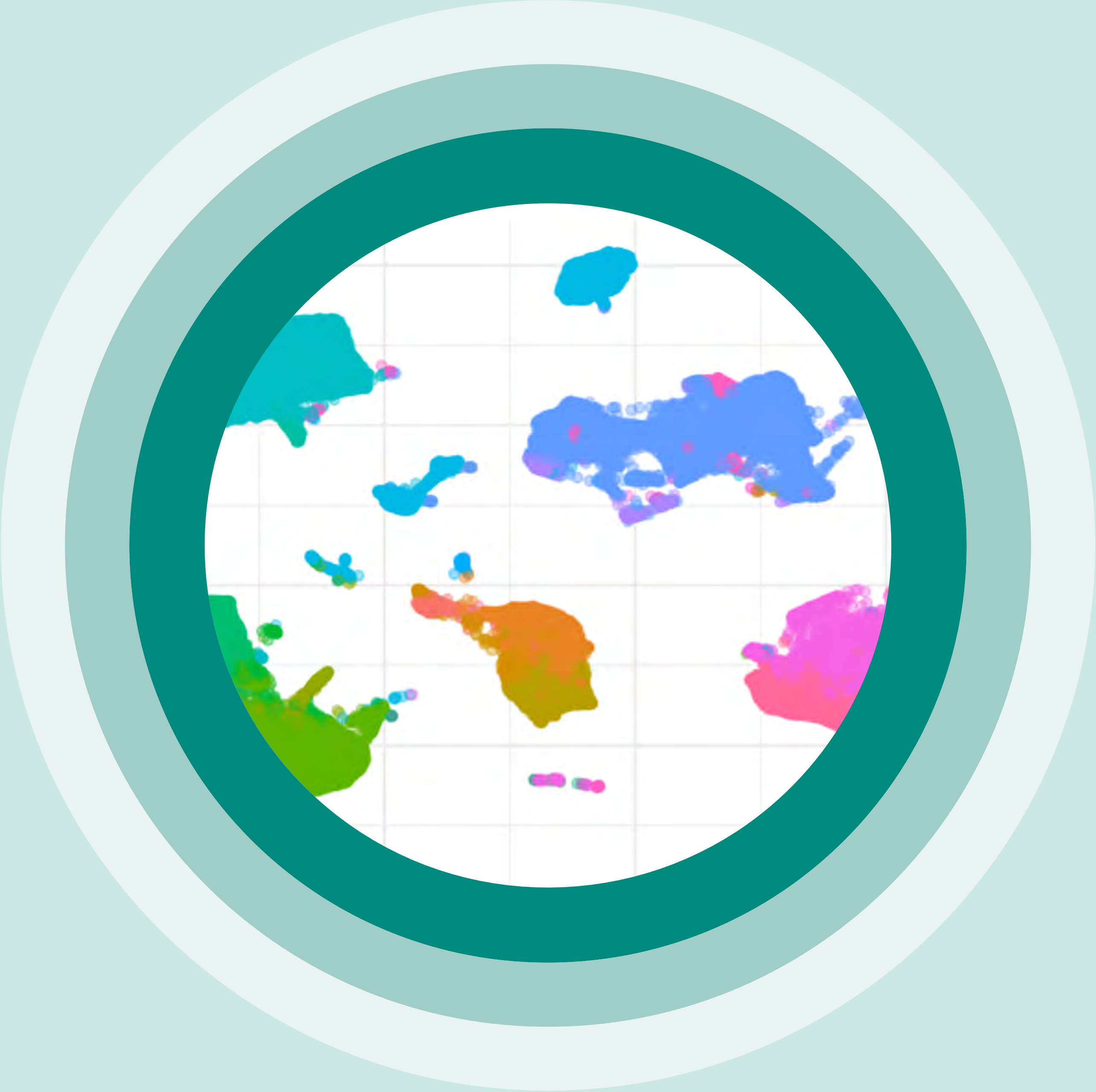
The future of flow cytometry data analysis: Navigating a new world with algorithms and machine learning

A CRO perspective



ICONplc.com/labs

Contents

Introduction

Leveraging algorithms and machine learning for unsupervised flow cytometry data analysis

The evolution of flow cytometry technology has significantly expanded the complexity of assays. While conventional instruments enabled incremental increases in fluorophore usage for over a decade, the recent adoption of spectral flow cytometry has accelerated this trend dramatically. Assays have shifted from 8-color configurations to high-parameter panels, with most clinical trial workflows now utilising 18- to 25-color instruments. These high-complexity panels deliver deeper biological insights and support a broader range of reportables, but they also introduce substantial challenges for data analysis. With spectral flow cytometry now integrated into ICON's global capabilities, the volume and dimensionality of data per sample continue to grow. To address this complexity, we are actively developing and implementing semi-automated analysis pipelines that incorporate advanced algorithms. These approaches enable unsupervised clustering, dimensionality reduction, and overall improvements in data quality, ensuring robust and scalable workflows for high-dimensional datasets.

