Modeling of cytometry data in logarithmic space: when is a bimodal distribution not bimodal?

Amir Erez[1*], Robert Vogel[2], Andrew Mugler[3], Andrew Belmonte[1,4], Grégoire Altan-Bonnet[1]

1 Immunodynamics Group, Cancer and Inflammation Program, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20814, USA
2 IBM T. J. Watson Research Center, Yorktown Heights, New York 10598, USA
3 Department of Physics and Astronomy, Purdue University, West Lafayette, Indiana 47907, USA
4 Department of Mathematics, Pennsylvania State University, University Park, Pennsylvania 16802, USA

* erezam@bgu.ac.il

Abstract

Recent efforts in systems immunology lead researchers to build quantitative models of cell activation and differentiation. One goal is to account for the distributions of proteins from single-cell measurements by flow cytometry or mass cytometry as a readout of biological regulation. In that context, large cell-to-cell variability is often observed in biological quantities. We show here that these readouts, viewed in logarithmic scale may result in two easily-distinguishable modes, while the underlying distribution (in linear scale) is uni-modal. We introduce a simple mathematical test to highlight this mismatch. We then dissect the flow of influence of cell-to-cell variability using a graphical model and its effect on measurement noise. Finally we show how acquiring additional biological information can be used to reduce uncertainty introduced by cell-to-cell variability, helping to clarify whether the data is uni- or bi-modal. This communication has cautionary implications for manual and automatic gating strategies, as well as clustering and modeling of single-cell measurements.

Author summary

Populations of cells are often composed of distinct sub-populations, each performing a unique biological function. A major tool to identify such populations is antibody staining followed by flow- and mass-cytometry. These technologies boast high acquisition speed, resolution and sensitivity, measuring $\sim 10^2 - 10^5$ molecules per cell in more than 15 different labels. With these data, identification of populations typically amounts to manually selecting clusters of cells with distinct molecular abundances. While such a strategy is sufficient for clearly distinct groups, our increasingly refined definitions of cell populations require objective criteria to partition a population. To establish the number of unique sub-populations, we apply both Hartigan's dip test and a new method which examines the statistical mode of abundance distributions. We demonstrate the necessity of these criteria both mathematically and experimentally. We find that the number of unique populations depends on whether the tag abundances are scaled linearly or logarithmically, an often overlooked fact. Interestingly, theoretical

models to explain such distributions are usually treated in linear space, whereas the experimental data is usually treated logarithmically. This mismatch between the two representations has the potential to mislead researchers, more so as technology advances and with it a growing reliance on automatic tools to distinguish populations in high-dimensional space. We detail an approach relying specifically on higher multiplexing in measurements that will be useful to circumvent this mismatch.

Introduction

Flow cytometry data typically stretches across several orders of magnitude, with fluorescence intensity I readily spanning values between 10^2 and 10^5. As such, when binning cytometry data to create histograms or distributions, it is natural to let bin sizes increase as a geometric progression, namely, to evenly bin the logarithm of the fluorescence intensity. As a result, instead of the distribution $Q(I)$ of fluorescence intensity I, one usually analyzes the distribution of $\log I$, which we denote $P(\log I)$. Indeed, $P(\log I)$ has many advantages: easy display of many orders of magnitude in I, easy to model as a two-component log-normal mixture model (as in [1]), and easy to intuitively understand the effect of changing the voltage gain on the flow-cytometer detector photo-multiplier. While such data presentation has been widely adopted in the field of cytometry out of these practical reasons, a rigorous assessment of this log-transformation reveals unwarranted features.

After estimating $P(\log I)$, by logarithmically binning or using a kernel-density method, one can formally derive $Q(I)$ as [2],

$$\begin{aligned} Q(I) &= P(\log I)\left|\frac{d}{dI}\log I\right| \\ &= \frac{1}{I}P(\log I) = e^{-y}P(y)\,, \end{aligned} \tag{1}$$

with $\log I \equiv y$.

Simply plotting $Q(I)$ vs. I is impractical as most of the data inevitably appears crowded against the $I = 0$ axis. Thus, it is common practice to plot $P(\log I)$ or variants thereof which deal with small and negative I values introduced by fluorescence compensation (*e.g.* "Logicle" [3], "VLog" [4] and other transformations [5]). Displaying faithfully flow-cytometry data is not easy, as the logarithmic scale and fluorescence compensation introduce problems that are easy to miss [6] leading to uncertainty in the number of distinct populations present in the data. Previously, attention has been given to the possibility of effects produced by logarithmic binning [7], contrasting the difference between plotting logarithmic histograms $P(\log I)$ vs. $\log I$ as opposed to rescaling the x-axis by plotting $Q(I)$ vs. $\log I$. However, an additional, potentially confusing situation seems to have been overlooked: the possible appearance of a second mode in $P(\log I)$, rendering $P(\log I)$ bi-modal, while for the same data only one mode exists in $Q(I)$. This is the focus of this work.

When considering biological measurements, I is proportional to the *actual* copy number of RNA or proteins. When theoretical considerations are applied to biological systems (such as biochemical dynamics [8–12, 15], mass-action chemical equilibria, cell-cycle measurements [13] and Hill dose-reponse curves [14]), it is the copy number itself that is under consideration. Despite that, the logarithm of copy number is an appealing quantity because of its approximately Gaussian statistics, yielding insight into details easily lost if the data were to be analyzed only in linear scale. This leads to a mismatch, where for instance models posed in linear space and data plotted in logarithmic space seem unable to be reconciled without invoking additional effects such

as stochastic gene expression noise [15] and cell-to-cell variability [16–18]. Even so, typically one must resort to approximations to analyze noise propagation linearly [8].

The difference between the convenient consideration of the logarithm of abundances and the theoretically-accurate analysis of the linear copy number renders the question of whether $Q(I)$ has one or two modes (3 extrema) relevant in the following ways: (i) the existence of 1 or 3 extrema is often used to infer the fixed points of a dynamic stochastic biochemical network [1, 9, 15] and other *in silico* methods [19] (ii) extrema are used to define cell-types in automatic (density based) gating and clustering algorithms [20–25]; (iii) The existence of a clearly bi-modal distribution is used for manual gating (*e.g.* discerning between activated and un-activated cells) in a way that appears more robust and compelling than it might truly be. It is this potential for confusion when the data is viewed in log-space, which we will elaborate on.

