Robust inference of expression heterogeneity from simultaneous single- and k-cell profiling

Manikandan Narayanan[1], Andrew J. Martins[1], and John S. Tsang[1,*]

***Short Abstract* — Cell-to-cell gene-expression heterogeneity is a pervasive phenomenon, but its accurate quantification remains challenging as the level of many transcripts in single cells falls near/below the detection limit of even the most sensitive measurement technologies currently available. We present a combined experimental and computational strategy for inferring cellular heterogeneity parameters through Bayesian integration of simultaneously obtained expression profiles from single and random pools of k (e.g., 10) cells. Simulations/experiments show our strategy combines the direct interpretability of single-cell data with enhanced sensitivity of pooled-cell measurements to enable quantitative comparison of cell-to-cell variations across cellular states or conditions using modern multiplexed technologies.**

***Keywords* — cell-to-cell heterogeneity, single-cell data analysis, k-cell data, Bayesian inference, stochastic gene expression, macrophage activation.**

I. Introduction

QUANTIFYING heterogeneity in gene expression across individual cells could help identify novel cell types in tissues and address fundamental questions such as how cellular fluctuations in gene expression propagate along the gene regulatory network. Despite rapid technological advances, accurate measurement of single-cell expression is a major challenge, particularly because many mRNAs are expressed at levels close to or below the detection limit of current profiling technologies [1]. Indeed, typical single-cell gene-expression profiles obtained by quantitative PCR (qPCR) or RNA-Seq contain a substantial number of zero or non-detected measurements, which are unlikely to be entirely attributable to cells expressing zero transcripts and instead may arise from technical factors such as missed capture/amplification of mRNA transcripts [1-2]. Measuring randomly sampled pools of a small number of cells (with the number of cells per pool denoted by k, such as k=10) offers more robust detection due to the increased amount of input mRNA and has been used to assess cell-to-cell heterogeneity within the sampled population, such as to infer whether expression distributions are bimodal [3]. However, information on single-cell variations using data from such k-cell pools is nonetheless indirect and the lack of measurements on individual cells would hinder applications such as novel cell type identification.

II. Results

Here we present a strategy for quantifying cellular heterogeneity that combines simultaneous expression profiling of single and k-cell samples from a cell population with a newly developed statistical model and computational method for Bayesian inference of heterogeneity parameters. Our method (called QVARKS) **q**uantifies the degree as well as the statistical uncertainty of expression **var**iation across cells by integrating **k**- and **s**ingle-cell data under explicit models of technical detection limits. Across diverse simulation scenarios representative of modern multiplexed technologies, we show that our approach allows robust inference of cellular heterogeneity parameters of difficult-to-detect transcripts even when technical noise or incomplete single-cell information hinder robust inference from either data alone.

When applied to single/10-cell expression data generated from human macrophages in resting vs. inflammatory conditions, we show our approach is able to effectively disentangle condition-specific biological cell-to-cell variation from detection limit induced technical noise. In addition, our analysis helped reveal several distinct modes of gene-specific responses upon cellular activation involving significant changes in the fraction of "on" cells, or in the average expression level in "on" cells, or both.

Thus QVARKS offers a promising way forward for statistically rigorous assessments of cellular heterogeneity, and can lead to compelling hypotheses on condition-dependent regulation of gene expression and cellular heterogeneity as demonstrated for an important immune cell type here.

Acknowledgements: We thank Yong Lu, Thorsten Prustel, Vinod Prabhakaran and Katherine Wendelsdorf for valuable inputs, all members of our lab and Ronald Germain for critical feedback, and Cal Eigsti and Kevin Holmes for help with cell sorting. This work is supported by the Intramural Research Program of NIAID, NIH.

[1]Systems Genomics and Bioinformatics Unit, Laboratory of Systems Biology, NIAID, NIH.

[*]Correspondence E-mail: john.tsang@nih.gov

References

[1] Brennecke P, et al. (2013) Accounting for technical noise in single-cell RNA-seq experiments. *Nat Methods* **10**: 1093–1095.

[2] McDavid A, et al. (2013) Data exploration, quality control and testing in single-cell qPCR-based gene expression experiments. Bioinformatics **29**:461–467.

[3] Janes KA, et al. (2010) Identifying single-cell molecular programs by stochastic profiling. *Nat Methods* **7**:311–317.