LINK TO ORIGINAL ARTICLE
LINK TO INITIAL CORRESPONDENCE

REPLY

Response to Orlova *et al.* "Science not art: statistically sound methods for identifying subsets in multi-dimensional flow and mass cytometry data sets"

Yvan Saeys, Sofie Van Gassen and Bart Lambrecht

"Science is built up of facts, like a house is with stones. But a collection of facts is no more science than a heap of stones is a house."

Henri Poincare

Unsupervised learning techniques such as clustering and dimensionality reduction have been widely used in many high-dimensional biological settings where they shed light on the internal problem structure. In their correspondence on our Review (Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nat. Rev. Immunol.* **16**, 449–462 2016)[1], Orlova *et al.* argue against the use of these techniques to identify cell populations in high-dimensional flow and mass cytometry data, based on arguments related to the curse of dimensionality (Science not art: statistically sound methods for identifying subsets in multi-dimensional flow and mass cytometry data sets. *Nat. Rev. Immunol.* http://dx.doi.org/10.1038/nri.2017.150-c1)[2].

The curse of dimensionality states that the number of samples needed to fit a model to an arbitrary degree of precision increases exponentially as the number of parameters that describe the data increases. This in itself might not be problematic for cytometry data, as the number of parameters are still relatively low (a few tens of markers) and sample sizes are large (up to millions of cells). Other high-dimensional biological settings, such as transcriptomics, measure many more parameters (for example, 10,000 transcripts) for fewer samples (typically a few tens or hundreds of samples), thus resulting in far more challenging situations from a statistical point of view. Nevertheless, even in these situations clustering techniques have proved useful to highlight grouping structures in such high-dimensional, low-sample settings.

Two other aspects of high-dimensional spaces have a more profound impact on unsupervised clustering techniques applied to cytometry data: the fact that high-dimensional spaces are inherently sparse (empty space phenomenon), and the fact that the notion of distance becomes increasingly meaningless as dimensionality increases. Since clustering methods crucially depend on the notion of distance and/or similarity, choosing the right variables to include in the analysis, and choosing the right distance metric is of utmost importance. In addition, as unsupervised learning is by definition more difficult than supervised learning, many clustering methods need to make additional assumptions on the data distribution and it is the scientist's responsibility to match these assumptions to the problem at hand. However, it does not mean that clustering techniques are fundamentally flawed for analysing high-dimensional data, as many of these techniques are based on sound mathematical formulations that enable one to analyse in an unbiased way grouping structure in high-dimensional data sets, as long as the methods' assumptions hold true.

In cytometry data analysis, unsupervised clustering techniques have been widely used for automated population identification, and recent benchmarks have shown that these techniques are indeed able to retrieve populations with great accuracy[3,4]. In addition, these automated techniques ensure that all cells in an experiment can be analysed, while also examining whether the markers that would not be checked in a sequential analysis for all individual populations show any further structure. Furthermore, they scale much better to larger and high-dimensional data sets, and facilitate the finding of novel and unexpected populations. This only makes their results stronger, without reducing their validity.

The main point demonstrated in the correspondence by Orlova *et. al.* is the difficulty in determining the correct number of clusters[2]. This indeed remains a hard problem, in both computer science and immunology research. However, most clustering techniques handle this gracefully by overclustering the data. This assures that all the main structures will be captured, even if they are further split up into smaller populations. These smaller populations might just capture some technical variance in the cell measurements, but they might also turn out to be unexpected populations of interest. As demonstrated in figure 1b in the correspondence[2], the four algorithms tested all correctly separate the two artificial populations created in this data set, even though the largest population is split further into additional populations. Further statistical analysis of results like this will still be able to indicate what is changing between, for example, different groups of patients, one of the main goals of most cytometry experiments.

Yvan Saeys, Sofie Van Gassen and Bart Lambrecht are at the VIB Inflammation Research Center, Technologiepark 927, Ghent B-9052, Belgium.

Correspondence to Y.S.
yvan.saeys@ugent.be

doi:10.1038/nri.2017.151
Published online 22 Dec 2017

1. Saeys, Y., Gassen, S. V. & Lambrecht, B. N. Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nat. Rev. Immunol.* **16**, 449–462 (2016).
2. Orlova, D. Y., Herzenberg, L. A. & Walther, G. Science not art: statistically sound methods for identifying subsets in multi-dimensional flow and mass cytometry data sets. *Nat. Rev. Immunol.* http://dx.doi.org/10.1038/nri.2017.150-c1 (2017).
3. Aghaeepour, N. *et al.* Critical assessment of automated flow cytometry data analysis techniques. *Nat. Methods* **10**, 445–445 (2013).
4. Weber, L. M. & Robinson, M. D. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry A* **89**, 1084–1096 (2016).

Competing interests statement
The authors declare no competing interests.