Single cell analysis

Dataset

We embarked on an analysis of our dataset using the Seurat Toolkit, a powerful tool designed for quality control and processing of single-cell RNA sequencing data[1]. This data is organized as a Seurat Object, which acts as a comprehensive container for both the data and the analysis of our single-cell dataset [2].
Our Seurat Object comprises 27,271 features spread across 43,266 samples, all contained within a single RNA assay that includes only one count layer. The assay is organized into seven distinct slots: 'layers', 'cells', 'features', 'default', 'assay.orig', 'meta.data', 'misc' and 'key'.
Diving into the metadata, we examined several quality control metrics:

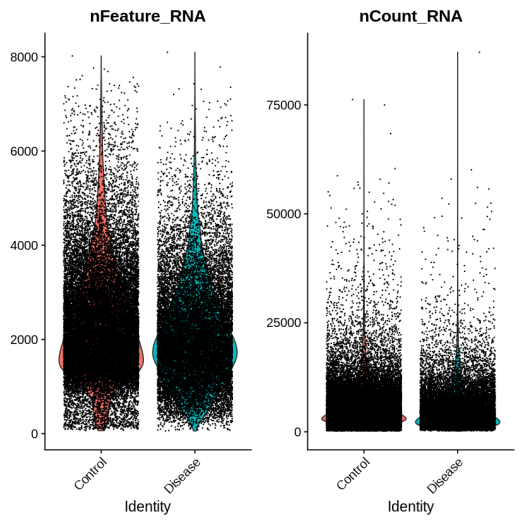
1. **orig.ident**: This identifier indicates the original sample or batch from which each cell originated, helping to determine whether a gene belongs to the control or disease group.
2. **nCount_RNA**: This metric represents the total number of RNA counts (unique molecular identifiers or UMI counts) detected within a cell, effectively measuring the total RNA content and reflecting overall transcript abundance.
3. **nFeature_RNA**: This number indicates the count of unique RNA features (genes) detected in each cell, representing the transcriptional diversity by counting the number of genes with non-zero expression in that cell.
4. **Barcode**: Each cell in our dataset is assigned a unique identifier, which in this context signifies the disease type and genome from which it derives.
5. **cell_type**: This biological label classifies the cells based on type—such as immune cells, neurons, etc.—often inferred from the expression of marker genes. The categories include:
    - Myeloid
    - Dividing Peripheral Immune Cells
    - Microglia
    - NK (Natural Killer) Cells
    - T Cells
    - B Cells
    - Newly Formed Oligodendrocytes (NFOL)
    - Myelin-Forming Oligodendrocytes (MFOL)
    - Mature Oligodendrocytes (MOL)
    - Oligodendrocyte Precursor Cells (OPC)
    - Vascular Endothelial Cells
    - Vascular Cells
    - Astrocytes
    - Ependymal Cells
    - Dorsal-Ventral Cells
6. **label**: This denotes whether the cell originates from a healthy or diseased sample, providing essential context for our analysis.

7. **replicate**: This identifier distinguishes between different biological replicates, ensuring reproducibility and helping us confirm that findings are not exclusive to any single batch or sample.

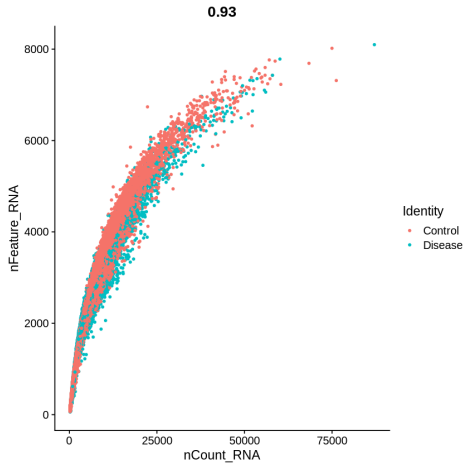
Below is the the top 10 rows of the metadata dataframe

