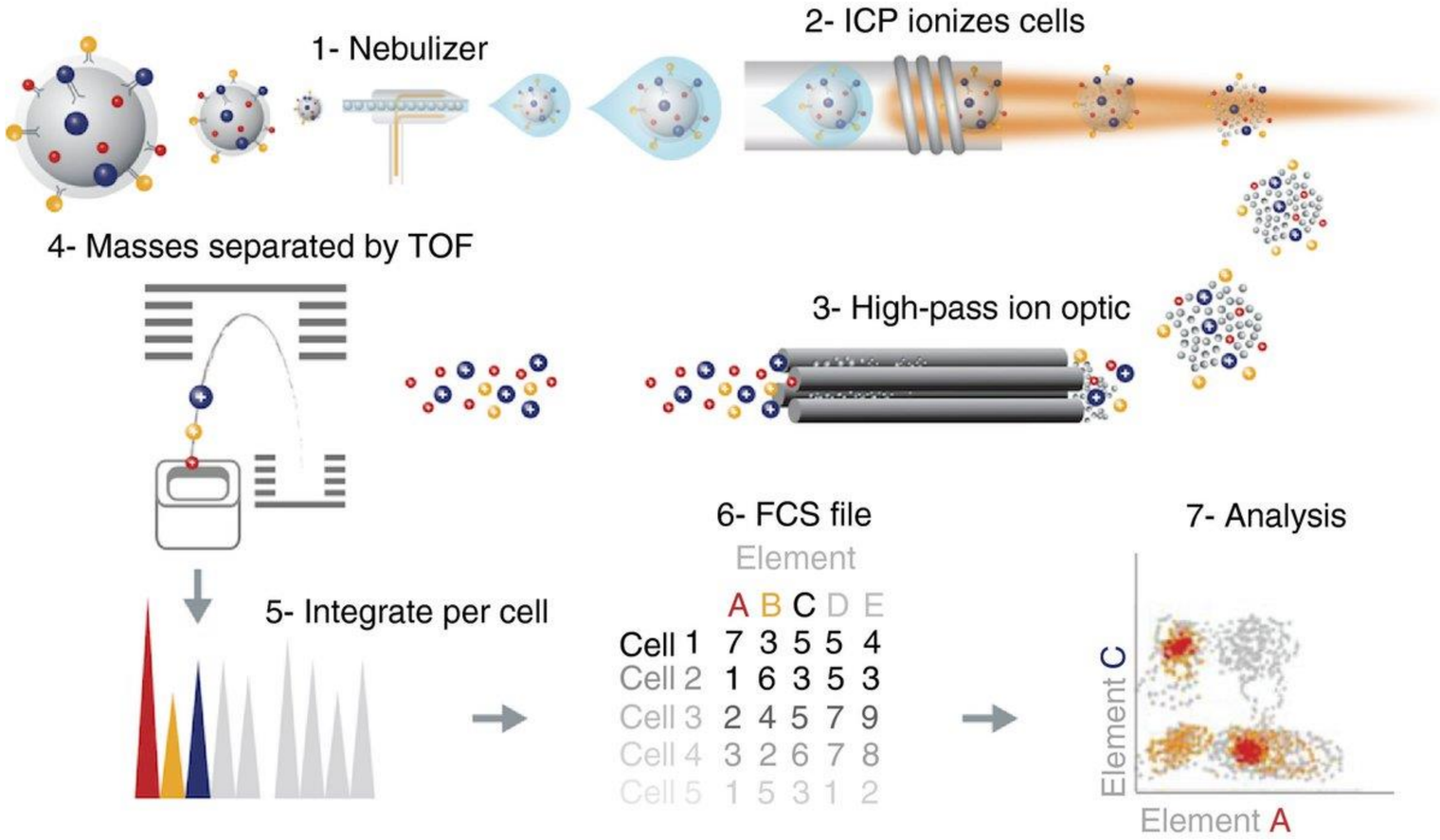
Silesian University of Technology

RESEARCH UNIVERSITY
EXCELLENCE INITIATIVE
Ministry of Science and Higher Education

COMPARISON OF BATCH EFFECT REMOVAL METHODS FOR HIGH DIMENSIONAL MASS CYTOMETRY DATA

Aleksandra Suwalska, Nelita du Plessis-Burger, Gian Van der Spuy, Joanna Polanska



Mass cytometry

(CyTOF)

Mass spectrometry

with inductively coupled plasma

+

Time-of-Flight detector (ToF)

CyTOF data analysis

- Normalization with calibration beads
- Preprocessing (data transformation; pre-gating)
- **Batch effect removal**
- Cell subtypes identification (dimensionality reduction, clustering)

Mass cytometry data can contain millions of cells and dozens of markers.

