Genomics Proteomics Bioinformatics 14 (2016) 21–30

HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

REVIEW

Single-cell Transcriptome Study as Big Data

Pingjian Yu [a], **Wei Lin** [*,b]

Genomics and Bioinformatics Lab, Baylor Institute for Immunology Research, Dallas, TX 75204, USA

Received 17 November 2015; revised 9 January 2016; accepted 10 January 2016
Available online 11 February 2016

Handled by Hongxing Lei

KEYWORDS

Single cell;
RNA-seq;
Big data;
Transcriptional heterogeneity;
Signal normalization

Abstract The rapid growth of **single-cell RNA-seq** studies (scRNA-seq) demands efficient data storage, processing, and analysis. **Big-data** technology provides a framework that facilitates the comprehensive discovery of biological signals from inter-institutional scRNA-seq datasets. The strategies to solve the stochastic and heterogeneous **single-cell** transcriptome signal are discussed in this article. After extensively reviewing the available **big-data** applications of next-generation sequencing (NGS)-based studies, we propose a workflow that accounts for the unique characteristics of scRNA-seq data and primary objectives of **single-cell** studies.

Introduction

Multi-institutional collaborative omics studies on the next-generation sequencing (NGS) platform have generated petabytes of data that constitute 'big data' from the perspective of scale and complexity [1–6]. Particularly, transcriptomics studies using the RNA-seq technique have become revolutionary and powerful [7–9]. Scientists have now moved one step forward to single-cell RNA sequencing (scRNA-seq) by employing new protocols for single cell isolation, low-input RNA extraction, reverse transcription, and unbiased amplification [9–13]. Given the high anticipated value of single-cell transcriptomics, explosive growth of scRNA-seq data is expected in the next 5–10 years. Consequently, uncovering the hidden pattern, connectivity, and interactions of such huge and heterogeneous data will be a major challenge.

Without a doubt, the detailed and extremely-valuable information that single-cell technology provides is at a significant cost due to sophisticated data acquisition, large data-storage requirements, as well as challenging data processing and management. Big data incorporate a body of technologies including computational parallelization and distribution, data visualization, and data integration that are used to reveal the hidden associations within large datasets that are diverse, complex, and of a massive scale. Data-intensive scientific discovery has been proposed as the 4th paradigm of scientific research [14], following and interacting with the other three paradigms – theory, experimentation, and simulation modeling. In 2001, Doug Laney defined characteristics of big data in three dimensions, *i.e.*, increasing volume (amount of data), velocity (speed of data I/O), and variety (range of data types and sources) [15]. While agreeing that volume, variety, and velocity are the quantitative characteristics of big data, Ivanov et al. [16] added that variability (the inconsistency the data can show over time) and veracity (the quality of captured data) are the qualitative characteristics of big data.

* Corresponding author.
E-mail: Wei.Lin@BaylorHealth.edu (Lin W).
[a] ORCID: 0000-0002-8422-7645.
[b] ORCID: 0000-0002-7506-3466.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

http://dx.doi.org/10.1016/j.gpb.2016.01.005

Big-data technology has many applications in biomedical research [17–20]. Particularly, high-throughput molecular and functional profiling of patients using NGS or single-cell technology is the key driving force of precision medicine [21–24]. By examining the annual growth of scRNA-seq datasets uploaded to the NCBI Gene Expression Omnibus (GEO) database [25] and the increasing number of new articles in PubMed over the past 7 years that involve scRNA-seq and big-data (**Figure 1**), we expect the extensive integration of big data and scRNA-seq technologies.

In the following sections, we will discuss the characteristics of single-cell transcriptomics, especially scRNA-seq, data as examples of big data. We will discuss how to adapt single-cell transcriptomics study to big-data infrastructure such as Hadoop and MapReduce.

Transcriptional stochasticity and cellular heterogeneity

scRNA-seq is always compared to bulk RNA-seq in terms of signal profile and noise level. In addition to the descriptive keyword like high resolution, stochasticity and heterogeneity are also frequently used to feature the single-cell transcription [26–29]. Most of the scRNA-seq investigators have experience with zero-inflation transcriptional signals. Some of them tend to regard this phenomenon as technical dropout. We prefer to use the phrase "bimodality" to delineate the signal distribution, since recent results have shown that the low transcriptional values are biologically meaningful signals rather than technical dropout. Shalek et al. have revealed the bimodality

