Genomics 105 (2015) 137–144

Contents lists available at ScienceDirect

Genomics

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

The reduction of gene expression variability from single cells to populations follows simple statistical laws

Vincent Piras, Kumar Selvarajoo *

Institute for Advanced Biosciences, Keio University, 14-1 Baba-cho, 997-0035 Tsuruoka, Japan
Systems Biology Program, Graduate School of Media and Governance, Keio University, 5322 Endo, 252-0882 Fujisawa, Japan

ARTICLE INFO

Article history:
Received 6 November 2014
Accepted 19 December 2014
Available online 29 December 2014

Keywords:
Single cells
Gene expression
Transcriptomics
Noise analysis
Central limit theorem
Law of large numbers

ABSTRACT

Recent studies on single cells and population transcriptomics have revealed striking differences in global gene expression distributions. Single cells display highly variable expressions between cells, while cell populations present deterministic global patterns. The mechanisms governing the reduction of transcriptome-wide variability over cell ensemble size, however, remain largely unknown. To investigate transcriptome-wide variability of single cells to different sizes of cell populations, we examined RNA-Seq datasets of 6 mammalian cell types. Our statistical analyses show, for each cell type, increasing cell ensemble size reduces scatter in transcriptome-wide expressions and noise (variance over square mean) values, with corresponding increases in Pearson and Spearman correlations. Next, accounting for technical variability by the removal of lowly expressed transcripts, we demonstrate that transcriptome-wide variability reduces, approximating the law of large numbers. Subsequent analyses reveal that the entire gene expressions of cell populations and only the highly expressed portion of single cells are Gaussian distributed, following the central limit theorem.

1. Introduction

The issue of cell to cell variability is taking center stage in recent years. Single cell studies relying on individual molecule imaging have shown stochastic fluctuations in the transcriptional machinery, thereby causing molecular constituents, such as proteins concentrations in flow cytometry, to vary among cells of the same type or within clonal populations [1–5]. The development of single cell RNA sequencing techniques are providing further evidences of variability on a transcriptome-wide scale across diverse cell types [6–9].

Molecular variability is not necessarily an unwanted feature or nuisance for living systems. It has been shown on numerous occasions that variable or noisy characteristics play crucial roles in the adaptation of species to environmental conditions or for cell fate decisions [10]. For example, the intestinal cell fate process from early embryonic lineage in *Caenorhabditis elegans* was shown to be regulated by the variability in *end-1* expression, providing the basis for incomplete penetrance [11], or the stochastic regulation in the levels of *comK* alone was necessary to control competent cell fate decision under nutrient-deficient conditions in *Bacillus subtilis* [12]. Thus, variability in molecular expressions has profound effect on species survival. Nevertheless, it will be intriguing to realize how heterogeneous single cells are able to execute well-defined deterministic cell responses required for cellular growth or the collective action of innate and adaptive immune cells against invading pathogens.

To probe into the question, we investigated gene expression distributions of single cells and cell populations in 6 cell types (prostate cancer LNCaP, embryonic kidney HEK293T, lymphoblastoid GM12878 cell lines in human, and Embryonic Stem (ES), Primary Endoderm (PE), Bone Narrow-derived Dendritic Cells (BMDC) in mouse) based on RNA-Seq datasets [7,9,13–15]. Firstly, we compared pairwise expressions scatter between two samples (cell ensembles). Next, we computed Pearson and Spearman's rank correlations and, subsequently, evaluated noise, based on expression variance over mean square values [16,17]. Finally, we compared the gene expression distributions by dividing the transcriptome into gene clusters from low to high mean values. The results clearly show that simple statistical law of large numbers (LLN) and the central limit theorem (CLT) arise when comparing transcriptome-wide data of single cells to increasing cell population, after the removal of transcripts that are biased by technical variability.

2. Results and discussion

2.1. Transcriptome-wide variability in single cells and cell populations

We analyzed gene expressions variability of single cells and cell populations by investigating RNA-Seq datasets obtained from trustworthy sources across diverse cell types (LNCaP, HEK293T, GM12878, ES, PE and BMDC, see Materials and methods) [7,9,13–15]. The datasets were specifically chosen as they provided both single cell and cell population

* Corresponding author at: Institute for Advanced Biosciences, Keio University, 14-1 Baba-cho, 997-0035 Tsuruoka, Yamagata, Japan. Fax: +81 235 29 0830.
E-mail address: kumar@ttck.keio.ac.jp (K. Selvarajoo).

http://dx.doi.org/10.1016/j.ygeno.2014.12.007
0888-7543/

data with different population sizes, to ensure consistency in experimental protocols and to reduce the effect of technical biases when comparing the transcriptomes of single cell and populations of the same cell type.

To visualize gene expressions variability between samples, we plotted the expression values of all genes with pairwise samples consisting of single cells and various cell population sizes ($p = 1, 5, 10,\ldots, 10\,000$) depending on the availability (Fig. 1). The resultant scatterplots generally showed linear distribution along the x = y diagonal, especially for the highly expressed genes from single cells to populations for all cell types. However, for the lowly expressed genes (<100 units), the distributions are largely variable for all types of single cells. In short, these data show that gene expression variability reduces with population size, especially for the lowly expressed ones.

