Robust, doublet-free, and low-cost molecular profiling of biological systems.

INTRODUCTION

The complex behaviors and functions of biological systems are encoded in their mRNA expression patterns, leading to significant interest in unsupervised methods for quantitative transcriptomic profiling. However, the field of single cell sequencing is associated with technical and analytical challenges- in particular, sensitivity, scalability, and the potential for artifactual 'doublets'. In principle, a method that overcomes these challenges could enable robust, high-sensitivity, and routine molecular profiling to be universally applied to any biological system.

We reasoned that the vast majority of technical challenges for scRNA-seq arise from a dogmatic insistence on performing separate mRNA quantifications in individual cells. This severely restricts the available pool of mRNA molecules for profiling while introducing inherent scalability challenges, namely, that the size of the dataset increases linearly with the number of cells profiled.

After careful consideration, we concluded that the primary limitations in scRNA-seq originate from two steps: an initial round of cellular barcoding, and the subsequent multiplexing of multiple cells together. We therefore propose an alternative workflow, where all molecules in a sample are simultaneously profiled, without regard for their specific cell-of-origin. We refer to our approach as Barcode-Free, Unmultiplexed, Low-throughput mRNA Quantification (BULQ-seq). We identify substantial improvement across a wide variety of metrics when comparing BULQ-seq to previously developed scRNA-seq technologies. We conclude that BULQ-seq represents a powerful approach for mRNA profiling.

RESULTS

BULQ-seq can tell the difference between a human and a mouse

We first benchmarked BULQ-seq by performing a 'barnyard' (species-mixing) experiment. We performed BULQ-seq on human HEK293 and Mouse 3T3 cells, followed by next-generation sequencing, and analyzed the results using a principal components analysis (Figure 1). Reassuringly, we observed that the two samples clearly separated along the first principal component, demonstrating that BULQ-seq can unambiguously identify the species-of-origin of each sample.

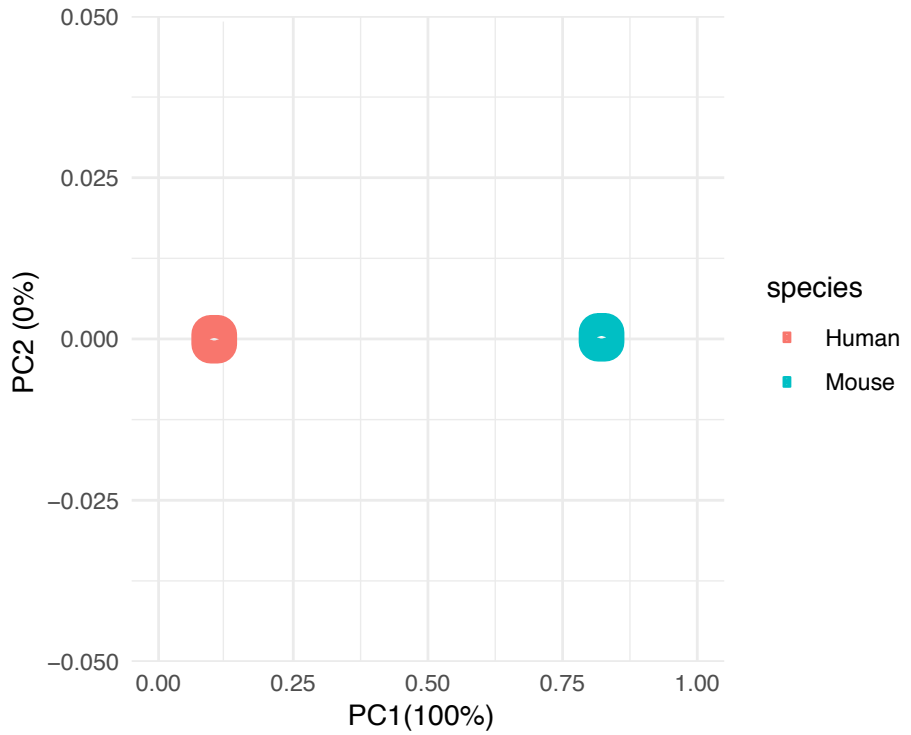

Figure 1: BULQ-seq correctly classifies human and mouse as separate organisms

We obtained similar, but slightly lower, species classification accuracy after training a standard architecture neural network with 42 convolutional layers, ReLU activation functions, max-pooling and batch normalization. Notably, we did not observe evidence of 'doublets' that represent potential artifacts for scRNA-seq data.

