CellPress

Assessing Inequality in Transcriptomic Data

Lan Jiang,[3,4,5] Daphne Tsoucas,[1,2] and Guo-Cheng Yuan[1,2,*]
[1]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, MA 02215, USA
[2]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA
[3]Howard Hughes Medical Institute, Boston Children's Hospital, Boston, MA 02115, USA
[4]Program in Cellular and Molecular Medicine, Boston Children's Hospital, Boston, MA 02115, USA
[5]Division of Hematology/Oncology, Department of Pediatrics, Boston Children's Hospital, Boston, MA 02115, USA
*Correspondence: gcyuan@jimmy.harvard.edu
https://doi.org/10.1016/j.cels.2018.02.007

Two studies in this issue of *Cell Systems* use the Gini index from economics to benchmark and quantify gene expression heterogeneity in single-cell or bulk RNA-seq datasets.

Massive contemporary efforts are yielding atlases of gene expression across thousands of individual cells and across many human tissues and cell lines. In this issue of *Cell Systems*, two papers by Torre et al. and O'Hagan et al. demonstrate the utility of analyzing these atlases using a metric of inequality that originated in economics, the Gini index. Torre et al. use the Gini index to compare single-cell RNA sequencing (RNA-seq) data generated by droplet-based and microfluidic technologies to reference data from single-molecule fluorescence *in situ* hybridization (smFISH) experiments (Torre et al., 2018). O'Hagan et al. use the Gini index to analyze bulk RNA-seq data from the Human Protein Atlas project to assess the inequal distribution of transporter genes across tissues and to identify ubiquitously expressed candidate housekeeping genes (O'Hagan et al., 2018).

Single-cell RNA-seq is a promising approach for identifying novel cell types and cell states associated with development and disease. There are many examples of small sets of cells that play critical roles in driving organism development (such as stem and progenitor cells) or clinical outcome (such as cancer stem cells and treatment-resistant cells). Yet our understanding of these cell types remains limited. One of the challenges in identifying rare cell states is the lack of known gene markers. Without prior biological knowledge, statistical methods are needed to identify candidate genes that are likely to serve as informative gene markers. For common cell types, ranking genes based on the variance or Fano factor of their expression level distributions is often sufficient for identifying these markers. However, these metrics are not suitable for rare cell detection because their values are insensitive to the presence of a small number of cells.

The Gini index (Gini, 1912) is a non-parametric metric commonly used in economics to describe the income inequality within a community or country. Briefly, it quantifies the deviation of the cumulative income from an absolutely equal community (Figure 1A). In previous work, our group adapted the Gini index (Jiang et al., 2016) to identify rare cell-type-associated genes from single-cell gene expression data. Recently, some of the

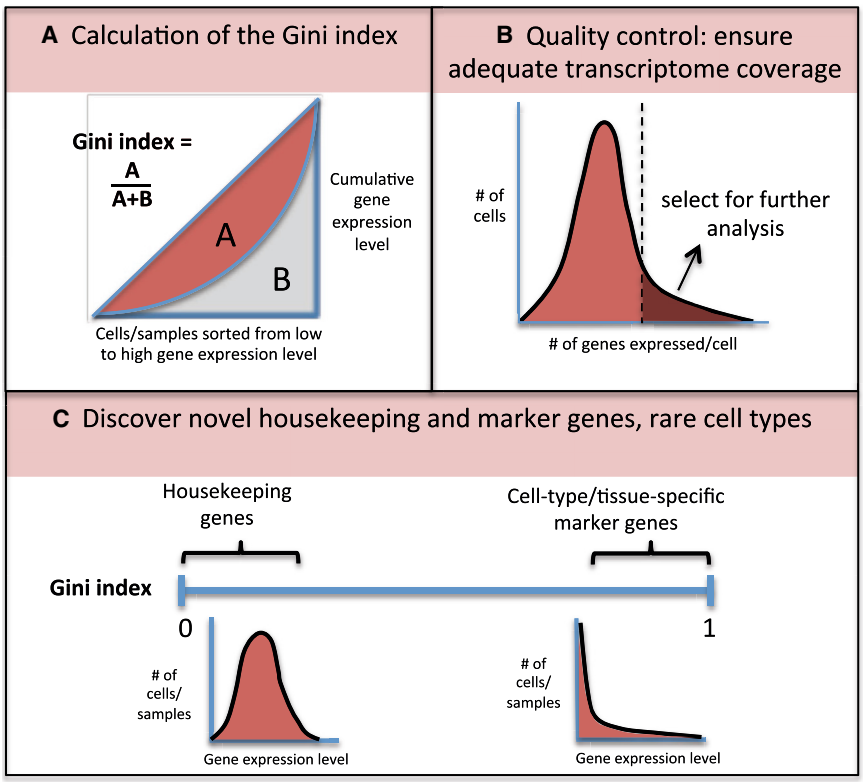

Figure 1. Guidelines for Using the Gini Index for Making Biological Discoveries from Gene Expression Data
(A) The Gini index is calculated as a function of the area under the Lorenz curve, representing the cumulative fraction of transcripts in cells (samples) sorted by the gene expression level.
(B) Care needs to be taken to ensure that the Gini index does not capture false positives. For single-cell gene expression data, Torre et al. suggest a sequencing-depth standard to prevent this.
(C) Low Gini indices can be indicative of housekeeping genes, as in O'Hagan et al. Genes with high indices are often marker genes for rare cell types or tissues, and their joint feature space can be used in conjunction with clustering to discover novel rare cell types.

authors from Torre et al. extended this idea to the identification of treatment-resistant cancer cells (Shaffer et al., 2017).

