“…and then I ran it using the default settings”: Exploring the Impact of Changing your settings in FlowSOM Clustering

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

- FlowSOM is a widely used clustering tool. It is the most common choice for users in OMIQ
- There are many settings you can change and these effect the results of the algorithm
- RLEN is the number of training iterations the algorithm takes to build the original SOM. The algorithm needs enough iterations, but too many may be of limited benefit
- The number of nodes for the initial SOM needs to be enough to accommodate the final metacluster number whilst also enough to convey the detail of the desired clusters.

Methods

- FlowSOM was run as part of a typical high dimensional analysis in OMIQ
- RLEN or the XY dimensions of the training SOM were varied in otherwise identical runs.
- Task logs, UMAP embeddings and statistical outputs were used to assess the effect of the different settings
- The same 40 input parameters were used to produce 10 metaclusters in every task

Publicly available data used from Flow Repository (ID: FR-FCM-Z2QV*).*

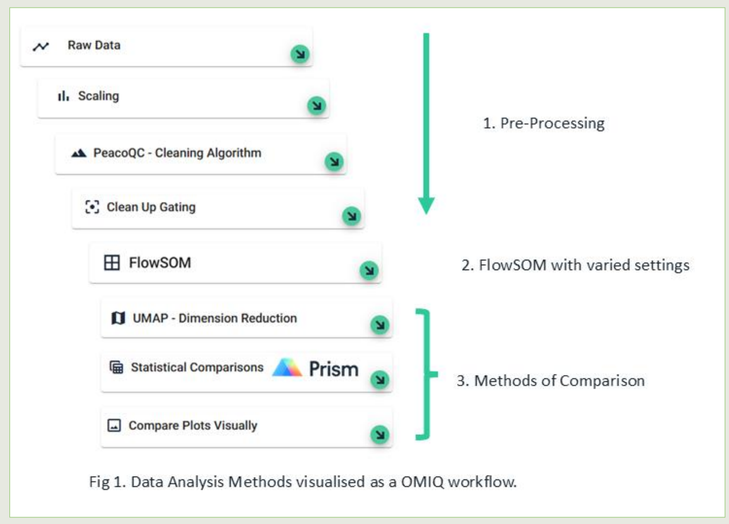

Fig 1. Data Analysis Methods visualised as a OMIQ workflow.

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

Results

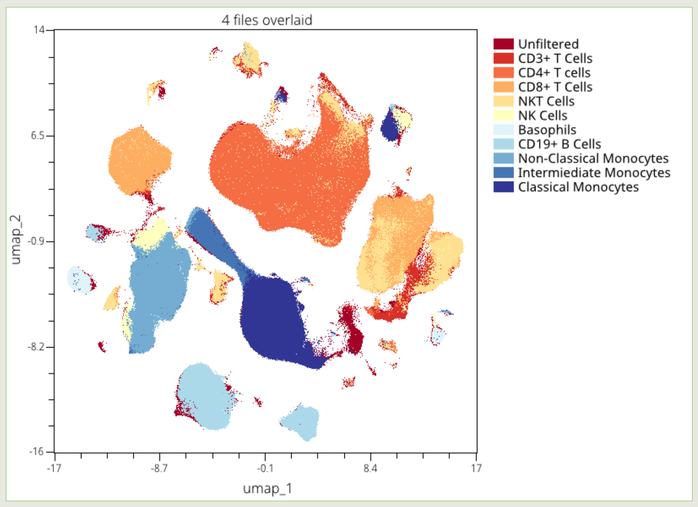

Fig 2. **Overlay of Manually Gated Immune Populations.**
Manually gated Immune populations are visualised on a UMAP Embedding. This will serve as a point of reference when viewing FlowSOM clusters.
FlowSOM clusters can be mapped back to equivalent locations on the UMAP.

