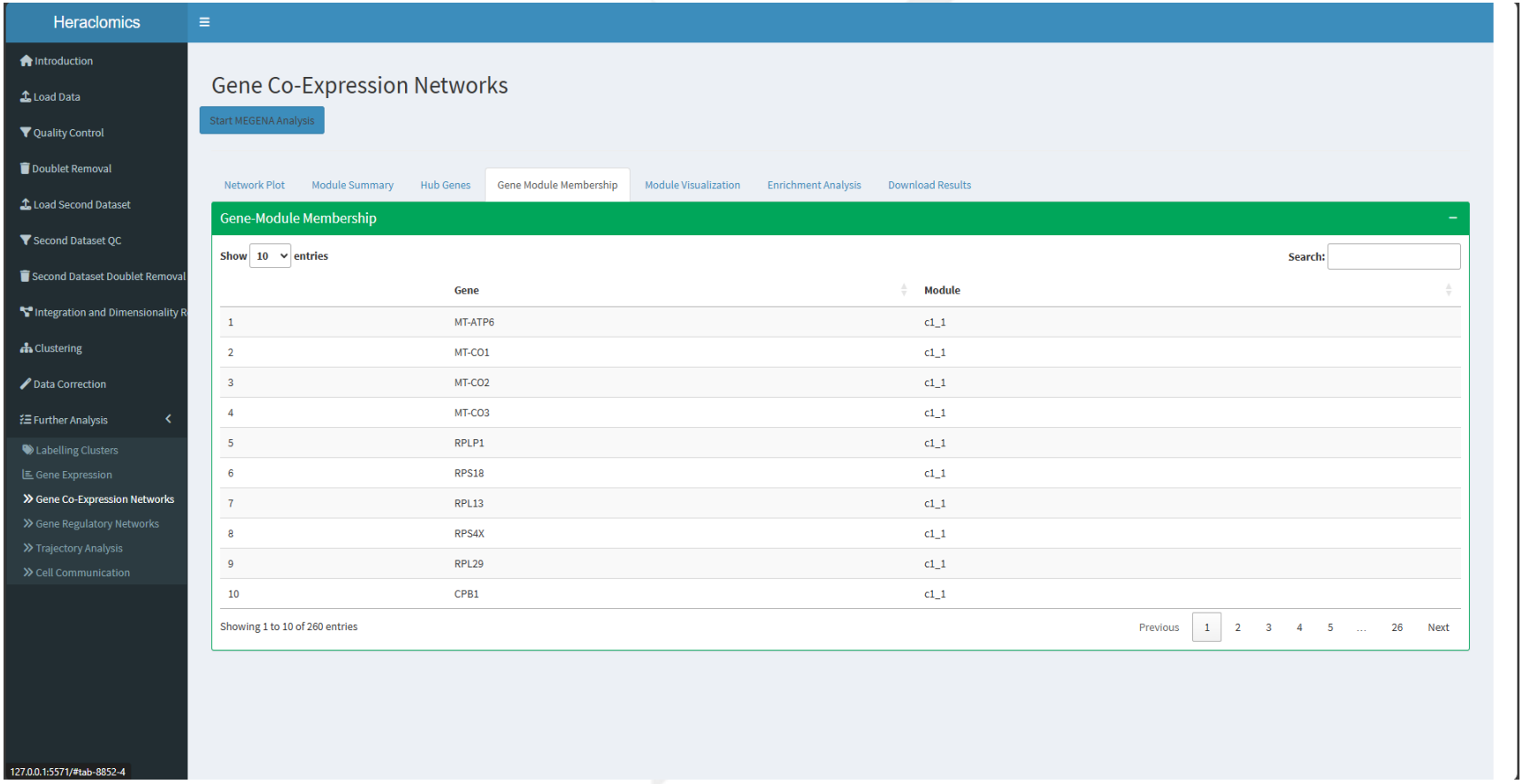


Figure 3.25 Gene co-expression networks gene-module membership table first page

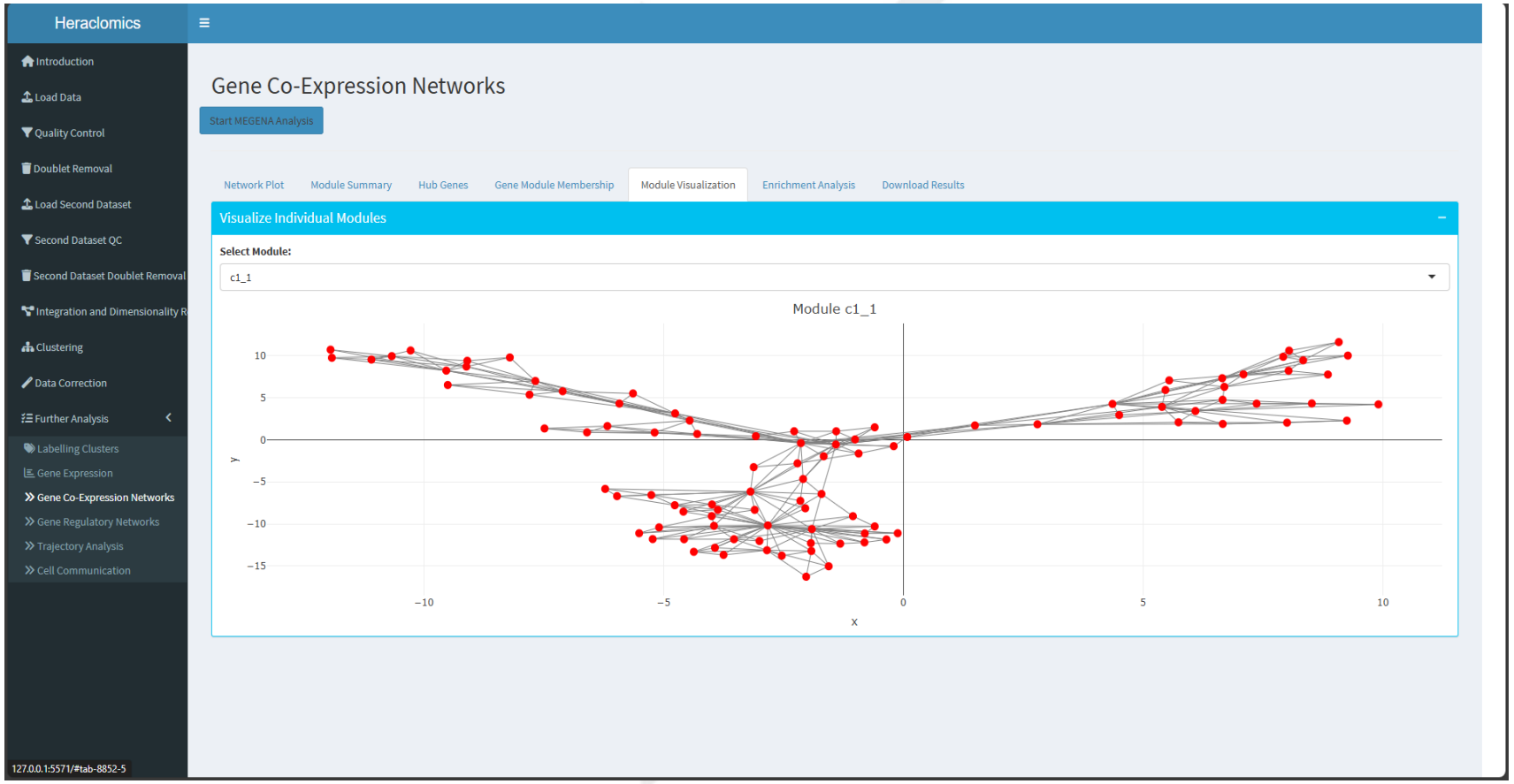
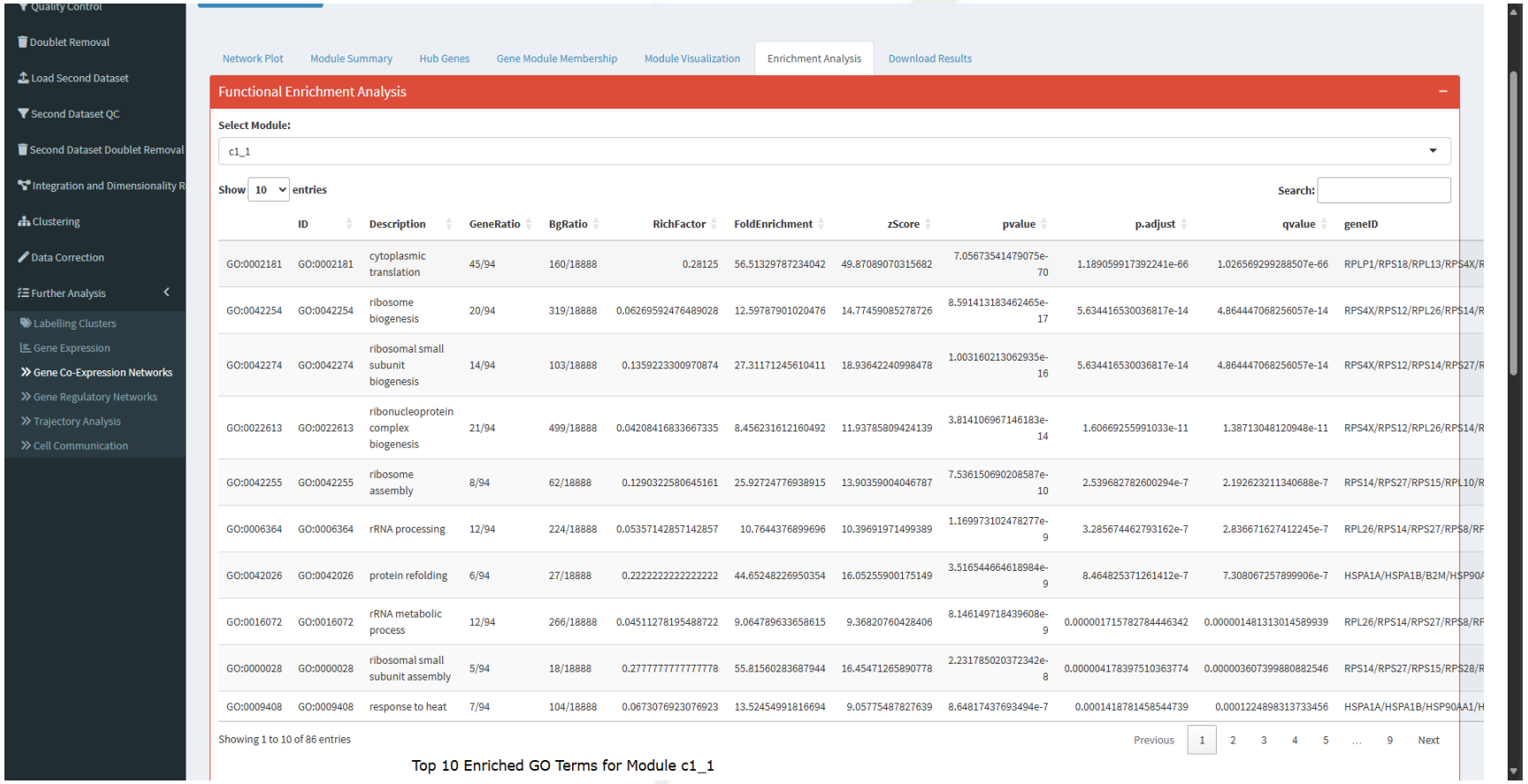


Figure 3.26 Gene co-expression networks individual module network and interactive module selection



	ID	Description	GeneRatio	BgRatio	RichFactor	FoldEnrichment	zScore	pvalue	p.adjust	qvalue	geneID
GO:0002181	GO:0002181	cytoplasmic translation	45/94	160/18888	0.28125	56.51329787234042	49.8708907031568 2	7.05673541479075e-70	1.189059917392241e-66	1.026569299288507e-66	RPLP1/RPS18/RPL13/RPS4X/R
GO:0042254	GO:0042254	ribosome biogenesis	20/94	319/18888	0.0626959247648902 8	12.59787901020476	14.7745908527872 6	8.5914131834624 65e-17	5.634416530036817e-14	4.864447068256057e-14	RPS4X/RPS12/RPL26/RPS14/R
GO:0042274	GO:0042274	ribosomal small subunit biogenesis	14/94	103/18888	0.1359223300970874	27.31171245610411	18.9364224099847 8	1.003160213062935e-16	5.634416530036817e-14	4.864447068256057e-14	RPS4X/RPS12/RPS14/RPS27/R
GO:0022613	GO:0022613	ribonucleoprotein complex biogenesis	21/94	499/18888	0.0420841683366733 5	8.456231612160492	11.9378580942413 9	3.814106967146183e-14	1.60669255991033e-11	1.38713048120948e-11	RPS4X/RPS12/RPL26/RPS14/R