MATERIALS

Data

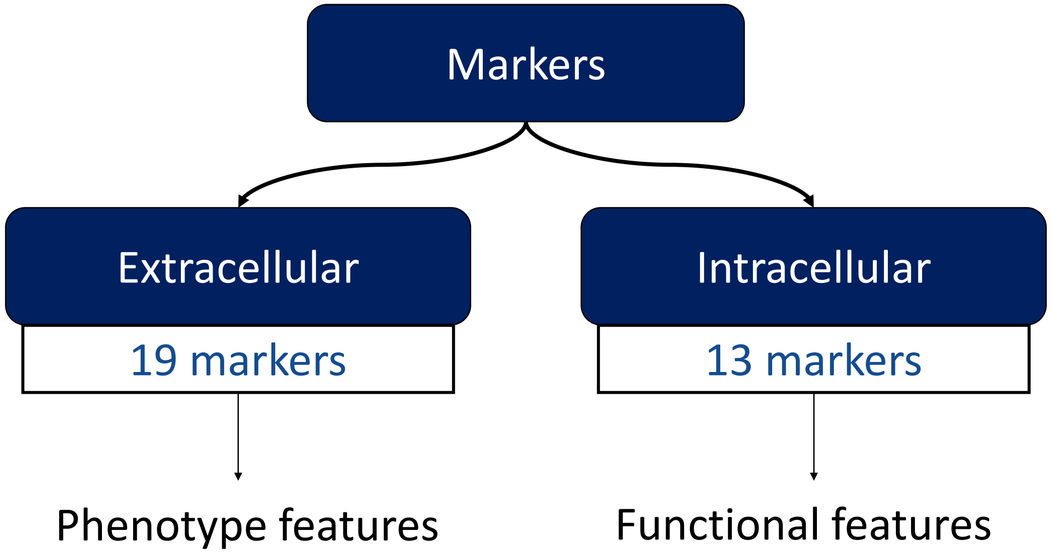
UNIVERSITEIT STELLENBOSCH UNIVERSITY

- 7 samples of healthy patients of bronchoalveolar lavage cells (BALC) from studies on drug-resistant tuberculosis,
- Bronchoscopies were performed in the bronchoscopy theatre in Tygerberg Hospital (TBH) from Cape Town, South Africa,
- Samples were normalized with MATLAB Normalizer v0.3 software,
- Samples were filtered during manual gating to discard debris, dead cells, beads or doublets from the analysis.

Batch number	Number of cells
Batch 1	761,230
Batch 2	598,492
Batch 3	205,958
Batch 4	329,228
Batch 5	341,007
Batch 6	1,449,084
Batch 7	460,713
Total	4,145,712

MATERIALS

Markers used



Total number of markers: 32

Extracellular Antibodies			Intracellular Antibodies	
Antibody Target	**Isotope Label**	**Cell Types Represented by Ab Lineage Marker**	**Antibody Target**	**Isotope Label**
CD3	Er170	T-cells	INF-γ	Gd158
CD14*	Eu151	Monocytes/macrophages, granulocytes	IL-17A	Nd148
CD172	Lu175	Dendritic cells, monocytes/macrophages	IL-4	Nd142
CD19	Ho165	B-cells, dendritic cells	IL-10	Er166
CD33*	Tm169	Dendritic cells, monocytes/macrophages, granulocytes	IL-1β	Yb176
CD45	Y89	Leukocytes	IL-6	Sm147
CD326	Pr141	Epithelial cells	iNOS	Yb171
CD11b*	Nd144	T-cells, B-cells, dendritic cells, NK cells, granulocytes, monocytes/macrophages	Arg-1	Dy164
CD4	Nd145	CD4 T-cells	TGF-β	Dy163
CD36*	Gd155	Dendritic cells, monocytes/macrophages	IDO-1	Gd160
CD56	Sm149	NK cells, T-cells	S100A8	Yb173
Cav-1	Nd146	Ubiquitous expression	TNF-α	Sm152
Lox-1	Gd156	B-cells, dendritic cells, macrophages, MDSC	Mtb (PPD)*	Eu153
CD15	Yb172	Monocytes/macrophages, granulocytes		
CD206	Er168	Alveolar macrophages		
HLA-DR	Yb174	Leukocytes		
CD11c*	Tb159	T-cells, B-cells, dendritic cells, NK cells, granulocytes, monocytes/macrophages		
CD1c	Dy161	T-cells, B-cells, dendritic cells, monocytes/macrophages		
CD47	Bi209	Ubiquitous expression		

* indicates antibodies within the designed panel which did not produce any staining signal throughout optimization, and were thus presumed to be ineffective in staining the desired epitopes.

Batch effect

Batch effect – technical variance introduced to the data during experimenting and it makes it difficult to reveal the biological variance.

Methods that require technical replicates:

- CytofBatchAdjust (R. P. Schuyler *et al*., Frontiers in Immunology, 2019, pp. 2367),
- CytoNorm (S. Van Gassen *et al*., Cytometry Part A, 2020, vol. 97(3), pp. 268-278),
- CytofRUV (M. Trussart *et al*., bioRxiv, 2020).

