Science not art: statistically sound methods for identifying subsets in multi-dimensional flow and mass cytometry data sets

Darya Y. Orlova, Leonore A. Herzenberg and Guenther Walther

Automated approaches that cluster high-dimensional flow and mass cytometry data simultaneously in multiple dimensions, such as those discussed in Saeys *et al.* (Computational flow cytometry: helping to make sense of high-dimensional immunology data. *Nat. Rev. Immunol.* **16**, 449–462 2016)[1], are currently coming into routine use in biomedical settings. However, the simultaneous clustering approach underlying these methods is fundamentally flawed. This is due to what statisticians call the 'curse of dimensionality' (REF. 2), which is well known to compromise both the statistical validity and the computational performance of clustering methods that operate on multiple dimensions at once.

Although the curse of dimensionality is a well-known problem, the statistical component of this problem, which renders clustering outcomes invalid, has not been properly recognized in flow and mass cytometry. This crucial problem arises from the marked increase in statistical uncertainty that occurs as the number of dimensions for which data are being considered increases (even three dimensions can be problematical[3]).

That is, as the number of dimensions increases: one, data become increasingly sparsely distributed; two, definitions of density and distance between points become increasingly meaningless; and three, fitting a mathematical model to the data set becomes infeasible because the number of combinations of possible parameters to be considered increases dramatically as the number of dimensions increases above three or four. These problems compromise high-dimensional clustering algorithms that rely on estimation of density and/or distance, or on fitting of mathematical models. Here, we show directly how the curse of dimensionality leads to invalid conclusions by some commonly used clustering methods (FIG. 1, Rphenograph[4], X-shift[5] and flowMeans[6]).

t-distributed stochastic neighbour embedding (t-SNE)[7] has recently been introduced into high-dimensional flow cytometry analyses as a preprocessing step intended to reduce data dimensionality before clustering. However, when t-SNE is applied to high-dimensional data with intrinsically high dimensional structure (that is, when N dimensional data cannot be closely approximated by some combination of $n<<N$ dimensions), it becomes subject to the curse of dimensionality[7]. We used Maximum Likelihood Estimation of Intrinsic Dimension (MLE) proposed by Levina *et al.*[8] to estimate the intrinsic dimensionality of a typical flow cytometry data set. MLE revealed four intrinsic dimensions for a 12-parameter flow cytometry sample (10-colour + side and forward scatter) shown in FIG. 1d. However, even three dimensions can be problematical and the severity of the curse of dimensionality problem increases sharply thereafter.

t-SNE also does not preserve either distances or density very well. It only preserves nearest-neighbours, and only to some extent. This means that distance or density-based clustering algorithms are not usable with t-SNE maps (FIG. 1). Furthermore, these properties of t-SNE, in addition to the curse of dimensionality are the primary causes of the lack of reproducibility illustrated in FIG. 1c,d.

The curse of dimensionality thus clearly mitigates against the use of high-dimensional simultaneous clustering methods for flow and mass cytometry data analysis. In contrast, automation[9] of sequential analysis methods that have been used for years offers statistically robust clustering and readily usable tools for flow cytometry and other technologies.

Darya Y. Orlova and Leonore A. Herzenberg are at the Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA.

Guenther Walther is at the Department of Statistics, Stanford University, Stanford, California 94305, USA.

Correspondence to D.Y.O. and G.W.

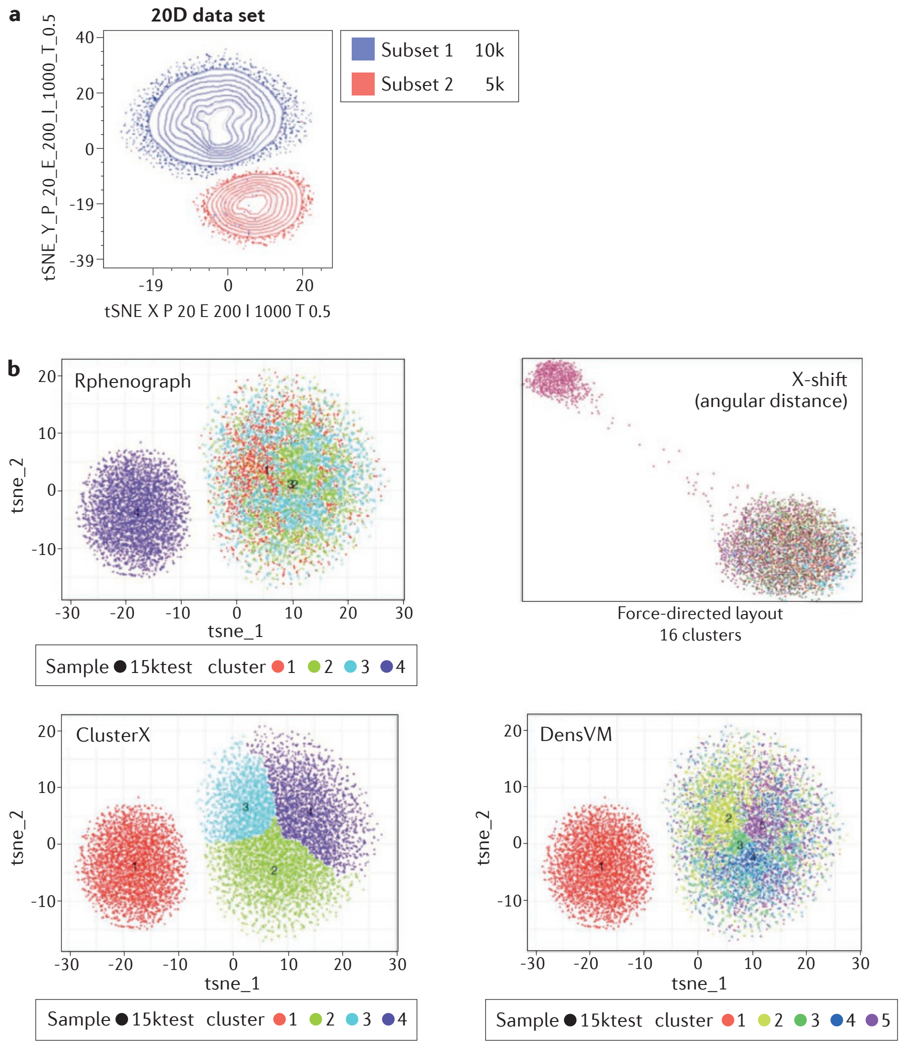
orlova@stanford.edu; dyorlova@gmail.com; gwalther@stanford.edu