Flow cytometry is a powerful technology for cellular analysis that combines laser-based detection with fluorophore-conjugated antibodies to identify cell subsets and characterise their properties. Since its inception, the number of detectable parameters has grown dramatically—from just two colors in early instruments to more than 40 today—driven by advances in cytometer hardware and the development of novel fluorophores. Spectral flow cytometry now enables panels with 45 or more colors (e.g., OMIP-102 and OMIP-109), significantly increasing the dimensionality of data generated. In practice, this translates into hundreds of reportable parameters per sample; for example, panels of 15–20 colors have already produced over 200 reportables, and spectral flow cytometry is expected to push these numbers even higher.

Traditionally, data analysis relies on manual gating, where cell subsets are identified by drawing gates on two-dimensional plots (dot plots) of marker expression - such as CD4 versus CD8. While templates and expert analysts help standardise this process, it remains time-consuming, subjective, and prone to variability. As panel complexity increases, the number of required plots and gates grows exponentially, making conventional gating increasingly impractical for high-parameter assays.

Fortunately, the evolution of flow cytometry has been accompanied by significant advances in computational algorithms and bioinformatics tools that support and partially automate data analysis. These algorithms not only facilitate clustering and dimensionality reduction but also improve data quality through automated QC and cleanup. Many of these tools are integrated into commercial software platforms widely used in pharmaceutical and CRO environments, such as FCS Express and OMIQ, which allow users to build analysis pipelines combining multiple algorithms. However, to fully leverage these capabilities- especially for custom workflows- programming expertise is often required, as most advanced flow cytometry analysis tools are developed in R.

A data analysis pipeline may include the following steps:

- Removal of detector-based margin events
- Data clean-up based on parameters such as dynamic range and flow rate
- Data normalisation to mitigate batch effects
- Doublet removal
- Clustering and dimensionality reduction
- Visualisation and export of processed data

The potential of such pipelines is substantial. They not only deliver significant time savings but can also enhance the overall quality, consistency, and reproducibility of the reported data. By reducing manual intervention and variability, these approaches support more robust and scalable workflows—critical for high-dimensional assays in regulated environments.

In this paper, we aim to provide an overview of current capabilities and share our experience and perspective on the use of algorithms and unsupervised data analysis tools. We present a case study that illustrates the potential of these approaches while addressing the challenges of implementing (semi-)automated pipelines within a regulated framework.