Methods that do not require technical replicates:

- iMUBAC (M. Ogishi *et al*., The Journal of Immunology, 2021, vol. 206(1), pp. 206-213),
- cyCombine (C. B. Pedersen *et al*., bioRxiv, 2021).

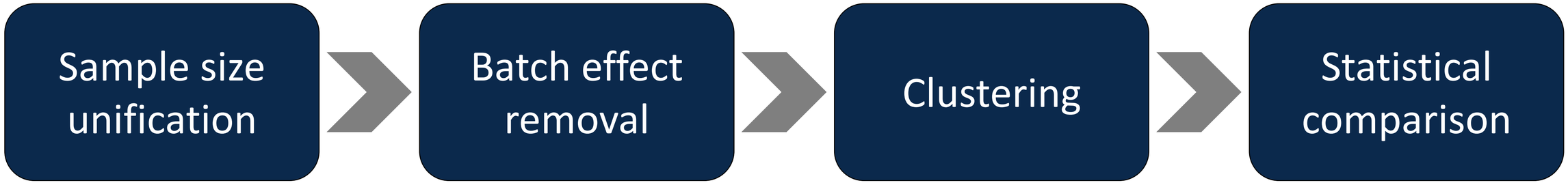
Batch effect

Questions:

- Which batch effect removal method to choose?
- Which method is the best for our data?
- How to evaluate which method is the best?
- How do these methods affect the results of cell subtypes identification?

METHODS

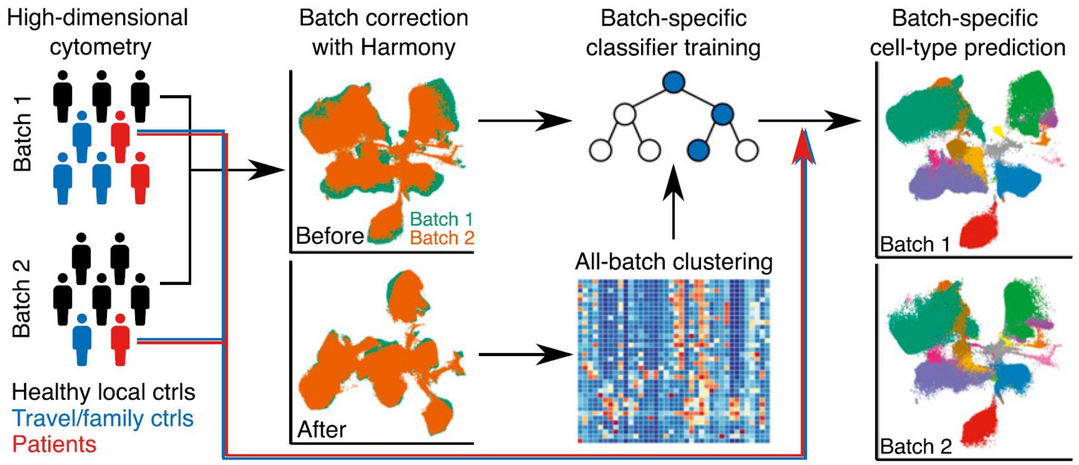
Pipeline



Batch effect removal

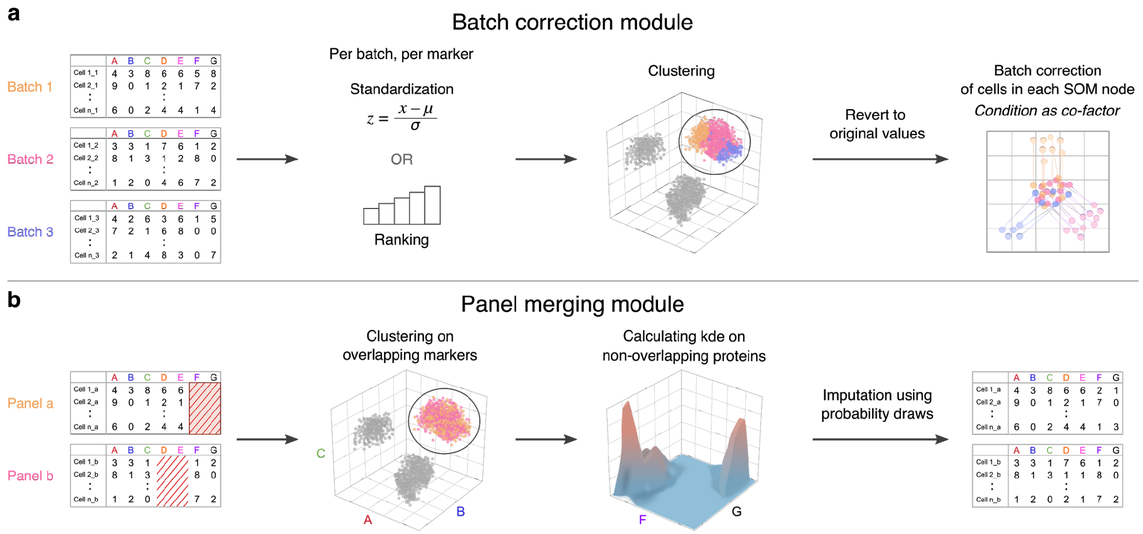
N cells = 2,005,958

iMUBAC



Ogishi, M., Yang, R., Gruber, C., Zhang, P., Pelham, S. J., Spaan, A. N., ... , Casano-va, J. L.: Multibatch cytometry data integration for optimal immunophenotyping. The Journal of Immunology, 206(1), 206-213 (2021).

