Organisms
Journal of Biological Sciences

Vol. 2, 2 (December 2018)
ISSN: 2532-5876
DOI: 10.13133/2532-5876_4.6

RESEARCH HIGHLIGHTS

Statistical Distribution as a Way for Lower Gene Expressions Threshold Cutoff

Bui Thuy Tien [a], *Alessandro Giuliani* [b] *and Kumar Selvarajoo* [a]

[a] *Biotransformation Innovation Platform (BioTrans), Agency for Science, Technology and Research A*STAR, Proteos, Biopolis 138673, Singapore*

[b] *Environment and Health Dept. Istituto Superiore di Sanità, Roma, Italy*

Corresponding author: Kumar Selvarajoo kumar_selvarajoo@biotrans.a-star.edu.sg

Abstract

While in mathematics (and in logic) the basic divide is between 'true' and 'false', in experimental science the frontier is between 'relevant' and 'irrelevant' and this is a much more tricky border. The classical way to track this frontier builds upon inferential statistics (signal analysis is a synonymous more popular among engineers) and is based on the definition of what we intend for 'randomness' in a given situation. Here we comment on the setting of the threshold between 'informative' and 'random' territories in the case of gene expression data where the definition of randomness is not only a 'statistical' but a 'biological' affair.

Citation: Thuy Tien, B, Giuliani, A, Selvarajoo, K, 2018, "Statistical Distribution as a Way for Lower Gene Expressions Threshold Cutoff", *Organisms. Journal of Biological Sciences*, vol. 2, no. 2, pp. 55- 58. DOI: 10.13133/2532-5876_4.6

1. Matter of concern

Large-scale gene expression studies using microarray or RNA-Seq techniques have gained tremendous momentum since the beginning of the millennium. Although fascinating to have such data handy, and to use numerous statistical tools to make sense of the data deluge, it is always a guess as to where to draw a line to say which is information and which is not. Biologists, as well as statisticians alike, have predominantly used a man-made arbitrary threshold cut-off way to draw the difference between what is useful and noisy data [e.g. FPKM > 5 (Koso *et al.*, 2016) or > 1.5 fold (Cromie *et al.*, 2017)]. This is no doubt a good starting point not to include unwanted non-informative data, however, is this the best way forward?

Recent works that investigated the distributions of gene expressions across diverse living cells have pointed to underlying statistical structures (Furusawa *et al.*, 2003; Bengtsson *et al.*, 2005; Beal, 2017). From these works, gene expressions are shown to follow power-law or lognormal distribution. We have also previously studied gene expression distribution across cells from human, mouse and bacteria and achieved similar results (Piras & Selvarajoo, 2015; Simeoni *et al.*, 2015). Here, we tested the statistical distribution of more recent RNA-Seq data of two microorganisms (*Saccharomyces cerevisiae*, *Escherichia coli*) and two higher organisms (*mus musculus*, *Homo sapiens*) (Figure 1, black). This time, we investigated lognormal, loglogistic, power law or Pareto, Burr, Weibull and Gamma distributions for the best high-throughput gene expressions fitting (Figure 1, colors). Notably, we observed all distributions fitted very well above a certain threshold level with the lognormal performing the best in all four cases (Figure 1, vertical lines).

The distribution of gene expressions at the lower range is dominated by two important constraints, due to *i*) technical errors such as protocol or sample sensitivity variance, and *ii*) binary transcriptional toggle switches that are commonly encountered (Xu *et al.*, 2016).