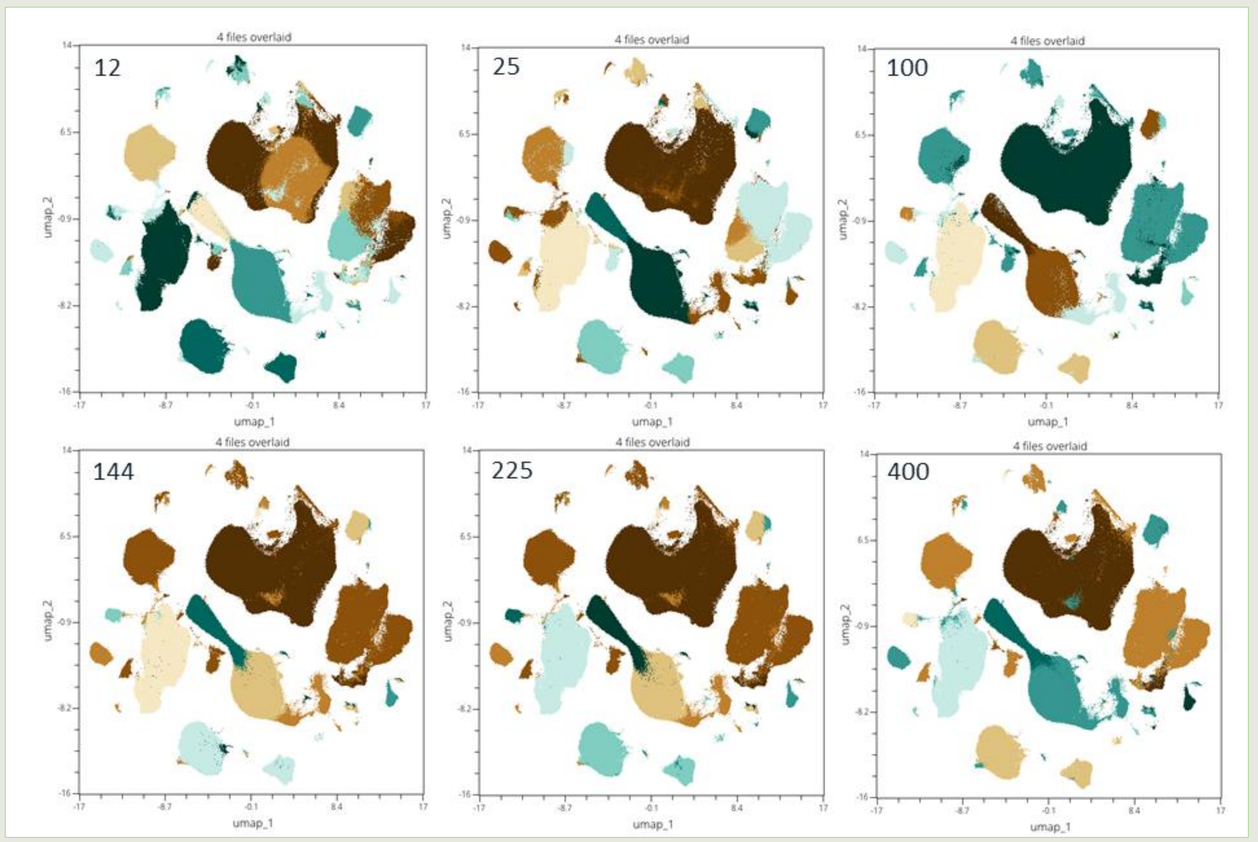

Fig 3. **FlowSOM Cluster Overlay for Different Initial SOM Sizes.**
FlowSOM was performed with varying initial SOM sizes. This was achieved by varying the X and Y dimensions of the SOM. (number of nodes is found in the top left of the embedding).
This FlowSOM was producing 10 Metaclusters. Therefore the minimum SOM size was 12 nodes (3 by 4), this was increased up to 400 nodes (20 by 20).
The clusters were overlaid on a UMAP embedding. They could then be visually inspected. There is variation in between the 12, 25 and 100 training nodes output, from 100 nodes upwards, there are less obvious visual differences.

