Distinct Characteristics of Correlation Analysis at the Single-Cell and the Population Level

Guoyu Wu (wuguoyu@gdpu.edu.cn)
Guangdong Pharmaceutical University https://orcid.org/0000-0002-4906-0659
Yuchao Li
Max-Planck-Institut fur molekulare Genetik

Research article

Keywords: Correlation analysis, Mathematical modeling, Single-cell level, Population level, Measurement errors

Posted Date: August 19th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-42825/v1

Distinct characteristics of correlation analysis at the single-cell and the population level

Guoyu Wu*#[1], Yuchao Li#[2]

[1] Clinical pharmacy of The First Affiliated Hospital of Guangdong Pharmaceutical University, Guangzhou, Guangdong Province, China

[2] Max Planck Institute for Molecular Genetics, Berlin, Germany

# Equal contributors

* To whom correspondence should be addressed. Email: wuguoyu@gdpu.edu.cn.

Abstract

Background:

Correlation analysis is widely used in biological studies to infer molecular relationships within biological networks. Recently, single-cell analysis has drawn tremendous interests, for its ability to obtain high-resolution molecular phenotypes. It turns out that there is little overlap of co-expressed genes identified in single-cell level investigations with that of population level investigations. However, the nature of the relationship of correlations between single-cell and population levels remains unclear. In this manuscript, we aimed to unveil the origin of the differences between the correlation coefficients at the single-cell level and that at the population level, and bridge the gap between them.

Results:

Through developing formulations to link correlations at the single-cell and the population level, we illustrated that aggregated correlations could be stronger, weaker or equal to the corresponding individual correlations, depending on the variations and the correlations within the population. When the correlation-within is weaker than the individual correlation, the correlation at the population level is stronger than that at the single-cell level. Through a bottom-up approach to model interactions between molecules in a signaling cascade or a multi-regulator controlled gene expression, we surprisingly found that the existence of interaction between two components could not be excluded simply based on their low correlation coefficients, suggesting a reconsideration of connectivity within biological networks which was derived solely from correlation analysis. We also investigated the impact of technical random measurement errors on the correlation coefficients for the single-cell level and the population level. The results indicate that the aggregated correlation is relatively robust and less affected.

Conclusions:

Because of the heterogeneity among single cells, correlation coefficients calculated based on data of the single-cell level might be different from that of the population level. Depending on the specific question we are asking, proper

sampling and normalization procedure should be done before we draw any conclusions.

Keywords

Correlation analysis; Mathematical modeling; Single-cell level; Population level; Measurement errors

Background

Correlation analysis is widely used to identify closely related components and interactions within biological networks[1-4]. Traditionally, cellular behaviors were investigated by experimental measurements derived from bulk-averaged assays[3]. In recent years, cellular heterogeneity within genetically identical populations has been well-recognized and there has been an explosion of interest in single-cell analysis[5, 6]. Comparative analysis of co-expressed genes between multi-levels suggested that correlation profiles obtained with single-cell data are not always consistent with those obtained with bulk-sample measurements[7-9]. Strikingly, in a previous study, researchers compared the glioblastoma expression profiles between bulk-tissue and single-cell data, and the result suggested that less than 10% of the co-expressed genes were shared, while a majority of gene pairs were highly correlated either at the bulk-tissue or the single-cell level[7].

Previous population-averaged measurements, such as western blotting, HPLC and microarray, were able to capture numerous protein-protein interactions within signaling pathways or regulatory networks[10]. Still, co-expression analyses at the single-cell level, which take intercellular variability into consideration, could give discrepant correlation coefficients[7, 11]. Studies suggested distinct co-regulatory mechanisms underlying the co-expression relationships in the bulk samples and the single cells [7, 12]. However, it remains to be fully addressed why a correlation based on individual-level (single-cell) data could be much stronger or weaker than those based on population-level (bulk-sample) data. Does a weak correlation indicate no interaction between two components? Which level of correlation is more robust and less affected by technical artifacts?

In this study, these fundamental questions were addressed by a bottom-up approach[13]. We bridged the gap between the single-cell level and the population-level correlations by deriving formulations, illustrating that their relationship depends on the relative variations or correlations within populations. Further, we developed mathematical models to mimic signaling cascades or multi-regulated gene expressions. Interestingly, the results indicated that the correlation between two components in the networks could be weak at the single-cell level, though a strong correlation existed at the population level. Thus, a strong correlation at the population level cannot demonstrate a strong correlation at the single-cell level, and connections within the biological networks should be clearly noted whenever the components are correlated at either level.

Results

Aggregated correlation could be stronger, weaker or equal to individual correlation

The relationship between correlations across aggregated and individual levels has been studied in sociology[14]. The derived formulations could also be applied to biological analyses to describe correlations at the population and single-cell level. Definitions of the components of the formulations are shown in Table 1.

