HybridQC: Machine Learning-Augmented Quality Control for Single-Cell RNA-seq Data

Kaitao Lai[1]

2025-09-03

[1] University of Sydney

1 Summary

HybridQC is an R package that streamlines quality control (QC) of single-cell RNA sequencing (scRNA-seq) data by combining traditional threshold-based filtering with machine learning-based outlier detection. It provides an efficient and adaptive framework to identify low-quality cells in noisy or shallow-depth datasets using techniques such as Isolation Forest (Liu et al. 2018), while remaining compatible with widely adopted formats such as Seurat objects.

The package is lightweight, easy to install, and suitable for small-to-medium scRNA-seq datasets in research settings. HybridQC is especially useful for projects involving non-model organisms, rare samples, or pilot studies, where automated and flexible QC is critical for reproducibility and downstream analysis.

2 Statement of Need

scRNA-seq experiments often suffer from technical noise, dropout events, and variability in sequencing depth. Traditional quality control relies on static cutoffs for metrics such as gene count, UMI count, and mitochondrial content, which may be suboptimal for non-standard datasets (Malte D. Luecken and Theis 2022).

HybridQC fills this gap by integrating machine learning methods—specifically unsupervised outlier detection with Isolation Forest (Liu et al. 2018)—to improve filtering precision and robustness. This dual-level approach can better preserve informative but unconventional cell types and adapt dynamically to diverse datasets. No existing R packages provide this hybrid QC strategy as a standalone tool with seamless integration into Seurat-based pipelines.

3 Features

- Computes standard QC metrics: `nFeature_RNA`, `nCount_RNA`, `percent.mt`
- Supports Isolation Forest outlier detection via `reticulate` and `pyod` (Liu et al. 2018)
- Filters cells using a hybrid decision rule
- Works on Seurat objects
- Lightweight and suitable for quick prototyping or small studies

4 Example Usage

We demonstrate HybridQC on a synthetic single-cell RNA-seq dataset derived from 10x Genomics-style PBMC data, consisting of 2,000 cells and 1,000 genes. The dataset is loaded into R as a Seurat object and

subjected to a two-stage quality control workflow.

First, we compute basic quality metrics such as gene count, UMI count, and mitochondrial content per cell. Then, an unsupervised outlier detection model (Isolation Forest) is applied to capture multivariate anomalies not addressed by static thresholds alone. Cells flagged by either criterion are filtered out, resulting in a cleaner, more reliable dataset for downstream analysis.

```
library(Seurat)
library(HybridQC)

pbmc <- LoadPBMC2k()  # Load example data
qc_basic <- run_basic_qc(pbmc)
ml_scores <- run_isolation_forest_qc(pbmc)
filtered <- filter_cells(pbmc, qc_basic, ml_scores)
```

5 UMAP Visualization of Outlier Scores

To visualize Isolation Forest–based anomaly scores in a low-dimensional embedding, HybridQC supports `FeaturePlot()` using UMAP projections (McInnes, Healy, and Melville 2018). The resulting plot highlights which cells are predicted as outliers in UMAP space based on their multivariate QC metrics.

Below is an example UMAP colored by Isolation Forest score:

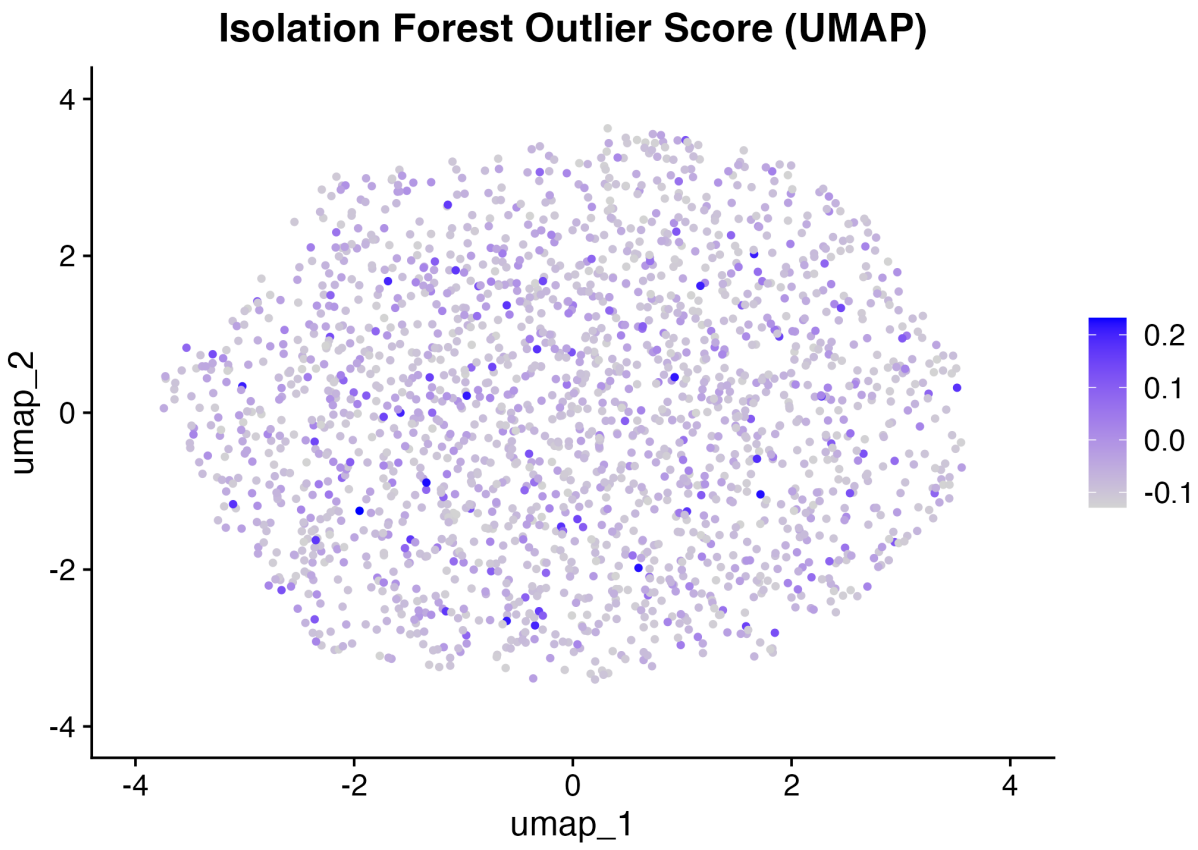

Figure 1: UMAP visualization of Isolation Forest outlier scores

6 Software Repository

The source code for HybridQC is freely available on GitHub at:
https://github.com/biosciences/HybridQC

7 Acknowledgements

The author thanks collaborators at University of Sydney for feedback on early concepts.

References

Liu, Yue, Zheng Li, Hao Xiong, Jian Pei, and S. Yu Philip. 2018. "PyOD: A Python Toolbox for Scalable Outlier Detection." *arXiv Preprint arXiv:1901.01588*. https://arxiv.org/abs/1901.01588.

Malte D. Luecken, K. Chaichoompu, M. Büttner, and Fabian J. Theis. 2022. "Benchmarking Atlas-Level Data Integration in Single-Cell Genomics." *Nature Methods* 19 (1): 41–50. https://doi.org/10.1038/s41592-021-01336-8.

McInnes, Leland, John Healy, and James Melville. 2018. "UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction." *arXiv Preprint arXiv:1802.03426*. https://arxiv.org/abs/1802.03426.