The rest of this paper is composed of two parts. In the first part, we point out and analyze the situation where a mismatch between the two representations can happen: we formally state the problem using theoretical modeling of cytometry data as a mixture of two log-normal distributions (colloquially the "negative" and "positive" modes), explicitly show situations where two modes appear in $P(\log I)$ while only one mode exists in $Q(I)$, and demonstrate this confounding effect on experimental data. In the second part, we analyze the role of cell-to-cell variability in experimental data and show how by measuring a suitable extra dimension one can factor out some of this variability, thus reducing the broadness of the modes sufficiently so as to reduce the mismatch between the two representations. Thus we provide a prescription to design experiments and analyze them so as to resolve the uni-modal *vs.* bi-modal discrepancy.

Theoretical method

Cytometry data is often amenable to modeling as a log-normal mixture (*e.g.* [1]). To demonstrate the log/linear mismatch we consider a mixture of two populations, characterized by the distribution of intensity. We define $P(\log I)$ as follows:

$$P(\log I) = \frac{(1-\alpha)}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(\log I - y_1)^2}{2\sigma_1^2}} + \frac{\alpha}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(\log I - y_2)^2}{2\sigma_2^2}}, \tag{2}$$

with $y_{1,2} = \log I_{1,2}$ which are the loci of the centers of the left and right Gaussians in log-space, respectively, and $\sigma_{1,2}$ the log-space standard deviations. We then define $Q(I) = \frac{1}{I} P(\log I)$ as in Eq. 1.

In Fig. 1, to illustrate with typical measurement values, we set $I_1 = 100$ and $I_2 = 1000$ (arbitrary units) and $\alpha = 0.5$, while varying $\sigma_1 = \sigma_2$. This figure presents the three cases we wish to contrast: on the left column, both $P(\log I)$ and $Q(I)$ are bi-modal; in the central column, $P(\log I)$ is bi-modal whereas $Q(I)$ is uni-modal; on the right column, both $P(\log I)$ and $Q(I)$ are uni-modal, a situation which we examine in more detail in Eq. 7. Moreover, we see an example where even when both are bi-modal (left column), the loci of the modes are different.

The question of uni/multi modality has been investigated before, in the context of modeling flow-cytometry data [19], by using Hartigan's dip test for uni-modality [26]. Briefly, Hartigan's dip statistic measures the maximum difference between the empirical distribution and the uni-modal distribution that minimizes that maximum difference. This is compared to the appropriate null distribution which is, in this case, the uniform distribution, to give p_u, a p-value for uni-modality. In Fig. 1, we report p_u for the data, according to Hartigan's test, by simulating 10^4 events drawn from the distributions under consideration [27]. We note that for the central column, Hartigan's test quantifies

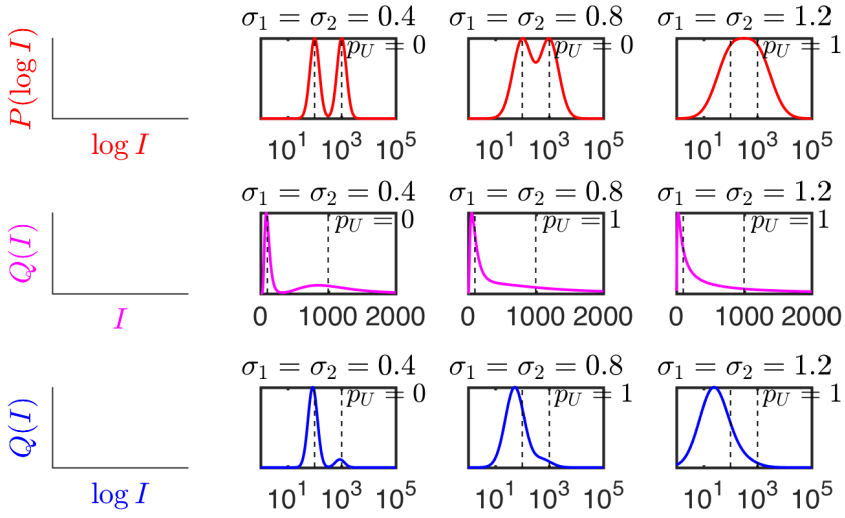


Fig 1. Log-normal mixture showing the mismatch in the number of peaks. Top row, red: $P(\log I)$ vs. $\log I$ normalized to max. **Middle row, magenta:** $Q(I)$ vs. I normalized to max, and plotted on a narrower range. **Bottom row, blue:** $Q(I)$ vs. $\log I$ normalized to max, rescaling the x-axis as in Ref. [7]. In the central column where $\sigma_1 = \sigma_2 = 0.8$, $P(\log I)$ shows explicit bi-modality whereas $Q(I)$ is uni-modal. **Right column:** in this case, the variance σ^2 is large enough (see Eq. 7) that $P(\log I)$ has only one mode even though it is modeled as a mixture. p_u is Hartigan's dip-test p-value for uni-modality. In the middle column plots, Hartigan's test further agrees that $P(\log I)$ is bi-modal whereas $Q(I)$ isn't. In all cases: $I_1 = 100$, $I_2 = 1000$, $\alpha = 0.5$ varying $\sigma_{1,2} = \{0.4, 0.8, 1.2\}$.

and concurs with a visual inspection of the data, *i.e.*, whereas $P(\log I)$ is not uni-modal ($p_u = 0$), for the corresponding $Q(I)$ it is uni-modal ($p_u = 1$).