Table 1. The definitions of components of the formulations

Component	Definition
$\boldsymbol{x}$	Value of variable x in single cells
$\boldsymbol{y}$	Value of variable y in single cells
$\boldsymbol{u}$	Aggregated value of x at bulk level in one population
$\boldsymbol{v}$	Aggregated value of y at bulk level in one population
$\boldsymbol{i_x}$	Deviation from the mean of x in each individual cells in one population ($x = u + i_x$)
$\boldsymbol{i_y}$	Deviation from the mean of y in each individual cells in one population ($y = v + i_y$)
$\boldsymbol{\sigma_x}$	Standard deviation of x
$\boldsymbol{\sigma_y}$	Standard deviation of y
$\boldsymbol{\sigma_u}$	Standard deviation of u
$\boldsymbol{\sigma_v}$	Standard deviation of v
$\boldsymbol{\sigma_{ix}}$	Standard deviation of i_x (σ_{ix}^2: Variance within one population for x; $\sigma_x^2 = \sigma_u^2 + \sigma_{ix}^2$)
$\boldsymbol{\sigma_{iy}}$	Standard deviation of i_x (σ_{iy}^2: Variance within one population for y; $\sigma_y^2 = \sigma_v^2 + \sigma_{iy}^2$)
$\boldsymbol{\rho_{xy}}$	Correlation between x and y (individual correlation)
$\boldsymbol{\rho_{uv}}$	Correlation between u and v (aggregated correlation)
$\boldsymbol{\rho_{ixiy}}$	Correlation between i_x and i_y (Correlation within)

Here, we calculated and compared the Pearson's correlation coefficients between the single cell and the population level. Assuming $\rho_{xy} \neq 0$, the ratio of the

aggregated correlation to individual correlation was defined by Equation 1 (Supplemental materials).

$$\frac{\rho_{uv}}{\rho_{xy}} = \sqrt[2]{\left(1+\frac{\sigma_{ix}^2}{\sigma_u^2}\right)\cdot\left(1+\frac{\sigma_{iy}^2}{\sigma_v^2}\right)} - \frac{\sigma_{ix}}{\sigma_u}\cdot\frac{\sigma_{iy}}{\sigma_v}\cdot\frac{\rho_{ixiy}}{\rho_{xy}} \quad (1)$$

The relationship between correlations at the single-cell and the population level depends on values of the standard-deviation-within relative to the aggregated-level standard deviations ($\frac{\sigma_{ix}}{\sigma_u}$ and $\frac{\sigma_{iy}}{\sigma_v}$) and the correlation-within relative to the individual correlation ($\frac{\rho_{ixiy}}{\rho_{xy}}$). Depending on varied values of these components, aggregated correlation could be stronger, weaker or equal to individual correlation (Figure 1). Notably, when the correlation-within (ρ_{ixiy}) is weaker than individual correlation (ρ_{xy}), or the signs of correlation-within and individual correlation differ from each other (one is positive and the other is negative), the correlation at the population level is stronger than it at the single-cell level (Figure 1C and Supplemental materials).

Interactions could not be excluded based on weak correlations at the single-cell level

Correlation analyses are widely used in discovering functional modules and exploring biological relationships[15, 16]. Typically, the correlated-components in a regulatory network were defined by a Pearson/Spearman correlation coefficient greater than |±0.5|[17, 18], and the others were filtered out as unrelated components. However, one question that is often overlooked was whether the weakly-correlated components were indeed unrelated at all? Single-cell analysis has become extremely popular in recent years and improved our understanding in many areas that have been traditionally studied at the population level. Does the correlation at the single-cell level provide novel insight into connections between the components?

To address these questions, we employed mathematical modeling to investigate the correlations between components in regulatory networks. Two toy models were developed to represent two typical types of biological regulatory systems respectively: Model1 described a multi-step signaling cascade (Figure 2A) and Model2 characterized a multi-regulator controlled gene expression (Figure 2B). X*(X1*, X2* and X3*) were the intermediate regulators and Y* was the responder ('*' indicates active form). Strikingly, although X* and Y* did interact with each other and they were indeed significantly correlated at the population level, the correlation coefficient between them at the single-cell level could be very weak (Figure 2C-F). Specifically, we generated population-level data by randomly sampling the cells based on their proximity of X*, and then averaging the respective values (See Methods for a detailed description). There're overlaps between the neighboring bulk samples in our simulations, which was intended to mimic the possible overlaps of population-level measurements in the biological experiments.