To statistically examine the relationships between the samples, we evaluated Pearson (R) correlation coefficients [18–22] (Materials and methods). Table 1A shows that R increases gradually when population size is increased. Notably, BMDC show an R value of 0.583 for single cells, which increases to 0.998 for 10,000 cell population. Thus, the variations in gene expressions between samples are progressively reduced as the number of cells increases in populations, indicating that the gene expressions converge to their cell-type-characteristic population mean. This is, especially, evidently shown for the lowly expressed genes.

For considering non-linear monotonic dependence, we computed Spearman's rank (ρ) correlation coefficients and found that, although values are generally lower than Pearson correlations, a comparable increase is observed when the population size is increased (Table 1B). Overall, these analyses reveal the emergence of correlated structure of transcriptomes for all investigated cell types when single cells form into populations.

2.2. Transcriptome-wide noise reduces as single cells form into populations

Correlation analyses are a measure of association (linear for Pearson and monotonic for Spearman) [18–22] and, therefore, do not quantify gene expressions variability or noise. The most widely adopted methodology to compute gene expression noise is the squared coefficient of variation, η^2, i.e. the variance in expressions among the total number of cells (σ^2) divided by the squared mean expression (μ^2) [16,17]. This noise

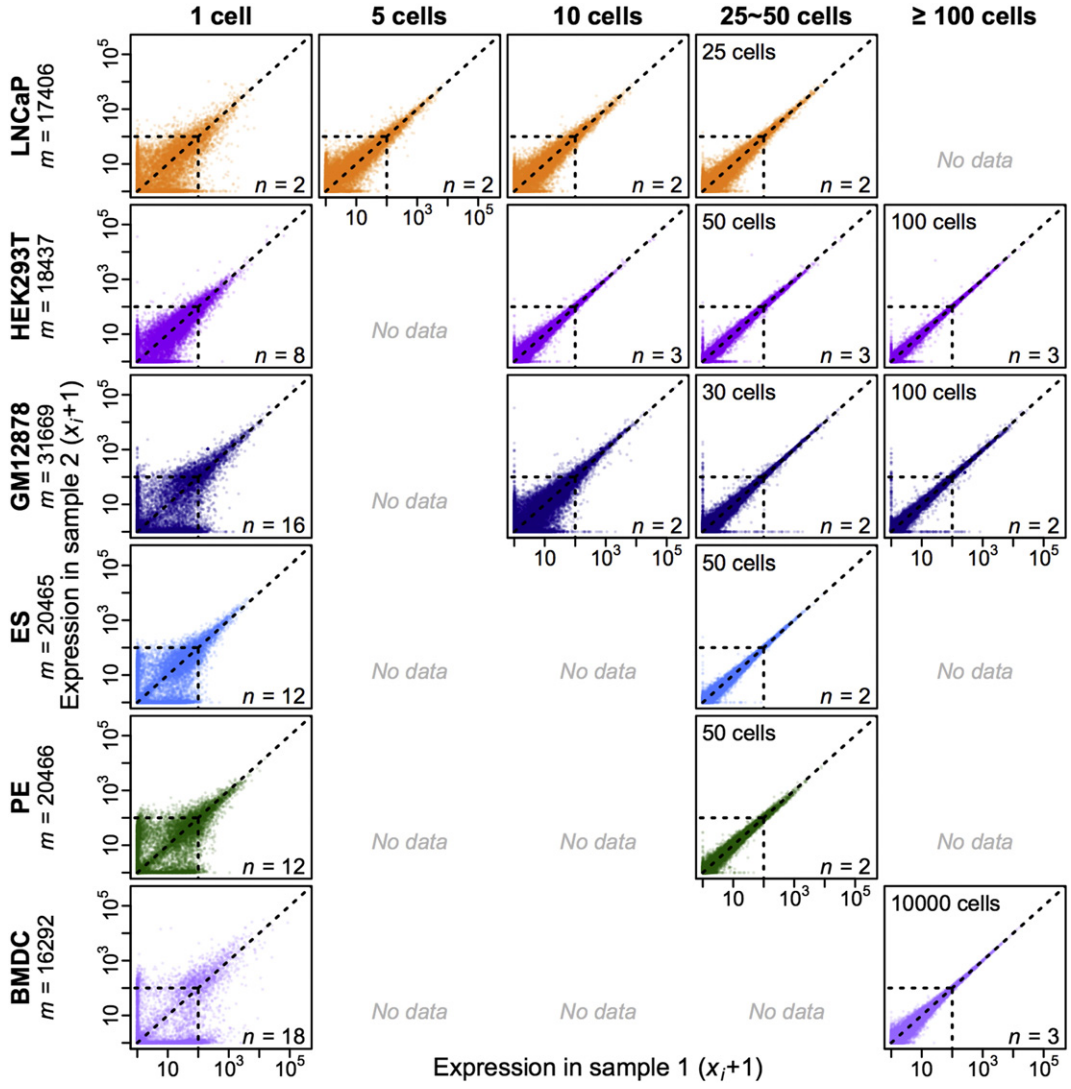

Fig. 1. Expressions scatter reduces as cell population size increases. Gene expression scatterplots between 2 representative samples of single cells and cell populations in increasing sizes for all cell types. Each dot represents the expression values ($x + 1$ to avoid infinite values on the log–log plot) of each gene in both samples. Expression values were downloaded in the format with normalization (RPKM, FPKM or TPM) provided by the original studies. n indicates the available number of samples for single cells or populations, and m, the number of genes (transcriptome size) for each cell type. Dotted squares indicate expressions lower than 100 units.