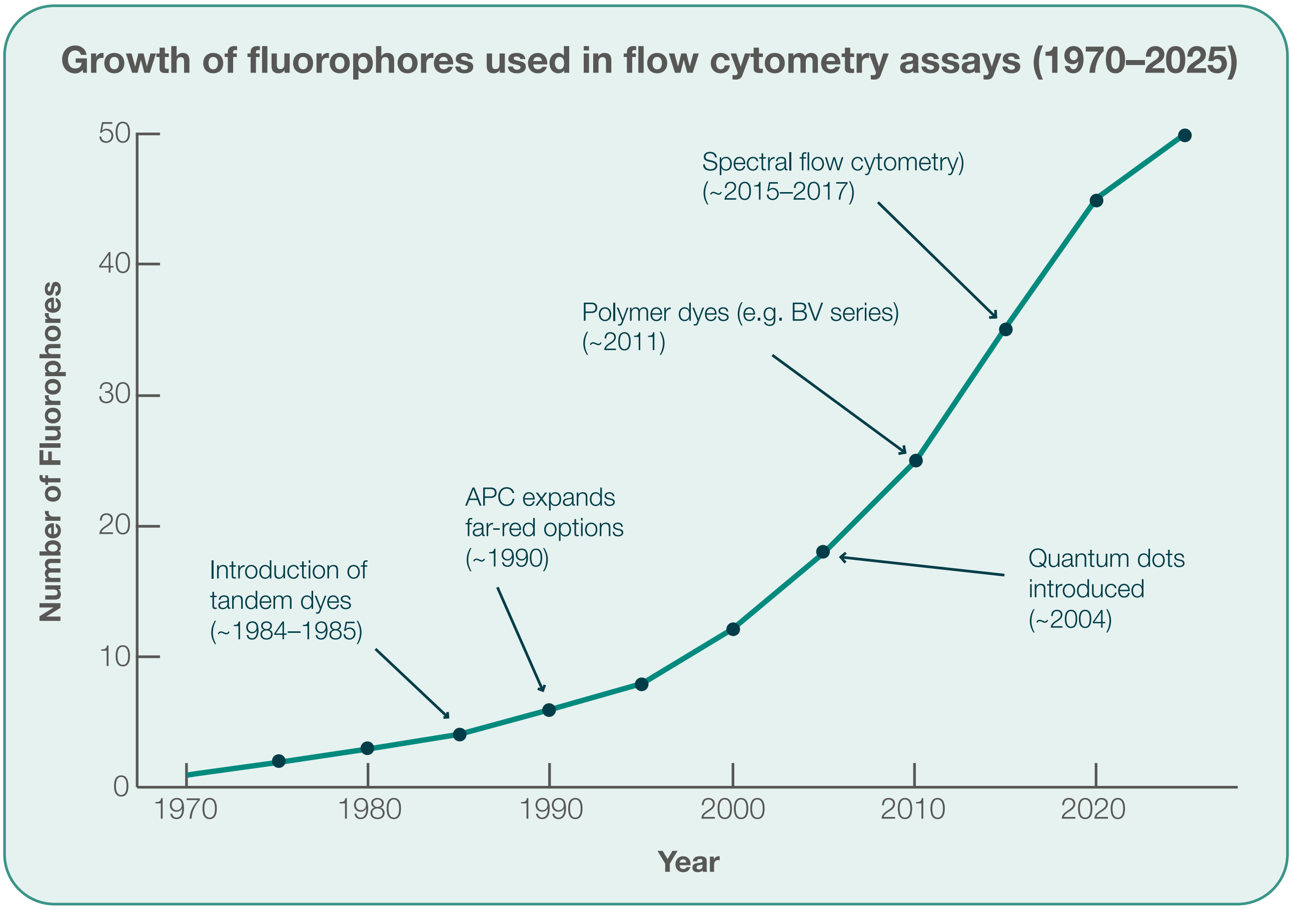


Figure 1. Evolution of fluorophore use in flow cytometry assays since the early use of flow cytometry

Implementing analysis pipelines for flow cytometry data analysis

As flow cytometry assays become more complex, there is growing interest in building structured analysis pipelines that go beyond manual gating. A wide range of algorithms have been developed specifically for flow cytometry data processing, and many are available within analysis software such as FCS Express. These algorithms can be grouped into categories like quality control, data normalisation and clean-up, clustering, and dimensionality reduction, each addressing a different challenge in managing high-dimensional datasets. In this section, we provide an overview of selected tools and approaches, highlighting how they could fit into future workflows.

Quality control and margin removal: Cleaning up before analysis

Before performing clustering or visualisation, it's essential to make sure the data is clean and reliable. Quality control tools help identify and remove anomalies that can creep in during acquisition—issues like clogs, bubbles, fluctuations in flow rate, or electronic noise. These problems aren't always obvious and can distort downstream analyses, which is why computational approaches are so valuable. Popular tools include FlowAI, FlowCut, and PeacoQC, each using slightly different strategies but sharing the same goal: leveraging raw data characteristics to flag and remove problematic regions in the data. Most of these algorithms assume that a well-acquired dataset should have stable fluorescence levels and flow rates throughout the acquisition. By dividing data into segments (or "buckets") and comparing patterns across them, they can detect irregularities and either remove or flag outlier events. A few tools are summarised below:

- **PeacoQC** (Peak Extraction And Cleaning Oriented Quality Control) was introduced by Emmaneel et al. in 2022 and works on transformed, compensated/unmixed data. It identifies density peaks in marker expression and filters out peaks with aberrant values, making it particularly useful for complex panels.

- **FlowAI**, published by Monaco et al. in 2016, evaluates three properties in the data: flow rate, signal acquisition, and dynamic range. It removes regions with unstable flow or signal fluctuations and trims events outside the dynamic range, ensuring only high-quality data remains.

- **FlowCut**, detailed in Meskas et al., takes a slightly different approach by segmenting events along the time axis and applying statistical tests to detect abrupt shifts or irregular fluorescence patterns. This makes it effective for cleaning datasets affected by clogs or instrument instability.

As panels grow and datasets scale, automated QC becomes indispensable—not just for saving time but for ensuring reproducibility and confidence in the results. By starting with clean data, we set the stage for accurate clustering, dimensionality reduction, and all downstream analyses.

Data normalisation: Keeping variability in check

After clean-up and quality control, the next step in many analysis pipelines is data normalisation. The goal here is to: remove non-biological variability so that differences in the data truly reflect biology, not technical noise. This is especially important in clinical studies where samples may be processed across different runs, instruments, or laboratories. Without normalisation, these technical differences can mask real biological signals or negatively impact clustering or gating.