We also investigated the correlations in a previous published regulatory network of mammalian cell cycle, the parameters of which were experimentally derived[19]. The discrepancy of correlation coefficients between the single-cell and the population level were shown. For the aggregated-level significantly correlated molecules, their correlation coefficients at the individual level could be much lower (Figure S1). In sum, we could not exclude the possibility of interactions between components by their single-cell level correlation coefficient only.

Aggregated correlation is robust and less affected by technical random error

Technical random measurement errors are often unavoidable in experimental assays. When comparing correlations at the population and single-cell level, random measurement errors should be taken into consideration before drawing any conclusions. By taking the random errors into account, the measured values of variables at the single-cell or the population level are represented by

$$x_m = x + e_x \tag{2}$$

$$y_m = y + e_y \tag{3}$$

$$u_m = u + e_u \tag{4}$$

$$v_m = v + e_v \tag{5}$$

where x_m, y_m, u_m, and v_m are the measured values, which are composed of x, y, u, or v and some random errors; e_x and e_y are the random measurement errors for variable x and y, respectively; e_u and e_v are the averaged random measurement errors for variable u and v, respectively.

Theoretically, for the measurements at the population level, the mean of random error should be a constant ($e_u = k1$ and $e_v = k2$) and thus the variance of random error is zero ($\sigma_{eu}^2 = 0$ and $\sigma_{ev}^2 = 0$). Therefore, the correlations at the population level are unaffected by the random measurement errors (Equation 6).

$$\rho'_{uv} = \frac{cov(u,v) + cov(e_u, e_v)}{\sqrt[2]{(\sigma_u^2 + \sigma_{eu}^2) \cdot (\sigma_v^2 + \sigma_{ev}^2)}} = \frac{cov(u,v) + cov(k1,k2)}{\sqrt[2]{(\sigma_u^2 + 0) \cdot (\sigma_v^2 + 0)}} = \frac{cov(u,v)}{\sigma_u \cdot \sigma_v} = \rho_{uv} \tag{6}$$

where cov denotes covariance, ρ'_{uv} is the population-level correlation with random measurement errors; σ_{eu}^2 and σ_{ev}^2 are the variances of e_u and e_v, respectively.

In contrast, measurements at the single-cell level, as well as their correlations, are affected. By assuming the random errors (e_x and e_y) are uncorrelated with each other, and they are independent of x, y, u, v, i_x or i_y, the individual correlation with measurement errors is represented by Equation 7.

$$\rho'_{xy} = \frac{cov(u,v) + cov(i_x, i_y)}{\sqrt[2]{(\sigma_u^2 + \sigma_{ix}^2 + \sigma_{ex}^2) \cdot (\sigma_v^2 + \sigma_{iy}^2 + \sigma_{ey}^2)}} \tag{7}$$

where ρ'_{xy} is the individual correlation with measurement error; σ^2_{ex} and σ^2_{ey} denote the variances of e_x and e_y, respectively.

Compared with the individual correlation without measurement errors (Equation 8)

$$\rho_{xy} = \frac{cov(u,v) + cov(i_x, i_y)}{\sqrt[2]{(\sigma_u^2 + \sigma_{ix}^2) \cdot (\sigma_v^2 + \sigma_{iy}^2)}} \quad (8)$$

since $\sigma^2_{ex} > 0$ and $\sigma^2_{ey} > 0$, we get $\rho_{xy} > \rho'_{xy}$, which demonstrates that the individual correlation is reduced by measurement errors.

These theoretical conclusions could also be confirmed and visualized by model simulations. We added random error to the same dataset of Figure 2C and 2D, and then calculated the corresponding correlation coefficients of the data with random error at the single-cell (Figure 3A) or the population level (Figure 3B). As expected, the individual correlations were largely reduced while the aggregated correlations were unaffected. Notably, although in theory the aggregated correlations should stay the same (Eq.6), in practice, the mean of random error might most probably be a small number approximate to, but not equal to zero. As a consequence, the aggregated correlations could be also slightly or markedly reduced (Figure S2), especially for the situation when the values of output (Y*) would not significantly change along with the inputs (X*).

Correlation within could not guide sampling

In many biological studies, researchers are exploring how a specific molecule E (effector) is positively or negatively regulate another molecule R (responder). Generally, molecule E might quantitatively affect molecule R only in a certain dose range. For instance, after reaching the saturation point, the amount of responder R does not change along with the effector E. It's usually essential to cover the entire responsive range in the experimentation design. But in the situation when we haven't clearly explored that range, it could be very tricky to identify the proper responsive range efficiently.

To find a method for efficiently locate the responsive range, within which the two molecules present mutual correlation, we were curious if the correlation-within sheds light on it. We proposed a hypothesis that the correlation-within of two samples would be similar if they located closely on the responsive range, while the correlation-within varied a lot among samples covering a large responsive space.