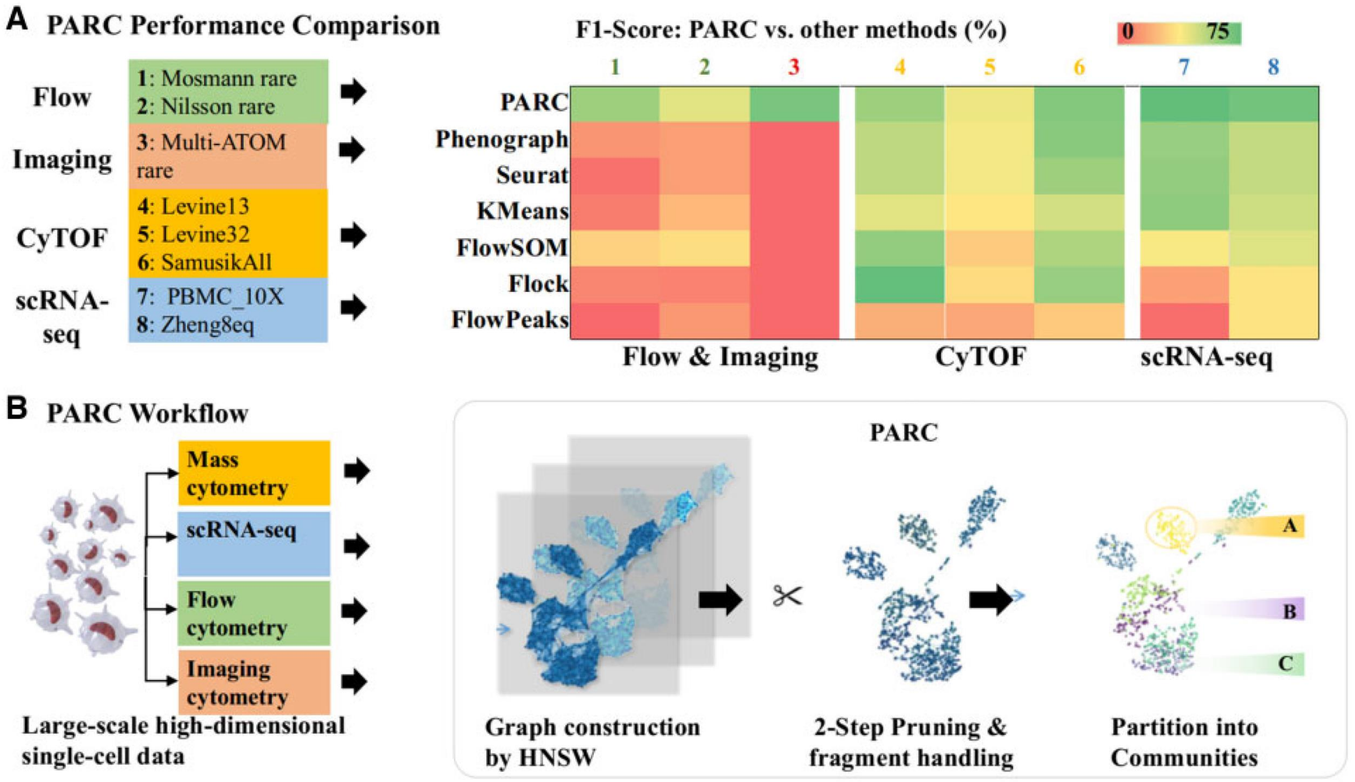
cyCombine



Pedersen, C. B., Dam, S. H., Barnkob, M. B., Leipold, M. D., Purroy, N., Rassenti, L. Z., ..., Olsen, L. R.: Robust integration of single-cell cytometry datasets. bioRxiv. (2021).

METHODS

Clustering



Stassen, S. V., Siu, D. M., Lee, K. C., Ho, J. W., So, H. K., Tsia, K. K.: PARC: ultra-fast and accurate clustering of phenotypic data of millions of single cells. Bioinformatics, 36(9), 2778-2786 (2020).

