Improved NK cell markers from single-cell RNA-seq of PBMC populations with the new ROC-driven combiroc R package

UNIVERSITÀ DEGLI STUDI DI MILANO

I. Ferrari[1 2], S. Mazzara[3], A. Gobbini[1], N. Di Marzo[4], M. Crosti[1], S. Abrignani[1 5], R. Grifantini[1 4], M. Bombaci[1]*, R.L. Rossi[1]*

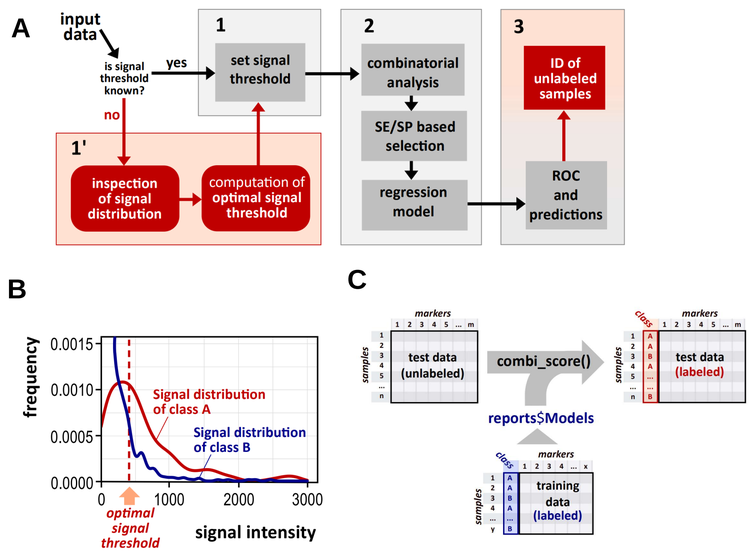
[1] National Institute of Molecular Genetics, Milan, Italy. [2] Department of Biosciences, University of Milan, Milan, Italy. [3] Department of computing Sciences and Bocconi Institute for Data Science and Analytics (BIDSA), Bocconi University, Milan, Italy, [4] CheckmAb Srl, Milan, Italy, [5] Department of Pathophysiology and Transplantation, University of Milan, Milan, Italy.
* Corresponding authors

INGM ISTITUTO NAZIONALE GENETICA MOLECOLARE "Romeo ed Enrica Invernizzi"

Abstract

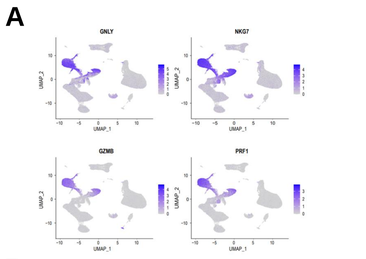
In this study, we introduce the combiroc R package, a tool for refining signatures in high-throughput omics data. Leveraging a ROC-driven combinatorial selection approach, this new package was designed to facilitate the identification of potent sub-signatures from single-cell RNA-seq experiments, enabling more efficient cell annotation using a reduced set of markers. By applying *combiroc* to Peripheral Blood Mononuclear Cells (PBMC) datasets, we identified non-canonical marker combinations for Natural Killer (NK) cells that aligned with the Human Protein Atlas (HPA). We further validated these combinations through cytometry staining and functional assays. **The single-cell workflow presented in this work significantly impacts marker signature research in general and transcriptomic gene signatures in particular.** It demonstrates that the top differentially expressed genes are not necessarily the most specific ones and that smaller signatures can be more powerful, regardless of the differential expression ranking of individual markers. This principle of "less is more" has the potential to re-evaluate existing gene signatures and bring forth new markers that may have gone unnoticed so far. https://ingmbioinfo.github.io/combiroc/ (also on CRAN).