We define $y = \log I$ and proceed to the number of extrema for $P(y)$ and $Q(I)$. It is possible to discern between one or three extrema of the distribution $P(\log I)$, corresponding to one or two modes (respectively) by counting the number of solutions for $\frac{d}{d\log I}P(\log I) = 0$. Similarly, $\frac{d}{dI}Q(I) = 0$ can have either one or three solutions. This raises the possibility of there being three extrema (two modes) for $P(\log I)$ whereas only one mode in $Q(I)$. We explicitly evaluate the extrema of the mixture of log-normal distributions by solving,

$$\begin{aligned}\left.\frac{dP(y)}{dy}\right|_{y=y_*} &= -\frac{(y_*-y_1)}{\sigma_1^2}\frac{(1-\alpha)}{\sqrt{2\pi}\sigma_1}e^{-\frac{(y-y_1)^2}{2\sigma_1^2}} - \frac{(y_*-y_2)}{\sigma_2^2}\frac{(\alpha)}{\sqrt{2\pi}\sigma_2}e^{-\frac{(y-y_2)^2}{2\sigma_2^2}} \\ &= 0\end{aligned}$$

which by algebraic rearrangement,

$$\underbrace{\frac{(y_*-y_1)/\sigma_1^2}{(y_2-y_*)/\sigma_2^2}}_{S_3(y_*)} = \underbrace{\frac{\sigma_1}{\sigma_2}\left(\frac{\alpha}{1-\alpha}\right)e^{\frac{(y_*-y1)^2}{2\sigma_1^2}}e^{-\frac{(y_*-y_2)^2}{2\sigma_2^2}}}_{F(y_*)}, \tag{3}$$

provides a more transparent form.

Here we refer to the LHS and RHS as $S_3(y)$ and $F(y)$, respectively. In like, computing the extrema of the linear scale distribution amounts to $\left.\frac{dQ(I)}{dI}\right|_{\log I=y_*} = 0$,

which by change of variable is equivalent to,

$$P(y_*) = \left.\frac{dP(y)}{dy}\right|_{y=y_*}, \tag{4}$$

and by substitution,

$$\underbrace{\frac{\frac{y_*-y_1}{\sigma_1^2}+1}{\frac{y_2-y_*}{\sigma_2^2}-1}}_{S_1(y_*)} = F(y_*) \tag{5}$$

We refer to the quantity on the LHS, $S_1(y)$. Thus we have introduced the following functions,

$$\begin{aligned} F(y) &= \left(\frac{\alpha}{1-\alpha}\right)\left(\frac{\sigma_1}{\sigma_2}\right) e^{\frac{(y-y_1)^2}{2\sigma_1^2}-\frac{(y-y_2)^2}{2\sigma_2^2}} \\ S_1(y) &= \frac{\frac{y-y_1}{\sigma_1^2}+1}{\frac{y_2-y}{\sigma_2^2}-1} \\ S_3(y) &= \frac{(y-y_1)/\sigma_1^2}{(y_2-y)/\sigma_2^2}\,. \end{aligned} \tag{6}$$

$F(y)$ is the ratio of the two Gaussians in Eq. 2 and is therefore always non-negative; this implies that any extremum y_* must satisfy $S_1(y_*) \geq 0$ and $S_3(y_*) \geq 0$. We note that in Eq. 4 the condition for extrema in $Q(I)$ requires that $\frac{d}{dy}P(y) = P(y)$ whereas, of course, extremizing $P(\log I)$ sets its derivative to zero, demonstrating the fact that the loci of the modes for $P(\log I)$ and $Q(I)$ are manifestly different. The region where the log-space distribution shows a second mode occurs when Eq. 3 for S_3 admits three solutions whereas Eq. 5 for S_1 admits only one. Given that Eq. 3 and 5 are transcendental, a graphical way to asses the number of solutions is to plot $F(y), S_1(y), S_3(y)$ and count the number of times S_1 and S_3 intersect F.

In Fig. 2, we present an example of this graphical method. The mismatch between the number of extrema of $P(\log I)$ and $Q(I)$ is apparent whenever (red curve) $S_3(y)$ intersects F at 3 points, whereas (blue curve) $S_1(y)$ only intersects F once.

In the plots along the diagonal, we have $\sigma_1 = \sigma_2$ (as in Fig. 1) which simplifies $F(y)$ since the quadratic (Gaussian) terms cancel, leaving only an exponential. This leads to a simple criterion to determine whether $P(\log I)$ itself admits one or two modes - previously in Fig. 1(right) we saw an example where $P(\log I)$ is uni-modal despite being generated from a mixture. Graphically, we see that for $S_3 = F$ to have 3 solutions, $\log S_3(y)$ has to have a slope less than $\log F(y)$ about the extremum y_*. In other words, $\left.\frac{d}{dy}\log S_3(y)\right|_{y_*} \leq \left.\frac{d}{dy}\log F(y)\right|_{y_*}$, with equality as the threshold between 1 and 3 extrema, similarly to the way Landau theory defines the critical point in second order phase transitions [28]. This leads to the following intuitive criterion,

$$(y_* - y_1)(y_2 - y_*) \geq \sigma^2 \implies 3 \text{ extrema for } P(\log I)\,, \tag{7}$$

which states that for $P(\log I)$ to appear bi-modal, it must have an extremum (y_*) such that the variance of the individual Gaussian components of $P(\log I)$ must be smaller than the distance between y_* and the Gaussian centers. Substituting for $y_* \approx \log 316$, $y_1 = \log 100$ and $y_2 = \log 1000$ and $\sigma^2 = 1.44$ we see that the criterion in Eq. 7 is not satisfied and indeed in Fig. 1(right) and Fig. 2(bottom right) we see that $P(\log I)$ has only one mode. A similar condition can be derived for $Q(I)$, that is,