GO:0042255	GO:0042255	ribosome assembly	8/94	62/18888	0.1290322580645161	25.92724776938915	13.9035900404678 7	7.536150690208587e-10	2.539682782600294e-7	2.192623211340688e-7	RPS14/RPS27/RPS15/RPL10/R
GO:0006364	GO:0006364	rRNA processing	12/94	224/18888	0.0535714285714285 7	10.7644376899696	10.3969197149938 9	1.169973102478277e-9	3.285674462793162e-7	2.836671627412245e-7	RPL26/RPS14/RPS27/RPS8/RF
GO:0042026	GO:0042026	protein refolding	6/94	27/18888	0.2222222222222222	44.65248226950354	16.0525590017514 9	3.516544664618984e-9	8.464825371261412e-7	7.308067257899906e-7	HSPA1A/HSPA1B/B2M/HSP90A
GO:0016072	GO:0016072	rRNA metabolic process	12/94	266/18888	0.0451127819548872 2	9.064789633658615	9.36820760428406	8.146149718439608e-9	0.000001715782784446342	0.000001481313014589939	RPL26/RPS14/RPS27/RPS8/RF
GO:0000028	GO:0000028	ribosomal small subunit assembly	5/94	18/18888	0.2777777777777778	55.81560283687944	16.4547126589077 8	2.231785020372342e-8	0.000004178397510363774	0.000003607399880882546	RPS14/RPS27/RPS15/RPS28/R
GO:0009408	GO:0009408	response to heat	7/94	104/18888	0.0673076923076923	13.52454991816694	9.05775487827639	8.64817437693494e-7	0.0001418781458544739	0.0001224898313733456	HSPA1A/HSPA1B/HSP90AA1/H

