The Influence of Clustering Quality on Cell Type Prediction Accuracy

Marta Benegas Coll[1], Ana Conesa Cegarra[2], Stefan Götz[1]

[1]BioBam Bioinformatics S.L., Valencia, Spain ,[2]I2SysBio, CSIC, Valencia, Spain

Abstract

Accurate cell type prediction is a crucial step in the interpretation of single-cell RNA-seq data, as downstream biological insights strongly depend on these predictions. However, most annotation strategies rely on an initial **unsupervised clustering** step that is sensitive to **parameter choices**, thus leading to substantial variation in cell grouping.

While it is widely acknowledged that **clustering quality** influences downstream analyses, the extent to which "good quality" clusterings truly translate into **better annotation outcomes** remains insufficiently characterized. The question of weather researchers should trust clustering metrics alone to select the "best" clustering for downstream analysis is yet unanswered, as well as weather robust annotation tools can compensate for suboptimal clustering.

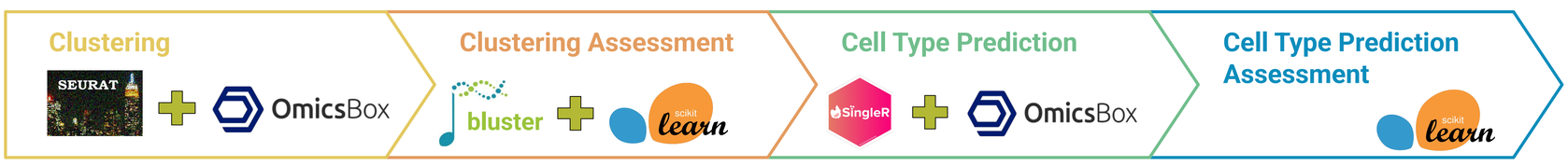
This study explores the relationship between clustering quality and cell type prediction accuracy. By comparing multiple clustering outputs of varying quality against ground-truth annotations, we evaluate whether commonly used **clustering metrics align with annotation performance**. Our findings aim to guide **best practices** in single-cell analysis by shedding light on the interplay between clustering and annotation, and by identifying which quality metrics are most informative when no ground truth is available.

Dataset

Human PBMC
~ 162K Cells
31 cell types
8 broad cell types

- ✓ Curated
- ✓ Well-studied tissue
- ✓ Ground-truth

Methods



- 9 parameter combinations:
  - Nº of **PCA Dimensions** ≃ Amount of data.
  - **Resolution** ≃ Granularity.

	Nº Dimensions	Resolution
5D_06R	5	0.6
5D_08R	5	0.8
5D_1R	5	1
15D_06R	15	0.6
15D_08R	15	0.8
15D_1R	15	1
20D_06R	20	0.6
20D_08R	20	0.8
20D_1R	20	1

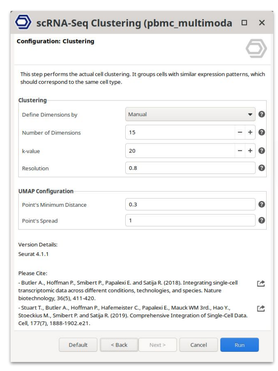
- Unsupervised metrics:
  - **Silhouette & Purity**: Assess intra-cluster cohesion and inter-cluster separation.
  - **RMSD**: Quantifies dispersion or compactness of cells within clusters.
- Supervised metrics:
  - **ARI**: Evaluates alignment between clustering and ground-truth cell type labels.

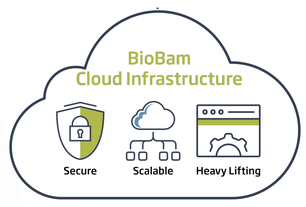
- Annotation performed **per-cluster**.
- We used the ScaleBio reference dataset, which is well-curated, and with matching gene IDs and cell type naming conventions.
- Predictions were performed using both granular and broad cell type definitions.