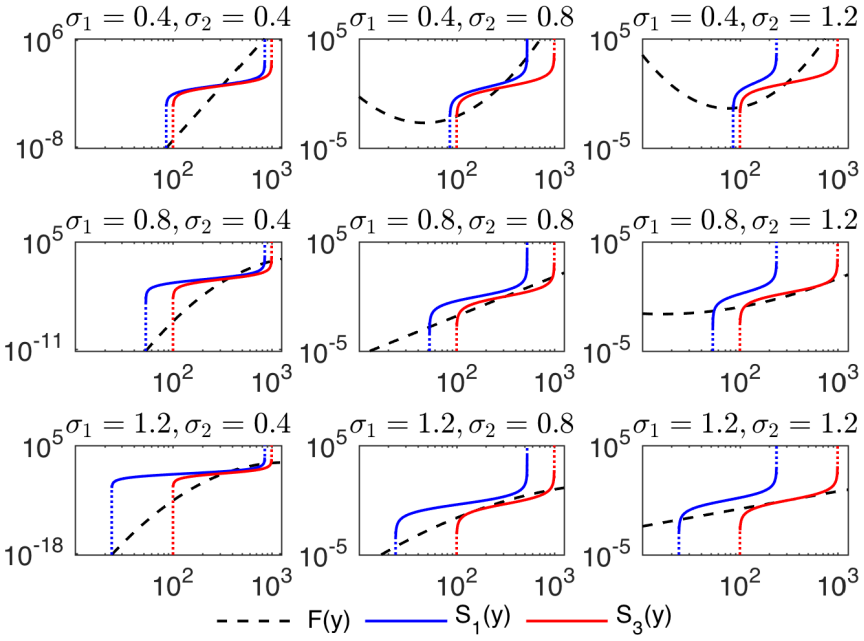


Fig 2. Graphical solution to count the number of extrema. When the red (blue) curves intersect the dashed black line, $P(\log I)$ ($Q(I)$) are extremized. Dashed black: $F(y)$, blue: $S_1(y)$ and red: $S_3(y)$. The mismatch between the number of extrema of $P(\log I)$ and $Q(I)$ is apparent when the red curve intersects F at 3 points, whereas the blue curve only intersects F once. In both cases, the loci of the extrema are different for the two distributions. The plots along the diagonal (which correspond to the cases in Fig. 1) show the case $\sigma_1 = \sigma_2$ which simplifies $F(y)$ to a straight line (the general case being a parabola) in these axes.

$\frac{d}{dy}\log S_1(y)\Big|_{y_*} \leq \frac{d}{dy}\log F(y)\Big|_{y_*}$, such that,

$$(y_* - y_1 + \sigma^2)(y_2 - y_* - \sigma^2) \geq \sigma^2 \implies 3 \text{ extrema for } Q(I)\,, \tag{8}$$

it is, however, hard to compare the two bounds analytically because the y_* which extremizes $P(\log I)$ is different from the y_* which extremizes $Q(I)$.

As a check for the predictive power of Hartigan's test with regards to experimental data, we apply it on a log-normal mixture comparing its predictive power as a function of the number of tests and number of events in each test [27]. In Fig. 3, we test it on the situation in Fig. 2(top,middle) in which $Q(I)$ is weakly bi-modal, meaning that its bi-modality is nearly marginal ($\sigma_1 = 0.4$ and $\sigma_2 = 0.8$). The Hartigan probability of unimodality (p_u) is not sensitive to the number of bootstrap tests in a reasonable range but becomes strongly predictive of the (weak) bi-modality only when there are more than 10^5 events. Such an abundance of cells may not always be available in typical flow cytometry data, especially for sub-populations which have been selected (gated) and may comprise only a small fraction of all the cells acquired. Fig. 3 supports that in such weakly bi-modal situations, Hartigan's p-value should be treated cautiously, a situation which we will encounter in Fig. 6.

We conclude this theoretical section of the manuscript by making a more intuitive argument. Above, we have demonstrated the theoretical existence of a situation where $P(\log I)$ has two modes when $Q(I)$ has only one. But how does this *come about* ?

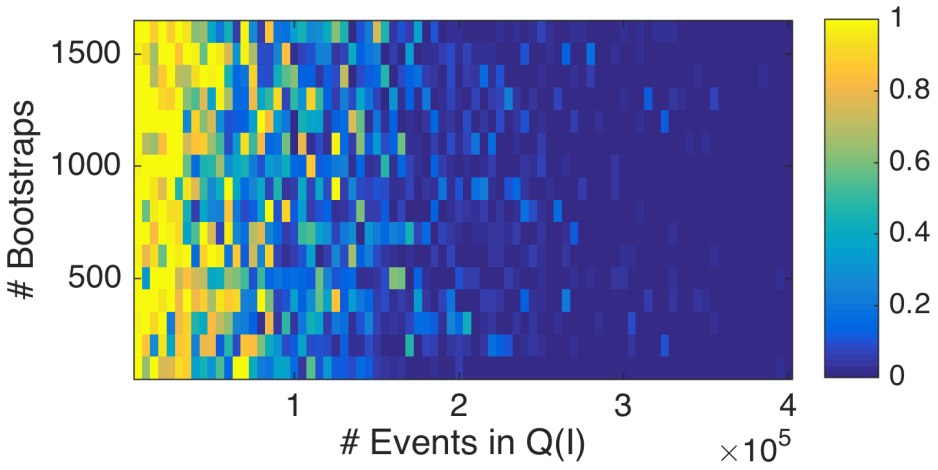


Fig 3. Testing Hartigan's p-value p_u for uni-modality on a weakly bi-modal log-normal mixture. The heat-map shows p_u as a function of the number of bootstrap tests and number of events in each test, for $\sigma_1 = 0.4$ and $\sigma_2 = 0.8$ as in Fig. 2(top,middle). With less than $\sim 10^5$ events, Hartigan's test for this case may misidentify the number of peaks.

Intuitively, by binning in logarithmic scale, we are effectively making the bin sizes grow as I increases. A larger bin can only lead to a higher count in that bin and so we might stumble upon a regime where this creates a quasi-mode. Probing further, we established a simple criterion (Eq. 7) where by making the underlying Gaussian mixture composed of too-close-together Gaussians, even in $P(\log I)$ there exists only one peak.

Experimental results

We now apply our analysis to experimental data, namely the measured distribution of Extracellular Signal-regulated Kinase (ERK) phosphorylation (ppERK) signaling in CD8+ primary mouse T-cells responding to antigens and inhibited by the SRC inhibitor Dasatinib. These data were acquired in exactly the same manner as the experiments in Ref. [1], for brevity we refer the reader there for all experimental details. In Fig. 4(A) we see how a commercially-available analysis software (FlowJo [29]) plots the distribution of ppERK in such an experiment, which clearly shows a bi-modal structure. Fig. 4(B) plots those same data when subjected to logarithmic binning $P(\log I)$, giving the two modes as in FlowJo (red dots) whereas $Q(I)$ has a single mode (blue dots). We fit $P(\log I)$ as a Gaussian mixture. This is followed in Fig. 4(C) by the same extrema analysis as in Fig. 2, revealing that indeed $Q(I)$ has a single maximum.