Figure 3.27 Gene co-expression networks functional enrichment analysis selected module output table

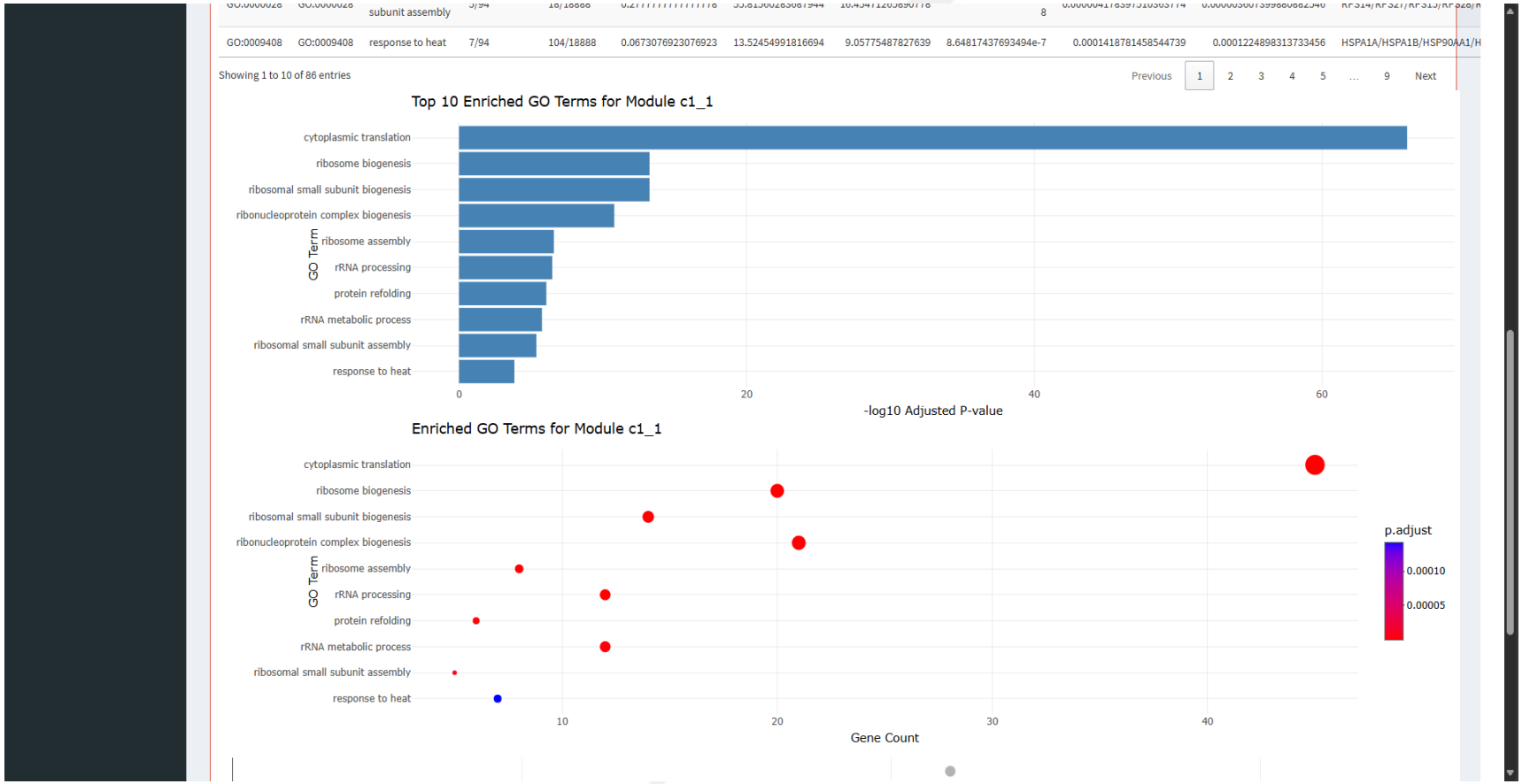


Figure 3.28 Gene co-expression networks bar plot and bubble plot for enriched GO terms of selected module

3.1.9 Cell Communication

The Cell Communication Analysis module in Heraclomics is a specialized step designed to explore and quantify intercellular communication networks in single-cell RNA sequencing (scRNA-seq) data. The interactive interface, as depicted in Figure 3.29, allows users to select specific cell clusters and signaling pathways for analysis. Users can initiate the analysis by choosing clusters (e.g., 0, 1, 2, etc.) and signaling pathways of interest. Figure 3.30 illustrates the resulting Global Communication Network Visualization, where nodes represent cell clusters, and edges represent ligand-receptor-mediated signaling interactions. The thickness of the edges reflects the strength of communication, enabling users to identify key signaling hubs. The module also provides a Specific Signaling Pathway Visualization (Figure 3.30), which focuses on selected pathways, such as "MIF," and their interactions across cell clusters. This pathway-specific plot highlights the directionality and intensity of interactions, allowing users to explore how particular pathways contribute to cellular communication. To further analyze signaling patterns, a Heatmap of Outgoing Signaling Patterns (Figure 3.31) summarizes the activity levels of different pathways across cell clusters. Users can evaluate which clusters act as dominant signal senders or receivers, providing a comprehensive overview of communication dynamics. Additionally, the Bubble Plot of Interactions (Figure 3.31) visualizes interaction probabilities and pathway enrichment, with bubbles color-coded by adjusted p-values and communication strengths. Finally, the module includes a Significant Ligand-Receptor Interactions Table (Figure 3.32), which lists detailed information about source and target clusters, ligand-receptor pairs, interaction probabilities, pathway annotations, and supporting evidence from external databases such as KEGG and PubMed. This table can be exported for downstream analyses. Heraclomics employs cutting-edge tools like CellChat to calculate and visualize intercellular communication patterns. By uncovering intricate cellular signaling dynamics, the Cell Communication Analysis module empowers researchers to generate testable hypotheses and gain deeper insights into the complex interactions shaping biological systems.

