Getting the Most from Your High-Dimensional Cytometry Data

Olsen, Lars R.; Pedersen, Christina B.; Leipold, Michael D.; Maecker, Holden T.

Published in:
Immunity

Link to article, DOI:
10.1016/j.immuni.2019.02.015

Publication date:
2019

Document Version
Peer reviewed version

Getting the most from your high-dimensional cytometry data

Olsen LR[1,2*], Pedersen CB[1,2*], Leipold MD[3,*], Maecker HT[3]

[1]Department of Bio and Health Informatics, Technical University of Denmark
[2]Center for Genomic Medicine, Copenhagen University Hospital
[3]Institute for Immunity, Transplantation, and Infection, Stanford University School of Medicine
[*]These authors contributed equally to this work

As mass cytometry (CyTOF®) and high-dimensional flow cytometry have become increasingly popular, many immunologists find themselves frequently encountering large single-cell data sets with little idea of how to fully analyze them. Here we propose some guidelines for immunologists looking to get the most from such high-dimensional cytometry data sets.

Main Text

The number of algorithms for dealing with high-dimensional single-cell data has multiplied dramatically, including some which are available with graphical user interfaces (eg, SPADE (Qiu et al., 2011), viSNE (Amir et al., 2013), and Citrus (Bruggner et al., 2014)). Unfortunately, with more tools come more questions. Which one to use? How to set the input parameters? What pre-processing is necessary? And, very importantly: Can I really trust the results?

Therefore, many immunologists turn to a local bioinformatics expert for help. Unfortunately, in most quarters, these experts were trained on genomics and transcriptomics data. They expect a flat file with samples in rows and expression data in columns. Or, they may start with raw reads, compile, normalize, etc., and eventually generate such a flat file in the process. But given a set of FCS files (the standard output of flow and mass cytometry), may leave them perplexed. *There is a need to train the next generation of bioinformaticians on high-dimensional cytometry data, and this training should start with immunologists.* With this Letter, we provide a kind of manifesto for immunologists seeking to get the most out of high-dimensional cytometry data, and to work most productively with computational biologists in this area.

1. Immunologists need to plan and execute their experiments so as to minimize unwanted variance. In other words, those of us supervising bench science could do a better job of designing robust experiments, starting with good panel design. Fortunately, there are many excellent high-dimensional flow cytometry panels published (e.g., search for “OMIP” in PubMed to find well-optimized panels published in Cytometry A). Mass cytometry also requires some attention to panel design, as there can be elemental impurities and oxidation products, causing spillover between mass channels (Takahashi et al., 2017). Since compensation matrices are

not usually applied to mass cytometry data (but see (Chevrier et al., 2018)), it is incumbent on the experimentalist to avoid significant spillover primarily through good panel design.

However, creating a good panel is only half the battle. "Batch effects", the general term given to unwanted variance between samples run at different times or on different instruments, are an enemy of virtually all assay technologies, and cytometry is no exception. Changes over time in reagents, processing and staining, instrument setup, etc., can all introduce unwanted variance (Maecker et al., 2012). If these variables are well-controlled, it's true that comparable results can be obtained, even across sites (Leipold et al., 2018), and normalization beads can reduce *some* sources of unwanted variance (Finck et al., 2013). But we still need to guard against batch effects and minimize the chances that they will confound our results. We should:

- Reduce batch effects by standardizing protocols and instrument setup, allocating a single batch of reagents to a study, and acquiring all data in as short a span of time as reasonably possible;
- Remove some of the remaining batch effects by use of normalization beads (Finck et al., 2013); and
- Protect against confounding results from the batch effects still present, by balancing our batches to include all outcome groups of interest in the same batch, or for longitudinal studies, all samples from the same patient in the same batch.

Lastly, the desired outcome of a study determines the eventual choice of analysis tools, sample size to achieve appropriate statistical power and so on. In order to be on top of these potential issues straight out of the gate, it is beneficial to include your favorite data scientist early in the planning of your project.

2. Immunologists need to educate themselves, and then their bioinformatics colleagues, about the structure and peculiarities of single-cell data. Stated simply, we could do a better job of understanding the important technical aspects of our flow and mass cytometry data, so that we don't unwittingly set traps for bioinformatician trying to navigate them. For example, anyone doing fluorescence flow cytometry should have carefully checked the compensation of all files using "NxN" dot plots to look for artifacts of under- or over-compensation. Time versus scatter plots should be examined to look for acquisition artifacts that may need to be excluded. Intensity normalization should be done where applicable. Pre-gating on live, intact, singlet events is a prerequisite for any unsupervised clustering/analysis algorithm. We can further screen our data files for outlier samples and/or outlier batches, either by visual examination of dot plots or by use of a dimension reduction algorithm like principal components analysis (PCA) that allows one to see overall differences between files and groups of files.