doi:10.1038/nri.2017.150
Published online 22 Dec 2017

There is a reply to this Correspondence by Saeys, Y., Van Gassen, S. and Lambrecht, B. *Nat. Rev. Immunol.* http://dx.doi.org/10.1038/nri.2017.151-c1 (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. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* (Springer-Verlag, 2009).
3. Scott, D. W. *Multivariate Density Estimation — Theory, Practice and Visualization* (Wiley, 1992).
4. Chen, H. *et al.* Cytofkit: a bioconductor package for an integrated mass cytometry data analysis pipeline. *PLoS Comput. Biol.* **12**, e1005112 (2016).
5. Samusik, N. *et al.* Automated mapping of phenotype space with single-cell data. *Nat. Methods* **13**, 493–496 (2016).
6. Broad Institute. Flow cytometry gating and clustering. *GenePattern* http://software.broadinstitute.org/cancer/software/genepattern/flow-cytometry-gating-and-clustering (2017).
7. van der Maaten, L. Hinton, G. Visualizing data using t-SNE. *J. Machine Learn. Res.* **9**, 2579–2605 (2008).
8. Levina, E. & Bickel, P. in *Advances in Neural Information Processing Systems 17 (NIPS 2004)* (eds Saul, L. K., Weiss, Y. and Bottou, L.) (MIT Press, 2004).
9. Meehan, S. *et al.* AutoGate: automating analysis of flow cytometry data. *Immunol. Res.* **58**, 218–223 (2014).
10. Orlova, D. *et al.* Earth Mover's Distance (EMD): a true metric for comparing biomarker expression levels in cell populations. *PLoS ONE* **11**, e0151859 (2016).

Competing interests statement
The authors declare no competing interests.



c **Number of identified clusters**

Sample ID	Rphenograph	X-shift	ClusterX	DensVM
1	4	16	4	5
2	4	13	2	2
3	5	21	3	3
4	6	56	3	3
5	4	6	6	6

d **Number of identified clusters**

Sample		Rphenograph	X-shift	ClusterX	DensVM	flowMeans
BALBc #1	Total 10k cells	18	22	29	23	3
	First half 5k cells	16	10	24	13	5
	Second half 5k cells	16	10	24	10	3
BALBc #2	2a 7k cells	18	11	22	17	2
	2b 7k cells	17	14	23	11	2
	2c 7k cells	17	44	26	9	2
C57 #1	1a 7k cells	15	40	28	8	2
	1b 7k cells	19	25	24	17	3
	1c 7k cells	18	15	32	5	2

Figure 1 | **Commonly used high-dimensional clustering methods yield irreproducible results and may report populations that do not exist. a** | We simulated a mixture of two 20-dimensional (20D) Gaussian distributions with unit variance in each dimension and the following means: M1, [0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]; M2, [2.3 2.3 2.3 2 2 2 1 1 2.3 2.3 2 3 5 1 3 2.3 2 4 5 2]. One distribution (Subset 1) consists of 10,000 events; another distribution (Subset 2) consists of 5,000 events. The t-distributed stochastic neighbour embedding (t-SNE) map for this data set is shown in the figure. **b** | Algorithms that work directly on the high-dimensional data (Rphenograph[4] and X-shift[5]) and algorithms that are applied to the t-SNE embedded map (ClusterX[4] and DensVM[4]) were run on the 20D data from part **a**. The output was colour-coded and presented in t-SNE parameter space (Rphenograph, ClusterX and DensVM) and in a force-directed layout (for X-shift only). The results of these clustering methods show no connection to the actual population structure in the data. **c** | We repeated (five times) the simulations shown in part **a**. The table shows the number of clusters for each simulation that each of the tested clustering algorithms reported. **d** | The table shows the number of clusters identified by the five distinct clustering algorithms applied to the two halves of the same sample (even and odd rank numbers of a single flow cytometry run) and applied to technical replicates (separate flow cytometry runs) of the same sample. We used a 10-colour data set previously published (see figure 6b in REF. 10). Data were compensated, Logicle transformed and pre-gated for live singlets using AutoGate (www.cytoGenie.org). We used the default input parameters provided by each clustering algorithm but omitted the data transformation as the data were already Logicle transformed. Clustering results are available here: https://drive.google.com/open?id=0B1SkmBF14Q2lOVhuclhDWldOVEU and https://www.dropbox.com/sh/4xbl0k5fb5qpk5s/AAAVEefS3rTUbPu9uJqDpc9Ba?dl=0