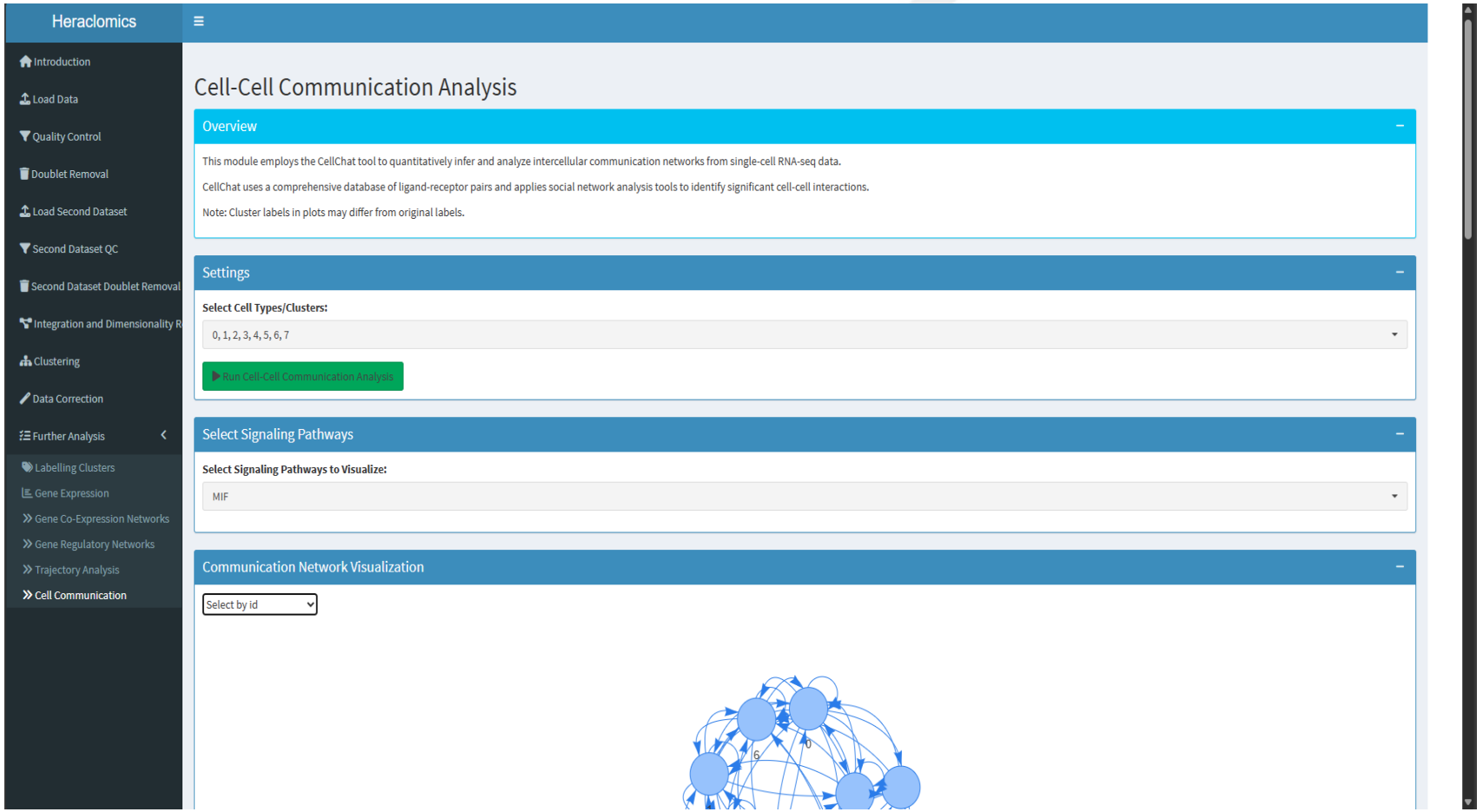


Figure 3.29 Cell-cell communication cluster and signaling pathways selection with interface