For those using mass cytometry, a recent review covers much of "what you need to know about your data" (Olsen et al., 2018). Furthermore, an online "data scientist's primer to mass cytometry data" has been established at http://cytof.biosurf.org/, collecting many useful tools and providing tutorial assistance for the steps going from raw to analyzed mass cytometry data. These types of resources should be read by bioinformaticians, of course. But the burden starts

with immunologists, who need to more fully understand their own data, in order to teach computational biologists what's important in the analysis of these data sets.

Should the newly enlightened immunologist "go it alone", and analyze his or her data with existing tools, without the aid of a bioinformatician? That depends on the complexity of the study, and the immunologist's analytical skills. Many bench scientists stop immediately when they see a command line, opting only for tools with a graphical user interface. Unfortunately, the universe of such tools is much smaller than those accessible to someone who can write and run R packages, for example. In the end, the idea is not to put bioinformaticians out of business (though they are currently in short enough supply that they may welcome a reduction in their business). Rather, the intent is to make the best use of the computational biologist's limited time, by presenting him or her with well-designed studies, with appropriate pre-processing of data, and a strong background knowledge of the data's structure and potential pitfalls. There may even be a partitioning of work, with some analyses performed by the immunologist, and some by the bioinformatician.

This gets us to that last and very important question, "Can you believe the results?" All other factors being equal, we would advise that if two or more analytical approaches reach the same or similar conclusions, the results are much more believable than if it can only be shown by one analytical method. Fortunately, there is now no shortage of analytical tools to choose from (Mair et al., 2016). Aimed at mortal immunologists, http://cytof.biosurf.org/ has been extended to also comprise "the bench cytometrist's primer to mass cytometry data analysis" - a tutorial style encyclopedia of layman's term descriptions of most common analysis tools. The goal is to facilitate efficient communication to bridge the gap between bench, biology and bioinformatics.

In whatever way the workflow is divided, the end result should be an efficient and productive relationship between immunologist and bioinformatician. Gone should be the days when data sets are tossed over a figurative wall to the domain of the computational biologist, who then toss "the findings" back some weeks later. Immunologists need to teach bioinformaticians the features of their data, and they also need to think more like computational biologists themselves, planning and processing their data sets with the aim of the best possible analysis in mind. Only then will we truly get the most from these wonderful and complex cytometry data sets.

References

Amir, E.D., Davis, K.L., Tadmor, M.D., Simonds, E.F., Levine, J.H., Bendall, S.C., Shenfeld, D.K., Krishnaswamy, S., Nolan, G.P., and Pe'er, D. (2013). viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nat. Biotechnol. *31*, 545–552.

Bruggner, R.V., Bodenmiller, B., Dill, D.L., Tibshirani, R.J., and Nolan, G.P. (2014). Automated identification of stratifying signatures in cellular subpopulations. Proc Natl Acad Sci USA *111*, E2770-7.

Chevrier, S., Crowell, H.L., Zanotelli, V.R.T., Engler, S., Robinson, M.D., and Bodenmiller, B. (2018). Compensation of signal spillover in suspension and imaging mass cytometry. Cell Syst. *1*.

Finck, R., Simonds, E.F., Jager, A., Krishnaswamy, S., Sachs, K., Fantl, W., Pe'er, D., Nolan, G.P., and Bendall, S.C. (2013). Normalization of mass cytometry data with bead standards. Cytometry A *83*, 483–494.

Leipold, M.D., Obermoser, G., Fenwick, C., Kleinstuber, K., Rashidi, N., McNevin, J.P., Nau, A.N., Wagar, L.E., Rozot, V., Davis, M.M., et al. (2018). Comparison of CyTOF assays across sites: Results of a six-center pilot study. J. Immunol. Methods *453*, 37–43.

Maecker, H.T., McCoy, J.P., and Nussenblatt, R. (2012). Standardizing immunophenotyping for the Human Immunology Project. Nat. Rev. Immunol. *12*, 191–200.

Mair, F., Hartmann, F.J., Mrdjen, D., Tosevski, V., Krieg, C., and Becher, B. (2016). The end of gating? An introduction to automated analysis of high dimensional cytometry data. Eur. J. Immunol. *46*, 34–43.

Olsen, L.R., Leipold, M.D., Pedersen, C.B., and Maecker, H.T. (2018). The anatomy of single cell mass cytometry data. Cytometry.

Qiu, P., Simonds, E.F., Bendall, S.C., Gibbs, K.D., Bruggner, R.V., Linderman, M.D., Sachs, K., Nolan, G.P., and Plevritis, S.K. (2011). Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nat. Biotechnol. *29*, 886–891.

Takahashi, C., Au-Yeung, A., Fuh, F., Ramirez-Montagut, T., Bolen, C., Mathews, W., and O'Gorman, W.E. (2017). Mass cytometry panel optimization through the designed distribution of signal interference. Cytometry A *91*, 39–47.