July 31, 2017

Cytometry A Editorial Office

Dear Editor,

Please find enclosed our manuscript "Modeling of cytometry data in logarithmic space: when is a bimodal distribution not bimodal?" submitted for consideration in *Cytometry A*.

Flow- and mass-cytometry have become essential tools in immunology and beyond. These measurement devices give abundant, highly sensitive single-cell data yielding distributions of cellular markers. These distributions are typically analyzed in logarithmic scale on a bi-axial plot where one selects ("gates") phenotypic subsets, thus defining distinct populations. As the technology progressed, it is now possible to measure a large number of dimensions, making manual analysis more difficult, as a 15-marker panel involves over 100 pairwise plots. Accordingly, the field moves to rely on automatic clustering and dimensional-reduction algorithms, sometimes rendering the analysis steps opaque. At the same time, advances in computational biology have yielded interesting mathematical models that are applied on such data, without always checking how logarithmically transforming the data might affect the models.

Here, we point out and analyze a mathematical phenomenon whereby distributions may appear bi-modal when binned in logarithmic scale, while appearing uni-modal in linear scale. Importantly, linear scale is where both mathematical models and actual DNA transcription and translation occurs. Thus we are faced with an inconvenient situation where models in linear scale and analysis in logarithmic scale may rely on density peaks which exist only in the logarithmic representation. Density-based automatic gating and clustering algorithms may be similarly confused by this mismatch. Indeed, technical changes in the experiment (eg. swapping one fluorescent dye with another, brighter one) might yield qualitatively different results despite being biologically equivalent, thus potentially causing reproducibility issues.

Our study advances the analysis of flow-and mass-cytometry data in several specific ways:

1. We explicitly analyze the occurrence of mismatch in the number of density peaks between the logarithmic and linear representations, and derive a mathematical test that indicates these conditions. We apply this test on experimental data, showing that the situation we describe may not be uncommon.
2. We compare our test with an existing statistical test, Hartigan's dip test, which has been employed previously in computational biology studies to determine bi-modality of cytometry distributions. Our test is more reliable when ran on a small subset of cells, less than approximately 10^5, a typical situation for sub-populations of flow data.
3. We explore the effect of cell-to-cell variability on creating conditions where such a mismatch is likely. We then propose a way to design experiments and analyze data so as to circumvent the mismatch between the two representations, thus increasing the robustness and reproducibility

of conclusions drawn.

We believe this manuscript is quite appropriate for *Cytometry A*, and will be particularly useful to its readership, and more broadly to the growing community concerned with quantitatively correct and robust analysis of cytometry data and experimental design. Our manuscript highlights, dissects and ameliorates a situation that is not uncommon in many experiments, although we do not address any specific biological question. Moreover, as the readership of *Cytometry A* is more mathematically inclined than the biological community in general, they are likely the precise audience to fully grasp the implications of the mismatch we present, as well as to be involved in the development of models, automatically gating procedures, or cluster high-dimensional data in ways susceptible to these effects.

We look forward to your reply,

Amir Erez