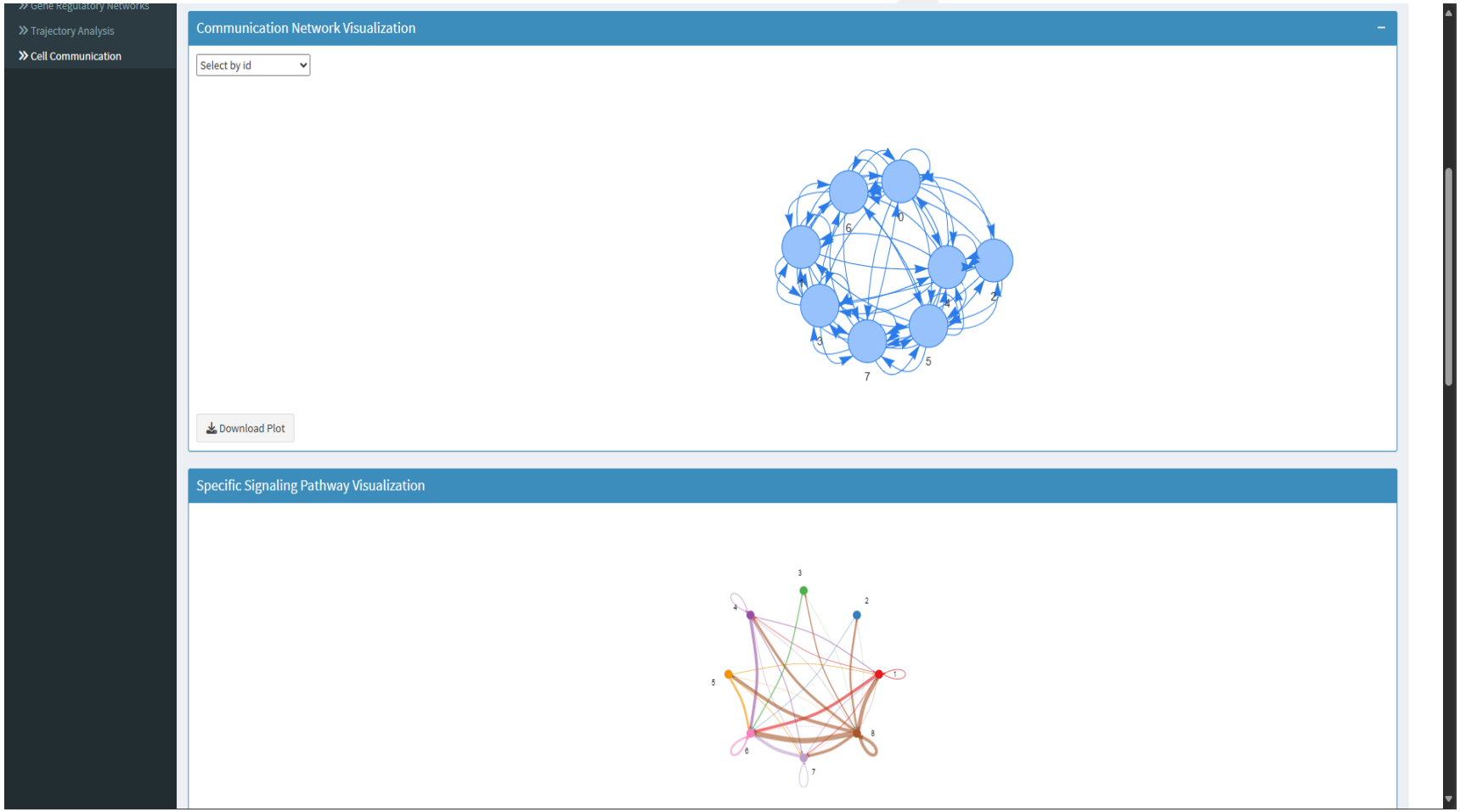


Figure 3.30 Cell-cell communication network visualization and pathway visualization according to cell clusters

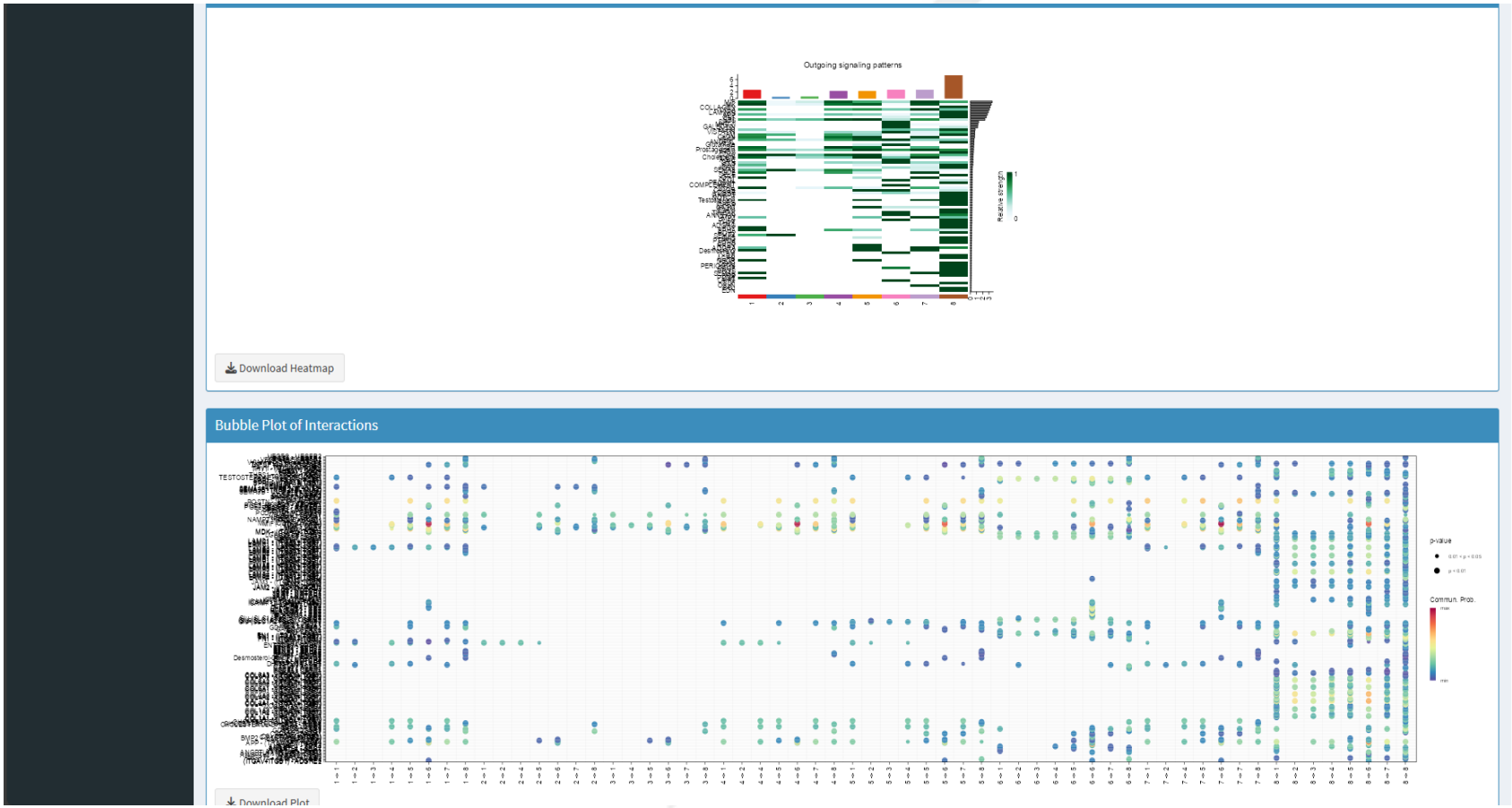


Figure 3.31 Cell-cell communication heatmap of signalling pattern between cell and interaction buble plot of communication probabilities