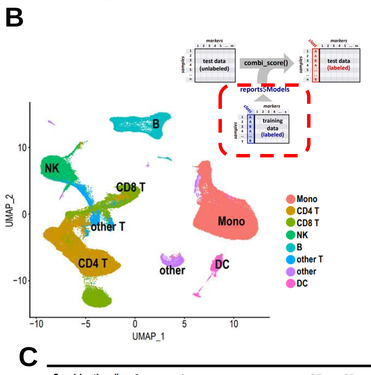
1. Combiroc automates signal threshold setting and creates models to annotate unlabelled samples



(A) *combiroc*'s *workflow*. Red boxes are new features introduced by the combiroc package: alternative phase 1' for computation of signal threshold and part of phase 3 with labeling of unknown samples.
(**B**) Display of the calculated optimal signal threshold (dashed line) on overlapping signal intensity distributions.
(**C**) Samples of unlabeled data (left) can be associated with a class (right, red annotations), using regression models generated from labeled training data (blue annotations). "A" and "B" refer to the binary annotation of classes in the datasets.

2. *Combiroc* unlocks hidden markers sub-signatures

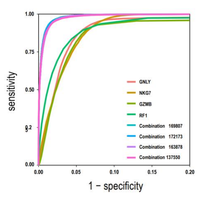




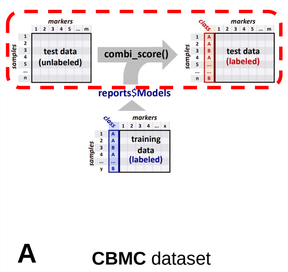
(**A**) Expression of the top four NK cells DEGs in the pbmc CITE-seq dataset: they are also highly expressed in other non-NK clusters (e.g. CD8 T). (**B**) Dataset's cluster annotations for eight different cell types according to gene expression and CITE-seq results in the original work. This is the training dataset on which we trained the *combiroc* model for NK cells (upper right inset) (**C**) Gene combinations found with *combiroc* model and their parameters: SE, sensitivity; SP, specificity; AUC, area under the curve; Youden index. (**D**) ROC curves for the top four individual NK markers and for the top four marker combinations.

C

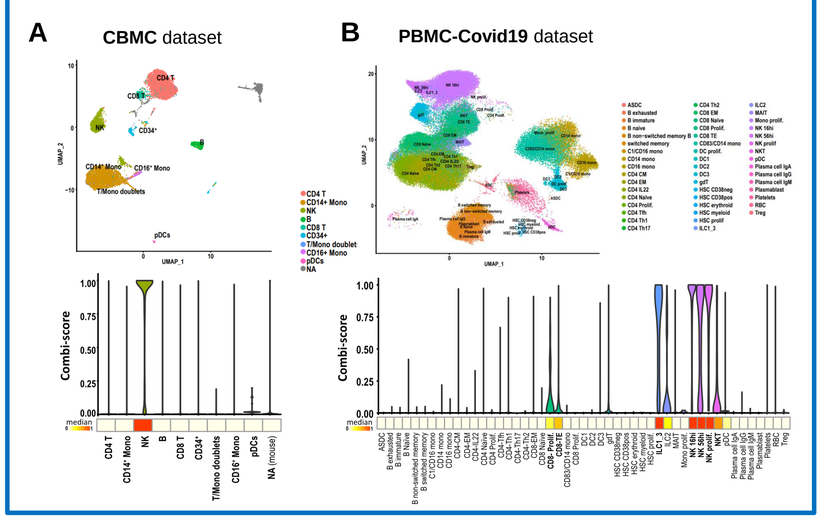
Combination #	Gene members	SE	SP	AUC	Youden
169807	GNLY, IL2RB, KLRF1, SPON2, TRDC	0.982	0.968	0.995	0.970
172173	GZMB, IL2RB, KLRF1, SPON2, TRDC	0.977	0.970	0.995	0.971
163878	FCGR3A, GNLY, IL2RB, KLRF1, MYOM2	0.985	0.962	0.995	0.965
137550	CLIC3, FCGR3A, IL2RB, KLRD1, KLRF1	0.976	0.966	0.995	0.967