In this issue, Torre et al. go one step further and examine the accuracy of Gini index estimation from single-cell RNA-seq data (Torre et al., 2018). Using a highly sensitive smFISH dataset as a reference, they compare Gini index estimates from droplet (Drop-seq)- and microfluidic (Fluidigm)-sequencing-based technologies. They find that the results from these sequencing-based assays tend to significantly overestimate the Gini index: the lack of sensitivity mainly prevents the distinction between genes that are uniformly expressed at a low level from those that are expressed only in a small number of cells. This deviation is especially large in datasets with low transcriptome coverage. Interestingly, estimation accuracy is significantly improved after filtering out cells with fewer than 2,000 detectable genes (Figure 1B), even though the resulting estimates are still too high. On the surface, this result may sound counterintuitive, as dropping a large number of cells (86%) means throwing away information. However, the important message here is that data quality is at least as important as data quantity, and the exact criterion for data quality is dependent on the specific biological question.

In another paper in this issue, O'Hagan et al. investigate the expression patterns of transporter genes across tissues and cell lines using the Gini index (O'Hagan et al., 2018). Similar to the aforementioned studies (Jiang et al., 2016; Shaffer et al., 2017; Torre et al., 2018), O'Hagan et al. find that high Gini indices are indicative of highly specific marker genes—in this case, certain solute carrier membrane transporters and ABC efflux transporters. They also find a unique use for the Gini index in discovering candidate housekeeping genes, whereby genes with low Gini indices tend to be ubiquitously expressed across tissues and cell types (Figure 1C). These candidate housekeeping genes can then be used as references for the normalization of transcriptomic data.

It is worth noting that while the Gini index is useful for detecting rare-cell-associated genes, it is not informative for distinguishing major cell types (Jiang et al., 2016). This limitation can be partially resolved by using an ensemble clustering approach, combining clustering results based on different gene sets (Tsoucas and Yuan, 2018). Additional investigations are needed to fully address this challenge. More generally, these new studies, especially the one by Torre et al. (2018), have raised an interesting question about how to balance the need for data quality versus quantity in single-cell analysis. As larger and more deeply sequenced single-cell datasets become available, these types of analyses will likely lead to exciting new discoveries.

REFERENCES

Gini, C. (1912). Variabilità e Mutabilità: Contributo allo Studio delle Distribuzioni e delle Relazioni Statistiche (Tipografia di Paolo Cuppini).

Jiang, L., Chen, H., Pinello, L., and Yuan, G.C. (2016). GiniClust: detecting rare cell types from single-cell gene expression data with Gini index. Genome Biol. *17*, 144.

O'Hagan, S., Muelas, M.W., Day, P.J., Lundberg, E., and Kell, D.B. (2018). GeneGini: assessment via the gini coefficient of reference "housekeeping" genes and highly heterogeneous human transporter expression profiles. Cell Syst. *6*, this issue, 230–244.

Shaffer, S.M., Dunagin, M.C., Torborg, S.R., Torre, E.A., Emert, B., Krepler, C., Beqiri, M., Sproesser, K., Brafford, P.A., Xiao, M., et al. (2017). Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. Nature *546*, 431–435.

Torre, E., Dueck, H., Shaffer, S., Gospocic, J., Gupte, R., Bonasio, R., Kim, J., Murray, J., and Raj, A. (2018). Rare cell detection by single-Cell RNA sequencing as guided by single-molecule RNA FISH. Cell Syst. *6*, this issue, 171–179.

Tsoucas, D., and Yuan, G.C. (2018). A cluster-aware, weighted ensemble clustering method for cell-type detection. bioRxiv. https://doi.org/10.1101/246439.

Merged Map of the Yeast Proteome

Alexander Schmidt[1,*]
[1]Biozentrum, University of Basel, Klingelbergstrasse 50/70, 4056 Basel, Switzerland
*Correspondence: alex.schmidt@unibas.ch
https://doi.org/10.1016/j.cels.2018.02.006

A comprehensive reference map of protein abundances in budding yeast is generated by combining the 21 largest quantitative proteome datasets currently available for this model organism.

For any organism, knowing whether and how much of a protein is expressed is crucial for understanding molecular and functional cellular mechanisms, particularly on a system-wide level. Yet such extensive quantitative proteome datasets are currently only available for a handful of organisms, mostly bacteria. In this issue of *Cell Systems*, Ho et al. use computational approaches to exploit the vast data resource of 21 different large-scale proteome datasets that exist for the model eukaryote budding yeast, *Saccharomyces cerevisiae* (Ho et al., 2018). The authors combine these data to generate the most extensive quantitative proteome abundance reference map of yeast to date. Covering 92% of annotated open reading frames, the resulting dataset allowed Ho et al. to explore transcriptional and translational control on protein abundance, one of the central dogmas in molecular biology, with unprecedented detail and provides a