The challenge and requirement is to apply normalisation in a way that it corrects technical variance while preserving biological differences. Several tools have been developed for this purpose, including gaussNorm, CytoNorm, and cyCombine, each with its own strengths and limitations. For a high-level comparison of these algorithms, see Table 1 below. In the case study presented in this paper, we applied CytoNorm to a 25-color phenotyping assay focused on reporting subsets, and observed that the algorithm very effectively aligns marker intensities between batches, thereby supporting clustering reproducibility.

As panels grow and studies scale across multiple sites, normalisation becomes more than a technical detail—it's a critical step for reproducibility and confidence in the data. By standardising this process, we can ensure that downstream analyses like clustering and dimensionality reduction start from a level playing field, making insights more reliable and easier to interpret.

Table 1

Tool	Description	Control samples required	Cell-type aware	Cross-batch integration
gaussNorm	Per-channel landmark-based normalisation aligning signal distributions; fast and simple for flow cytometry	Optional – predefined landmarks or automatic peak detection	No – treats all cells uniformly	Yes – aligns channel distributions across batches (assumes comparable biology)
CytoNorm	Learns normalisation per cell subset using FlowSOM clusters and control samples; well suited for clinical studies	Yes – requires control samples in each batch	Yes – cluster-aware (FlowSOM-based, not biologically supervised)	Yes – model-based batch correction using learned cluster-specific transformations
cyCombine	Empirical Bayes–based batch correction and integration across batches and technologies, without shared controls	No – works without technical replicates	Yes – optionally cell-type aware via clustering or metadata	Yes – harmonises feature space across batches and modalities

Doublet removal: An unsupervised approach

Doublets - events where two or more cells pass through the interrogation point simultaneously - can introduce artifacts that distort marker expression and compromise clustering accuracy. While doublet exclusion is straight-forward in conventional gating, automated solutions are available, providing ease of use, and consistency.

Algorithmic approaches detect doublets based on signal patterns and statistical anomalies rather than predefined gates. For example, the RemoveDoublets() function in the PeacoQC package applies predefined Median Associated Distances (MAD) to the SSC-parameters to exclude doublets effectively. This helps keep the dataset clean and ensures that subsequent analysis focuses on true single-cell events.

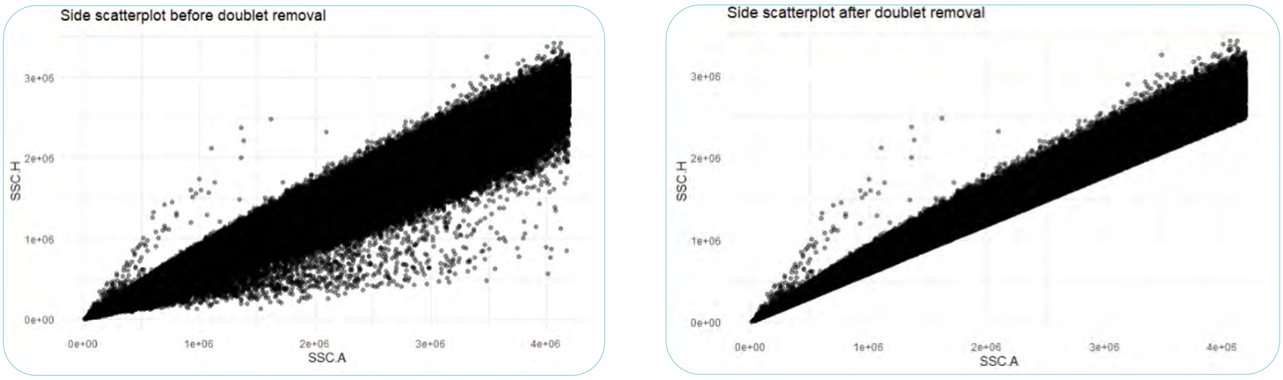


Figure 2. Example of doublet removal using PeacoQC RemoveDoublets(). The left image shows a sample pre-removal, and the right image shows the same sample after the automated doublet removal algorithm.

Clustering: Moving beyond manual gating

Clustering is all about finding patterns in the data without telling the algorithm what to look for—a big shift from traditional manual gating. Instead of drawing gates by hand in the analysis software, clustering automatically groups cells based on similarities across all markers, saving time and reducing bias. For example, FlowSOM uses self-organising maps to handle large, complex datasets efficiently, while K-means offers a simple, fast way to partition data into clusters. Another tool, Phenograph, uses graph-based approaches to uncover rare populations and subtle differences that manual gating might miss. Together, these algorithms make it possible to explore cellular diversity at a scale and depth that was previously out of reach. In short, clustering is a critical component of automated workflows - turning complexity into clarity and freeing scientists to focus on interpretation rather than working through the data manually. For a quick comparison of these algorithms, see Table 2 below.