Immunologists have relied on cytometry for over forty years to identify new cell populations, often based on manual gating of cytometry data. The success of this method (with validation by identification of new transcription factors) stands in contrast with the danger of generating modes in the distributions of log-transformed data $P(\log I)$, as presented above. This is particularly cogent to recent efforts at clustering single-cell measurements in high-dimensional space by mass cytometry. Hence, we wondered whether gating in logarithmic scale could be justified *a posteriori*, based on biological knowledge. We aimed to include additional information in our analysis such that the biological significance of the distributions in our single-cell measurements is better captured.

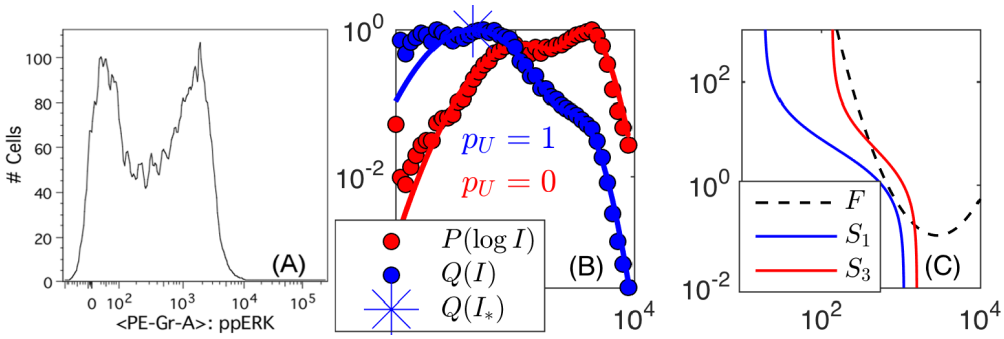


Fig 4. Analysis of experimental data reveals the effect we describe in a real scenario. (A) Histogram of ppERK as plotted by FlowJo [29]; **(B)** Histograms for $P(\log I)$ (red,dots) and $Q(I)$ (blue,dots) vs. $\log I$ as estimated from the data binned logarithmically. Note that plotting $Q(I)$ vs. $\log I$ is somewhat unusual but allows both P and Q to be plotted on the same axis. The red line shows the result of fitting $P(\log I)$ to a gaussian mixture model (Eq. 2), and the blue line is the estimate for $Q(I)$ from $P(\log I)$ according to Eq. 1. The blue star indicates the location of the only maximum for $Q(I)$ obtained from Eq. 5, despite the obvious two maxima in $P(\log I)$ (red). Hartigan's uni-modality p-values for $\log I$ (red) and I (blue) are taken directly from the data without binning, corroborating that $\log I$ is bi-modal whereas I is unimodal. **(C)** Graphic solution of the extrema conditions as in Fig. 2 explicitly reveals the three solutions for $P(\log I_*)$ (red line intersects black dashed) as opposed to the single solution for $Q(I_*)$ (the blue line intersects the black dashed line up beyond the plotted area, solution also plotted as blue star in the middle plot), indicating that $Q(I)$ has only one mode.

In Fig. 5(A), we show the experimental data we will use in our proposed solution. Here we returned to our single-cell measurements of ERK phosphorylation in primary mouse T cells in Ref. [16]. We show a heat map of the *joint* distribution of ppERK (I_{ppERK}) and total ERK1 (I_{ERK1}) expression in mouse CD8+ T-cells. Notably, whereas the two modes of ppERK significantly overlap when plotted in the marginal distribution $P(\log I_{ppERK})$, ppERK expression correlates with total ERK1 levels in their joint distribution $P_2(\log I_{ppERK}, \log I_{ERK1})$. Each cell's state encodes another *latent* variable, its activation status - which tells if the cell has been successfully activated by the stimulus. To deduce the activation status, it is common practice to use manual gating of the data by drawing of a boundary between the active and inactive states. To account for the correlation between ERK1 and ppERK we consider two manual gating strategies: (i) perpendicular gating (dashed red) according to $P(\log I_{ppERK})$ with $I_{ppERK} > ppERK_*$ considered an activated cell, and (ii) diagonal gating according to the apparent correlation in P_2 (dashed grey). We set the diagonal gate with a slope of unity, meaning that we take the dividing line, reflecting proportionality $ppERK \propto ERK1$, as a good way to partition the two states. We define "Inactive" to the left of the dashed line, and "Active" to the right of it.

To understand the structure of these data, it is important to characterize explicitly the dependency structure of our observables (ERK1, ppERK), the latent activation status, and the influence of external factors on these three. The existence of two peaks, in ppERK which appear distinct from each other but correlated with ERK1 levels, guides us to use a Bayesian network to capture these features in the data as a graphical model. First - we test whether ERK1 and the cell's activation status are independent. In S1 Fig we see that whereas independence implies that $P(\log ERK1) = P(\log ERK1|Activation)$, in fact there is a weak dependence between

them regardless of whether we employ a vertical (red) or a diagonal (grey) gate (the diagonal gate showing a weaker dependence). The weak dependence between activation state and ERK1 levels is reasonable, if we account for cell-to-cell variability, since for a given stimulus some cells inevitably respond differently from the typical cell [30]. We summarize the causal structure for this system in Fig. 5(B) which depicts a probabilistic graphical model [31] of the flow of influence from cell-to-cell variability and activation signal, to ERK1 levels and activation status, and finally to the distribution of ppERK. We depict the pair ERK1-ppERK in a template, to suggest to the reader the existence of multiple other pairs. Importantly, in what follows we show how to better resolve the log-space peak; this recipe, together with the model in Fig. 5(B) can be used *a priori* in automatic gating and clustering algorithms to prevent some of the mismatch between logarithmic and linear binning strategies. For stochastic modeling, such a structure presents an opportunity to analyze the structure and propagation of noise in the system [8, 32].