Statistical comparison

- For the pairwise comparison of marker's expression between clusters an effect size was calculated:

$$d_{AB} = \frac{mA - mB}{SE * \sqrt{N_{ps}}} = \frac{mA - mB}{\sqrt{\frac{SS_{within}}{N-k} * \frac{1}{N_{ps}}} * \sqrt{N_{ps}}} = \frac{mA - mB}{\sqrt{\frac{SS_{within}}{N-k}}}$$

$$N_{ps} = \frac{2}{\frac{1}{N_A} + \frac{1}{N_B}}$$

- The pairwise comparison resulted in a set of dAB effect size values for each marker and the median value was calculated as the global effect size.
- The effect sizes were then compared with the Wilcoxon signed-rank test to check if the heterogeneity of markers between the clusters after correction is greater than before correction at a 5% significance level.

Jacob Cohen: Statistical Power Analysis for the Behavioral Sciences, Routledge, 1988

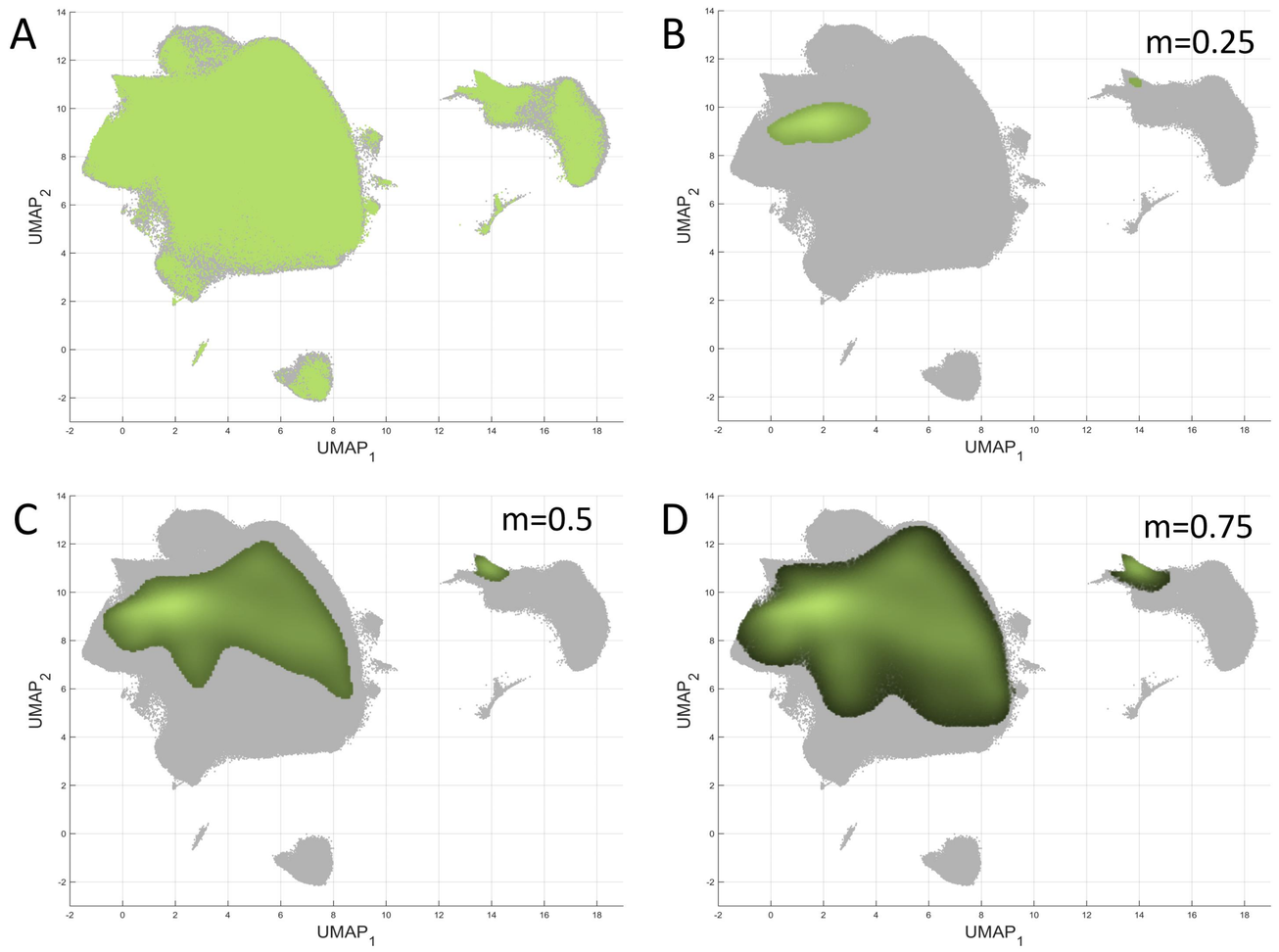
METHODS

mISO plots

We introduce **median isoline plot (mISO)** that can be superimposed on UMAP plots.

It is based on isolines that determine the density of the points.