As datasets grow larger and panels become more complex, clustering is no longer just an efficiency tool—it's a necessity for reproducibility and scalability. By reducing human bias and standardising the analysis, it helps CROs and research teams deliver consistent, high-quality results across high-complexity flow cytometry studies.

Table 2

Algorithm	Core approach	Key parameters	Strengths	Limitations	Runtime / scalability
FlowSOM	Self-organising map to project data onto a grid, followed by meta-clustering	Grid size, metacluster number	Very fast on large datasets; captures hierarchy; good visualisation	Sensitive to grid/ meta-cluster choice; may miss rare subsets	Fast, near-linear; scales to millions of cells
K-means	Iteratively partitions data into k clusters by minimising within-cluster variance	k, initialisation	Simple, very fast; widely available	Assumes spherical clusters; poor on non-convex; sensitive to k	Very fast; handles large n easily
Phenograph	Builds k-NN graph and detects communities using Louvain/Leiden	k, metric, resolution	Detects complex and rare populations; shape agnostic	Computationally heavy; memory intensive; sensitive to k	Moderate-heavy; practical up to ~1–2M cells; may need subsampling

Dimensionality reduction: Making complexity visible

Building on clustering, dimensionality reduction takes things a step further by turning complex, high dimensional data into something we can actually see and interpret. Dimensionality reduction algorithms create intuitive two-dimensional maps that reveal patterns and relationships at a glance. t-SNE has been a favorite for years because it does a great job showing local structure—helping us spot subtle differences between cell populations. UMAP builds on that idea but adds speed and scalability, while also preserving more of the global picture, which makes it ideal for larger datasets. Together, these tools transform raw numbers into visual maps of cellular diversity, making interpretation of the data much easier. In modern workflows, dimensionality reduction isn't just a nice visualisation trick—it's a critical bridge between complex data and meaningful insights. For a side-by-side look at t-SNE, UMAP, and PCA, see Table 3 below.

As panels expand and algorithms become more integrated, these visual maps are increasingly used not only for exploration but also for communication—helping teams and stakeholders quickly grasp complex findings. In many ways, dimensionality reduction turns data into a narrative, making it easier to share discoveries and drive decisions.

Table 3

Algorithm	Core approach	Strengths	Limitations	Runtime / scalability
t-SNE	Nonlinear embedding preserving local neighborhoods	Excellent local structure; widely used for visualisation	Poor global structure; slow on large datasets; non-deterministic	Slow for >100k cells; often minutes to hours
UMAP	Manifold approximation + graph layout	Preserves local & global structure; faster than t-SNE; scalable	Sensitive to parameters; can distort distances; stochastic results	Fast, handles millions; near-linear scaling
PCA	Linear projection maximising variance	Very fast; interpretable; good for initial reduction	Misses nonlinear patterns; limited for visualisation	Very fast, scales easily to millions of events

The CRO perspective

Although clinical trials do not typically include the most complex flow cytometry panels, we do experience a clear trend toward higher complexity in the requested analyses over recent years. This shift means that manual gating, while historically the gold standard, is becoming increasingly impractical as marker counts and event reportable numbers grow. Unsupervised clustering and dimensionality reduction are no longer just “nice to have”—they’re becoming essential for efficiency, consistency, and while providing deeper insights. For a CRO, adopting these approaches isn’t just about picking an algorithm; it’s about finding the right balance between accuracy, speed, and practicality. We need to define optimal settings for these tools, ensure they run within reasonable computation times, and have the digital infrastructure to support them. Large datasets can push local hardware to its limits, so scalable solutions like cloud-based platforms are increasingly attractive for timely delivery. At the same time, assay validation remains central, and data analysis—whether manual or automated—must meet regulatory expectations and fit-for-purpose principles. The challenge is deciding how to validate the unsupervised workflows, where algorithm choices and parameter sensitivity can influence results. Standardising processes for algorithm selection, tuning, and quality control will help reduce variability and build confidence. Transparency in computational method application and clear documentation of validation steps will be key for regulatory acceptance. Ultimately, moving toward automation and advanced analytics isn’t just a technical upgrade - it’s a strategic step that positions CROs to deliver high-quality, reproducible insights in an era of increasingly complex flow cytometry data. And while the journey requires investment and planning, the payoff is clear: faster, smarter, and more reliable data analysis for clinical trials.

Case study: Developing a pipeline for unsupervised data processing of a 25-color phenotyping panel

Challenge

ICON has validated a 25-color Immunoprofiling Assay (Cytek Biosciences) for off-the-shelf use in whole blood and isolated PBMCs. As part of the implementation and validation, we aimed to develop an analysis pipeline that is optimised to perform unsupervised processing of this high-parameter phenotyping panel.