3. Sub-signatures accurately identifies NK cells in unlabelled single-cell datasets

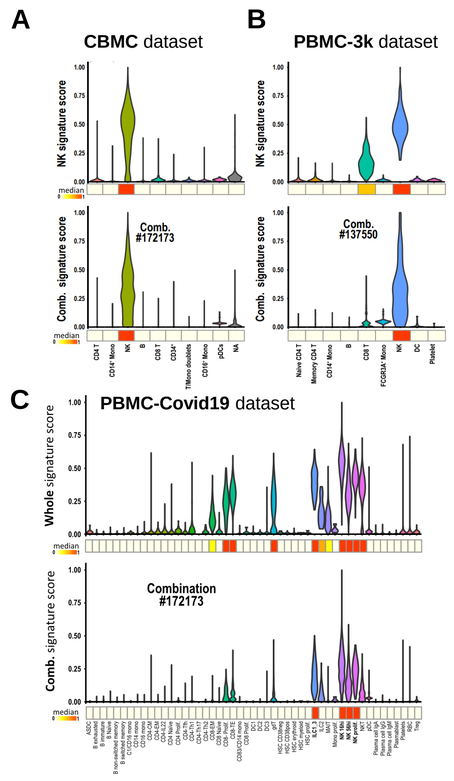


Testing the *combiroc*'s NK cells models on other single-cell datasets: cell cluster view and combi-scores (probabilities of being NK cells) computed across cell clusters of test datasets using the gene combinations found with *combiroc* for (**A**) CBMC dataset; (**B**) PBMC-Covid19 dataset, using combination #172173 (also made on PBMC-3K dataset with combination #137550, *not shown here*).



4. Revolutionary 5-gene cocktail better discriminates NK cells than 30-genes parent signature

Gene signature scores can be calculated to determine the descriptive power of any gene signature (Della Chiara et al., 2021): scores derived from *combiroc*'s 5-genes combinations proved to be just as effective in identifying NK cells as those obtained from the entire 30-genes signature as observed in CBMC dataset (**A**), and even better in PBMC-3k dataset (**B**) and in PBMC-Covid19 dataset (**C**).



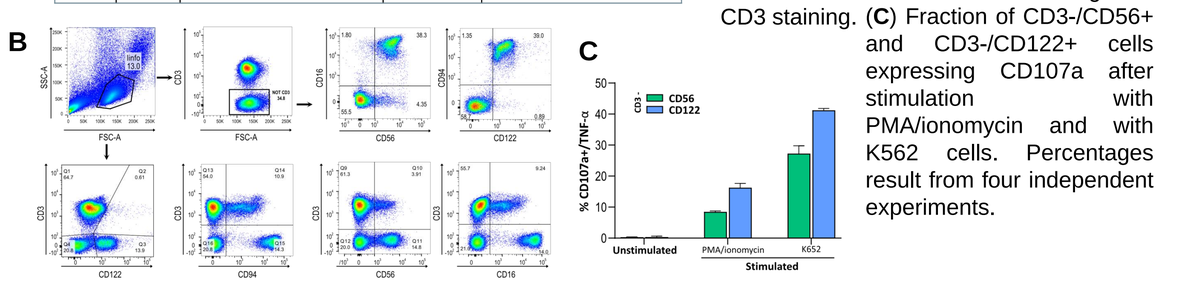
Less is More!

