Batch to the Future: How normalisation tools help reduce batch effect in high parameter cytometry

Jamie S. McCarthy[1], Geoffrey Kraker[1], JF Tirados[1], Julia E. Prier[1], Chris Ciccolella[1].
1: OMIQ, Dotmatics Ltd.

Dotmatics OMIQ
www.omiq.ai

Introduction

Batch effects are an important consideration in the analysis of large longitudinal biological datasets, including high-dimensional flow and mass cytometry data. Experimental and instrument design can help to reduce these effects, but it is important to understand and account for their presence.

Unaccounted-for batch effects can lead to misinterpretation of your data or cloaking of potential key outcomes in the noise of any batch-to-batch differences.

Prior to analysis of a multi-batch experiment, each batch can be normalized to other batches to attempt to address any discrepancies. There are many computational methods employed in the flow cytometry field to achieve this and in this work we will be comparing the behaviour of two of the more elegant options, cytoNorm and cyCombine.

OMIQ is a modern cloud-based analysis platform built for interrogating cytometry and other data types. During analysis, a user can choose which algorithms to employ and while flexibility is often beneficial, the field is still working out which methods are best for a given situation. When undertaking the task of normalisation on OMIQ, users most often select either cytoNorm or cyCombine and as such, we wanted to communicate the scope and best use of these two tools. In this work, we test both algorithms by applying them as part of a typical high parameter workflow.

Methods

To test normalisation we used healthy control PBMCs from a Mass Cytometry (CYToF) sample in a time course experiment. There are 57 files across 18 batches, including a repeat donor across all batches. There were 35 parameters available for normalisation.

Within OMIQ data was Arc Sinh scaled, underwent the PeacoQC cleaning algorithm followed by manual clean up gating.

Different normalisation methods were then performed, and the resulting normalised channels could be assessed. This was done visually via dimension reduction embeddings and histogram overlays. In additional a statistical assessment of the Variance per channel and per gated population using GraphPad Prism.

Publicly available data used from Flow Repository (ID: FR-FCM-Z2YR).

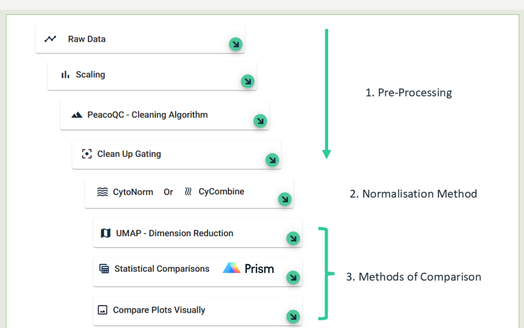
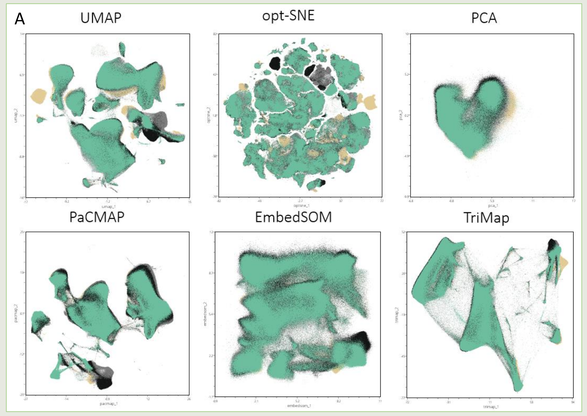


Fig 1. Data Analysis Methods visualised as a OMIQ workflow.

Results



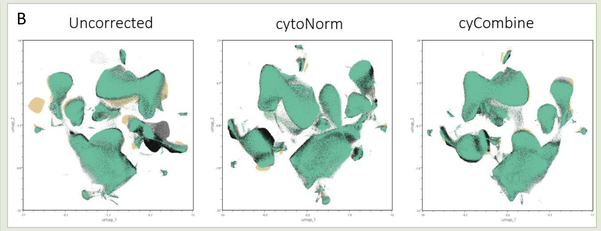


Fig 2. Dimension Reduction embeddings from uncorrected and normalised parameters.

Each batch was virtually concatenated and overlaid as a single colour. Differences in the coloured layers could indicate technical differences between batches. In A. different dimension reduction algorithms were compared, all on the uncorrected data. In B. UMAP was used to assess the effect of cytoNorm and cyCombine on reducing seen batch effects.

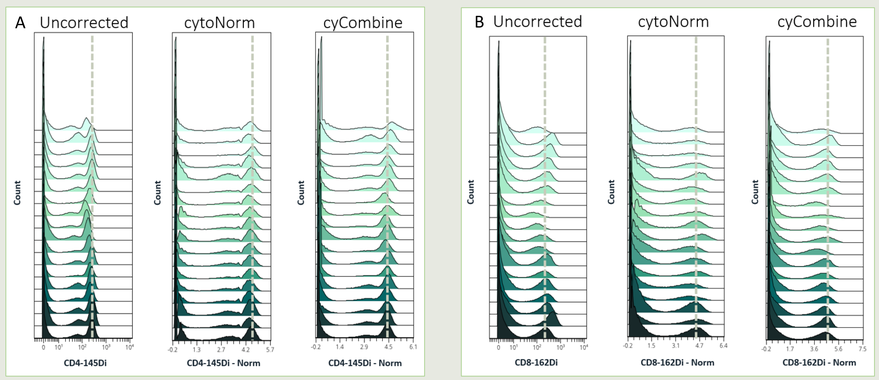


Fig 3. Histogram overlay of uncorrected and normalised parameters.

