Finding the cells amidst the data

Hejblum Boris P[1,2], Commenges Daniel[1,2*], Alkhassim Chariff[1,2],
Caron François[3], Gottardo Raphael[4], Thiébaut Rodolphe[1,2*]

[1]Univ. Bordeaux, Inserm U1219 BPH ; Inria SISTM, France
[2]Vaccine Research Institute (VRI), Creteil, France
[3]Department of Statistics, University of Oxford, UK
[4]Fred Hutchinson Cancer Research Center, Seattle, WA
[*]*rodolphe.thiébaut@u-bordeaux.fr*

Introduction

Flow and mass (CyTOF) cytometry are high-throughput technologies quantifying multiple surface and intracellular markers at the level of a single cell. Improvements of these technologies allow to describe **millions of individual cells** from a single blood sample according to several dozens of markers. Hence, up to 2^{30} cell populations could be identified with 30 dichotomous markers.

Such data are usually analyzed by human operators, giving results that are :

- **poorly reproducible**
- **expensive** (highly time-consuming)
- possibly **missing out new cell populations of potential interest**

In **research studies**, multiple samples can be collected for each patient, with measurements across several time points. This results in **very large datasets** whose full analysis becomes **impossible manually** and requires **automated solutions**.

Methods

1. Automatic gating without human intervention

We have developed **two new machine learning approaches** to perform **unsupervised clustering of cell populations**:

`NPflow` a **Bayesian nonparametrics modeling** that **directly estimates the number of clusters (cell populations) from the data** and allows to use **prior knowledge** in the model. To accommodate **repeated flow-cytometry measurements**, such as in a clinical trial, a sequential strategy relying on a parametric approximation of the posterior is also proposed (`NPflow seq`) [1].

`cytometree` a **binary tree** construction, whose nodes represents cellular sub-populations. Each node **optimally splits the cells into different cellular populations** according to the normalized AIC between a Gaussian mixture models considering either one or two sub-populations. Post-processing automatically provide a complete annotation of the derived populations [2].

2. Datasets

- The **FlowCAP I challenge** provides 5 experimental benchmark datasets to compare automatic gating against a consensus of 8 manual operators [3]:

GvHD	Graft-versus-Host Disease: 12 human samples with ∼14,000 cells quantified on 6 dimensions
HSCT	Hematopoietic Stem Cell Transplant: 30 mice samples with ∼10,000 cells quantified on 6 dimensions
DLBCL	Diffuse Large B-Cell Lymphoma: 30 human samples with ∼5,000 cells quantified on 5 dimensions
WNV	Diffuse Large B-Cell Lymphoma: 13 human samples with ∼100,000 cells quantified on 8 dimensions
ND	Normal Donors: 30 human samples with ∼17,000 cells quantified on 6 dimensions

- The **HIPC T-cell panel** [4] was used in the FlowCAP III challenge [3]: three different cryo-preserved PBMC samples were stained and manually gated by seven different laboratories (each lab processed three replicates of each original sample). Each of the 63 resulting data matrices had ∼125,000 cells quantified on 7 dimensions.

- The **DALIA-1 Trial** is a phase I trial evaluating a therapeutic vaccine against HIV [5]. We analyzed flow-cytometry measurements from 12 HIV positive patients at two consecutive time-points and looked especially at the CD4+ effector T-cells population from the ∼40,000 cells quantified on 10 dimensions.

Conclusion

The **constant increase of the number of markers** available to characterize cell populations leads to an **intractable situation with manual gating**. However, improvements of **machine learning approaches allow for the automatic analysis of cytometry samples**, yielding the relative count for all possible cell populations.

R packages `NPflow` & `cytometree` are both available on the *Comprehensive R Archive Network* (CRAN).

References

1. Hejblum BP, Alkhassim C, Gottardo R, Caron F, Thiébaut R, Sequential Dirichlet process mixture of skew t-distributions for model-based clustering of flow cytometry data, Annals of Applied Statistics, (in press), 2019.
2. Commenges D, Alkhassim C, Gottardo R, Hejblum BP, Thiébaut R, cytometree: a binary tree algorithm for automatic gating in cytometry analysis, Cytometry Part A, 93(11):1132-1140, 2018.
3. Aghaeepour N *et al.* (2013). Critical assessment of automated flow cytometry data analysis techniques. *Nat Methods*, 10(3): 228-38.
4. Finak G *et al.*. Standardizing flow cytometry immuno-phenotyping analysis from the Human ImmunoPhenotyping Consortium. *Sci Rep*, 6:20686, 2016.
5. Lévy Y, Thiébaut R, Montes M, *et al.* (2014). Dendritic cell-based therapeutic vaccine elicits polyfunctional HIV-specific T-cell immunity associated with control of viral load. *Eur J Immunol*, 44: 2802-10.

Results

1. FlowCAP I challenge benchmark

Compared to other available approaches, the new method `cytometree` performed at the **top position** with the **shortest runtime**. Its $\mathcal{F}$-measures > 0.90 demonstrate the validity of the new methods in comparison with the gold standard (consensus of 8 experimentalists) [1]. The $\mathcal{F}$-measure is an indicator between 0 and 1 to quantifying an agreement between two clustering of the same data (the closer to 1 the better).