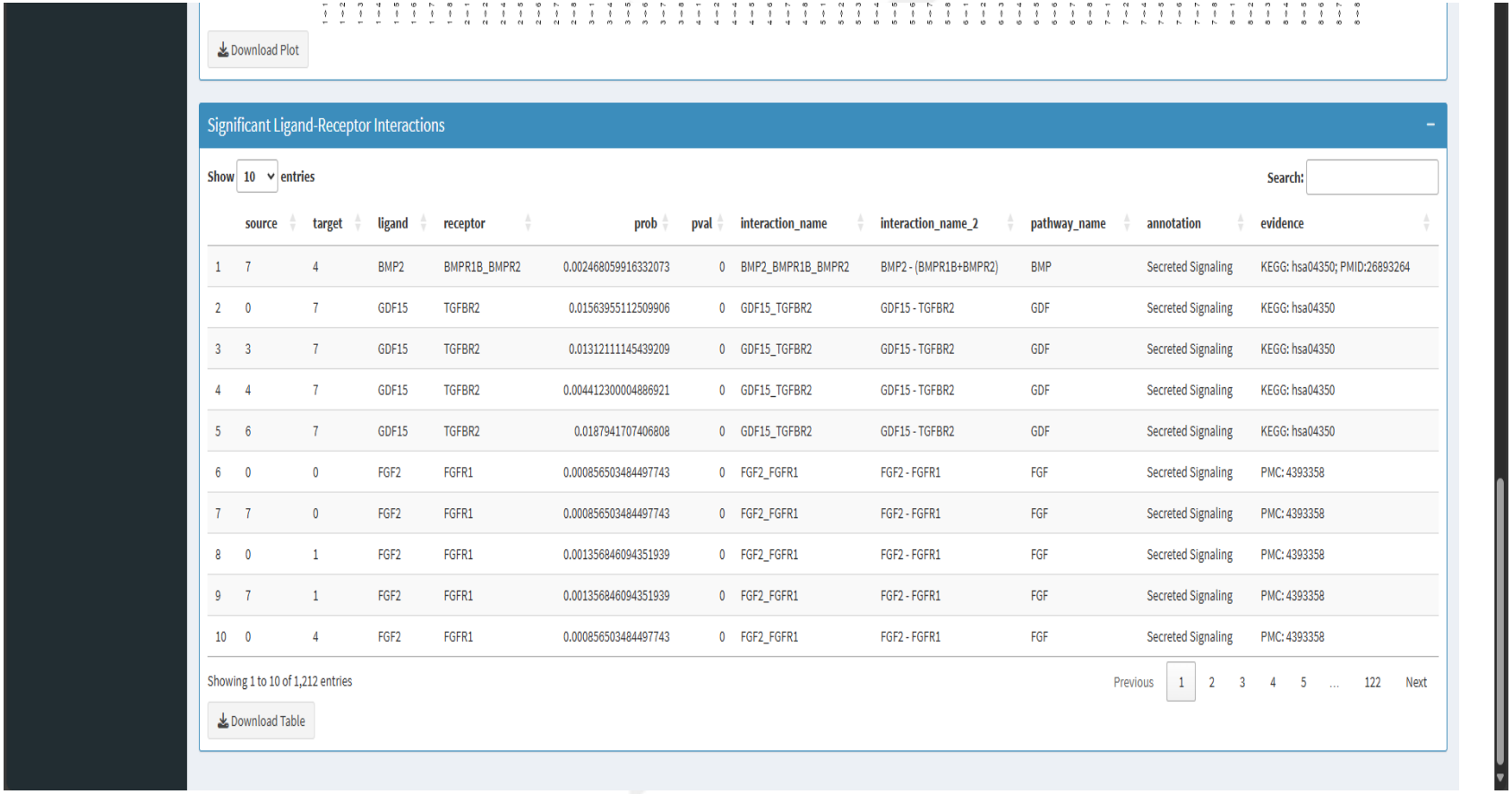


Figure 3.32 Cell-cell communication significant ligand-receptor interaction table with evidence references

3.1.10 Differential Expression

Differential gene expression (DE) analysis is a crucial step in single-cell RNA sequencing (scRNA-seq) workflows, enabling the identification of genes that are significantly upregulated or downregulated between different cell clusters or experimental conditions. In Heraclomics, multiple approaches were implemented to perform differential expression analysis, providing flexibility depending on the dataset size and research objectives.

Seurat's built-in methods, including the Wilcoxon rank-sum test, ROC analysis, and MAST modeling, were utilized for quick and interactive cluster-level comparisons. These methods allow for the identification of marker genes specific to different cell populations. In addition, DESeq2 was integrated for more statistically rigorous analyses, especially in scenarios where pseudobulk comparisons or grouped sample analyses were appropriate. DESeq2 models gene expression using a negative binomial distribution, providing robust normalization and variance estimation. Furthermore, edgeR was employed to offer additional flexibility for complex experimental designs, particularly those involving biological replicates. edgeR also uses a negative binomial model and supports TMM normalization, making it well-suited for datasets requiring precise normalization and dispersion handling. The outputs generated from differential expression analysis include volcano plots, MA plots, heatmaps, violin plots, and dot plots, which collectively allow users to visualize the most significantly altered genes and their expression profiles across clusters or conditions. Through these analyses, Heraclomics enables users to define cell identities, explore biological processes, and prioritize genes for downstream functional enrichment studies. By integrating multiple statistical approaches, Heraclomics ensures both speed and analytical robustness, providing a comprehensive and adaptable platform for differential expression analysis in single-cell research.

3.2 Validation and Benchmarking

Validation and Benchmarking for Heraclomics evaluates its functionality, accuracy, and scalability by analyzing publicly available scRNA-seq datasets and

comparing its performance to two established platforms: ICARUS and ezSingleCell. Validation was performed using breast cancer and immune cell datasets sourced from the 10X Genomics repository. These datasets provided diverse cell types, batch effects, and varying levels of heterogeneity, allowing a comprehensive assessment of Heraclomics across its analytical modules. Heraclomics was benchmarked against ICARUS (Jiang et al., 2022), a web-based scRNA-seq analysis platform known for its user-friendly interface and visualization capabilities, and ezSingleCell (Sethi et al., 2024), which offers a streamlined workflow for scRNA-seq data analysis. The comparison focused on the breadth of analytical steps, library integration, and output interactivity provided by each platform. Heraclomics utilizes Seurat for clustering, offering flexible resolution adjustments and UMAP/t-SNE visualizations in 2D and 3D. Clustering accuracy was validated using ARI and NMI, with Heraclomics providing comparable results to ICARUS and slightly outperforming ezSingleCell in datasets with high cellular heterogeneity. ICARUS provides clustering results based on Seurat but lacks interactive resolution adjustments. ezSingleCell simplifies clustering but offers limited customization and interactivity. Heraclomics integrates Harmony for batch effect correction, achieving seamless integration of datasets with diverse experimental conditions. ICARUS offers batch correction but uses fewer integration tools, resulting in less flexibility for large, complex datasets. ezSingleCell does not support robust batch correction, making it less suitable for datasets with significant batch effects. Heraclomics employs monocle3 and UMAP-based trajectory visualizations, allowing users to infer pseudotime with high accuracy. Its pseudotime correlation scores were consistent with ICARUS. ICARUS provides trajectory analysis using UMAP but lacks pseudotime-specific gene expression visualizations. ezSingleCell does not include a trajectory analysis module. Heraclomics integrates CellChat for cell-cell communication analysis, offering network visualizations, signaling pathway plots, and ligand-receptor interaction tables. It provided results comparable to ICARUS but with superior interactive features. ICARUS includes cell communication analysis but does not support pathway-specific visualizations or interactivity. ezSingleCell lacks a dedicated cell communication module. Heraclomics uses DESeq2 for robust differential expression analysis, offering interactive volcano plots, MA plots, and

heatmaps. Its results aligned with ICARUS and ezSingleCell, with added interactivity for exploring gene-specific results.