To test our hypothesis, we investigated the correlations between Y* and X3* in Model2 by simulation. 30000 cells were simulated and divided into 30 bulk samples according to their close proximity of X3* (Figure S3A and S3B). The correlations-within were calculated (Figure S3C) and every five samples closely located were merged into one close-sample group. For each close-sample group, we calculated the correlation coefficients of the five bulk-sample within it, and then derived the standard deviation of these coefficients (Figure 4A). Next, we generated a sparse-sample group by choosing one sample from each of the close-sample groups, and then performed similar calculations. Interestingly, the

standard deviations of the sparse-sample group (covering a large space) could be higher or lower than those of the close-sample groups (Figure 4B and 4C). As a conclusion, our primary hypothesis was rejected by the model simulation, and the correlation-within could not offer us a better guess for the responsive range. It is important to perform sampling covering the responsive space as large as possible without prior knowledge, in particular for investigations on digital responses[20]. Unfortunately, correlation-within could not provide additional information about it.

Discussion

A central challenge in the biological research is to unveil the connections in the regulatory networks and understand them in a quantitative way. Correlation analysis, traditionally based on bulk-averaged assays, is widely used to identify interactions within metabolic, signaling or transcriptional networks. In recent decades, as advanced technologies make more single cell measurements accessible, one of the key questions arising is whether the conclusions based on population-averaged assays could be applied to individual cell behaviors.

Despite a well-recognized inconsistency of correlation analysis between the single-cell level and the population level, the underlying mechanism remains to be fully addressed. Our formulations illustrated that the individual-correlation could be stronger, weaker or equal to the corresponding aggregated-correlation, depending on the variations and the correlation within the population. By modeling a signal cascade and a parallel regulation respectively, we found that the correlation between two interactive components could be weak at the single-cell level, though they were strongly correlated at the population level. Therefore, the connection possibility between components could not be excluded solely by their correlation coefficient. Existence of connection between two molecules should be considerated whenever the components are highly correlated at either level.

Our results demonstrated that the variance-within plays a crucial role in the consistency of the correlation coefficients across the levels, which describes the heterogeneity within samples. Statistical artifacts, biased estimates or elimination of error variance might somewhat all contribute to the higher correlations at the aggregated level than those at the individual level[14]. Biological activities are often much more complex than just several definite interactions among a few molecules, one specific response often results from a combination of factors. Though most of these factors might not impact a lot, however, on the one hand, these impacts could constantly accumulate, on the other hand, the chaos theory suggests that even a tiny interference could induce a largely deviant response. All of these would attenuate the correlation between two variables at the individual level. However, these deviances are often averaged out at the aggregated level, which help to expose the correlation. Basically, the more deviant the individual cells, the larger the ratio of the aggregated correlation to the individual correlation. Thus, it was not surprising to find a low correlation coefficient for a pair of molecules at the single cell resolution while a strong correlation for the same pair has been reported in a previous population-based study. Variability in protein levels is a common

phenomenon even in genetically identical cells. Across the population, proteins were log-normally distributed, with an average coefficient of variation (CV = standard deviation/mean) ranging from 0.12 to 0.28[21, 22]. Therefore, the discrepancy between the aggregated correlation and the individual correlation is always observed.

Our results illustrated an unavoidable difference of correlation coefficient between the single-cell and the population level, which raises the question of which level of correlation analysis weights more when discrepancy occurs. Depending on the purpose of an investigation, individual correlation, aggregated correlation or correlation-within should be taken into consideration to make proper conclusions wisely. Studies have revealed distinct biological insights of the co-expressed genes at the population or the single-cell level: the population level co-expressed genes share the same biological functions, while the single-cell level co-expression indicates interactions[7]. Specifically, considering the existence of technical random measurement errors, our data show that individual correlation is less robust. Therefore, higher accuracy is required for single-cell level investigation and if possible, proper normalization procedures for individual-level measurements should be done before drawing any conclusions.

Conclusions

Measurements at the single-cell level indeed proved us with a fantastic tool to increase the resolution of our exploration into the biological activities. When taking a dive from the population level into the single-cell level, discrepancy could occur for correlation analysis because of the heterogeneity within samples. Since distinct biological insights could be revealed from either the individual or aggregated perspective analysis, we should always be aware of the level of analysis we are at, or choose the proper level of data to explore in our research.

Methods

Model development and simulation

Two toy models were developed to represent two typical types of biological regulatory systems: Model1 described a multi-step signaling cascade and Model2 characterized a multi-regulator controlled gene expression. The illustrations of the two models are shown in Figure 2.

Model1 comprises three steps of signaling transduction and one feedback loop, described by eight ordinary differential equations. Model2 comprises three regulators and three feedback loops, described by eight ordinary differential equations. X1*, X2*, X3* and Y* are the active forms of X1, X2, X3 and Y, respectively.