We treat the broadness of ppERK modes as generated by cell-to-cell variability in total ERK1 content - a reasonable assumption since the noise in the phosphorylation of ERK is negligible in comparison [33]. We further neglect the indirect influence between activation status and ERK1 levels due to its weakness (checked in S1 Fig). Thus we approximate that the conditional independence between ERK1 and activation status (given that both are influenced by cell-to-cell variability) is true independence. This implies an approximately linear relation $I_{ppERK} \propto I_{ERK1}$ given activation status. We define the normalized intensity $\tilde{I} = I_{ppERK}/I_{ERK1}$ as the ratio of ppERK to ERK1 intensity, thereby eliminating the linear dependence of ppERK on ERK1 levels and reducing uncertainty due to cell-to-cell variability. The resulting $P(\log \tilde{I})$ may boast a sufficiently reduced noise in ppERK such that a clear bi-modal signature appears regardless of logarithmic or linear binning of $\tilde{I}$. In Fig. 6 we show such an example, where in Fig. 6(A,B), $P(\log I)$ and $Q(I)$ do not agree on the number of modes, whereas in Fig. 6(C,D) $\tilde{I} = I_{ppERK}/I_{ERK1}$ do agree. These data have order 25,000 events and so, similarly to Fig. 3, Hartigan's test may not identify the number of peaks correctly, as is indicated in the p_u values.

Thus we demonstrate how by suitably accounting for cell-to-cell variability one can reduce the measured noise so as to circumvent the mismatch in the number of modes between the logarithmic and linear treatment. For testing bi-modality, whereas our method relies on fitting a Gaussian mixture, Hartigan's test requires no fitting yet may lack statistical power when applied to typical experimental situations.

Conclusion

The scale-dependent bi-modality as demonstrated in Fig. 4 and Fig. 6(A,B) may be not uncommon. Specifically, one must take extra care when attempting to manually gate, automatically cluster or build dynamical models which rely on an apparent bi-modal structure, as it might depend on whether the data was log-transformed or not. This becomes increasingly relevant as cytometry moves forward to higher dimensional measurements which become tractable only with automatic gating schemes. Instead, one might consider plotting $Q(I)$ on the log-log scale, a presentation which preserves the number of maxima, at the expense of the measure of the distribution. It is possible to ameliorate the mismatch between the two scales, as we demonstrate in Fig. 6(C,D), if one can simultaneously measure correlated observables (in our example, ppERK and ERK1). This allows to control for cell-to-cell variability, increasing the resolution of the data. Recently, this favorable scenario has become more attainable with the introduction of mass cytometry - where one can rely on a large number of channels

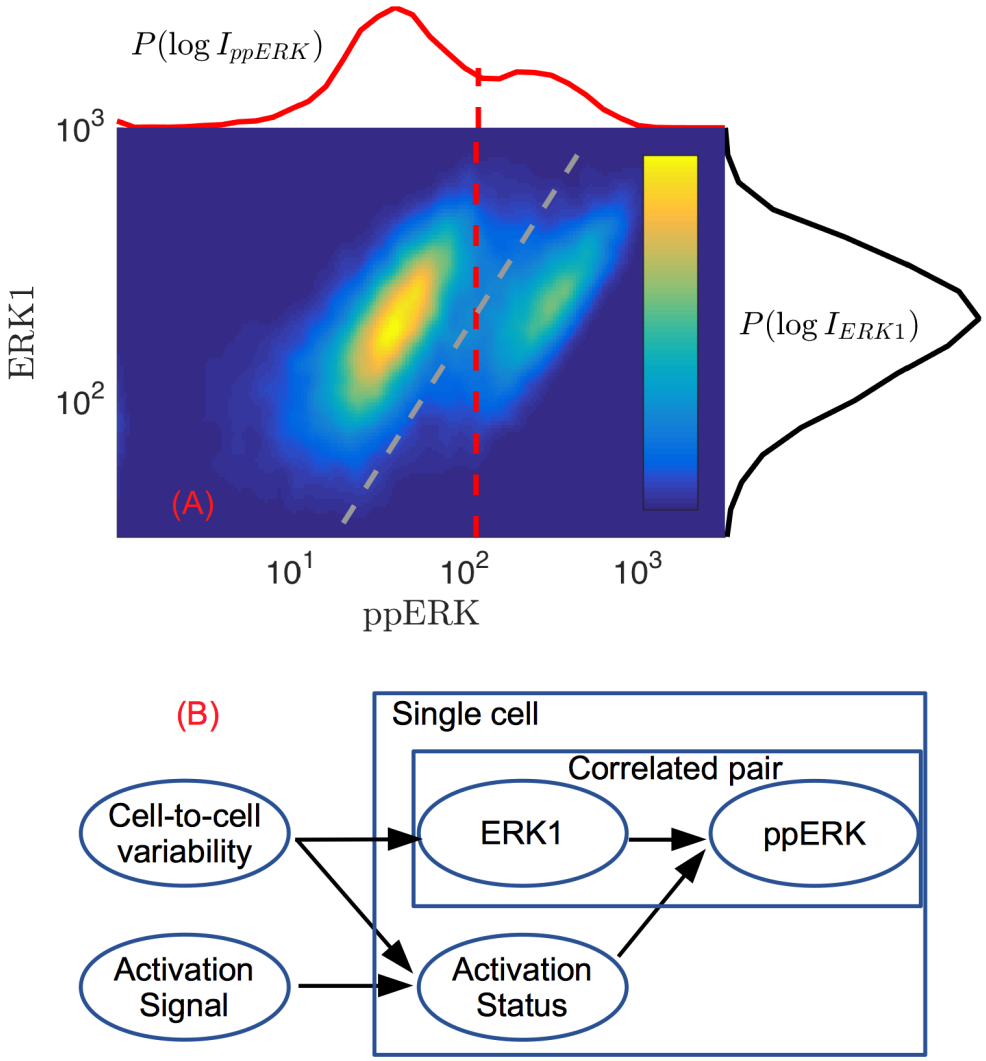


Fig 5. Analysis of experimental data with two correlated measurements. **(A)** The joint distribution $P_2(\log I_{ppERK}, \log I_{ERK1})$ as a heat map with its marginals plotted on the top and on its right. The correlation between ppERK and ERK1 levels is clear in the data. Dashed red (grey) lines are proposed manual gates according to the marginal (joint) distributions $P(\log I_{ppERK})$ ($P_2(\log I_{ppERK}, \log I_{ERK1})$). **(B)** Bayesian network depicted as a graphical model to show the flow of influence on the measurement of ppERK. The pair ERK1 and ppERK are in a template to suggest that there exist other pairs of correlated observables that depend on activation status and cell-to-cell variability.

