SUMMARY

Single-cell RNA-seq gives access to gene expression measurements for thousands of cells, allowing discovery and characterization of cell types. However, the data is noise-prone due to experimental errors and cell type-specific biases. Current computational approaches for analyzing single-cell data involve a global normalization step which introduces incorrect biases and spurious noise and does not resolve missing data (dropouts). This can lead to misleading conclusions in downstream analyses. Moreover, a single normalization removes important cell type-specific information. We propose a data-driven model, BISCUIT, that iteratively normalizes and clusters cells, thereby separating noise from interesting biological signals. BISCUIT is a Bayesian probabilistic model that learns cell-specific parameters to intelligently drive normalization. This approach displays superior performance to global normalization followed by clustering in both synthetic and real single-cell data compared with previous methods, and allows easy interpretation and recovery of the underlying structure and cell types.

KEYWORDS

single-cell; RNA-seq; clustering; normalization; imputing; dropouts

INTRODUCTION

Recent advances in single-cell RNA-seq (scRNA-seq) technologies has enabled measurements of gene expression at the resolution of individual cells [1–3]. As opposed to traditional bulk RNA-seq where expression is measured by an average readout across a bulk of cells, scRNA-seq allows characterizing mixtures of cell types and functions in various tissues through the observed heterogeneity in gene expression [4–8]. This can lead to exciting opportunities for building an atlas of cells similar to how DNA sequencing technologies allowed mapping genomes.

However, scRNA-seq involves significant technical variation that can be mistaken for biological heterogeneity and there is a strong need for robust algorithms that distinguish technical variation from biological signal in scRNA-seq data.

We can view single-cell data as sampling the transcriptome, and current scRNA-seq technologies capture 5-20% of total transcripts per cell which therefore leads to high chance of missing low expression genes, termed as dropouts (see example in Figure 1). Hence, imputing dropouts is essential for downstream analysis of gene expression. Furthermore, generating adequate library for sequencing requires multiple rounds of exponential amplification which can have different effects on different cell types and create a heavy-tail distribution of library size. Finally, the data are also prone to technical variation due to e.g. different enzyme activity or lysis efficiency. Resolving these problems and correcting for this variation are thus some of the key challenges in analyzing scRNA-seq data [9–11].

In order to characterize cell types, current analysis methods applied to scRNA-seq data perform a global normalization by median or mean library size [12], which is an approach adapted from bulk RNA-seq analysis methods. This normalization however is invariant to cell types (Figure 1), would not resolve dropouts and the added noise can lead to spurious differential expression (Figure 1). This is especially problematic in samples with heterogeneity in cell sizes, and leads to undesired grouping of cells due to technical artifacts rather than true biology. Thus, normalization prior to clustering fails to consider cell type dependent variation. The cell types however are not known a priori, to be used for normalization or imputing dropouts, hence the normalization and clustering of single-cell data becomes a circular problem.

Computational tools that can separate biological from technical variation in single-cell data are very limited. A mixture model has been presented by Kharchenko et al. [11] for dropouts and detected genes. Also,
down-sampling approaches have been suggested for normalizing cells e.g. in Jaitin et al. [2], which have the downside of discarding a large portion of data. BASICS [14] presents a Bayesian model for normalization in cases where spike-in genes are available. These models however, do not perform any clustering of cells and hence do not account for cell-type variations.

A more general cell-type dependent normalization is desirable that does not necessarily require spike-ins, mainly because the use of spike-ins in recent scalable single-cell technologies such as in-drop [15] and drop-seq [16] in not straightforward and some technical variations such as lysis efficiency occur prior to adding spike-ins. This motivated our approach BISCUIT (Bayesian Inference for Single-cell ClUstering) [13] which integrates clustering and normalization into one model. This is done through incorporating cell-specific parameters into a Hierarchical Dirichlet Process mixture model (HDPMM) allowing simultaneous inference of clusters according to similarity in gene expression patterns as well as co-expression patterns and identifies technical variations per cell (Figure 2). BISCUIT is the first method to simultaneously correct technical artifacts (without requiring spike-ins) and infer clusters of cells based on gene co-expression patterns.

**DISCUSSION**

We have shown superior performance of BISCUIT on both synthetic and scRNA-seq data. We compared the performance to the naive HDPMM [18] along with normalization to median library size as well as Generalized Linear Model-based normalization followed by clustering, which showed significant improvement with BISCUIT. We also showed the robustness of our inference with a model mismatch experiment, where synthetic data was generated from models other than the log-Normal model [13].

We then evaluated the performance of BISCUIT on single-cell data from 3005 mouse cortex cells in [17]. This dataset and the identified cell types were ideal for this test due to the deep coverage. We showed BISCUIT can correctly identify the ground truth
cell types and our identifiability guarantees provide interpretability for inferred parameters. We also showed its capabilities in imputing dropouts and normalizing data. We demonstrated that this inference is superior to step-wise normalizing and then clustering in a decoupled manner [13].

BISCUIT can identify more refined sub-types of immune cells as shown in an example dataset of 6000 peripheral blood mononuclear cells (PBMCs) from a healthy donor from Zheng et al. [19]. Figure 3 shows the imputed and normalized data projected with TSNE and labeled with inferred clusters. We identified multiple sub-types of naive, memory and cytotoxic T cells, two different clusters of natural killer cells, and multiple myeloid clusters. These distinct clusters are not revealed with common clustering and normalization methods [19].

The scRNA-seq technology allows measurement of heterogeneous expression, however the data is also extremely sparse and noisy. Appropriate computational techniques are not yet fully developed for analyzing such data. We address a number of these problems with BISCUIT, that concurrently clusters cells and learns co-expression structures specific to each cluster while inferring normalization and imputing parameters.

BISCUIT significantly outperforms previous methods and provides a near accurate inference of ground truth clusters. BISCUIT can be utilized to understand tissue heterogeneity and discovery of novel cell types especially in cases involving significant diversity of cell types.

Figures 3: PBMC clusters identified by BISCUIT.

REFERENCES