Solution

A pipeline was built in R, leveraging well-established algorithms including PeacoQC, FlowSOM, and UMAP. These algorithms are available in FCS Express as well, and therefore most pipeline (components) are compatible with our current analysis software. The pipeline consists of the following steps:

- Margin event removal
- Spectral unmixing (AutoSpectral)
- Doublet removal (RemoveDoublets, PeacoQC)
- Quality control and clean-up (PeacoQC)
- Clustering (FlowSOM)
- Cluster identification algorithm (R)
- Visualisation (UMAP)
- Reporting

Each step was optimised for the 25-color panel, while maintaining flexibility for the pipeline to be adapted to other panels by implementing minor modifications. Optimisation of each algorithm turned out to be critical, and requires full understanding of the parameters that are used by each algorithm. Small changes/adaptations can have strong impact on the results, for example in the QC and clean-up steps, which then also impacts the outcome of the clustering. In addition, small unmixing errors that might be tolerated in manual gating can significantly impact the clustering results because closely related fluorophores may introduce noise across multiple markers. Another aspect taken into account was algorithm runtime and scalability. As summarised in this paper, multiple algortihms may be available for specific steps, and each algorithm has pro's and con's. The overall runtime of an analysis depends on the sum of the individual algorithm runtimes, and the size of the dataset. For large datasets as in high-parameter flow – aiming to provide insight into small subsets – the choice of algorithms is strongly impacting the overall analysis runtime. In a follow-up publication, we will elaborate further regarding the algorithm selections, and in depth parameter settings.

Post QC, clean-up and clustering, we implemented an R script that serves as cluster identification algorithm, and assigns individual clusters to subsets. The pipeline currently identifies the immune cell populations listed below:

- T cells: T-helper, T-cytotoxic, T-double positive/negative, T-regulatory, T-gamma-delta, NK-T
- B cells: Naïve, Memory and Marginal zone, Plasmablasts
- NK cells: Early, Mature, and Terminal NK cells
- Monocytes: Classical, Intermediate, Non-classical
- Basophils

While small subsets such as dendritic cells are clustered, the unsupervised cluster identification for these populations is not yet fully accurate. For these types of subsets, a manual gating step remains necessary, but we continue to work on improving the R script to also work with very small subsets in the near future as well.

The images below provide some insight into the functionality of the pipeline. Figure 3 represents T-cell subpopulations, as identified by colors, in a CD4-CD8-plot. Each subset is defined by specific rules on marker expression, and may contain multiple clusters. It is interesting to observe that the subsets in some occasions slighty overlap or mix, thereby highlighting that data is handled differently than manually gated, where we are tied to the hard borders as defined by our gates. This highlights the power of using clustering rather than gating, as gating is performed in 2D-plots only, whereas clustering evaluates all data dimensions in parallel.

Figure 4 is an example UMAP aiming to visualize all clustered subsets in a 2D overview. We chose to use UMAP for visualisation, because it can use non-linear data and handles large datasets faster than t-SNE. Since we acquire at least 500.000 white blood cells for this assay, UMAP provided significant time saving when running the full pipeline.