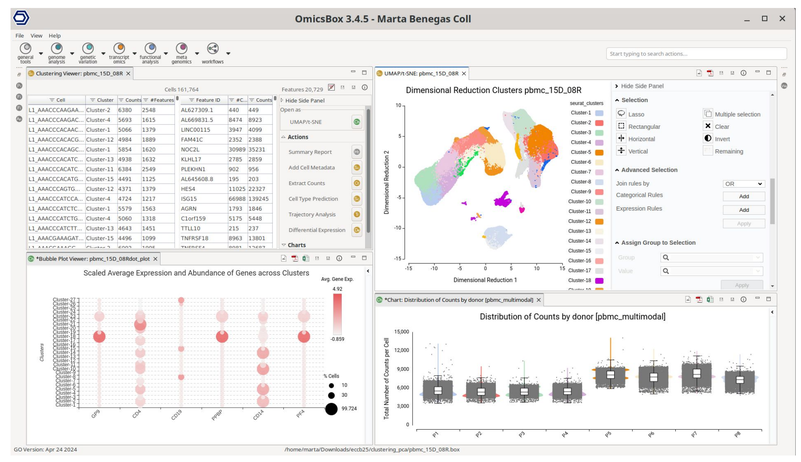
~ 685K cells
26 Cell Types
8 Broad Cell Types

- Comparison between the predicted and the original (ground-truth) cell types.
- **Accuracy**: Measures overall correctness of predictions.
- **Precision, Recall & F1-Score**: Assess class-wise prediction performance.
- **Cohen's Kappa & MCC**: Evaluate agreement beyond chance, accounting for class imbalance and prediction reliability.

OmicsBox







Results

Table 1. Summary of clustering (in purple) and cell type prediction (in blue) evaluation metrics for the different clusterings. **(a)** Unsupervised clustering metrics are computed without external evidence. **(b)** Supervised clustering metrics measure grouping agreement between cluster labels and ground-truth cell types. The overall clustering quality is assigned based on the combination of internal and external metrics. **(c)** Cell type prediction metrics measured across the entire dataset, while **(d)** macro-averaged and **(e)** weighted-averaged metrics summarize performance per cell type, either giving equal weight to all types or adjusting for their abundance. Thus, the macro-averaged metrics highlight the prediction performance on rare or infrequent cell types. The overall prediction quality, for both macro and weighted metrics, is interpreted in conjunction with the global metrics in (c). A detailed description of each metric can be found by scanning the QR.

				Unsupervised Clustering Evaluation Metrics[(a)]			**Supervised Clustering Evaluation Metrics[(b)]**			**Cell Type Prediction Evaluation Metrics[(c)]**		**Cell Type Prediction Evaluation Metrics Macro-Averaged[(d)]**					**Cell Type Prediction Evaluation Metrics Weighted-Averaged[(e)]**				
Clustering	# Clusters	# Predicted Granular Cell Types	# Predicted Broad Cell Types	Mean Silhouette	Mean Purity	Mean RMSD	ARI Granular Annot.	ARI Broad Annot.	Overall	Cohen's Kappa	MCC	Balanced Accuracy	Macro Precision	Macro Recall	Macro F1-Score	Overall	Accuracy	Weighted Precision	Weighted Recall	Weighted F1-Score	Overall
5D_06R	18	11	8	**0.1200**	0.8506	7.2450	0.3913	**0.3583**	Med-High	**0.4730**	**0.4838**	0.2742	0.1862	0.2742	0.2080	Med-Low	0.4768	0.4674	0.4768	0.4326	Med-High
5D_08R	21	12	8	0.1106	0.8008	7.6208	0.3372	0.2873	Med	0.4247	0.4415	0.2923	0.2000	0.2924	0.2194	Low	**0.5323**	0.4540	**0.5323**	0.4722	High
5D_1R	23	11	8	0.0901	0.7795	7.3104	0.2847	0.2535	Low	0.4522	0.4707	0.2935	0.1994	0.2936	0.2170	Low	0.5061	0.4719	0.5061	0.4524	Med
15D_06R	27	15	8	0.0958	**0.9121**	7.0688	0.3938	0.3010	High	0.4444	0.4634	0.3952	0.3132	0.3953	0.3146	Med-Low	0.4915	0.4989	0.4915	0.4473	Med-Low
15D_08R	27	16	8	0.1003	0.9113	7.1545	**0.4009**	0.3053	High	0.4574	0.4762	0.3755	0.3022	0.3756	0.2993	Med	0.5041	0.5412	0.5041	0.4689	Med
15D_1R	31	17	8	0.0872	0.9003	6.9917	0.3590	0.2782	Med	0.4708	0.4890	0.4033	0.3387	0.4034	0.3307	Med	0.5172	0.5529	0.5172	**0.4844**	High
20D_06R	29	15	7	0.0874	0.9019	6.7050	0.3704	0.2920	Med-High	0.4213	0.4488	0.3710	0.3227	0.3710	0.3085	Med-Low	0.4612	0.5217	0.4612	0.4249	Med-Low
20D_08R	36	17	8	0.0824	0.8874	6.7008	0.3380	0.2542	Med	0.4303	0.4529	0.4159	0.3702	0.4159	0.3593	High	0.4742	0.5633	0.4742	0.4561	Med-Low
20D_1R	42	18	8	0.0739	0.8763	**6.5552**	0.3200	0.2337	Med	0.4369	0.4616	**0.4354**	**0.3834**	**0.4355**	**0.3611**	High	0.4781	**0.5800**	0.4781	0.4546	Med