BULQ-seq sensitivity *is signif…, we mean, much better, than scRNA-seq*

We hypothesized that BULQ-seq might ameliorate the extensive false negative (dropouts) associated with scRNA-seq. When examining a scRNA-seq dataset produced on cell lines, we found that only 1% of the elements in the count matrix were non-zero. Remarkably, 100% of the measurements in BULQ-seq were non-zero. Following recent guidelines, we eschewed the use of statistical tests to compare the approaches. However, when we showed the data to a few people, everyone agreed the BULQ-seq was "much better" (Figure 2).

Unfortunately, we found that BULQ-seq data exhibited more non-zero values than could be modeled using a standard Zero-Inflated Negative Binomial (ZINB) distribution. We therefore propose the use of a Nonzero-inflated ZINB to model BULQ-seq data. The development of new statistical methods leveraging the NZIZINB distribution represents an important future analytical challenge.

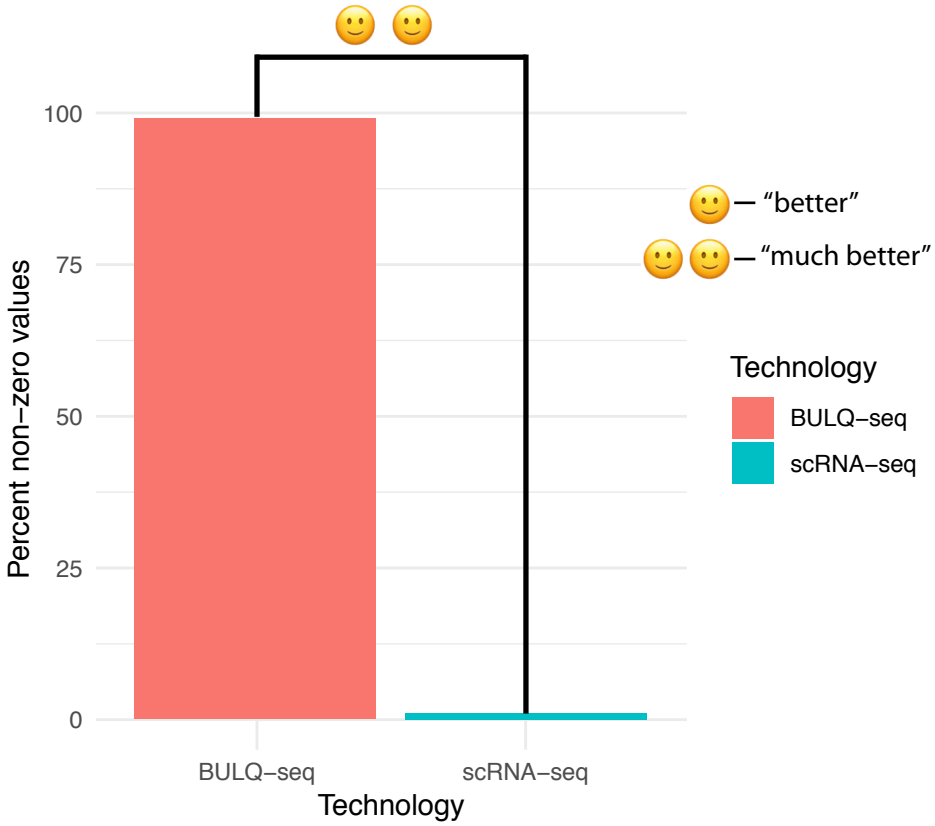

Figure 2: Sensitivity of BULQ-seq compared to scRNA-seq

BULQ-seq experiments and analysis, can scale to millions, and even billions, of cells