A data.frame: 10 × 7

	orig.ident	nCount_RNA	nFeature_RNA	barcode	cell_type	label	replicate
	<fct>	<dbl>	<int>	<chr>	<chr>	<chr>	<chr>
Disease_4-ACGGGTCCAATGCAGG	Disease	13027	3813	Disease_4-ACGGGTCCAATGCAGG	Peripheral immune cells	Disease	Disease_4
Disease_4-ACTATCTAGGGAGGTG	Disease	11683	3643	Disease_4-ACTATCTAGGGAGGTG	Peripheral immune cells	Disease	Disease_4
Disease_4-GCGAGAAAGGGTACAC	Disease	11141	3272	Disease_4-GCGAGAAAGGGTACAC	Peripheral immune cells	Disease	Disease_4
Disease_4-TCCCACACATGTCAGT	Disease	10822	3297	Disease_4-TCCCACACATGTCAGT	Peripheral immune cells	Disease	Disease_4
Disease_4-GACAGCCAGTTACGTC	Disease	10620	3269	Disease_4-GACAGCCAGTTACGTC	Peripheral immune cells	Disease	Disease_4
Disease_4-GTCTACCTCTGTTGGA	Disease	10587	3439	Disease_4-GTCTACCTCTGTTGGA	Peripheral immune cells	Disease	Disease_4
Disease_4-GACCTTCTCTCGCCTA	Disease	10489	3346	Disease_4-GACCTTCTCTCGCCTA	Peripheral immune cells	Disease	Disease_4
Disease_4-TGGTTAGGTGTACATC	Disease	10040	3147	Disease_4-TGGTTAGGTGTACATC	Peripheral immune cells	Disease	Disease_4
Disease_4-CAGAGCCAGCTTGTTG	Disease	9234	3302	Disease_4-CAGAGCCAGCTTGTTG	Peripheral immune cells	Disease	Disease_4
Disease_4-TAAGCACTCGTTCTAT	Disease	9332	3115	Disease_4-TAAGCACTCGTTCTAT	Peripheral immune cells	Disease	Disease_4



This violin plot illustrates the distribution of RNA features and counts across both the disease and control groups.
Left Side: The plot displays the number of unique genes detected in each cell. Higher values indicate increased gene expression. Both groups exhibit similar distributions, although the control group shows slightly higher median values, suggesting comparable gene expression features. Notably, the spread of values is wider, with some cells detecting a high number of features (approximately 6,000-8,000).
Right Side: This section represents the total RNA transcripts captured. The distributions of total RNA counts are also quite similar, but the disease group displays a broader range, with some cells exhibiting significantly higher RNA counts (ranging from about 50,000 to 75,000). While both groups have comparable distributions, the disease group shows more variability in RNA counts, which may indicate altered transcriptional activity linked to stress, inflammation, or other disease processes.



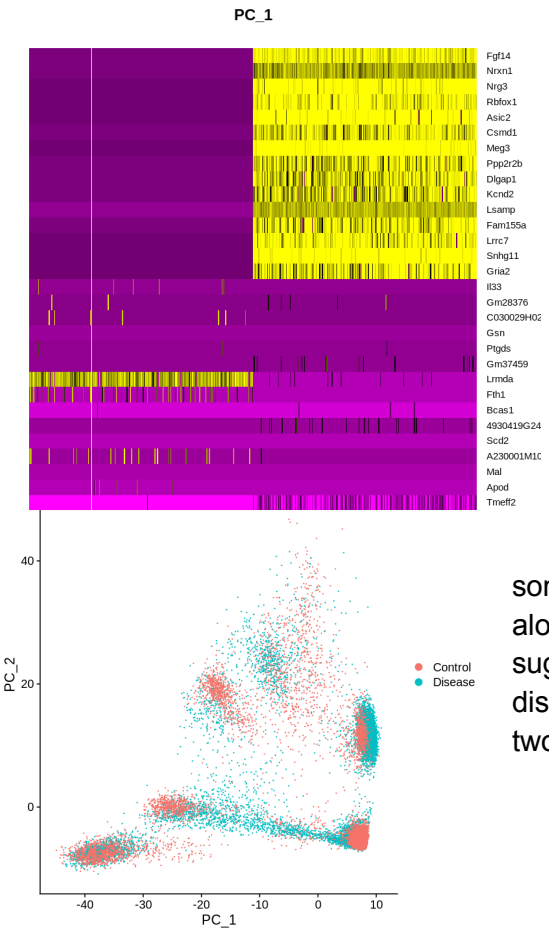
This scatter plot examines the relationship between RNA features and counts, revealing a strong positive correlation between nFeature_RNA and nCount_RNA, with a correlation coefficient of 0.93. This suggests that as RNA counts increase, so do the detected RNA features in both control and disease samples. The overlap between the two groups indicates no significant difference in RNA

feature and count relationships. However, as RNA counts rise, the spread of data points also increases, suggesting greater variability in RNA features detected for cells with higher RNA counts. Thus, this plot supports the findings from the violin plot, indicating similar RNA expression behaviors across both sample types.

