International Journal of Biochemistry and Cell Biology 90 (2017) 161–166

Contents lists available at ScienceDirect

International Journal of Biochemistry and Cell Biology

journal homepage: www.elsevier.com/locate/biocel

Delineating biological and technical variance in single cell expression data

Ángeles Arzalluz-Luque[b], Guillaume Devailly[a], Anna Mantsoki[a], Anagha Joshi[a,*]

[a] Division of Developmental Biology, The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush Campus, Midlothian, EH25 9RG, UK

[b] Genomics of Gene Expression Laboratory, Centro de Investigación Principe Felipe (CIPF), Carrer d'Eduardo Primo Yúfera 3, 46012, Valencia, Spain

ARTICLE INFO

Keywords:
Single cell
RNA-seq
Noise
Variability

ABSTRACT

Single cell transcriptomics is becoming a common technique to unravel new biological phenomena whose functional significance can only be understood in the light of differences in gene expression between single cells. The technology is still in its early days and therefore suffers from many technical challenges. This review discusses the continuous effort to identify and systematically characterise various sources of technical variability in single cell expression data and the need to further develop experimental and computational tools and resources to help deal with it.

1. Introduction

Next generation sequencing (NGS) technologies have revolutionized the way of approaching molecular biology to advance our understanding of the working principles of biological systems, including identification of the building blocks. Genome sequencing is now widely used across diverse fields in biology, ranging from medicine and population studies to animal breeding. However, the information encoded in the genome is static, an ensemble of the cell's potentialities manifest once the process of transcription is triggered. Therefore, studying the transcriptome is essential to understand how genome information is decoded in a particular cell specific context, as the cells ultimately constitute as adaptable and dynamic entities.

NGS evolved from a range of laboratory techniques developed for expression analysis over the years. Initial experimental approaches include the early Northern blotting (Alwine et al., 1977), which targets a single gene and measures its expression levels through hybridization of a labelled probe. Advances in increasing throughput of transcriptome studies came with microarrays (Schena et al., 1995), a technology that used a similar probing approach, but increased the number of quantified transcripts by using tens of thousands of probes on a chip, onto which the RNA sample is hybridized. Both approaches described above are limited by the fact that probe design requires previous knowledge of the transcript sequences. To this end, the use of sequencing technologies such as Sanger sequencing and its derivatives, including expressed sequence tag (EST), improved access to the diversity of the transcriptomic landscape by overcoming the probe design constraint. Currently, however, the most widely used application of NGS technologies to transcriptomics is RNA sequencing (RNA-Seq) (Mortazavi et al., 2008), by which -potentially- all mRNA molecules in a cell can be sequenced, and hence characterized and quantified.

The importance of RNA-Seq is not only founded in its ability to access unknown transcripts and spliced variants, but also to increase microarray's dynamic range (i.e. the lowly expressed transcripts could be successfully detected) and sensitivity (i.e. the expression level measurements show higher accuracy). RNA-Seq has therefore become the technology of choice to provide a high-throughput and fully quantitative approach to studying the transcriptome of a broad range of species, including the ones lacking full genome sequence availability. The technology has therefore been widely applied, replacing microarrays for the analysis of gene expression profile differences among cell populations, comparative transcriptomics and disease biomarker identifications (Wang et al., 2009). However, it has become apparent that not all cells within a population behave similarly when it comes to gene expression or splicing and, in this context, bulk RNA-Seq fails to address some important questions (Sandberg, 2014).

2. Single cell expression technologies and applications

Over the years, single-cell approaches have been developed in combination with microscopy to visualize gene expression patterns in individual cells. For example, single-molecule RNA fluorescence *in situ* hybridization (RNA FISH) technology combines probe hybridization with fluorescent labelling to resolve the location of a target transcript (Lubeck and Cai, 2012). The main disadvantage of RNA FISH is that, although parallelizable, it only allows access to a limited subset of genes. The implementation of single-cell microarrays (Iscove et al., 2002) presented itself again as a high-throughput alternative to RNA

* Corresponding author.
E-mail address: Anagha.Joshi@roslin.ed.ac.uk (A. Joshi).

http://dx.doi.org/10.1016/j.biocel.2017.07.006
Received 13 December 2016; Received in revised form 11 July 2017; Accepted 13 July 2017
Available online 15 July 2017

FISH, and although it helps overcome this main limitation, it suffers the drawbacks of bulk microarrays. Furthermore, the limited amount of starting material and the relatively low sensitivity of microarrays enforced high levels of pre-amplification, which can introduce significant biases.

In the light of these limitations, RNA Sequencing was implemented at the single-cell level, theoretically enabling access to the transcriptome of every individual cell in a population (Ramsköld et al., 2012; Tang et al., 2010). Essentially, single-cell RNA-Seq requires the following steps: single cell isolation, mRNA capture and reverse transcription to cDNA, cDNA amplification to improve the low transcript yields rendered by single cells, and sequencing (Picelli et al., 2014).