ICARUS and ezSingleCell both provide differential expression results but lack the interactive features and visualization options available in Heraclomics. Heraclomics excels in interactivity by leveraging libraries like plotly, ggplot2, and shinyWidgets. Users can dynamically adjust parameters and explore results in real time. ICARUS offers a static visualization approach, while ezSingleCell provides basic visualizations with minimal interactivity. Breadth of Analysis: Heraclomics offers a more comprehensive suite of modules than ezSingleCell and is comparable to ICARUS, with additional features like advanced batch correction and regulatory network analysis. Heraclomics outperforms ezSingleCell in providing dynamic visualizations and customizable workflows and comparable to ICARUS. Heraclomics matched or exceeded ICARUS in clustering accuracy and cell communication analysis while significantly outperforming ezSingleCell in batch correction and trajectory inference. Validation using breast cancer and immune cell datasets demonstrated Heraclomics' ability to uncover biologically meaningful insights, such as novel ligand-receptor interactions and pseudotime-specific gene expression changes. Its modular and interactive design makes it an ideal platform for researchers requiring a comprehensive yet accessible scRNA-seq analysis tool.

4
CONCLUSION

4.1 Conclusion

The Heraclomics platform represents a transformative step forward in the field of single-cell RNA sequencing (scRNA-seq) analysis, providing a comprehensive and interactive toolkit designed to address the challenges inherent in analyzing complex biological systems. By combining state-of-the-art computational tools, dynamic visualizations, and a user-friendly interface, Heraclomics empowers researchers to derive meaningful biological insights with precision and efficiency, regardless of their computational expertise.

Heraclomics bridges a critical gap by democratizing access to advanced scRNA-seq analysis for researchers, biologists, and scientists who may lack programming or bioinformatics skills. Its primary goal is to foster inclusivity and collaboration across disciplines by lowering the barriers to entry for scRNA-seq data analysis. The platform's intuitive and modular design streamlines the entire workflow, from data loading and quality control to advanced analyses like cell communication, differential expression, and co-expression network modeling.

Each module within Heraclomics is meticulously designed to provide flexibility and customization, allowing users to tailor parameters to the specific requirements of their datasets. For instance, steps such as data correction address unwanted technical and biological biases, while cluster labeling and pathway-specific analyses emphasize biological relevance. Advanced visualization tools ensure that users can intuitively explore and interpret results, facilitating hypothesis generation and data-driven discoveries. Throughout this process, Heraclomics integrates cutting-edge algorithms, real-time feedback, and adjustable parameters, enabling iterative refinement and exploration of data.

By enabling seamless integration of multiple datasets, uncovering intricate signaling pathways, and revealing gene regulatory and co-expression networks, Heraclomics accelerates research across diverse fields, including developmental biology, immunology, and disease pathology. The platform's scalability ensures that it caters to the needs of both novice users and experienced bioinformaticians, making it a versatile tool for a wide range of applications.

Heraclomics not only simplifies the technical complexities of scRNA-seq analysis but also aims to serves as a bridge between data analysis and experimental science. By allowing scientists to focus on answering critical biological questions rather than navigating technical challenges, the platform enhances the accessibility, efficiency, and collaborative potential of single-cell research. Ultimately, Heraclomics contributes to the acceleration of scientific discovery and the advancement of precision medicine by providing a robust, interactive, and accessible framework for scRNA-seq analysis.

In conclusion, Heraclomics equips researchers with the tools needed to unlock the potential of scRNA-seq data, transforming raw data into actionable biological insights. By addressing the needs of both technical and non-technical users, Heraclomics drives forward our understanding of cellular heterogeneity, intercellular interactions, and the underlying mechanisms of health and disease. This platform serves as a resource for advancing single-cell research and empowering scientists to make meaningful contributions in the era of precision medicine.

REFERENCES

10X Genomics. (n.d.). Human Breast Cancer Block A, Section 1-1.1 Standard. Retrieved January 2, 2025, from https://www.10xgenomics.com/datasets/human-breast-cancer-block-a-section-1-1-standard-1-1-0

Aibar, S., González-Blas, C. B., Moerman, T., Imrichova, H., Hulselmans, G., Rambow, F., ... & Aerts, S. (2017). SCENIC: Single-cell regulatory network inference and clustering. Nature Methods, 14(11), 1083–1086. https://doi.org/10.1038/nmeth.4463

Andrew Jiang, Klaus Lehnert, Linya You, Russell G Snell, ICARUS, an interactive web server for single cell RNA-seq analysis, Nucleic Acids Research, 50(W1), W427–W433. https://doi.org/10.1093/nar/gkac322

Aran, D., Looney, A. P., Liu, L., Wu, E., Fong, V., Hsu, A., ... & Butte, A. J. (2019). Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage. Nature Immunology, 20(2), 163–172. https://doi.org/10.1038/s41590-018-0276-y

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., ... & Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. Nature Biotechnology, 37(1), 38–44. https://doi.org/10.1038/nbt.4314

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., ... & Shendure, J. (2019). The single-cell transcriptional landscape of mammalian organogenesis. Nature, 566(7745), 496–502. https://doi.org/10.1038/s41586-019-0969-x

Carlson, M. (2023). org.Hs.eg.db: Genome wide annotation for Human. R package version 3.16.0. Retrieved from https://bioconductor.org/packages/release/data/annotation/html/org.Hs.eg.db.html

Castro, M. A. A., Colaprico, A., Barault, L., Cava, C., Olsen, C., Testa, G., ... & Cava, C. (2016). RTN: An R/Bioconductor package for regulatory transcriptional network analysis. Bioinformatics, 32(9), 1477–1479. https://doi.org/10.1093/bioinformatics/btv737

Chang, W., Cheng, J., Allaire, J., Xie, Y., & McPherson, J. (2021). Shiny: Web application framework for R. R package version 1.7.1. Retrieved from https://CRAN.R-project.org/package=shiny

Csardi, G., & Nepusz, T. (2006). The igraph software package for complex network research. InterJournal, Complex Systems, 1695.