Preprocessing

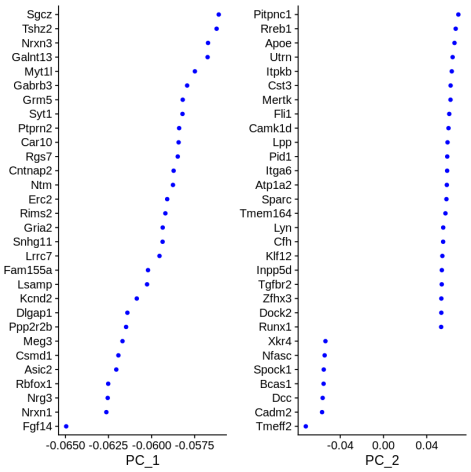
The dataset underwent standard preprocessing:

1. Outliers were removed by filtering out cells with unusually low (fewer than 200) or high (more than 2,500) gene counts.
2. Each feature was normalized for total expression, scaled by a factor of 10,000, and log-transformed to ensure comparability across cells.
3. The top 2,000 most variable genes were identified using variance-stabilizing normalization, capturing essential biological differences between cell types or states.
4. The data was centered and scaled to a mean of 0 and standard deviation of 1.
5. Principal Component Analysis (PCA) was performed to display data variance.



This heatmap correlates different marker genes across 500 genes in Principal Component 1, with columns representing cells and rows representing markers. Yellow indicates high expression levels, while dark purple reflects minimal to no expression.

The dimensionality reduction scatter plot demonstrates variance within the dataset, revealing some separation between the two groups, particularly along Principal Component 1 (PC1). While the overlap suggests that transcriptional profiles are not entirely distinct, clusters or regions exist where cells from the two groups are somewhat separated.



The last plot highlights the genes most responsible for separating the data along the two principal components. For PC1: Genes like Sgcz, Tshz2, Nrxn3, and Galnt13 are highly loaded, meaning they contribute significantly to the variation in PC1.
For PC2: Genes such as Pitpnc1, Rebp1, Apoe, and Utrn contribute to the variation in PC2.

Task 1

A data.frame: 6 × 5

	p_val	avg_log2FC	pct.1	pct.2	p_val_adj
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Scd2	0	2.634501	0.717	0.232	0
Fth1	0	1.797720	0.682	0.224	0
4930419G24Rik	0	3.386258	0.599	0.143	0
Gm11149	0	2.804186	0.557	0.133	0
Lrrtm3	0	2.066090	0.647	0.225	0
Gm37459	0	3.795326	0.492	0.073	0

Part 1

We first set the “orig.ident” as the identity class for the cells. This classification in Seurat defines how the data is categorized into "Control" and "Disease" groups. To confirm the naming format of the two classes, we executed unique(Idents(single_cell_filtered)). Next, we utilized the FindMarkers function to pinpoint genes that are differentially expressed between these two conditions. By specifying ident.1 = "Control" and ident.2 = "Disease", we instructed the function to focus on comparing these groups. The output yields a ranked list of differentially expressed genes (DEGs), along with key statistics like log-fold change, p-value, and adjusted p-value, which we then displayed using head(group.markers).

	p_val	avg_log2FC	pct.1	pct.2	p_val_adj
	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
Fmn1	0	2.250777	0.590	0.226	0
Abca1	0	3.323676	0.513	0.169	0
Abr	0	2.530906	0.481	0.154	0
Ptprj	0	2.022040	0.501	0.205	0
Nav2	0	1.443322	0.660	0.370	0
Man1c1	0	2.855571	0.383	0.101	0

Part 2

We proceeded to identify differentially expressed features between disease cells and all other cells. The FindMarkers function conveniently provides a ranked list of these features.

Task 2

Part 1

We began by splitting the dataset into training and testing sets. Using ncol(single_cell_filtered), we retrieved the total number of cells in the filtered dataset. We then defined a pivot point to divide the data into two halves, opting for this method instead of traditional 60/40 or 80/20 splits. The first half (designated as train_set) was allocated for training the classifier, while the second half (test_set) was reserved for evaluation.