Over the last few years, single-cell RNAseq has been proven useful to unravel biological phenomena that can only be understood in the light of differences in gene expression between single cells, including:

- **Studying early embryonic development**: In early stages of embryonic development, only a few cells contribute to activating the molecular machinery for cell differentiation. The characterisation of transcription changes in individual inner cell mass (ICM) cells of blastocysts was proven crucial to understand the complex transition from ICMs to embryonic stem cells (ESCs) (Tang et al., 2010). This approach set a precedent for subsequent studies of later and more complex stages in the process of cell commitment and differentiation into specific lineages. In this context, a spatial-temporal profiling of gene expression in embryonic development in *Caenorhabditis elegans* was used to study the evolution of the germ layers. The authors noted that the gene expression program of the mesoderm is induced after those of the ectoderm and endoderm and strikingly, the endoderm gene expression program activates earlier than ectoderm expression program, a phenomenon that is conserved across many species (Hashimshony et al., 2014).

- **Measuring diversity in cell populations**: Single cell analysis is the most powerful tool to study the diversity between individual cells treated as homogenous in a typical bulk RNA-seq experiment. It has proven potential of providing valuable insights in some of the key problems in biomedical field e.g. tumour heterogeneity, which poses substantial challenges in cancer treatment. For example, single cell analysis can unravel intra- and inter-tumour differences (Patel et al., 2014) as well as distinguishing between malignant and non-malignant cells (Tirosh et al., 2016).

- **Identification of new rare cell types**: Complex tissues often contain previously unidentified cell types that cannot be studied using bulk RNA-Seq, as it provides only an estimate of expression influenced by the abundance of the different cell types present. Single cell transcriptomics provides a promise to address this underlying diversity in order to assess meaningful differences in phenotype. Using this strategy, authors identified and characterised a rare population of dormant neural cells which were activated upon brain injury (Llorens-Bobadilla et al., 2015). Another example is the development of a computational approach (scLVM) to identify subpopulations of cells using latent variable models to account for hidden factors such as cell cycle. Namely, different sub-populations of cells corresponding to the differentiation stages during naive T cells to T helper 2 cells were identified (Buettner et al., 2015). Identification of rare cells is of high relevance, particularly characterisation of progenitor cells to understand vertebrate development. To this end, single cell RNA-Seq has been used to unravel transcription heterogeneity and lineage commitment in myeloid progenitors, in order to further demonstrate how Cebpe deletion results into diminishing of certain myeloid lineages (Paul et al., 2015).

- **Mapping developmental hierarchies**: transcription dynamics during development and disease can be studied in much greater details using single cell studies, as bulk RNA-seq, by averaging out signal from multiple cells, misses out on the signal from rare developmentally relevant cells. However, single cell transcriptome profiling over time is not feasible. Taking advantage of the fact that an experiment characterising hundreds of unsynchronised cells from a population typically provides a snapshot of cells at various stages during differentiation, various methods for pseudo-time inference form single cell RNA-seq data have recently been developed (Haghverdi et al., 2016; Reid and Wernisch, 2016; Trapnell et al., 2014) and reviewed (Bacher and Kendziorski, 2016). As an example of this, single cell expression data has successfully been used to reconstruct the developmental progression of cells and identify transient and terminal states together with the branching decisions (Treutlein et al., 2014).
- **Understanding diverse features of transcription control**: Single cell transcriptomics has facilitated unravelling mechanistic details of transcription control such as kinetics and bimodality, as well as studying other features such as allelic biases and transcription networks. Even though single cell transcriptomics does not measure expression changes in one gene over time, an overall rate of transcription between individual cells can be acquired and approximately represent the stochasticity of expression of a vast number of genes, facilitating estimation of kinetics of gene expression (Kim and Marioni, 2013). Recent studies have unravelled the stochastic modes of gene expression, which were not apparent at the population level. The functional implications of this stochasticity (i.e. changes on the phenotype of seemingly identical cells) can be explained by variation in gene regulation processes across individual cells (Munsky et al., 2012). Allelic biases in gene expression have also been investigated, including stochastic allelic expression in early embryogenesis (Tang et al., 2011) as a particularly relevant example. Finally, single cell transcriptome data is successfully used to reconstruct gene regulatory networks (Moignard et al., 2015).

In summary, single cell analysis has a huge potential to bring new insights into diverse fields of biological research. In the next sections, we will put this in context by discussing the technical challenges currently faced by single cell analysis to extract the 'biological' or functionally relevant variability from the data, which hinder its theoretical potential.

3. Technical variability in single cells

Despite the promise held by the approach, single-cell RNA-Seq is not free from biases. Quite contrarily, the low availability of starting material (i.e. RNA extracted from an individual cell) introduces high technical variability, making single-cell RNA-Seq data analysis especially challenging (Stegle et al., 2015). This typically results into many missing values (technical) or true absence of expression (biological) in typically lowly expressed transcripts, and discriminating both, although important, is not currently feasible. Furthermore, the necessary amplification of starting material introduces additional biases, such as 3′ end enrichment of signal and preferential amplification of some transcripts and/or mRNA fragments. Reassuringly, bulk RNA-Seq experiments can be recapitulated *in silico* by pooling 30 or more single cell transcriptomes *in silico* (Marinov et al., 2014), and used to estimate technical variability.

The technical variability in single-cell RNA-Seq can be divided into two categories: Inter-cell variability and within cell variability (Fig. 1).

3.1. Inter-cell variability

Inter-cell variability can appear as a result of the biological process under scrutiny, or can be due to unrelated phenomena, which can act as confounding factors. For example, the differences in cell cycle stage are