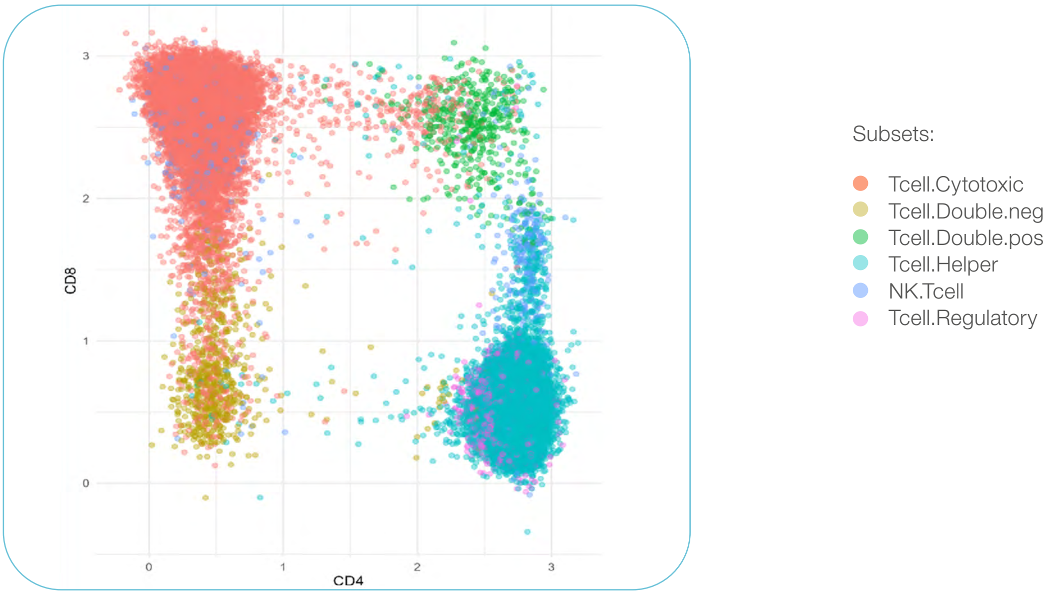


Figure 3. Representative example of unsupervised clustering of T-cell subsets. The dotplot contains all CD3+ T-cell events (as identified by FlowSOM), plotted for CD4 and CD8 expression. The T-cell populations as defined by the clustering and subsequent subset-identification-algorithm are represented by the individual colors.

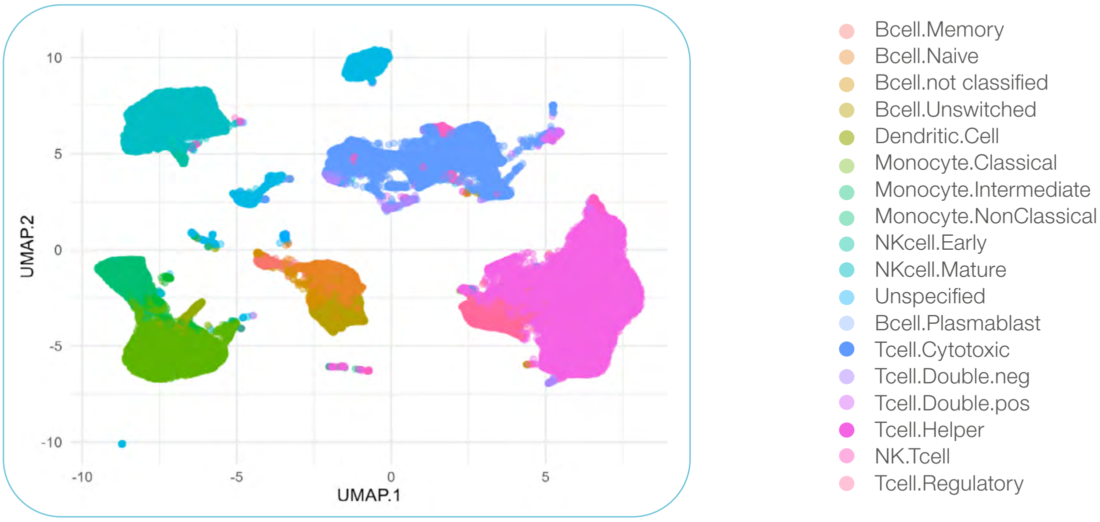


Figure 4. Representative example of a UMAP visualization from the 25-color analysis pipeline. The visualization provides a 2D-representation of the clustered subsets from a PBMC sample. Subsets as defined by the identification-algorithm are represented by individual colors, presented on the right-hand side.

Outcome and next steps

We successfully developed a pipeline that performs unsupervised data clean-up, spectral unmixing, and clustering for high-parameter flow cytometry panels. The pipeline is currently undergoing validation, comparing its outputs to manually gated results.

Future development will focus on refining subset annotations for rare populations such as dendritic cells and expanding the pipeline's capabilities for deeper phenotyping. A follow-up paper will detail the validation results and outline planned improvements and updates.

Conclusion

This paper provides an overview of current possibilities for applying unsupervised algorithms and automated pipelines in high-parameter flow cytometry analysis. We outline key steps such as data clean-up, spectral unmixing, doublet removal, clustering, and visualization, and discuss how these approaches can improve consistency and scalability compared to conventional gating. Our case study illustrates how such a pipeline can be applied for a 25-color phenotyping panel, highlighting both the potential for streamlined workflows and the importance of optimisation for reliable results. From a CRO perspective, these approaches offer opportunities to reduce variability, accelerate turnaround times, and deliver high-quality data for clinical trials. However, these pipelines must be appropriately qualified and validated, and careful consideration is required to define fit-for-purpose validation strategies. Looking ahead, further development will focus on refining rare subset identification, enhancing batch harmonisation, and expanding automation to meet regulatory expectations. Together, these developments mark an exciting advancement in clinical trial flow cytometry, positioning automated and unsupervised analysis pipelines to provide robust, reproducible data that delivers faster, reliable data, in support of clinical trials.