Choosing to split the train-test based on cells allows us to maximize the training data available for each class, especially in a large and diverse dataset. This approach may enhance the classifier's ability to generalize across unseen data, addressing concerns about potential imbalances across replicates that could skew predictions.

To focus on significant features, we selected the top 15 disease-related marker genes for the classifier. To facilitate better generalization, we opted not to make the classifier cell-type specific, as a specific classifier could yield more meaningful insights if the perturbation is known to affect different cell types differently.

We then employed the train_classifier function to develop a classifier using the train_set and the top 15 marker genes, setting a probability threshold of 0.5. This process utilized Support Vector Machines (SVM) with a linear kernel, cross-validated over 10 folds. Subsequently, we extracted the model for further evaluation using caret_model().

```
Loading required package: SingleCellExperiment

Warning message:
“replacing previous import ‘ape::where’ by ‘dplyr::where’ when loading ‘scAnnotatR’”
27648
An object of class scAnnotatR for Disease
* 15 marker genes applied: Ptprj, Fmn1, Ophn1, Abca1, Man1c1, Arhgap24, Etv6, Nav2,
Xylt1, Cacna1a, Abr, Cerk, Runx1, C4b, Piezo2
* Predicting probability threshold: 0.5
* No parent model
```

```
Support Vector Machines with Linear Kernel

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 12442, 12442, 12442, 12441, 12441, 12442, ..
Addtional sampling using down-sampling

Resampling results:

  Accuracy  Kappa
  0.860821  0.6533186

Tuning parameter 'C' was held constant at a value of 1
Current probability threshold: 0.5

               Positive        Negative        Total

Actual         1975            11849           13824

Predicted      9191            4633            13824


Accuracy: 0.459056712962963


Sensivity (True Positive Rate) for Disease: 0.933670886075949

Specificity (1 - False Positive Rate) for Disease: 0.379947674909275

Area under the curve: 0.7857027725461
```

Task 2

After testing the classifier on the test_set—the remaining cells not used in training—we calculated performance metrics. The cross-validation accuracy reached 0.860821, with Cohen's κ-coefficient at 0.6533186. However, the accuracy on the test set dropped to 0.459056712962963, indicating that the classifier struggles when faced with unseen data.

The sensitivity (True Positive Rate) for Disease was calculated at 0.933670886075949, demonstrating the classifier's proficiency in accurately identifying patients with the Disease condition. However, the specificity (1 - False

Positive Rate) for Disease was only 0.379947674909275, indicating frequent misclassification of Control samples as Disease, resulting in a high false positive rate. The area under the ROC curve was 0.7857027725461, suggesting that the model can distinguish between the two classes. However, this also reveals that while the model is capable of separating the classes, its overall performance is hindered by the imbalanced classification, as reflected in the sensitivity and specificity metrics.

Additionally, we computed the Receiver Operating Characteristic (ROC) curve, illustrating the trade-off between the true positive rate and the false positive rate.

p_thres	fpr	tpr
0.1	0.9940079	0.9974684
0.2	0.9862436	0.9888608
0.3	0.7839480	0.9751899
0.4	0.7010718	0.9610127
0.5	0.6200523	0.9336709
0.6	0.5597941	0.9058228
0.7	0.4976791	0.8724051
0.8	0.4138746	0.8192405
0.9	0.2904043	0.7053165

0.459056712962963
0.7857027725461

The confusion matrix further illustrates that while the classifier accurately identifies many Disease cases, the high number of false positives indicates significant misclassification of Control cases. This issue may stem from the SVM model with a linear kernel not adequately capturing the non-linear relationships between features, limiting its effectiveness in distinguishing between classes.

References

[1] Hao, Y., Stuart, T., Kowalski, M.H., Choudhary, S., Hoffman, P., Hartman, A., Srivastava, A., Molla, G., Madad, S., Fernandez-Granda, C. and Satija, R., 2024. Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nature biotechnology, 42(2), pp.293-304.
[2] Analysis, visualization, and integration of visium HD spatial datasets with Seurat (no date) Analysis, visualization, and integration of Visium HD spatial datasets with Seurat • Seurat. Available at: https://satijalab.org/seurat/articles/pbmc3k_tutorial.html (Accessed: October 2024).