De Jonghe, J., Opzoomer, J. W., Vilas-Zornoza, A., Nilges, B. S., Crane, P., Vicari, M., ... & Taylor-King, J. P. (2024). scTrends: A living review of commercial single-cell and spatial 'omic technologies. Cell Genomics, 4(12), 100723. https://doi.org/10.1016/j.xgen.2024.100723

Dowle, M., & Srinivasan, A. (2021). data.table: Extension of data.frame. R package version 1.14.2. Retrieved from https://CRAN.R-project.org/package=data.table

Guo, G., Huss, M., Tong, G. Q., Wang, C., Li Sun, L., Clarke, N. D., & Robson, P. (2010). Resolution of cell fate decisions revealed by single-cell gene expression analysis from zygote to blastocyst. Developmental Cell, 18(4), 675-685. https://doi.org/10.1016/j.devcel.2010.02.012

Haghverdi, L., Lun, A. T. L., Morgan, M. D., & Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nature Biotechnology, 36(5), 421–427. https://doi.org/10.1038/nbt.4091

Hafemeister, C., & Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biology, 20(1), 296. https://doi.org/10.1186/s13059-019-1874-1

Huynh-Thu, V. A., Irrthum, A., Wehenkel, L., & Geurts, P. (2010). Inferring regulatory networks from expression data using tree-based methods. PLoS ONE, 5(9), e12776. https://doi.org/10.1371/journal.pone.0012776

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., ... & Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. Nature Methods, 11(2), 163-166. https://doi.org/10.1038/nmeth.2772

Jin, S., Guerrero-Juarez, C. F., Zhang, L., Chang, I., Ramos, R., Kuan, C.-H., ... & Nie, Q. (2021). Inference and analysis of cell-cell communication using CellChat. Nature Communications, 12(1), 1088. https://doi.org/10.1038/s41467-021-21246-9

Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C., & Teichmann, S. A. (2015). The technology and biology of single-cell RNA sequencing. Molecular Cell, 58(4), 610-620. https://doi.org/10.1016/j.molcel.2015.04.005

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., ... & Raychaudhuri, S. (2019). Fast, sensitive, and accurate integration of single-cell data with Harmony. Nature Methods, 16(12), 1289–1296. https://doi.org/10.1038/s41592-019-0619-0

Lake, B. B., Ai, R., Kaeser, G. E., Salathia, N. S., Yung, Y. C., Liu, R., ... & Zhang, K. (2016). Neuronal subtypes and diversity revealed by single-nucleus RNA sequencing of the human brain. Science, 352(6293), 1586-1590. https://doi.org/10.1126/science.aaf1204

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P., & Tamayo, P. (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Systems, 1(6), 417–425. https://doi.org/10.1016/j.cels.2015.12.004

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology, 15(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., ... & Regev, A. (2015). Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell, 161(5), 1202-1214. https://doi.org/10.1016/j.cell.2015.05.002

Picelli, S., Faridani, O. R., Björklund, Å. K., Winberg, G., Sagasser, S., & Sandberg, R. (2013). Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nature Methods, 10(11), 1096-1098. https://doi.org/10.1038/nmeth.2639

Reyes, M., Filbin, M. R., Bhattacharyya, R. P., Billman, K., Eisenhaure, T., Hung, D. T., ... & Blainey, P. C. (2020). Single-cell RNA-seq reveals cell-specific resistance mechanisms to infectious agents. Nature Communications, 11(1), 312. https://doi.org/10.1038/s41467-019-13904-w

Sapub, S. (2024). Advances in bioinformatics for scRNA-seq. Journal of Bioinformatics, 12(1), 1–15. Retrieved from http://article.sapub.org/10.5923.j.bioinformatics.20241301.01.html

Sethi, R., Ang, K. S., Li, M., et al. (2024). ezSingleCell: An integrated one-stop single-cell and spatial omics analysis platform for bench scientists. Nature Communications, 15, 5600. https://doi.org/10.1038/s41467-024-48188-2

Sievert, C. (2020). Interactive web-based data visualization with R, plotly, and shiny. Boca Raton, FL: CRC Press.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W. M., ... & Satija, R. (2019). Comprehensive integration of single-cell data. Cell, 177(7), 1888–1902. https://doi.org/10.1016/j.cell.2019.05.031

Svensson, V., Natarajan, K. N., Ly, L.-H., Miragaia, R. J., Labalette, C., Macaulay, I. C., ... & Teichmann, S. A. (2017). Power analysis of single-cell RNA-sequencing experiments. Nature Methods, 14(4), 381–387. https://doi.org/10.1038/nmeth.4220

Tang, F., Barbacioru, C., Wang, Y., Nordman, E., Lee, C., Xu, N., ... & Surani, M. A. (2009). mRNA-Seq whole-transcriptome analysis of a single cell. Nature Methods, 6(5), 377-382. https://doi.org/10.1038/nmeth.1315

Tirosh, I., Izar, B., Prakadan, S. M., Wadsworth, M. H., Treacy, D., Trombetta, J. J., ... & Regev, A. (2016). Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science, 352(6282), 189-196. https://doi.org/10.1126/science.aad0501

Trapnell, C. (2015). Defining cell types and states with single-cell genomics. Genome Research, 25(10), 1491–1498. https://doi.org/10.1101/gr.190595.115

Vafadarnejad, E. (2021). Single-cell bioinformatics. University of Wurzburg. Retrieved from https://opus.bibliothek.uni-wuerzburg.de/opus4-wuerzburg/frontdoor/deliver/index/docId/26925/file/Vafadarnejad_Ehsan_SingleCellBioinformatics.pdf

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: A revolutionary tool for transcriptomics. Nature Reviews Genetics, 10(1), 57-63. https://doi.org/10.1038/nrg2484

Wickham, H. (2016). ggplot2: Elegant graphics for data analysis. Springer-Verlag New York. https://doi.org/10.1007/978-3-319-24277-4

Wilke, C. O. (2024). Cowplot: Streamlined plot theme and plot annotations for ggplot2 (R package version 1.1.3). Retrieved from https://cran.r-project.org/web/packages/cowplot/index.html

Yu, G., Wang, L.-G., Han, Y., & He, Q.-Y. (2012). clusterProfiler: An R package for comparing biological themes among gene clusters. OMICS: A Journal of Integrative Biology, 16(5), 284–287. https://doi.org/10.1089/omi.2011.0118

Zilionis, R., Nainys, J., Veres, A., Savova, V., Zemmour, D., Klein, A. M., & Fischbach, M. A. (2017). Single-cell barcoding and sequencing using droplet microfluidics. Nature Protocols, 12(1), 44-73. https://doi.org/10.1038/nprot.2016.154

A
APPENDIX CHAPTER HEADING

PUBLICATIONS FROM THE THESIS

Conference Papers

1. Uğun, D. (2024, April). HERACLOMICS: A User-Friendly Local Server App for scRNA-Seq Analysis. 17th The International Symposium on Health Informatics and Bioinformatics (HIBIT 2024). https://hibit2024.bezmialem.edu.tr/img/Hibit2024-abstract-book.pdf