10000 cells in total were simulated. The initial values of X1, X2, X3 and Y were randomly generated from lognormal distributions. The initial values of X1*, X2*, X3* and Y* were zeros. Correlation analyses were preformed based on the data

of the systems at steady state. See supplemental materials for a detailed description.

The sampling procedure for the population level analysis

To represent the possible overlaps of population-level samples in the biological experiments, aggregated values with overlaps between neighboring groups were obtained by the following steps: (1) 10000 single cells were grouped based on their proximity of X* (2000 cells per group); (2) For each group, chose the closest 500 cells from its two neighboring groups, respectively (1000 cells in total). Specifically, for group on the most left or most right side, chose only 500 cells from the neighboring group. Then combined these cells with the original group (obtained a 3000-cells group or 2500-cells group); (3) Randomly chose 1000 cells from it and then averaged the values of X* and Y*.

Abbreviations

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Competing interests

No conflict of interest exits in the submission of this manuscript.

Funding

This study was supported by the 12th five-year plan key discipline of clinical pharmacy and high level clinical key specialty in Guangdong province.

Authors' contributions

GW: Conceptualization, implementation, investigation, writing, editing and revising the manuscript. YL: Investigation, editing and revising the manuscript. All authors read and approved the final manuscript.

Acknowledgements

This study was supported by the 12th five-year plan key discipline of clinical pharmacy and high level clinical key specialty in Guangdong province.

Reference

1. Batushansky A, Toubiana D, Fait A: **Correlation-Based Network Generation, Visualization, and Analysis as a Powerful Tool in Biological Studies: A Case Study in Cancer Cell Metabolism**. *Biomed Res Int* 2016, **2016**:8313272.
2. Toubiana D, Xue W, Zhang N, Kremling K, Gur A, Pilosof S, Gibon Y, Stitt M, Buckler ES, Fernie AR *et al*: **Correlation-Based Network Analysis of Metabolite and Enzyme Profiles Reveals a Role of Citrate Biosynthesis in Modulating N and C Metabolism in Zea mays**. *Front Plant Sci* 2016, **7**:1022.
3. Maier T, Guell M, Serrano L: **Correlation of mRNA and protein in complex biological samples**. *FEBS Lett* 2009, **583**(24):3966-3973.
4. Lahtvee PJ, Sanchez BJ, Smialowska A, Kasvandik S, Elsemman IE, Gatto F, Nielsen J: **Absolute Quantification of Protein and mRNA Abundances Demonstrate Variability in Gene-Specific Translation Efficiency in Yeast**. *Cell Syst* 2017, **4**(5):495-504 e495.
5. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, Cahill DP, Nahed BV, Curry WT, Martuza RL *et al*: **Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma**. *Science* 2014, **344**(6190):1396-1401.
6. Bartlett TE, Muller S, Diaz A: **Single-cell Co-expression Subnetwork Analysis**. *Sci Rep* 2017, **7**(1):15066.
7. Wang J, Xia S, Arand B, Zhu H, Machiraju R, Huang K, Ji H, Qian J: **Single-Cell Co-expression Analysis Reveals Distinct Functional Modules, Co-regulation Mechanisms and Clinical Outcomes**. *PLoS Comput Biol* 2016, **12**(4):e1004892.
8. Dzamba D, Valihrach L, Kubista M, Anderova M: **The correlation between expression profiles measured in single cells and in traditional bulk samples**. *Sci Rep-Uk* 2016, **6**.
9. Yu T: **A new dynamic correlation algorithm reveals novel functional aspects in single cell and bulk RNA-seq data**. *PLoS Comput Biol* 2018, **14**(8):e1006391.

10. Albert R: **Scale-free networks in cell biology**. *J Cell Sci* 2005, **118**(Pt 21):4947-4957.

11. Crow M, Paul A, Ballouz S, Huang ZJ, Gillis J: **Exploiting single-cell expression to characterize co-expression replicability**. *Genome Biol* 2016, **17**:101.

12. Wang N, Zheng J, Chen Z, Liu Y, Dura B, Kwak M, Xavier-Ferrucio J, Lu YC, Zhang M, Roden C *et al*: **Single-cell microRNA-mRNA co-sequencing reveals non-genetic heterogeneity and mechanisms of microRNA regulation**. *Nat Commun* 2019, **10**(1):95.

13. Shahzad K, Loor JJ: **Application of Top-Down and Bottom-up Systems Approaches in Ruminant Physiology and Metabolism**. *Curr Genomics* 2012, **13**(5):379-394.

14. Ostroff C: **Comparing correlations based on individual-level and aggregated data**. *Journal of Applied Psychology* 1993, **78**(4):569-582.