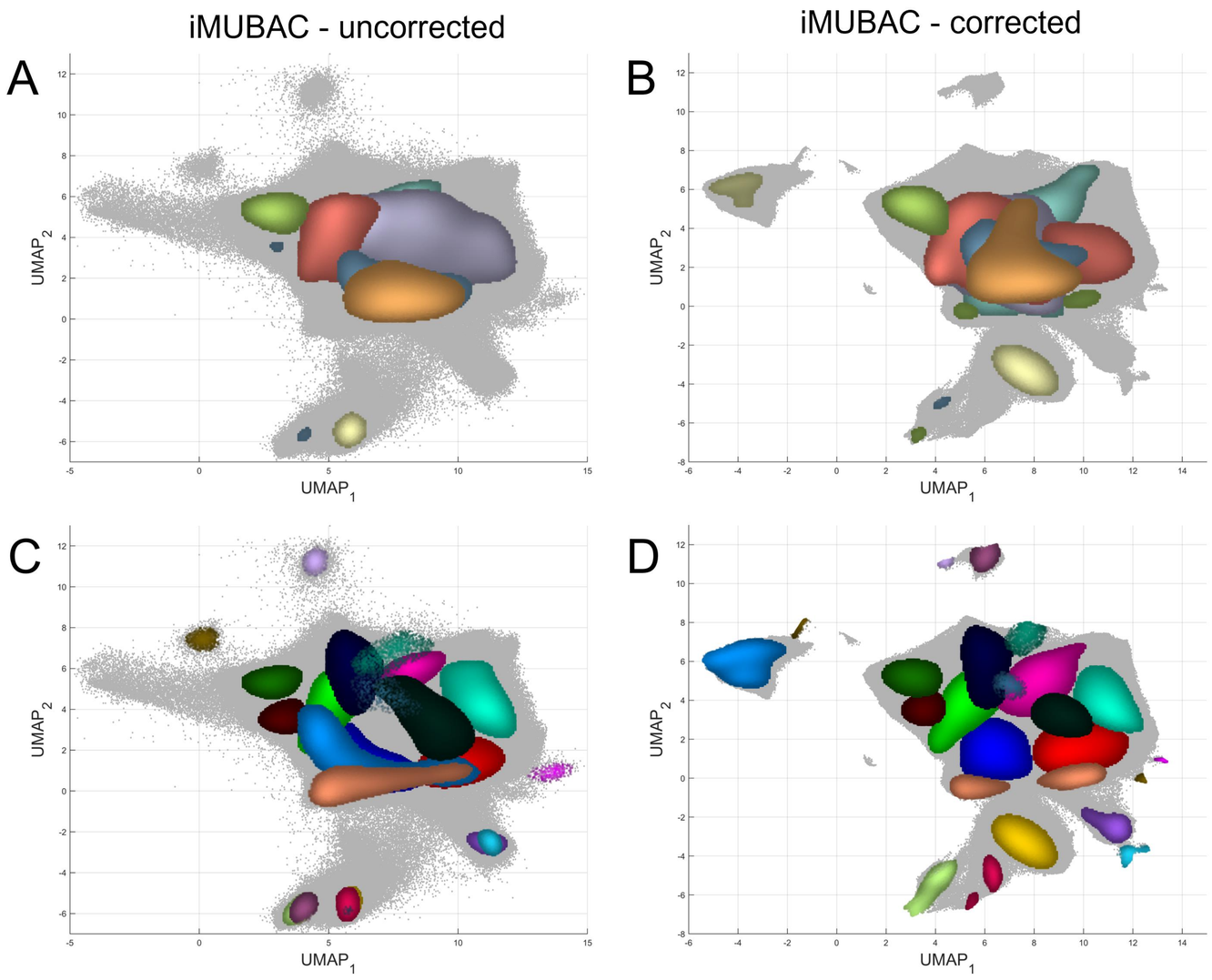
m – a parameter that defines the density level above which the data will be displayed.