TABLE 1: Mean $\mathcal{F}$-measures across the 12 samples from the GvHD benchmark dataset

Method	Dataset					Mean	Runtime
	GvHD	HSCT	DLBCL	WND	ND		per sample
`cytometree`	**0.94**	**0.95**	**0.93**	**0.90**	0.89	**0.92**	00:01:06
`NPflow`	0.85	–	–	–	–	–	∼ 12:00:00
`NPflow-seq`	0.89	–	–	–	–	–	∼ 12:00:00
ADICyt*	0.81	**0.93**	0.93	0.86	**0.92**	0.89	04:50:37
flowMeans*	0.88	0.92	0.92	0.88	0.85	0.89	00:02:18
FLAME*	0.85	0.94	0.91	0.80	0.90	0.88	00:04:20
FLOCK*	0.84	0.86	0.88	0.83	0.91	0.86	00:00:20

Estimates number from [1] are provided for comparison.

2. HIPC T-cell panel: an exemple of the excellent performance of `cytometree`

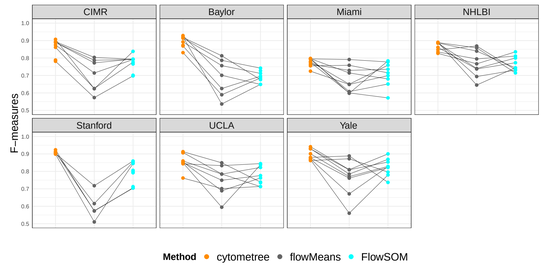


FIGURE 1: Comparison to manual gating with HIPC T-cell panel: `cytometree` leads to results closer to manual gating (F=1)

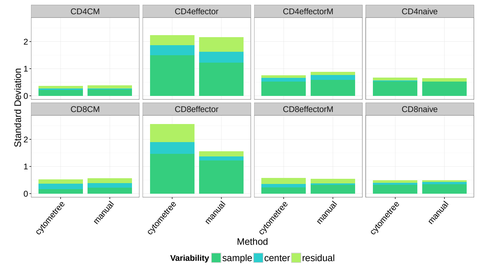


FIGURE 2: Center, biological, and residual variability are similar between `cytometree` results and manual gating

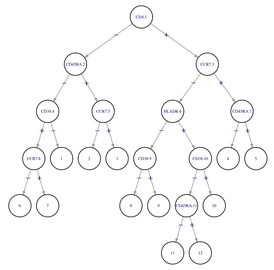


FIGURE 3: Exemple of a binary tree generated by `cytometree`

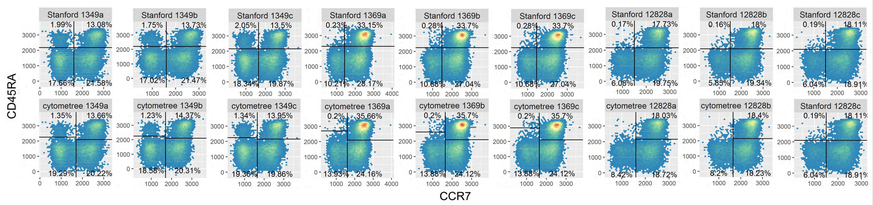


FIGURE 4: Comparison of `cytometree` results for CD4 T-cells to manual gating on the 9 samples from Stanford centre in HIPC T-cell panel

3. DALIA trial: Interest of the sequential approach with NPflow for T-effector population

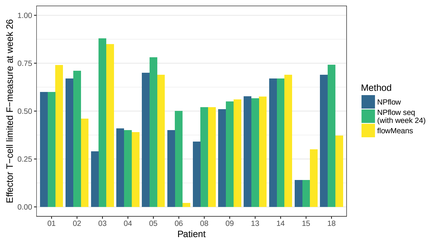


FIGURE 5: T-effector population in the DALIA trial

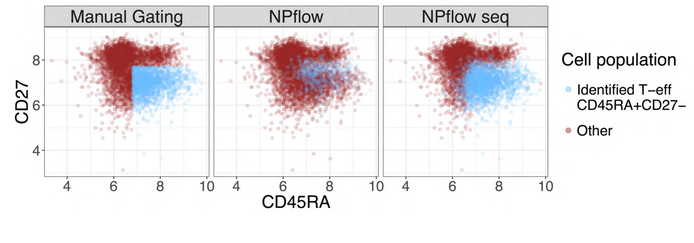


FIGURE 6: T-effector population from Patient 3 of the DALIA trial

Acknowledgements

The authors are extremely grateful to Jean-Louis Palgen for his time and efforts to manually gate the effector T-cells in the DALIA-1 trial at weeks 24 and 26. Part of this work has been supported by the BNPSI ANR project no ANR-13-BS-03-0006-01. Part of this work has been supported by the IMI2 grant EBOVAC2. Computer time for this study was partly provided by the computing facilities MCIA (Mésocentre de Calcul Intensif Aquitain) of the Université de Bordeaux and of the Université de Pau et des Pays de l'Adour. The DALIA-1 trial is funded by an ANRS grant.