While scRNA-seq technologies now enable the generation of large datasets, encompassing potentially millions of cells, these datasets introduce significant analytical challenges, particularly in the absence of high-memory servers. We therefore performed experiments on large cell populations to test the BULQ-seq scalability. Remarkably, we were able to cluster BULQ-seq samples in less than one second on a standard MacBook computer with minimal memory requirements, even on samples representing one million cells (Figure 3). Linear extrapolation suggests that even 1,000,000,000 cells can therefore be profiled and analyzed with BULQ-seq.

DISCUSSION

We have demonstrated that BULQ-seq represents a robust, low-cost, and scalable method for unsupervised molecular profiling. We benchmark our approach on cell lines, but hypothesize that future technological advances could enable BULQ-seq to be applied to *ex-vivo* tissue samples as

well. Such an advance would enable routine transcriptomic profiling to identify molecular differences across sample conditions, genotypes, time points, and even species.

We note that transcriptomic readouts represent only one of the assays with associated challenges at single-cell resolution. Multiple molecular modalities, including DNA mutations, epigenetic marks, chromatin accessibility, and protein levels also may benefit from being measured across multiple cells in bulq.

Acknowledgements

We acknowledge the wonderfully open, collaborative, and innovative single-cell genomics community which (all jokes aside), has been enormous fun to be a part of.

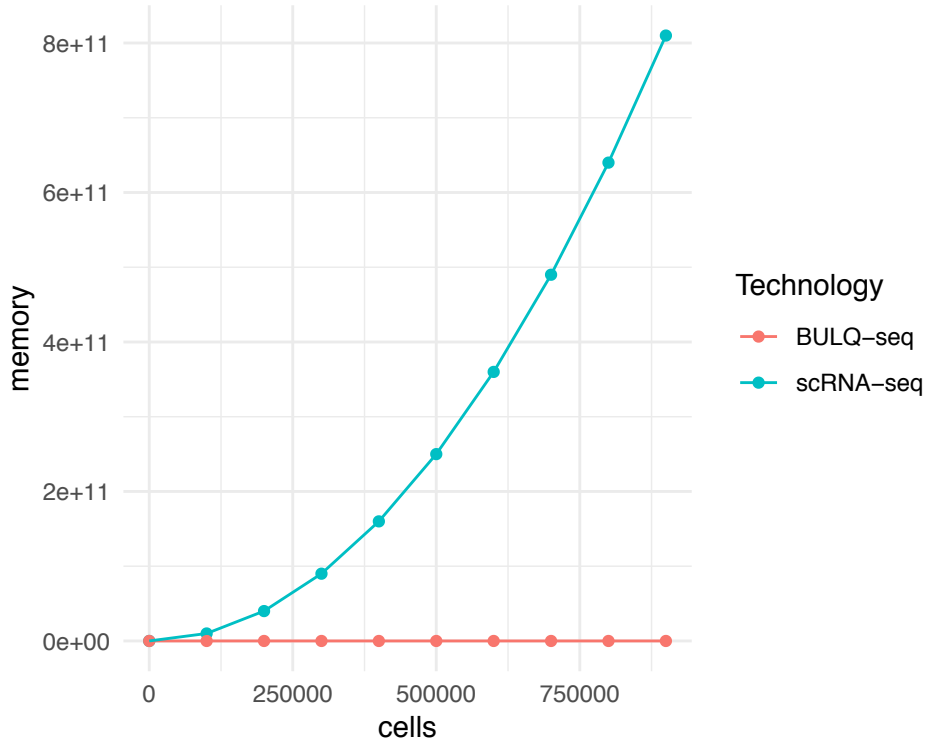

Figure 3: BULQ-seq can analyze millions to billions of cells in a single experiment, using a standard laptop computer.