



UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Effects of misclustering in multi-sample, multi-group scRNA-seq studies: a stability-based approach

Gabriele Dolfi, Davide Risso

- Multi-group scRNA-seq experiments aggregate single-cell counts into **pseudo-bulk profiles** across subjects.
- Inference is conditioned on partitions produced by upstream clustering algorithms.
- **Errors** occurring in the clustering step can have critical consequences in between-group inference.

Differential Misclustering

We define **differential misclustering** as the situation in which the probability of misclassification differs systematically across experimental conditions due to a difference in clustering error rate or compositional frequency between conditions.

Mechanism of Differential Misclustering

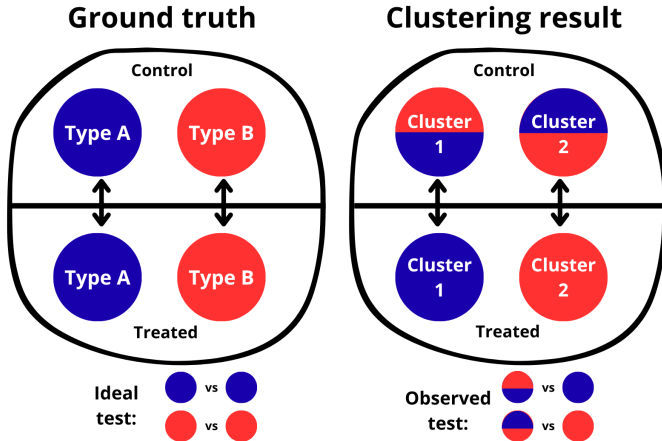


Figure 1: Visual representation of differential misclustering and its impact on pseudo-bulk inference.

Practical Consequences: FDR Inflation



Differential Misclustering: Impact on FDR Control

FDR inflation under increasing transcriptomic divergence between misclustered cell types

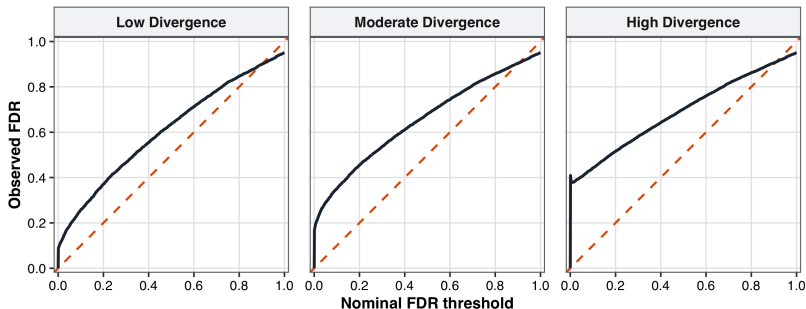


Figure 2: FDR inflation under increasing transcriptomic divergence between misclustered cell types.

Building on Yu et al. (2019), we repeated clustering under data perturbation schemes to quantify **observation-level stability** $\hat{S}_{\text{obs}}(j)$.

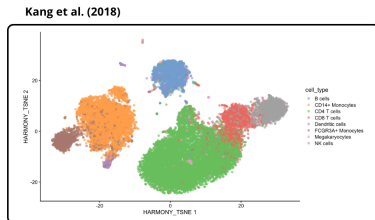
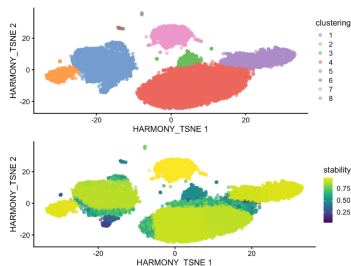


Figure 3: Bootstrap stability scores obtained in the PBMC multi-sample, multi-group dataset from Kang et al. (2018).

Downstream pseudo-bulk (PB) modeling can be corrected using stability-based **aggregation weights**:

$$y_i(r, k) = \sum_{j \in \hat{C}_{rk}}^{|C_{rk}|} w_j x_{ij}, \quad \hat{w}_j = F[\hat{S}_{\text{obs}}(j)]$$

Table 1: DEGs detected per cluster in a mock-comparison based on Kang et al. (2018); standard PB vs. weighted PB (WPB).

Cell type	Standard PB	Weighted PB (WPB)
1	610 (15%)	26 (0.6%)
2	0 (0%)	0 (0%)
3	13 (0.3%)	3 (0.07%)
4	0 (0%)	0 (0%)
5	1 (0.02%)	1 (0.02%)

- **Reducing misclustering:** Stability measures can be used to optimize the choice of clustering hyperparameters (e.g., resolution).
- **Uncertainty propagation:** Perturbation schemes can be extended beyond clustering to dimensionality reduction and cell-type annotation steps.
- **Multi-resolution frameworks:** Stability-based losses can be implemented to overcome single-resolution clustering.

Thank you for your attention!

- Yu, H., Chapman, B., Di Florio, A., Eischen, E., Gotz, D., Jacob, M., Blair, R. H. (2019). *Bootstrapping estimates of stability for clusters, observations and model selection: H. Yu et al. Computational Statistics*, 34(1), 349-372;
- Kang, H. M., Subramaniam, M., Targ, S., Nguyen, M., Maliskova, L., McCarthy, E., ... Ye, C. J. (2018). *Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nature biotechnology*, 36(1), 89-94.