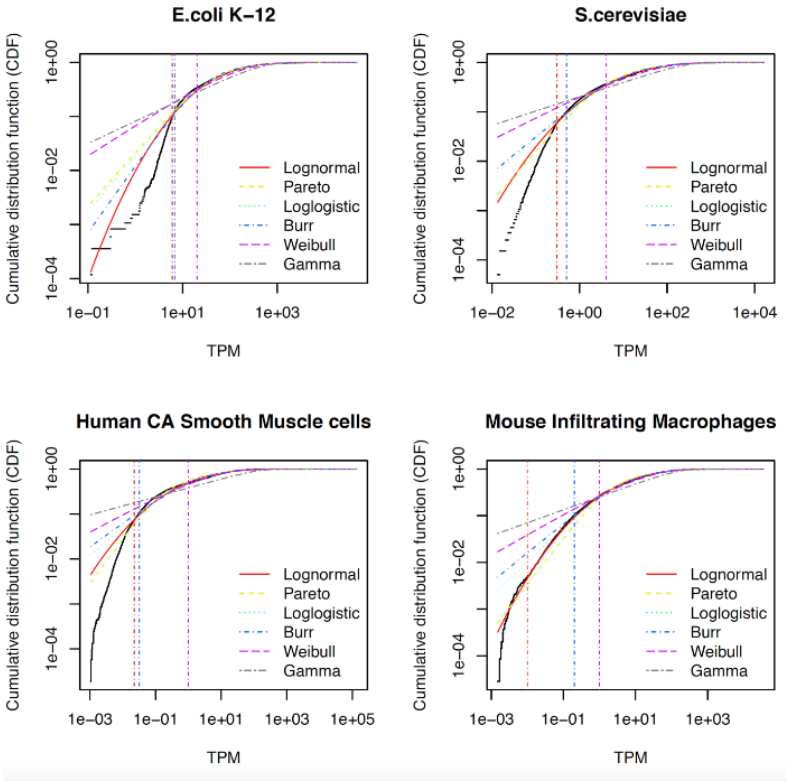

Figure 1. *Testing Statistical Distributions on Gene Expressions.* Cumulative density plot versus TPM values for A) E. coli (von Wulffen et al., 2016), B) S. cerevisiae [2], C) human smooth muscle cells (Li et al., 2018), and D) mouse macrophages (Koso et al., 2016). Lognormal (solid red), Pareto or power law (green dashed), loglogistic (dark green dotted), Burr (blue dotted), loglogistic (cyan dotted/dashed), Weibull (pink dashed) and Gamma (grey dashed). The vertical lines are drawn at the crossing between the actual gene expressions (black) and respective distribution curves. Only three vertical lines are shown for clarity.

The latter case is important, as it is points to biological relevance that cannot be simply discarded as unwanted or non-informative. The adherence to lognormal distribution points to the fact that the gene expression variability, or noise, has a 'multiplicative' and not an additive nature as is usually assumed (Elowitz *et al.*, 2002). That is, the noise is not (mainly) a result of external interference, but rather is an integral part of the gene regulatory network response, this comes from the simple fact multiplication (product) implies the correlative interaction among the different operators (gene expressions in this case).

To summarize, we highlight that gene expressions are governed by clear statistical distribution, here we show it is closer to lognormal. The statistical structure is highly conserved among different organisms and cell types, and is most probably due to the scale-free or fractal organization of gene regulatory networks (Albert, 2005). Therefore, instead of using arbitrary threshold cut-off, statistical distribution fitting-based cut-off could increase the resolution of high dimensional analyses.

Acknowledgments

The authors thank the support of BioTrans, A*STAR.

References

Albert R., 2005, Scale-free networks in cell biology. *J Cell Sci.* vol. 118, pp. 4947-4957

Beal J. 2017, Biochemical complexity drives log-normal variation in genetic expression. *IET Engineering Biol,* vol. 1(1), pp. 55-60.

Bengtsson M, Ståhlberg A, Rorsman P, Kubista M. 2005, Gene expression profiling in single cells from the pancreatic islets of Langerhans reveals lognormal distribution of mRNA levels. *Genome Res*, vol.10, pp. 1388-1392.

Cromie GA, Tan Z, Hays M, Sirr A, Jeffery EW, Dudley AM., 2017. Transcriptional Profiling of Biofilm Regulators Identified by an Overexpression Screen in Saccharomyces cerevisiae. *G3 (Bethesda),* vol. 7(8), p. 2845-2854.

Elowitz, MB, Levine AJ, Siggia ED, Swain PS, 2002, Stochastic gene expression in a single cell. *Science*, vol. 297 (5584), p. 1183–1186.

Furusawa C, Kaneko K. 2003, Zipf's law in gene expression. *Phys Rev Lett*, vol. 90(8), p. 088102.

Koso H, Tsuhako A, Lai CY, Baba Y, Otsu M, Ueno K, Nagasaki M, Suzuki Y, Watanabe S. 2016, Conditional rod photoreceptor ablation reveals Sall1 as a microglial marker and regulator of microglial morphology in the retina. *Glia*, vol. 64 (11), p. 2005-2024.

Li S, Chang Z, Zhu T, Villacorta L, Li Y, Freeman B.A., Chen Y.E, Zhang, J. , 2018, Transcriptomic sequencing reveals diverse adaptive gene expression responses of human vascular smooth muscle cells to nitro-conjugated linoleic acid, *Physiol Genomics*, vol. 50(4), p. 287-295.

Piras V, Selvarajoo K. 2015, The reduction of gene expression variability from single cells to populations follows simple statistical laws. *Genomics*, vol.105, pp. 137-144.

Simeoni O, Piras V, Tomita M, Selvarajoo K. 2015, Tracking global gene expression responses in T cell differentiation. *Gene*. vol. 569, pp. 259-266.

von Wulffen J, RecogNice-Team., Sawodny O, Feuer R, 2016, Transition of an Anaerobic Escherichia coli Culture to Aerobiosis: Balancing mRNA and Protein Levels in a Demand-Directed Dynamic Flux Balance Analysis. *PLoS One* vol. 11(7), e0158711.

Xu Y, Li Y, Zhang H, Li X, Kurths J., 2016, The Switch in a Genetic Toggle System with Lévy Noise. *Sci Rep*, vol. 6, p. 31505.