5. ILR2B (CD122) is specifically associated with highly functional NK cells

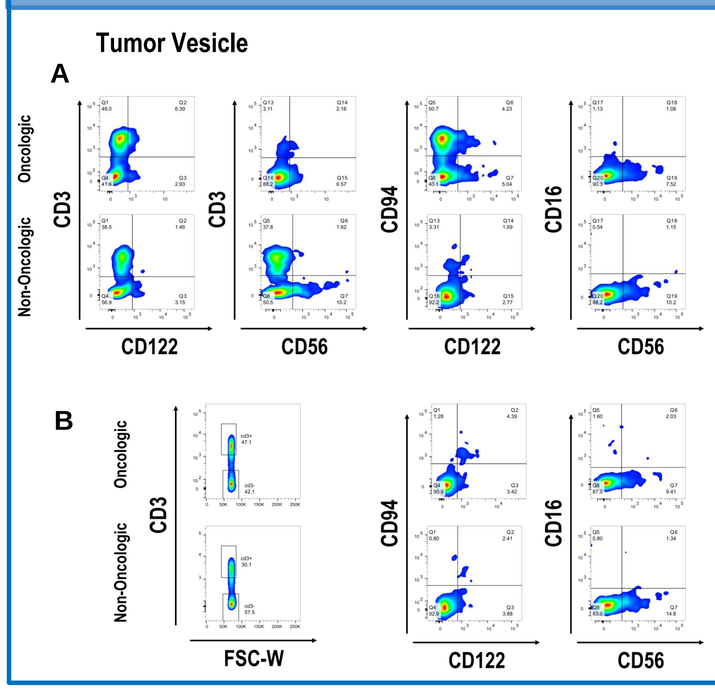
A

Gene	Alias	Description	Localization	Enriched cell type (Tau specificity score)
IL2RB	CD122	Interleukin 2 receptor subunit beta	Membrane	NK cells, T cells (0.89)
KLRD1	CD94	Killer Cell Lectin Like Receptor D1	Membrane	NK cells (0.83)
KLRF1	CLEC5C	Killer Cell Lectin Like Receptor F1	Membrane	NK cells (0.90)
SPON2	DIL1	Spondin 2	Intracellular	NK cells, Dendritic cells (0.66)
TRDC1	TCRD	T Cell Receptor Delta Constant	Membrane	NK cells, T cells (0.93)
GNLY	TLA519	Granulysin	Intracellular	NK cells (0.84)
FCGR3A	CD16	Fc Gamma Receptor IIIa	Membrane	Monocytes, NK-cells, Macrophages, Hofbauer cells, Kupffer cells (0.84)
MYOM2	TTNAP	Myomesin 2	Intracellular	NK cells (0.76)
GZMB		Granzyme B	Intracellular	Dendritic cells, NK cells (0.89)
CLIC3		Chloride Intracellular Channel 3	Intracellular	Dendritic cells, NK cells (0.70)

(**A**) The ten genes belonging to the four selected combination (from **panel 2C**) with cell type enrichment (Tau scores) as they are annotated in the "Blood and Immune" subset of Human Protein Atlas (HPA, https://www.proteinatlas.org/).
(**B**) Gating strategy to discriminate among the different NK cells subsets. Live lymphocytes were gated based on side and forward scatter dot plot display during the acquisition process. The NK specific markers were then measured along with CD3 staining. (**C**) Fraction of CD3-/CD56+ and CD3-/CD122+ cells expressing CD107a after stimulation with PMA/ionomycin and with K562 cells. Percentages result from four independent experiments.



6. CD94/CD122 expression identifies NK cell population in tumours



(**A**) CD56 expression in the CRC tumor is significantly higher than CD122, although it remains comparable when contrasted with the non-tumor region, where CD122 exhibits notable differences in expression.

(**B**) In CD3-negative cells revealed that while the expression of the CD56/CD16 combination remains consistent between the tumor and non-tumor regions, the CD94/CD122 combination is uniquely present in the tumor region and absent in the non-tumor region.

SUMMARY

***Combiroc* simplifies cell identification:** *combiroc* streamlines the single-cell RNA-seq data analysis, making it easier to pinpoint specific cell populations by reducing the number of marker genes to consider.
Smaller signatures for better insight: smaller marker combinations, identified by *combiroc*, offer more precise insights than traditional differential expression rankings.
Discovering overlooked markers: the "less is more" approach reveals potentially hidden markers, shedding light on previously unnoticed cell characteristics.
Translational potential: highly performant *combiroc*-identified marker combinations have diagnostic and therapeutic applications, enhancing our understanding of cell populations.

REFERENCES

- **Mazzara et al.** Sci. Rep. 2017**.** CombiROC: an interactive web tool for selecting accurate marker combinations of omics data.
- **Bombaci & Rossi.** Methods Mol. Biol. 2019. Computation and selection of optimal biomarker combinations by integrative ROC analysis using combiroc.
- **Satija et. al** Nat. Biotechnol. 2015. Spatial reconstruction of single-cell gene expression data.
- **Stephenson et al.** Nat. Med. 2021. Single-cell multi-omics analysis of the immune response in COVID-19.
- **Hao, Y. et al.** Cell. 2021. Integrated analysis of multimodal single-cell data.
- **Della Chiara et al.** Nat. Commun. 2021. Epigenomic landscape of human colorectal cancer unveils an aberrant core of pan-cancer enhancers orchestrated by YAP/TAZ.

WHAT WE'RE WORKING ON

We are currently working on improvements:

- Algorithm improvement by integration of feature selection and parallelization
- Porting and release of a Python package
- Application of the algorithm to elusive immunological populations

ferrari@ingm.org
bombaci@ingm.org
rossi@ingm.org