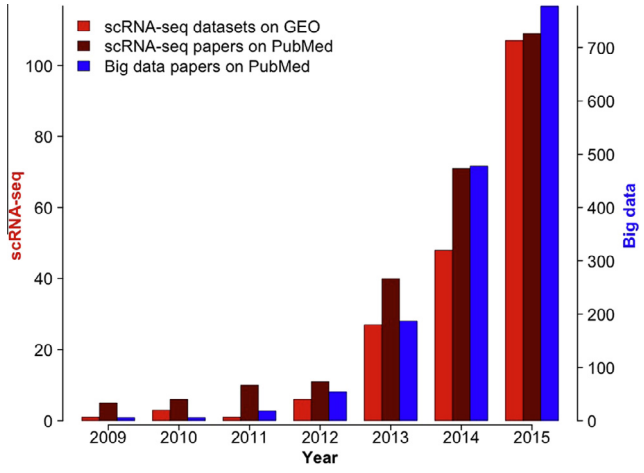

Figure 1 Number of papers/datasets addressing single-cell data and big data

Searches were performed on January 04, 2016 on http://www.ncbi.nlm.nih.gov/gds for datasets and http://www.ncbi.nlm.nih.gov/pubmed for papers. Data were obtained according to the search criteria as follows filtered by year: (1) for scRNA-seq datasets on GEO: "single cell"[All Fields] AND "Expression profiling by high throughput sequencing"[Filter]; (2) for scRNA-seq papers on PubMed: "single cell"[All Fields] AND ("rna-seq"[All Fields] OR "rna sequencing"[All Fields] OR ("sequencing"[All Fields] AND "transcriptome"[All Fields])); and (3) for big-data papers on PubMed: "big data"[All Fields] OR "hadoop"[All Fields].

of single-cell expression and splicing using both scRNA-seq and RNA fluorescence *in situ* hybridization (RNA-FISH) [30]. The two modes in an expression profile can be attributed to the "on" or "off" transcriptional status. **Figure 2** demonstrates two clusters of cells showing different expression level and the change of the ratio of on/off status of a marker gene *MYH2* over time during human myoblast cell differentiation using both scRNA-seq and RNA-FISH [31]. The aforementioned studies indicate that even from a seemingly homogeneous population, many genes are expressed in a stochastically-bursting fashion and their abundance exhibits a bimodal distribution in the cell population examined. The traditional RNA-seq analysis method rarely takes such transcriptional bimodality into account. Further investigation on co-bursting networks have validated the biological significance of the "bimodality" rather than just relegating it to technical dropout [31].

Several computational models have been proposed to analyze transcriptional stochasticity and cellular heterogeneity in scRNA-seq data in the context of zero-inflation or bimodality. Kim and Marioni [32] use a mixture of two Poisson distributions to model theoretical kinetics for 'bursty' gene expression. However, in the presence of massive variability, the model is compromised by excessive over-dispersion in read counts. Kharchenko et al. take the probability of "dropout" into consideration in their differential-expression algorithm [33]. Pierson and Yau proposed using zero-inflated factor analysis to perform dimensionality reduction [29]. Gu et al. use a mixture of two negative binomial distributions to model over-dispersed read counts generated from a gene's two distinct biological states: an 'on' component and an 'off' component [31]. All of these four studies acknowledge the fact that single-cell transcription signals cannot be solved by unimodal statistics. Gu et al. first introduced the statistics term "bimodal proportion" to measure the ratio of two signal modes in a single-cell population. The functional enrichment of co-bursting transcription supports the biological significance of transcriptional bursting over technical dropout. The value of "bimodal proportion" ranges from 0 to 1 and notably, it can be compared across different datasets without additional normalization.

The opportunities and challenges of scRNA-seq

Single-cell transcriptomics provides us unprecedented opportunity to understand the transcriptional stochasticity and cellular heterogeneity in great detail, which are crucial for maintaining cell functions and for facilitating disease progression or treatment response [34–38]. Such stochasticity and heterogeneity are always masked in bulk-cell studies [27]. Recent single-cell applications have utilized a broad range of tissues [28,39–42], stem cell lines [43,44] and cell populations with clinical backgrounds [45]. The cell types that have been interrogated using scRNA-seq in the GEO database are briefly summarized in **Table 1**.

scRNA-seq is one of the most promising technologies for single-cell transcriptomics [46,47]. Nevertheless, it also poses big challenges, largely stemming from the aforementioned big-data characteristics with regard to the data management, query, and analysis. There are five 'V's to consider for scRNA-seq data. (1) Volume. NGS data has become one of the largest big-data domains in terms of data acquisition,