15. Toubiana D, Fernie AR, Nikoloski Z, Fait A: **Network analysis: tackling complex data to study plant metabolism**. *Trends Biotechnol* 2013, **31**(1):29-36.

16. Seyfried NT, Dammer EB, Swarup V, Nandakumar D, Duong DM, Yin L, Deng Q, Nguyen T, Hales CM, Wingo T *et al*: **A Multi-network Approach Identifies Protein-Specific Co-expression in Asymptomatic and Symptomatic Alzheimer's Disease**. *Cell Syst* 2017, **4**(1):60-72 e64.

17. Buenrostro JD, Corces MR, Lareau CA, Wu B, Schep AN, Aryee MJ, Majeti R, Chang HY, Greenleaf WJ: **Integrated Single-Cell Analysis Maps the Continuous Regulatory Landscape of Human Hematopoietic Differentiation**. *Cell* 2018, **173**(6):1535-1548 e1516.

18. Chen PY, Zhang P, Cripps AW, West NP, Cox AJ: **Correlation-based network analysis for biomarkers in obesity**. *Ieee Int C Bioinform* 2018:1361-1366.

19. Yao G, Lee TJ, Mori S, Nevins JR, You L: **A bistable Rb-E2F switch underlies the restriction point**. *Nat Cell Biol* 2008, **10**(4):476-482.

20. Ferrell JE, Jr.: **Self-perpetuating states in signal transduction: positive feedback, double-negative feedback and bistability**. *Curr Opin Cell Biol* 2002, **14**(2):140-148.

21. Spencer SL, Gaudet S, Albeck JG, Burke JM, Sorger PK: **Non-genetic origins of cell-to-cell variability in TRAIL-induced apoptosis**. *Nature* 2009, **459**(7245):428-432.

22. Sigal A, Milo R, Cohen A, Geva-Zatorsky N, Klein Y, Liron Y, Rosenfeld N, Danon T, Perzov N, Alon U: **Variability and memory of protein levels in human cells**. *Nature* 2006, **444**(7119):643-646.

Figures

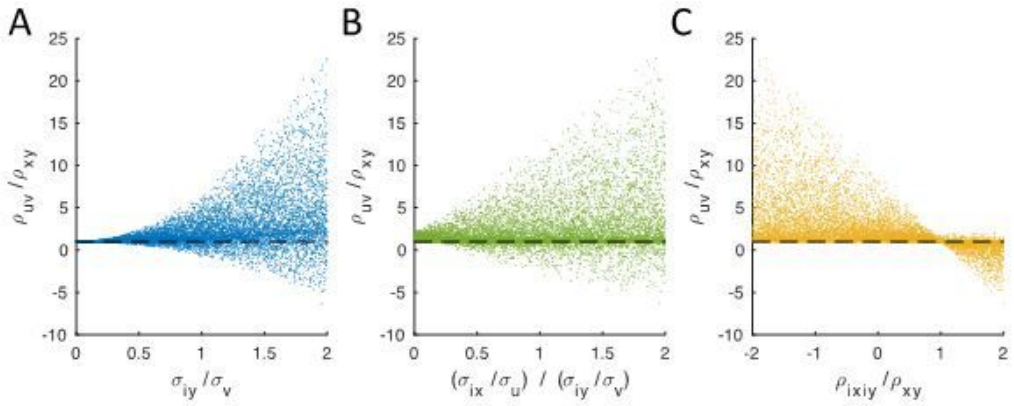

Figure 1. Ratio of aggregated correlation to individual correlation.

The values of $\frac{\sigma_{ix}}{\sigma_u}$, $\frac{\sigma_{iy}}{\sigma_v}$ and $\frac{\rho_{ixiy}}{\rho_{xy}}$ were taken from a Latin hypercube sample containing 10000 values on the interval [0, 2], [0, 2] and [-2, 2], respectively, then the values of $\frac{\rho_{uv}}{\rho_{xy}}$ were calculated according to Equation 1. (A) The value of $\frac{\rho_{uv}}{\rho_{xy}}$ in relationship to $\frac{\sigma_{ix}}{\sigma_u}$ or $\frac{\sigma_{iy}}{\sigma_v}$. (B) The value of $\frac{\rho_{uv}}{\rho_{xy}}$ in relationship to $(\frac{\sigma_{ix}}{\sigma_u}/\frac{\sigma_{iy}}{\sigma_v})$. (C) The value of $\frac{\rho_{uv}}{\rho_{xy}}$ in relationship to $\frac{\rho_{ixiy}}{\rho_{xy}}$.