Note: *Colours always refer to the same cluster designation MC1, MC2 etc. However, in each run different Cluster Numbers may be assigned to different populations.*

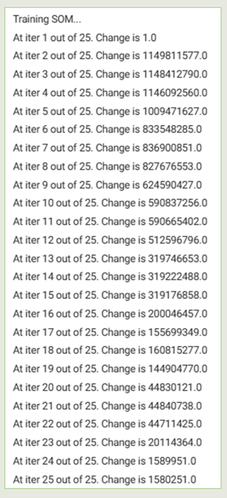

Fig4. Screenshot of a typical FlowSOM Task Log
Each training iteration results in a change statistic being generated. If *enough* iterations are completed this change will eventually stop decreasing.

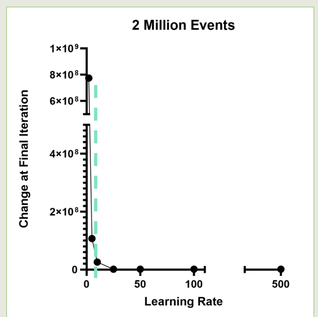

Fig5. **Measure of Final Change at Different RLEN iterations.**
The final change for different RLEN settings was noted. All other settings were kept the same. 10 RLEN is highlighted on the graph.
This was repeated for other subsample counts (not shown) with a similar patterns, although shape may shift to require more or less RLEN for the flattening of the graph.

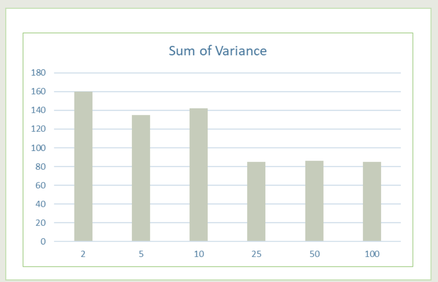

Fig6. **Total Variance of Clusters produced with different Training Iterations**
For each value of RLEN, the 10 clusters were taken, the variance of the parameters used to define the clusters were taken. This variance was added for every file and every cluster to give a total variance for the run.
Broadly, variance decreased until 25 training iterations, at which point it stabilised.

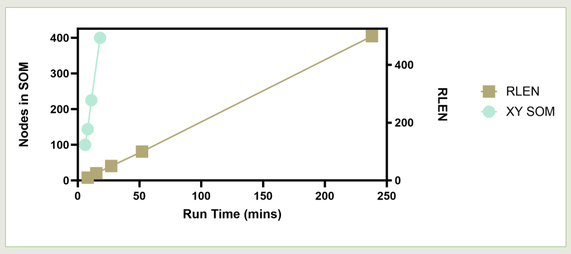

Fig7. **Effect on Different Settings on Total Time for FlowSOM runs**
The total time for different algorithm runs when different initial SOM size and RLEN iterations was recorded. Increasing complexity results in increased run time, with RLEN having a greater effect.

Conclusion and Discussion

- X and Y size of the initial training SOM must be at least larger than the final metacluster number. It seems there is also added benefit to having more than the minimum number of nodes. There appears to be limited return to continually increasing the node size
- RLEN has an effect on the final metaclusters produced. The change recorded by the algorithm will continue to decrease beyond 10 iterations in some datasets
- The exact RLEN required appears to alter between datasets and smaller datasets may require a higher RLEN
- There are many complex tools used to assess cluster quality or stability. This comparison focused on visual inspection and basic statistical calculations
- Other variables, such as Metacluster number, event count, parameter number and the data files themselves would also warrant consideration

References

- Van Gassen, S. *et al.* (2015) ‘Flowsom: Using Self-organizing maps for visualization and interpretation of cytometry data’, *Cytometry Part A*, 87(7), pp. 636–645. doi:10.1002/cyto.a.22625.
- Tao, W. *et al.* (2024) ‘Parameter optimization for stable clustering using FLOWSOM: A case study from cytof’, *Frontiers in Immunology*, 15. doi:10.3389/fimmu.2024.1414400.

Presented at CYTO CONNECT 2025