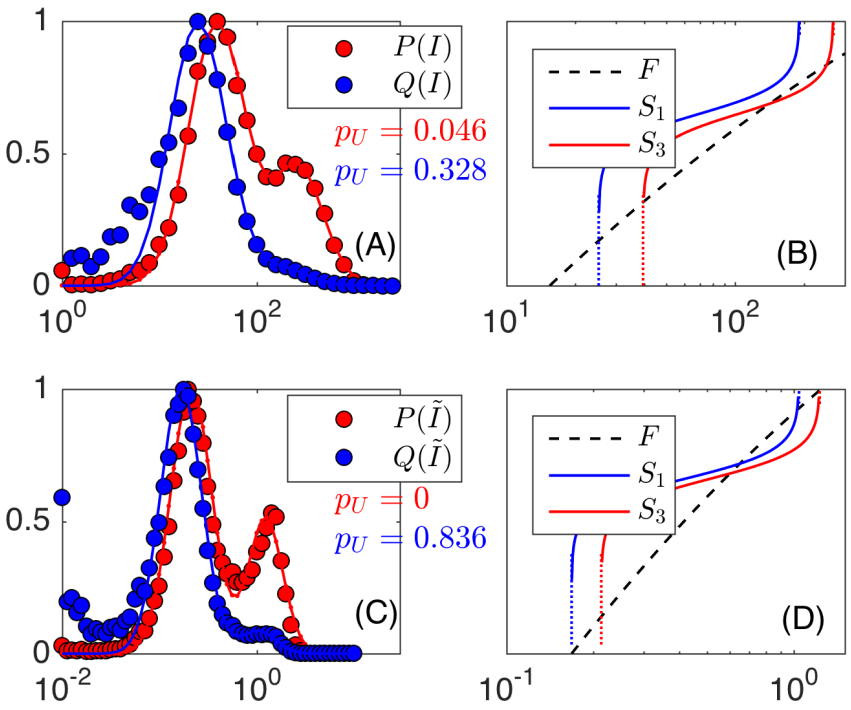
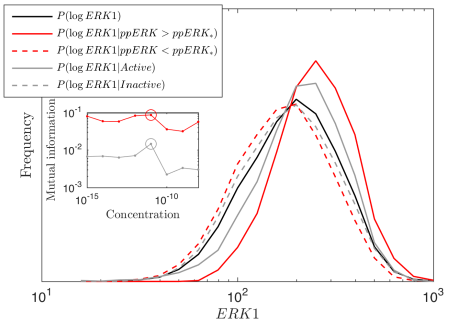


Fig 6. By dividing ppERK readings by ERK1, we can ameliorate the mismatch between the two representations. (A) $P(\log I)$ and $Q(I)$ vs. $\log I$ (dots: data, lines: gaussian mixture fit) together with **(B)** their extrema analysis, showing that the second mode in log ppERK does not exist if the data is linearly binned. **(C)** The same treatment but for $\tilde{I} = I_{ppERK}/I_{ERK1}$, **(D)** shows that both $\tilde{I}$ and $\log \tilde{I}$ have two modes, thus normalizing ppERK levels by total ERK1 maintains the bi-modal structure both in $P(\log \tilde{I})$ and in $Q(\tilde{I})$.

without compromising the flow-cytometry panel. Based on the analysis carried out in this paper, we conjecture that such extra channels, chosen wisely, can provide automatic clustering/gating algorithms the right information needed to make more reliable clustering and population defining. This is a simple way to introduce knowledge of the biological structure of the data into otherwise objective clustering algorithms, without compromising their objectivity. We propose and test some features of a graphical model that captures the structure of such dependencies in a way potentially useful for those interested in automatic gating and clustering algorithms. Though we caution on the use of $P(\log I)$, we find it remarkable how well the distribution of biological quantities can be modeled as a log-normal mixture. This highlights the deep and still little understood connection between distributions observed in living things and their relation to the logarithm of abundance, a subject likely to puzzle researchers for years to come.

Supporting information

S1 Fig Test for weak dependence of ERK1 and activation status. Whereas independence implies that $P(\log ERK1) = P(\log ERK1|Activation)$, in fact there is a weak dependence between them regardless of whether we employ a vertical (red, defined by threshold value $ppERK_*$) or a diagonal (grey) gate (the diagonal gate showing a weaker dependence); this is observed directly by noting that the different distributions in S1 Fig do not lie on top of each other. To quantify this difference, the inset shows the mutual information between $P(\log ERK1)$ and $P(\log ERK1|Activation)$, with the circle pointing out the particular concentration of stimulus (out of all concentrations used in Ref. [16]) we chose to plot in this example. The chosen concentration has the highest mutual information, *i.e.*, the lowest ability to discern between the active and inactive states.



Acknowledgments

This work was supported by Human Frontier Science Program grant LT000123/2014 (A. E.) and by the Intramural Research Program of the NCI, NIH.

References

1. Vogel RM, Erez A, Altan-Bonnet G. Dichotomy of cellular inhibition by small-molecule inhibitors revealed by single-cell analysis. Nature Communications. 2016;7:12428.

2. Mukhopadhyay N. Probability and Statistical Inference. 1st ed. New York: CRC Press; 2000.

3. Parks DR, Roederer M, Moore WA. A new Logicle display method avoids deceptive effects of logarithmic scaling for low signals and compensated data. Cytometry Part A. 2006;69A(6):541–551. doi:10.1002/cyto.a.20258.

4. Bagwell CB, Hill BL, Herbert DJ, Bray CM, Hunsberger BC. Sometimes simpler is better: VLog, a general but easy-to-implement log-like transform for cytometry. Cytometry Part A. 2016;89(12):1097–1105. doi:10.1002/cyto.a.23017.

5. Finak G, Perez JM, Weng A, Gottardo R. Optimizing transformations for automated, high throughput analysis of flow cytometry data. BMC Bioinformatics. 2010;11(1):546. doi:10.1186/1471-2105-11-546.

6. Herzenberg LA, Tung J, Moore WA, Herzenberg LA, Parks DR. Interpreting flow cytometry data: a guide for the perplexed. Nat Immunol. 2006;7(7):681–685.

