CellBender removes technical artifacts from single-cell RNA sequencing data

The conversion of biological molecules into digital signals through sequencing is a complex process that often generates substantial systematic background noise. This noise can obscure important biological insights. However, by precisely identifying and removing this noise, we can bring the true signal into focus and eliminate misleading results from downstream analyses.

This is a summary of:
Fleming, S. J. et al. Unsupervised removal of systematic background noise from droplet-based single-cell experiments using CellBender. *Nat. Methods* https://doi.org/10.1038/s41592-023-01943-7 (2023).

Published online: 17 August 2023

The problem

High-throughput assays are helping to answer fundamental questions in cell biology by enabling measurements on a scale that has never before been possible. Sequencing (that is, reading the nucleic acid sequence of individual molecules) enables the quantification of RNA in individual cells in a high-throughput fashion. In single-cell RNA sequencing (scRNA-seq), RNA molecules are counted by copying them to complementary DNA, amplifying their numbers using the polymerase chain reaction, sequencing them, mapping the sequences to genes, and counting the number of times each gene is detected in each cell[1]. Some trade-offs are necessary when moving from low-throughput assays to high-throughput ones, however, and it is important to understand and quantify these trade-offs if subsequent measurements are to be precise. In scRNA-seq, technical artifacts that are inherent in the experimental method add systematic noise to these measurements (Fig. 1a). Such systematic and structured noise can skew conclusions about cell functions, fates and perturbation responses.

The solution

We spent several years carefully studying scRNA-seq datasets to understand the phenomenology of systematic noise in these experiments and translating this understanding into mathematical models. Leveraging advances in machine learning — in particular, building on stochastic variational inference in generative Bayesian models[2] — we were able to construct principled models that captured the generative process for scRNA-seq data, both signal and noise. Performing inference[3] in the context of these models enabled us to calculate probabilities of both signal and noise components of the measured data. Distilling these probabilities down into an optimal estimate of gene expression, ready for use in downstream analyses, was the aim of much of our research toward the end of this project.

Our approach provides a rigorous way to denoise scRNA-seq data in an unsupervised manner — that is, without the need for sample-specific preprocessing or prior biological knowledge from an analyst. Rather, we rely on the phenomenology of how such experiments generate these data to determine which droplets in an experiment contain cells, so that the empty droplets can be excluded from downstream analysis. CellBender also provides an estimate of the profile of cell-free RNA, which is one of the main contributors to background noise. CellBender not only provides a 'best estimate' of the denoised gene expression matrix (meaning a denoised integer count matrix suitable for downstream use), but also enables use of the full inferred posterior probability distribution over the number of denoised counts in each cell, which can provide precise quantitative answers to interesting biological questions that would otherwise be difficult to answer, such as: "What is the probability that this cell contains a nonzero number of viral RNA counts?"

The implications

Removal of systematic noise from scRNA-seq datasets improves several aspects of downstream analysis and sharpens the biological inferences that can be drawn from such data. Single nucleus (sn) RNA-seq data, which is more prone than scRNA-seq data to systematic noise, shows particularly striking improvement after noise removal. Marker genes show increased specificity for their respective cell types (Fig. 1b), and proteins measured via antibody labeling (as in cellular indexing of transcriptomes and epitopes (CITE)-seq[4]), which are highly affected by background noise, likewise show increased cell type specificity and concordance with RNA measurements. Importantly, in experiments comparing case and control conditions (such as diseased and healthy states), we demonstrate that systematic background noise present in the raw data leads to detection of spurious differentially expressed genes and that these false discoveries can be removed by CellBender preprocessing of the datasets.

Although the principled model underlying CellBender is faithful to the phenomenology of scRNA-seq experiments, it cannot capture the full complexity of real datasets. The extent to which our inferred probability distributions over noise counts reflect reality will depend on the validity of the modeling assumptions, which we describe in detail in the research paper.

In future work, it will be interesting to explore the extent and phenomenology of systematic noise in other single-cell measurement modalities. Do assay for transposase-accessible chromatin (ATAC)-seq data[5] exhibit the same systemic noise seen in scRNA-seq and antibody-labeled protein measurements? If not, how is the noise different, and is there sufficient noise in ATAC-seq data to warrant its removal?

Stephen Fleming & Mehrtash Babadi
Broad Institute, Cambridge, MA, USA.