RESULTS

iMUBAC

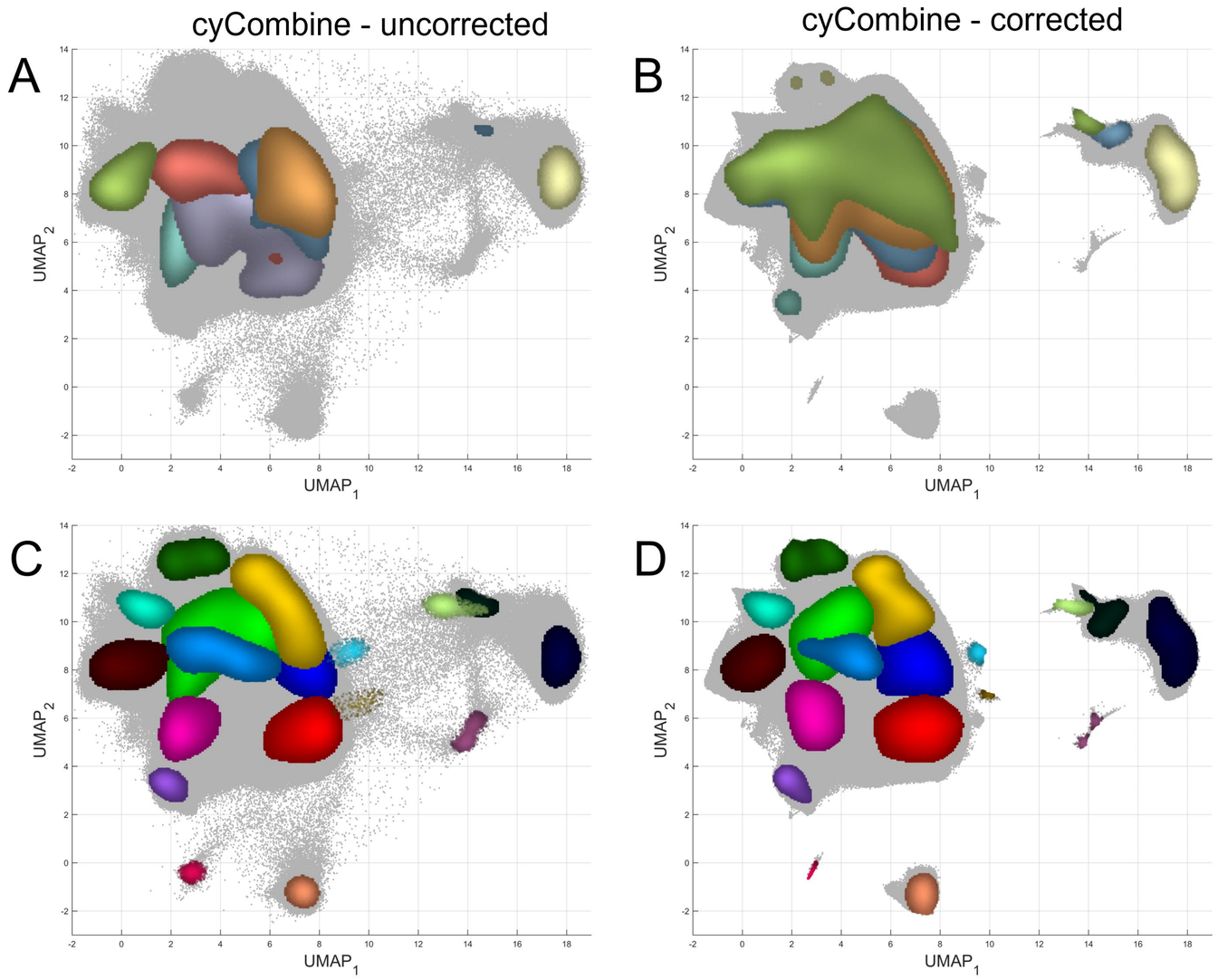
Number of clusters = 22



RESULTS

cyCombine

Number of clusters = 18



Samples (batches)