Key takeaways

- Unsupervised algorithms and automated pipelines offer scalable, consistent solutions for high parameter flow cytometry, reducing manual variability
- Our case study highlights how such a pipeline can streamline analysis of a 25-color panel while emphasizing the need for optimization and validation
- Future developments will focus on rare subset annotation, improved batch harmonization, and expanded automation to meet clinical trial and regulatory requirements
- Validation is essential to ensure automated pipelines are reliable and reproducible, but strategies are still evolving and consensus approach has yet to be established

Many sponsors choose to outsource customised assays and need a partner with the expertise to address their trial requirements. ICON has the scientific expertise to implement a broad range of flow cytometric assays in clinical trials. We cultivate a partnership with sponsors to provide scientific expertise, full-service assay development, and validation, followed by high- quality sample analysis to drive successful clinical trials forward.

For more information or to discuss your project requirements, please visit ICONplc.com/labs or email: globalflowcytometryrequests@iconplc.com.

Contributions to this article were made by members of ICON's global Flow Cytometry team:

Henko Tadema
Giel Bakker
Ymkje Brouwer
Than-Long Nguyen
Ron Suk

References

1. OMIP-102: 50-color phenotyping of the human immune system with in-depth assessment of T cells and dendritic cells. Konecny et al. Cytometry A. 2024 Jun;105(6):430-436. doi: 10.1002/cyto.a.24841. Epub 2024 Apr 18

2. OMIP-109: 45-color full spectrum flow cytometry panel for deep immunophenotyping of the major lineages present in human peripheral blood mononuclear cells with emphasis on the T cell memory compartment. Park et al. Cytometry A. 2024 Nov;105(11):807-815. doi: 10.1002/cyto.a.24900. Epub 2024 Oct 28

3. PeacoQC: Peak-based selection of high quality cytometry data. Emmaneel et al. Cytometry A. 2022 Apr;101(4):325-338. doi: 10.1002/cyto.a.24501. Epub 2021 Oct 3

4. FlowCore: data structures package for flow cytometry data. Bioconductor Project. Le Meur, N., Hahne, F., Ellis, B., & Haaland, P. (2007)

5. flowAI: automatic and interactive anomaly discerning tools for flow cytometry data. Monaco et al. Bioinformatics, Volume 32, Issue 16, August 2016, Pages 2473–2480, https://doi.org/10.1093/bioinformatics/btw191

6. flowCut: An R package for automated removal of outlier events and flagging of files based on time versus fluorescence analysis. Meskas et al. Cytometry A. 2023 Jan;103(1):71-81. doi: 10.1002/cyto.a.24670. Epub 2022 Jul 23

7. CytoNorm 2.0: A flexible normalization framework for cytometry data without requiring dedicated controls, Quintelier et al. Cytometry A. 2025 Feb;107(2):69-87. doi: 10.1002/cyto.a.24910. Epub 2025 Jan 28

8. cyCombine allows for robust integration of single-cell cytometry datasets within and across technologies. Pedersen et al. Nat Commun. 2022 Mar 31;13(1):1698. doi: 10.1038/s41467-022-29383-5

9. Per-channel basis normalization methods for flow cytometry data. Hahne et al. Cytometry A. 2010 Feb;77(2):121–131. doi: 10.1002/cyto.a.20823

10. FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. Van Gassen et al. Cytometry A. 2015 Jul;87(7):636-45. doi: 10.1002/cyto.a.22625. Epub 2015 Jan 8

11. flowPeaks: a fast unsupervised clustering for flow cytometry data via K-means and density peak finding. Ge, Sealfon. Bioinformatics. 2012 Aug 1;28(15):2052-8. doi: 10.1093/bioinformatics/bts300. Epub 2012 May 17

12. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. Weber, Robinson. Cytometry A. 2016 Dec;89(12):1084-1096. doi: 10.1002/cyto.a.23030

13. AutoSpectral improves spectral flow cytometry accuracy through optimised spectral unmixing and autofluorescence-matching at the cellular level. Burton et al. https://doi.org/10.1101/2025.10.27.684855

ICON plc Corporate Headquarters

South County Business Park
Leopardstown, Dublin 18
Ireland
T: (IRL) +353 1 291 2000
T: (US) +1 215 616 3000
F: +353 1 247 6260

ICONplc.com/contact

About ICON

ICON plc is a world-leading healthcare intelligence and clinical research organisation. From molecule to medicine, we advance clinical research providing outsourced services to pharmaceutical, biotechnology, medical device and government and public health organisations. We develop new innovations, drive emerging therapies forward and improve patient lives. With headquarters in Dublin, Ireland, ICON employed approximately 41,900 employees in 106 locations in 55 countries as at December 31, 2024. For further information about ICON, visit: www.iconplc.com.