7. Novo D, Wood J. Flow cytometry histograms: Transformations, resolution, and display. Cytometry Part A. 2008;73A(8):685–692. doi:10.1002/cyto.a.20592.

8. Prill RJ, Vogel R, Cecchi GA, Altan-Bonnet G, Stolovitzky G. Noise-Driven Causal Inference in Biomolecular Networks. PLOS ONE. 2015;10(6):e0125777. doi:10.1371/journal.pone.0125777.

9. Pal M, Ghosh S, Bose I. Non-genetic heterogeneity, criticality and cell differentiation. Physical Biology. 2015;12(1):016001.

10. Ridden SJ, Chang HH, Zygalakis KC, MacArthur BD. Entropy, Ergodicity, and Stem Cell Multipotency. Phys Rev Lett. 2015;115(20):208103.

11. Mojtahedi M, Skupin A, Zhou J, CastaÃ±o IG, Leong-Quong RYY, Chang H, et al. Cell Fate Decision as High-Dimensional Critical State Transition. PLOS Biology. 2016;14(12):e2000640. doi:10.1371/journal.pbio.2000640.

12. Erez A, Byrd TA, Vogel RM, Altan-Bonnet G, Mugler A. Criticality of biochemical feedback. ArXiv 170304194. 2017;.

13. Brown MR, Summers HD, Rees P, Smith PJ, Chappell SC, Errington RJ. Flow-Based Cytometric Analysis of Cell Cycle via Simulated Cell Populations. PLOS Computational Biology. 2010;6(4):e1000741. doi:10.1371/journal.pcbi.1000741.

14. Prinz H. Hill coefficients, dose-response curves and allosteric mechanisms. Journal of Chemical Biology. 2010;3(1):37–44.

15. Friedman N, Cai L, Xie XS. Linking Stochastic Dynamics to Population Distribution: An Analytical Framework of Gene Expression. Phys Rev Lett. 2006;97(16):168302–.

16. Feinerman O, Veiga J, Dorfman JR, Germain RN, Altan-Bonnet G. Variability and Robustness in T Cell Activation from Regulated Heterogeneity in Protein Levels. Science. 2008;321(5892):1081–1084. doi:10.1126/science.1158013.

17. Pelkmans L. Using Cell-to-Cell Variability—A New Era in Molecular Biology. Science. 2012;336(6080):425–426. doi:10.1126/science.1222161.

18. Cotari JW, Voisinne G, Dar OE, Karabacak V, Altan-Bonnet G. Cell-to-Cell Variability Analysis Dissects the Plasticity of Signaling of Common gamma Chain Cytokines in T Cells. Sci Signal. 2013;6(266):ra17–. doi:10.1126/scisignal.2003240.

19. Das J, Ho M, Zikherman J, Govern C, Yang M, Weiss A, et al. Digital Signaling and Hysteresis Characterize Ras Activation in Lymphoid Cells. Cell. 2009;136(2):337–351.

20. Cron A, Gouttefangeas C, Frelinger J, Lin L, Singh SK, Britten CM, et al. Hierarchical Modeling for Rare Event Detection and Cell Subset Alignment across Flow Cytometry Samples. PLOS Computational Biology. 2013;9(7):e1003130. doi:10.1371/journal.pcbi.1003130.

21. O'Neill K, Aghaeepour N, Å pidlen J, Brinkman R. Flow Cytometry Bioinformatics. PLOS Computational Biology. 2013;9(12):e1003365. doi:10.1371/journal.pcbi.1003365.

22. Anchang B, Do MT, Zhao X, Plevritis SK. CCAST: A Model-Based Gating Strategy to Isolate Homogeneous Subpopulations in a Heterogeneous Population of Single Cells. PLOS Computational Biology. 2014;10(7):e1003664. doi:10.1371/journal.pcbi.1003664.

23. Finak G, Frelinger J, Jiang W, Newell EW, Ramey J, Davis MM, et al. OpenCyto: An Open Source Infrastructure for Scalable, Robust, Reproducible, and Automated, End-to-End Flow Cytometry Data Analysis. PLOS Computational Biology. 2014;10(8):e1003806. doi:10.1371/journal.pcbi.1003806.

24. Saeys Y, Gassen SV, Lambrecht BN. Computational flow cytometry: helping to make sense of high-dimensional immunology data. Nat Rev Immunol. 2016;16(7):449–462.

25. Chen H, Lau MC, Wong MT, Newell EW, Poidinger M, Chen J. Cytofkit: A Bioconductor Package for an Integrated Mass Cytometry Data Analysis Pipeline. PLOS Computational Biology. 2016;12(9):e1005112. doi:10.1371/journal.pcbi.1005112.

26. Hartigan JA, Hartigan PM. The Dip Test of Unimodality. The Annals of Statistics. 1985; p. 70–84.

27. Price N. Hartigan's dip test MATLAB implementation;. Available from: `http://www.nicprice.net/diptest/`.

28. Pathria RK, Beale PD. Statistical Mechanics, Third Edition. 3rd ed. Amsterdam ; Boston: Academic Press; 2011.

29. FlowJo-LLC;. Ver 9.9 for OSX. Available from: `https://flowjo.com`.

30. 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.

31. Koller D, Friedman N. Probabilistic Graphical Models: Principles and Techniques. 1st ed. Cambridge, MA: The MIT Press; 2009.

32. Ching ESC, Tam HC. Reconstructing links in directed networks from noisy dynamics. Phys Rev E. 2017;95(1):010301.

33. Filippi S, Barnes C, Kirk PW, Kudo T, Kunida K, McMahon S, et al. Robustness of MEK-ERK Dynamics and Origins of Cell-to-Cell Variability in MAPK Signaling. Cell Reports. 2016;15(11):2524–2535. doi:10.1016/j.celrep.2016.05.024.

34. Marin JM, Mengersen K, Robert CP. Bayesian Modelling and Inference on Mixtures of Distributions. In: Dey DK, Rao CR, editors. Handbook of Statistics. vol. Volume 25. Elsevier; 2005. p. 459–507. Available from: http://www.sciencedirect.com/science/article/pii/S0169716105250162.