EXPERT OPINION

"CellBender has already been adopted by many research groups. It is distinguished by the sophistication of its computational model, which is based on unsupervised neural networks and probabilistic modeling and which includes many relevant experimental details (including ambient RNA transcripts, barcode swapping, per-cell and per-droplet sampling rates, and appropriate count-based noise models). The detailed comparison of different noise estimation strategies sets a strong example for computational methods developers." **Eran A. Mukamel, University of California San Diego, La Jolla, CA, USA.**

FIGURE

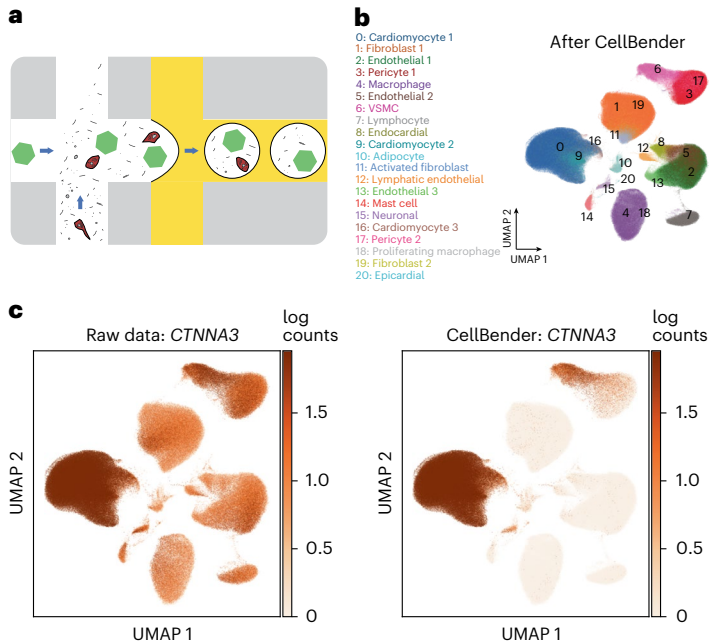

Fig. 1 | The effect of CellBender noise removal on RNA sequencing data. a, Single-nucleus RNA sequencing library preparation generates cell-free RNA (black), which is packaged into droplets along with cells (red) and barcode beads (green hexagons); this is a source of systematic noise. Uniform manifold approximation and projection (UMAP) plot shows cell types from a published 600,000-nucleus heart dataset. VSMC, vascular smooth muscle cell. **b**, UMAP plots show counts of *CTNNA3*, a protein-coding gene involved in cell–cell adhesion with known cell type expression, before and after CellBender noise removal. Gene expression assay results become much more cell-type specific. © 2023, Fleming, S. J. et al.

BEHIND THE PAPER

CellBender was the result of a fruitful partnership between the Data Sciences Platform and the Precision Cardiology Lab (PCL) at the Broad Institute. The PCL team, led by Patrick Ellinor, aimed to build a comprehensive cellular atlas of the healthy human heart using scRNA-seq technology. Despite the relentless efforts of wet-lab scientists to optimize the sample preparation steps, we kept getting substantial off-target expression of marker genes in the wrong cell types. Our 'eureka!' moment came when we realized that such noise is an unshakeable aspect of the technology and could not be fully mitigated experimentally. We also realized that the commonly ignored empty droplets could be used to learn and remove the background noise profile from cell-containing droplets. The solution was hiding within each dataset all along. We turned this key insight into robust software for ourselves and for the larger single-cell genomics community. **M.B.**

REFERENCES

1. Aldridge, S. & Teichmann, S. A. Single cell transcriptomics comes of age. *Nat. Commun.* **11**, 4307 (2020). **This review article presents an overview of the past decade of technological advances in single-cell RNA sequencing.**
2. Kingma, D. P. & Welling, M. Auto-encoding variational Bayes. Preprint at https://doi.org/10.48550/arXiv.1312.6114 (2013). **This preprint establishes the foundations for the use of stochastic variational inference in amortized Bayesian models using modern gradient-based optimization techniques.**
3. Bingham, E. et al. Pyro: deep universal probabilistic programming. *J. Machine Learning Res.* **20**, 1–6 (2019). **This paper establishes a general programming language to facilitate stochastic variational inference in Bayesian models using modern gradient-based optimization techniques.**
4. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017). **This paper introduces CITE-seq, a technique for quantifying antibody-labeled proteins along with RNA in single-cell experiments.**
5. Chen, X. et al. A rapid and robust method for single cell chromatin accessibility profiling. *Nat. Commun.* **9**, 5345 (2018). **This paper introduces the single-cell version of ATAC-seq, enabling chromatin accessibility measurements.**

FROM THE EDITOR

"An indispensable contributory factor to the success of single-cell analysis is the development of well-performing and robust computational methods that unleash its full potential in biological discovery. As a useful addition to this toolbox, CellBender models and removes systematic background noise from droplet-based single-cell assays, enhancing the biological signal and improving downstream analysis." **Lin Tang, Senior Editor, *Nature Methods*.**