Figure 1

Figure 1

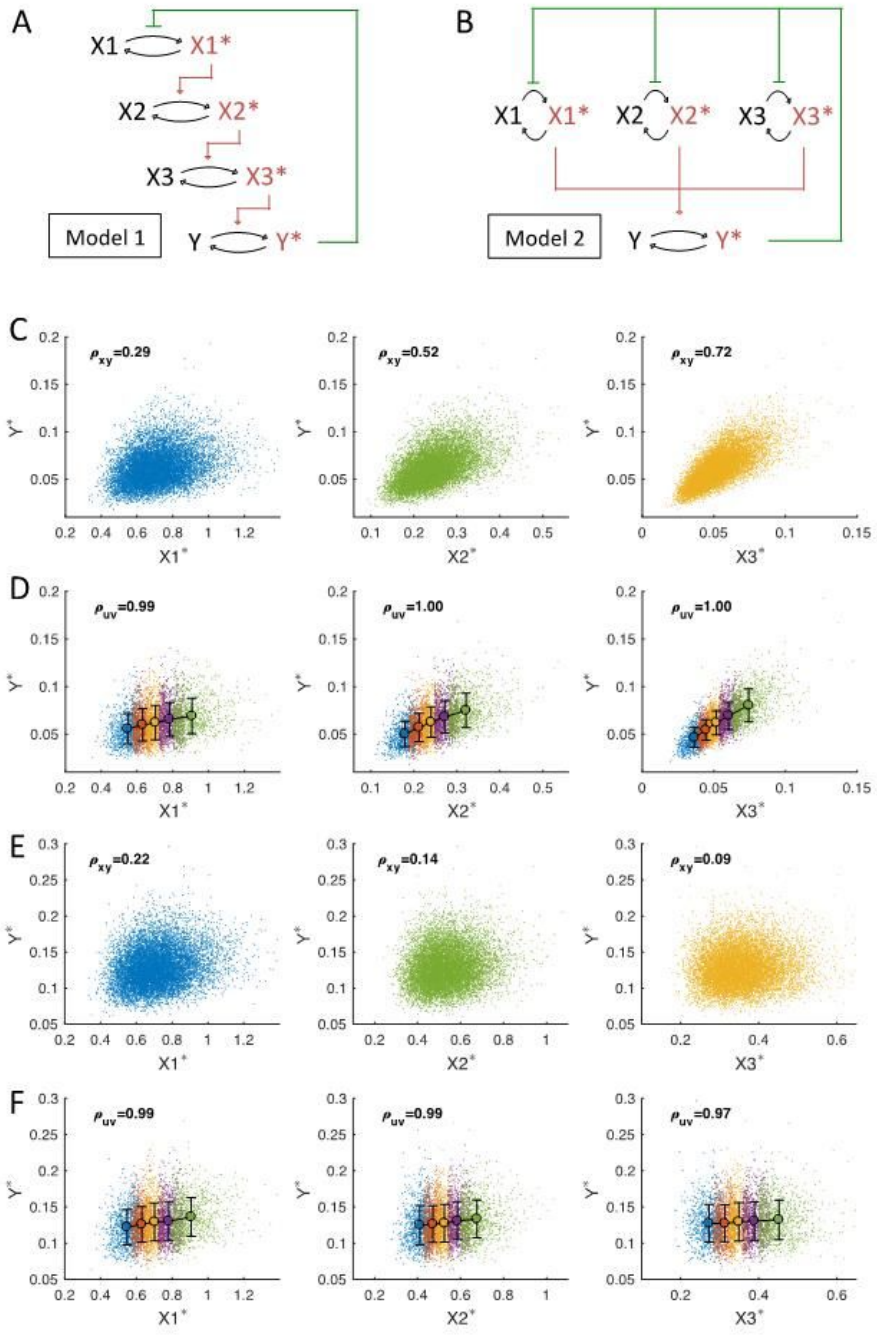

Figure 2

Discrepancy of correlation coefficients between the single-cell level and the population level. Two models were developed to represent typical biological regulatory systems. (A) Illustration of Model 1 describing a three-step signaling cascade. (B) Illustration of Model 2 characterizing a multi-regulator gene expression. (C) and (D) Pearson's correlation coefficients between the variable pairs X1* and Y*, X2* and Y*, X3* and Y* of the simulated 10000 cells in Model 1 at the single-cell level (C) or at the population level (D). (E) and

(F) Pearson's correlation coefficients between the variable-pairs X1* and Y*, X2* and Y*, X3* and Y* of the simulated 10000 cells in Model 2 at the single-cell level (E) or the population level (F). The values of the population level were generated by grouping the single cells based on their proximity of X*. See Methods for further detailed information.