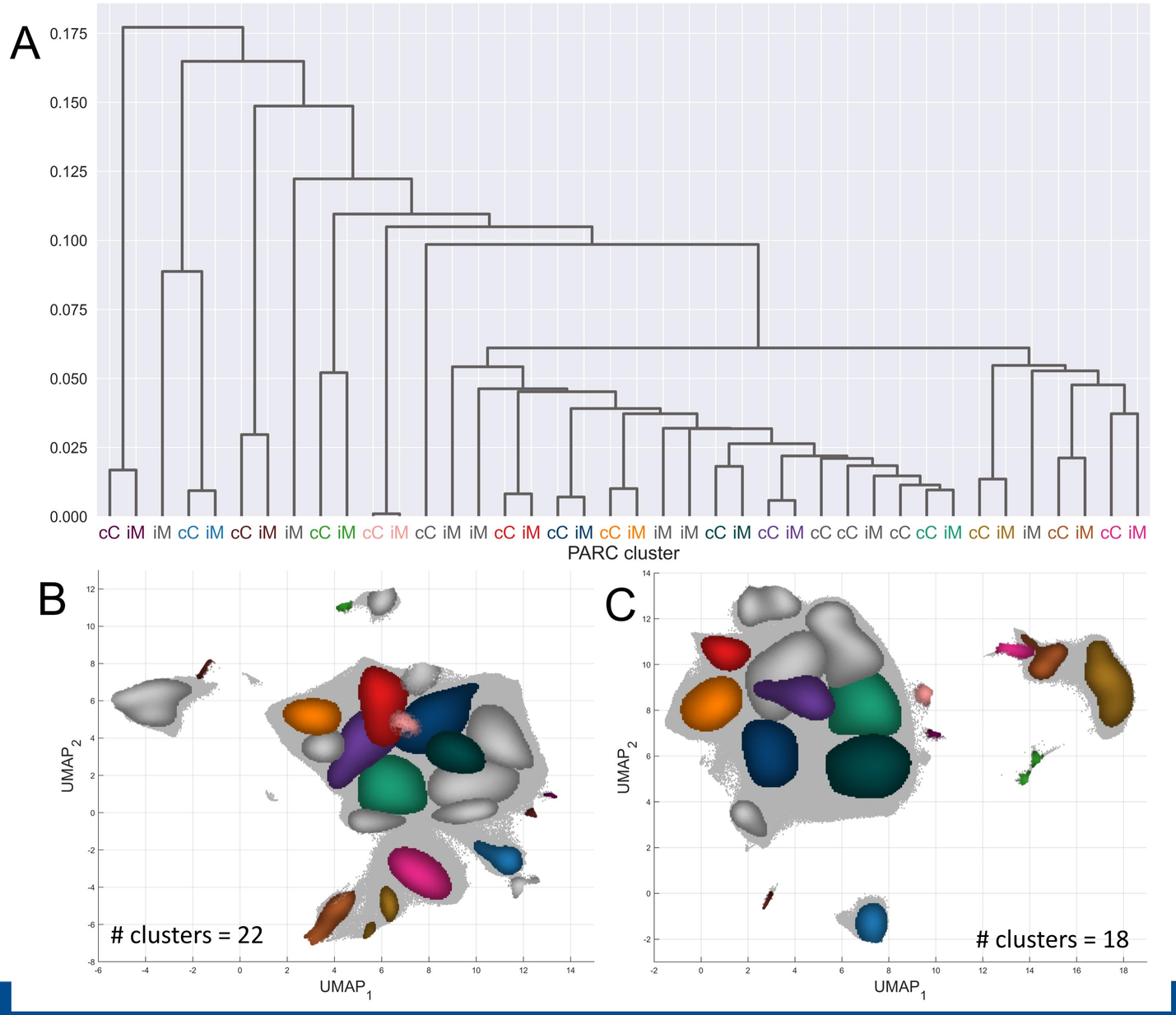
PARC Clusters

iMUBAC vs cyCombine

Dendrogram showing similar clusters (color-coded) between the two correction methods;

iM - cluster created after iMUBAC correction;
cC – cluster created after cyCombine correction.

The clusters that are most similar according to the dendrogram share the same color. Clusters that do not have a similar pair from the other experiment are presented in grey.

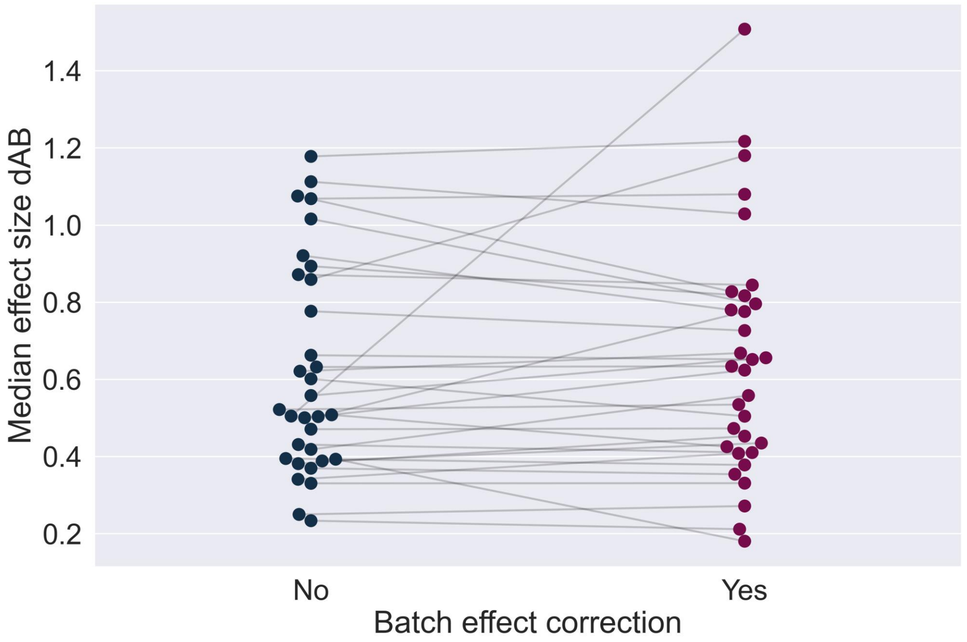


RESULTS

iMUBAC vs cyCombine

iMUBAC

Wilcoxon test p-value: 0.4628



cyCombine

Wilcoxon test p-value: 4.38e07

Median effect sizes from post-hoc ANOVA test before and after batch correction.

Summary

- There are several batch effect removal methods for mass cytometry data.
- Two of the methods were compared.
- Both methods decrease the batch effect.
- The homogeneity of cell clusters after cyCombine correction has increased significantly.
- The results indicate the outperformance of cyCombine over iMUBAC for our dataset.

CONTACT

Thank you
for your attention

KIAED
KATEDRA INŻYNIERII I ANALIZY EKSPLORACYJNEJ DANYCH

Faculty of Automatic Control, Electronics
and Computer Science
Department of Data Science and Engineering

Phone
+48 32 400 30 86

E-mail
Aleksandra.Suwalska@polsl.pl