Each batch was virtually concatenated and plotted as a histogram overlay. This demonstrates the observed batch effect in uncorrected data. Two representative examples given for A. CD4 and B. CD8 expression.

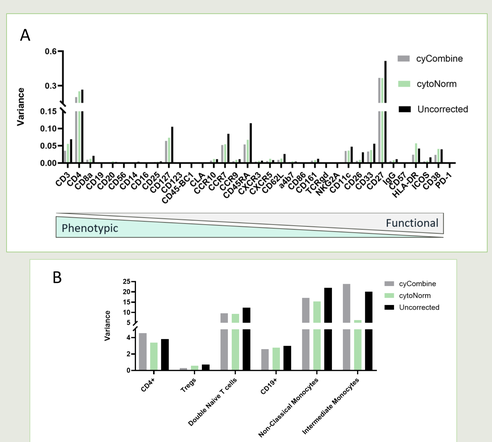


Fig4. Variance among uncorrected and normalised parameters and filters.

A. Median expression of each parameter was calculated for uncorrected and normalised (cytoNorm and cyCombine).

B. Typical populations were gated using both the uncorrected and normalised channels, the Percentage of Parent statistic was calculated.

In both instances, the variance between these per file values were calculated.

Total Events	CytoNorm	CyCombine
30 million	Completed	Completed
40 million	Completed	Failed
70 million	Completed	Failed
80 million	Failed	Failed

Fig5. Efficiency of algorithms based on input number of rows

Both cytoNorm and cyCombine were "titrated" to see when the algorithm would fail on the OMIQ platform. All other settings kept the same between runs, only subsampled number of rows changed.

Presented at CYTO 2025

Conclusion and Discussion

Dimension Reduction maps offer an overview of all your parameters. As such they can be useful to assess technical differences. The assumption is that if no-batch effects are present, then every batch should align/overlay via the embedding. If any islands exist differently between batches of the embeddings are "offset" then this indicates batch effects. We explored different dimension reduction algorithms, and batch effect is more prominently displayed in some embeddings when compared to others. The choice of algorithm when assessing batch effect is an important consideration. We can see the UMAP ran on uncorrected data shows an offset embedding, typical of batch effects, both the results of cytoNorm and cyCombine reduce this effect.

To drill into the specifics of a batch effect, histograms for every channel need to be assessed. Markers which do not align are typical of batch effects. Here we see the Histograms for CD4 and CD8 do not align across multiple batches in the uncorrected data. Both cytoNorm and cyCombine reduce this effect.

Intra-batch differences will contribute to a greater variance between samples. To assess this, we have taken the Median of every channel and then calculated the variance of these medians across every sample. Both cytoNorm and cyCombine reduce this variance. We can also look at how this would affect different populations, by looking at variance between different gated population, we can see the effect of normalisation is not uniform across all phenotypes.

As datasets in cytometry get bigger, efficiency of algorithms is an important consideration. cytoNorm and cyCombine were run with increasing number of events until the algorithm failed on the OMIQ platform. This was on a separate anonymous dataset (113 files, 39 parameters, 6 batches).

We set out to create a resource and open a discussion to help clarify common questions which arise during analysis of batched data. Whether cytoNorm, cyCombine or no normalisation at all is best for you will really depend on your data and the question at hand. We've summarised some of the clearer points in this decision process in our normalisation decision tree. As this space in cytometry develops it's important to consider how and when and whether to normalise, and how you can assess its success. At OMIQ we are always happy to have these conversations as you explore Data Science.

We would like to acknowledge the scope of this work was to consider algorithms currently available within OMIQ. There are other available methods, including an updated cytoNorm 2.0 which was not tested within this work.

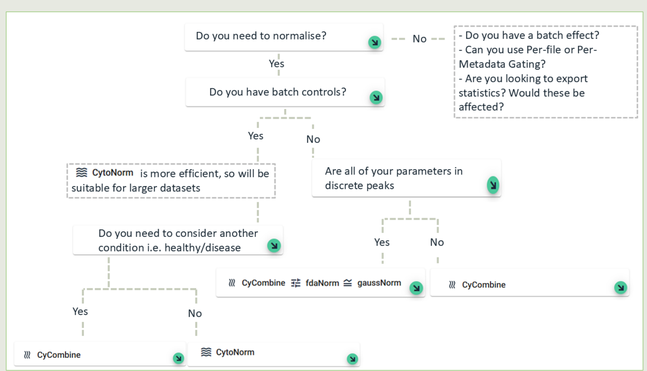


Fig 6. Decision Tree for Normalisation in OMIQ

References

1 -Van Gassen, S., Gaudilliere, B., Angst, M.S., et al. **CytoNorm: A Normalization Algorithm for Cytometry Data**. Cytometry, 97: 268-278 (2019). https://doi.org/10.1002/cyto.a.23904
2- Alvarez-Perez, M. P., et al. (2022). **cyCombine allows for robust integration of single-cell cytometry datasets within and across technologies.** Nature Communications, 13(1), 1698. doi: 10.1038/s41467-022-29383-5
3 - Schuyler RP, Jackson C, Garcia-Perez JE, Baxter RM, Ogolla S, Rochford R, Ghosh D, Rudra P, Hsieh EWY. **Minimizing Batch Effects in Mass Cytometry Data**. Front Immunol. 2019 Oct 15;10:2367. doi: 10.3389/fimmu.2019.02367. PMID: 31681275; PMCID: PMC6803429.