Scientific Report: Application and Rationale of the Kruskal–Wallis Test for Disease-associated Gene Expression Differences

1. Background and Research Objective

The goal of this analysis is to identify genes whose expression levels differ significantly across disease states in a single-cell RNA sequencing (scRNA-seq) dataset derived from colon tissues. The dataset, scIBD_Colon.h5ad, contains single-cell expression profiles from multiple inflammatory conditions (e.g., UC_inflamed, UC_non_inflamed, CD_inflamed, Colitis_inflamed, and Healthy).

Each cell is annotated by two categorical features: 'disease' (pathological or healthy state) and 'major_cluster' (refined cell-type label). The objective is to test whether gene expression distributions differ between diseases, both globally and within individual cell clusters.

For each gene g, we aim to determine whether its expression distribution differs among k disease groups. Formally, the null hypothesis (H0) states that all diseases share the same distribution, while the alternative hypothesis (H1) posits that at least one group differs. Due to the non-normal and zero-inflated nature of scRNA-seq data, parametric tests such as ANOVA are not appropriate, necessitating a rank-based non-parametric approach.

2. Why Analyze Cell Clusters Separately

Each cell cluster represents a unique biological population, characterized by its own gene expression program. Analyzing all cells together would mix signals from diverse cell types, obscuring real disease effects. For example, a gene upregulated in epithelial cells might appear non-significant if averaged together with immune cells. Performing statistical testing within each cluster ensures that we detect disease-associated expression changes that are specific to that cell type. This approach enhances biological interpretability and preserves the cellular context of disease-related transcriptional shifts.

3. Why the Kruskal–Wallis Test is Appropriate

The Kruskal–Wallis (KW) test is a nonparametric non-parametric analogue of one-way ANOVA test used to compare multiple independent groups. I It assesses whether samples from multiple groups originate from the same distribution, without assuming normality or homogeneity of variance. This makes it ideal for single-cell data, which are often skewed, sparse, and heteroscedastic. In this analysis, the groups correspond to different disease states (e.g., Healthy, UC_inflamed, CD_inflamed). The test ranks all expression values and compares the rank sums across groups, making it robust to non-normal and zero-inflated distributions common in scRNA-seq data.

Mathematically, the KW statistic is defined as:

$$H = (12 / (N * (N + 1))) * \Sigma[n_i * (\bar{R}_i - \bar{R})^2]$$

where N is the total number of observations, n_i is the number of observations in group i, $\bar{R}_i$ is the mean rank in group i, and $\bar{R}$ is the average rank across all groups. If all diseases have similar expression distributions, their average ranks will be similar, and H will be small. Larger deviations in average ranks result in higher H values, indicating significant differences between diseases. Under the null hypothesis, H follows approximately a chi-squared distribution with (k-1) degrees of freedom.

Key advantages of the Kruskal–Wallis test in this context include:

- It does not assume a normal distribution, handling zero-inflated data naturally.
- It accommodates unequal group sizes, as disease groups may contain differing cell counts.
- It is robust to outliers and non-linear effects.
- It allows simultaneous comparison of more than two diseases.
- It operates on ranked values, mitigating scaling differences in expression levels.

4. Why Check the Effect Size

A statistically significant difference does not always indicate a biologically meaningful change. With thousands of cells, very small differences in gene expression can become statistically significant. To ensure biological relevance, we measure the effect size for each gene within each cell type, defined as:

$$\text{Effect\_size} = \max(\text{mean across diseases}) - \min(\text{mean across diseases})$$

This metric captures the largest observed change in mean expression between any two disease groups. Genes with an effect size below 0.25 are classified as having weak changes, 0.25–0.5 as moderate, and above 0.5 as strong. This ensures that reported genes not only pass statistical significance but also exhibit meaningful biological shifts.

Kruskal–Wallis (KW) is **very sensitive** with thousands of cells, even a *tiny* difference ($\Delta = 0.05$) can yield $\text{FDR} < 0.001$.

But such small changes might be:

- random noise,
- not meaningful for cell function,
- or simply reflect sample imbalance.

That's why we add an **effect-size threshold**: to filter for genes whose expression difference is **large enough to matter biologically**.

Gene expression values are roughly **0–10** (normalized counts).
Let's interpret Δ thresholds on that scale:

Δ = max(mean) − min(mean)	**Interpretatio n**	**Biological meaning**
< 0.1	Minimal	noise or technical variation
0.1 – 0.25	Weak	small modulation, uncertain biological relevance
0.25 – 0.5	Moderate	clear expression shift; can be biologically relevant
> 0.5	Strong	gene clearly upregulated/downregulated in one or more diseases

So, because data is scaled 0–10, a difference of **0.25** means about a **2.5% absolute shift**, which is typically enough to distinguish one disease state from others in transcriptomics data.

5. Why Perform Pairwise Post-hoc Tests

A significant Kruskal–Wallis test only indicates that at least one disease group differs but not which ones. Therefore, we apply pairwise post-hoc testing using the Mann–Whitney U test (also known as the Wilcoxon rank-sum test). This nonparametric test compares two independent groups by ranking all values and assessing whether one group tends to have higher ranks than the other. It is appropriate here because, like the Kruskal–Wallis test, it does not assume normality and is robust to outliers.

By applying the Mann–Whitney U test to all disease pairs (e.g., UC_inflamed vs Healthy), we can identify which specific diseases differ significantly in gene expression. P-values from these pairwise tests are adjusted for multiple testing using the Benjamini–Hochberg FDR correction.

6. Why Use P-values, FDR, and Bonferroni Correction

In this study, tens of thousands of genes are tested across multiple cell types, leading to hundreds of thousands of statistical comparisons. Without correction, this would produce many false positives. To address this, three levels of significance control are applied:

1. P-value: The probability of observing the data assuming no real difference exists. It measures statistical significance but not the expected false discovery rate.

2. False Discovery Rate (FDR): Controlled using the Benjamini–Hochberg method. FDR limits the expected proportion of false positives among all significant results, making it ideal for genome-scale analyses.

3. Bonferroni correction: Provides an additional conservative control of the family-wise error rate. The adjusted significance level is α / N, where N is the total number of tests. In this case, with approximately 16,106 genes across 10 cell types (161,060 tests), the corrected threshold becomes $\alpha = 0.05 / 161{,}060 \approx 3.1\times10^{-7}$. This ensures that the probability of even a single false positive across the study is below 5%.

7. Results

8. Conclusion

By analyzing each cell type independently, applying robust nonparametric tests, quantifying effect sizes, and controlling false discovery rates with FDR and Bonferroni corrections, this framework ensures that the identified disease-associated genes reflect true biological signals rather than statistical artifacts. This approach is particularly well-suited to scRNA-seq data, where heterogeneity and non-normal distributions are the norm.