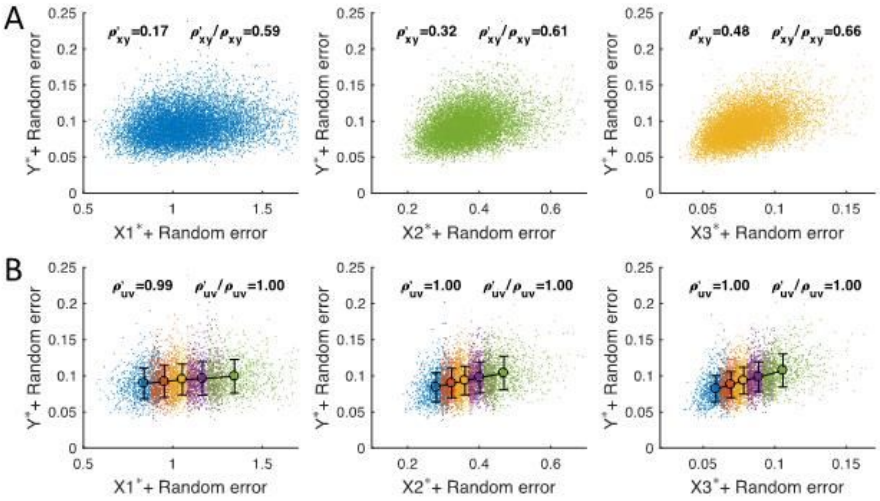

Figure 3

Aggregated correlation is less affected by technical random measurement errors than individual correlation. Pearson's correlation coefficients between variables X*(X1*, X2* or X3*)+random errors and Y*+random errors at the single cell level (A) or the population level (B) were shown. Random errors were introduced to the simulated 10000 cells in Figure 2 (C) and (D). The values of random error for each variable were generated from an independent lognormal distribution.

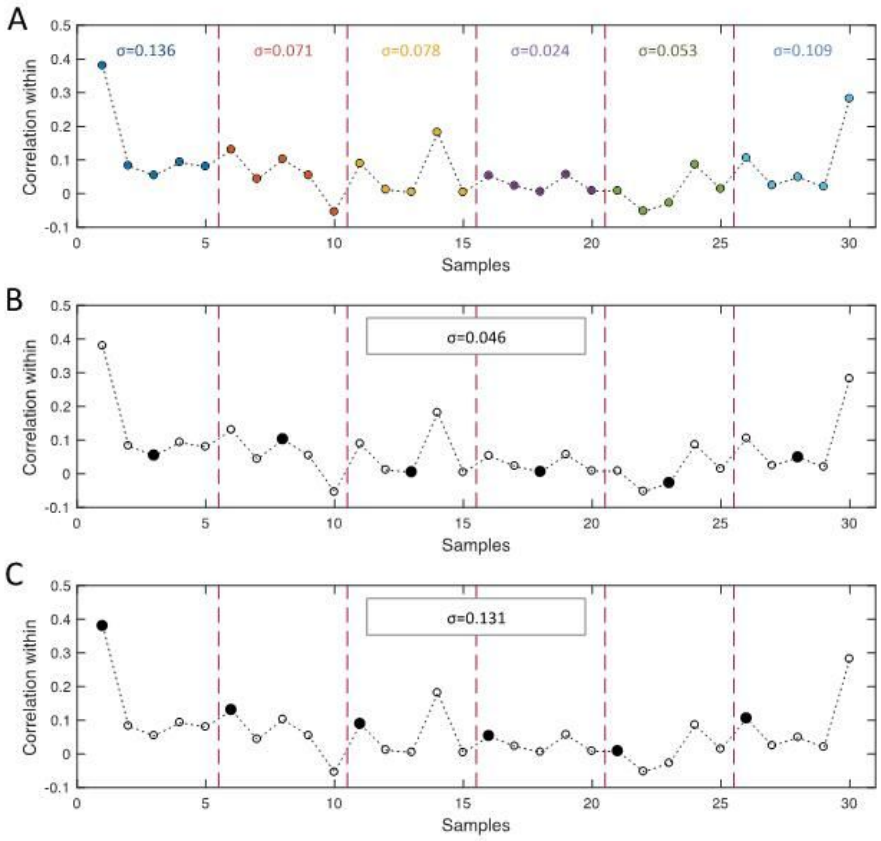

Figure 4

Comparing correlations-within between close-sample groups and sparse-sample groups. The correlation between Y* and X3* in Model 2 was investigated. 30000 cells were simulated and divided into 30 bulk samples according to their close proximity of X3*. (A) The correlations-within were calculated and every five samples closely located were merged into one close-sample group. For each close-sample group, we calculated the correlation coefficients of the five bulk-sample within it, and then derived the standard deviation of these coefficients. (B) and (C) The sparse-sample groups were generated by choosing one sample from each of the close-sample groups(as indicated by black filled circles). Then the standard deviation of the correlation coefficients in the sparse-sample group was calculated.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalFigures.docx
- Supplementalmaterialsrevised.docx