- Clusterings with a higher number of partitions:
  - Lower Silhouette, Purity, and ARI ➡ less defined clusters and poorer agreement with ground-truth.
  - Better RMSD ➡ clusters with lower substructure.
  - Improve detection of less frequent cell types, as shown by better macro-averaged metrics.
- Clusterings with lower number of partitions:
  - Higher Silhouette, Purity, and ARI ➡ more clearly separated clusters.
  - Capture broader cell-type structure, but miss finer distinctions.
- SingleR still miss many granular or rare cell types despite robust reference matching.

Table 2. Overall quality assignment for each clustering based on the performance of the different evaluation metrics.

	Overall Clustering Assessment Quality	Overall Cell Type Prediction Quality by Macro Metrics	Overall Cell Type Prediction Quality by Weighted Metrics
55D_06R	Med-High	Med-Low	Med-High
55D_08R	Med	Low	High
55D_1R	Low	Low	Med
15D_06R	High	Med-Low	Med-Low
15D_08R	High	Med	Med
15D_1R	Med	Med	High
20D_06R	Med-High	Med-Low	Med-Low
20D_08R	Med	High	Med-Low
20D_1R	Med	High	Med

Conclusions

- Clustering quality does not directly correlates with cell type prediction performance.
- Granular clusterings help uncover rare cell types.
  - ↪ Use RMSD to detect them.
- Coarser clusterings are more clearly delimited and defined.
  - ↪ Use Silhouette & Purity to detect them.

Key Takeaway

Suggested strategy:

- ✔ Select well-defined clustering based on Silhouette and Purity.
- ✔ Refine by sub-clustering and adding information obtained by low-RMSD clusterings.

STAY TUNED!

References

1. Gotz, S., García-Gómez JM, Terol J, et al. (2008) 'High-throughput functional annotation and data mining with the Blast2GO Suite', Nucleic Acids Research, 36(10), pp. 3420–3435. doi:10.1093/nar/gkn176.
2. Hao, Y., Stuart, T., Kowalski, M.H. et al (2024). 'Dictionary learning for integrative, multimodal and scalable single-cell analysis'. Nat Biotechnol 42, 293–304. doi:/10.1038/s41587-023-01767-y
3. Hao, Y., Hao S, Andersen-Nissen E, et al. (2021) 'Integrated Analysis of multimodal single-cell data', Cell, 184(13). doi:10.1016/j.cell.2021.04.048.
4. ScaleBio Single Cell RNA Sequencing of Human PBMCs [dataset]. (n.d.). CZ CELLxGENE Discover. Retrieved July 2025, from https://cellxgene.cziscience.com/collections/4a9fd4d7-d870-4265-89a5-ad51ab811d89
5. Lun A (2025). 'bluster: Clustering Algorithms for Bioconductor'. doi:10.18129/B9.bioc.bluster, R package version 1.18.0, https://bioconductor.org/packages/bluster.
6. Pedregosa et al. (2011). 'Scikit-learn: Machine learning in Python'. Journal of Machine Learning Research, 12, pp.2825–2830.
7. Aran, D., Looney AP, Liu L, et. al. (2019). 'Reference-based analysis of lung single-cell sequencing reveals a transitional profibrotic macrophage.' Nat. Immunol., 20, 163-172. doi:10.1038/s41590-018-0276-y.

POWERED BY: OmicsBox