Year : 2020

D-AREdevil: a novel approach for discovering disease-associated rare cell populations in mass cytometry data

Suffiotti Madeleine

Suffiotti Madeleine, 2020, D-AREdevil: a novel approach for discovering disease-associated rare cell populations in mass cytometry data

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive http://serval.unil.ch
Document URN : urn:nbn:ch:serval-BIB_C4EC0725191A5

UNIL | Université de Lausanne
Faculté de biologie et de médecine

Centre hospitalier universitaire vaudois
Département de Médicine
Service d'Immunologie et Allérgie

D-AREdevil: a novel approach for discovering disease-associated rare cell populations in mass cytometry data

Thèse de doctorat ès sciences de la vie (PhD)

présentée à la

Faculté de biologie et de médecine
de l'Université de Lausanne

par

Madeleine SUFFIOTTI

Master en Bioinformatique, Université de Lausanne

Jury

Prof. John-David Aubert, Président·e
Prof. Giuseppe Pantaleo, Directeur·trice de thèse
Dr Mauro Delorenzi, Co-directeur·trice de thèse
Prof. Mark Robinson, Expert·e
Prof. Didier Trono, Expert·e

Lausanne
2020

UNIL | Université de Lausanne
Faculté de biologie et de médecine

Ecole Doctorale
Doctorat ès sciences de la vie

Imprimatur

Vu le rapport présenté par le jury d'examen, composé de

Président·e	Monsieur	Prof.	John-David	**Aubert**
Directeur·trice de thèse	Monsieur	Prof.	Giuseppe	**Pantaleo**
Co-directeur·trice	Monsieur	Dr	Mauro	**Delorenzi**
Expert·e·s	Monsieur	Prof.	Mark	**Robinson**
	Monsieur	Prof.	Didier	**Trono**

le Conseil de Faculté autorise l'impression de la thèse de

Madame Madeleine Suffiotti

Maîtrise universitaire en Sciences moléculaires du vivant , Université de Lausanne

intitulée

D-AREdevil: a novel approach for discovering disease-associated rare cell populations in mass cytometry data

Lausanne, le 10 décembre 2020

pour le Doyen
de la Faculté de biologie et de médecine

Prof. Niko GELDNER
Directeur de l'Ecole Doctorale

Contents

Summary

Background: The advances in single-cell technologies such as mass cytometry provides increasing resolution of the complexity of cellular samples, allowing researchers to deeper investigate and understand the cellular heterogeneity and possibly detect and discover previously undetectable rare cell populations. The identification of rare cell populations is of paramount importance for understanding the onset, progression and pathogenesis of many diseases. However, their identification remains challenging due to the always increasing dimensionality and throughput of the data generated.

Aim: This study aimed at implementing a straightforward approach that efficiently supports a data analyst to identify disease-associated rare cell populations in large and complex biological samples and within reasonable limits of time and computational infrastructure.

Methods: We proposed a novel computational framework called D-AREdevil (disease-associated rare cells detection) for cytometry datasets. The main characteristic of our computational framework is the combination of an anomaly detection algorithm (*i.e.* LOF, or FiRE) that provides a continuous score for individual cells with one of the best performing and fastest unsupervised clustering methods (*i.e.* FlowSOM). In our approach, the LOF score serves to select a set of candidate cells belonging to one or more subgroups of similar rare cell populations. Then, we tested these subgroups of rare cells for association with a patient group, disease type, clinical outcome or other characteristic of interest.

Results: We reported in this study the properties and implementation of D-AREdevil and presented an evaluation of its performances and applications on three different testing datasets based on mass cytometry data. We generated data mixed with one or more known rare cell populations at varying frequencies (below 1%) and tested the ability of our approach to identify those cells in order to bring them to the attention of the data analyst. This is a key step in the process of finding cell subgroups that are associated with a disease or outcome of interest, when their existence and identification is not previously known and has yet to be discovered.

Conclusions: We proposed a novel computational framework with demostrated good sensitivity and precision in detecting target rare cell poopulations present at very low frequencies in the total datasets ($<$1%).

Résumé

Contexte: Les avancées en technologies sur cellules individuelles telles que la cytométrie de masse offrent une meilleure résolution de la complexité des échantillons cellulaires, permettant aux chercheurs d'étudier et de comprendre plus en profondeur l'hétérogénéité cellulaire et éventuellement de détecter et découvrir des populations de cellules rares auparavant indétectables. L'identification de populations de cellules rares est importante pour comprendre l'apparition, la progression et la pathogenèse de nombreuses maladies. Cependant, leur identification reste difficile en raison de la haute dimensionnalité et du débit toujours croissants de données générées.

But: Cette étude met en œuvre une approche simple et efficace pour identifier des populations de cellules rares associées à une maladie dans des échantillons biologiques vastes et complexes dans des limites de temps et d'infrastructure de calcul raisonnables.

Méthodes: Nous proposons un nouveau cadre de calcul appelé D-AREdevil (détection de cellules rares associées à une maladie) pour l'analyse de données de cytométrie de masse. La principale caractéristique de notre cadre computationnel est la combinaison d'un algorithme de détection d'anomalies (LOF ou FiRE) qui fournit un score continu pour chaque cellule avec l'une des méthodes de regroupement non-supervisé les plus performantes et les plus rapides (FlowSOM). Dans notre approche, le score LOF sert à sélectionner un ensemble de cellules candidates appartenant à un ou plusieurs sous-groupes de populations de cellules rares similaires. Ensuite, nous testons ces sous-groupes de cellules rares pour déterminer s'ils sont associées avec un groupe de patients, un type de maladie, un résultat clinique ou une autre caractéristique d'intérêt.

Résultats: Dans cette étude, nous avons rapporté les propriétés et l'implémentation de D-AREdevil, et présenté une évaluation de ses performances et applications sur trois jeux de données différents de cytométrie de masse. Nous avons généré des données mélangées contenant une ou plusieurs populations de cellules rares connues à des fréquences variables (inférieures à 1%) et nous avons testé la capacité de notre approche à identifier ces cellules afin de les porter à l'attention de l'analyste. Il s'agit là d'une étape clé dans le processus de recherche de sous-groupes de cellules qui sont associés à une maladie ou à un résultat d'intérêt qui est encore inconnu.

Conclusions: Nous proposons un nouveau cadre de calcul avec une bonne sensibilité et une bonne précision dans la détection de cellules rares qui sont présentes à de très

basses fréquences dans l'ensemble des données (<1%).

Table of Abbreviations

AML	acute myeloid leukemia
AUC	area under the (ROC) curve
BMMCs	bone marrow mononucler cells
CC	consensus clustering
CyTOF	cytometry by time-of-flight mass spectrometry
D-AREdevil	disease-associated rare cell detection
FCS	flow cytometry standard (file)
FDR	false discovery rate
FiRE	finder of rare entities
FP/FN	false positive/negative
GC	germinal center (B cells)
HIV	human immunodeficiency virus
iNKT	invariant natural killer T cells
k-NN	k-nearest neighbors
LOF	local outlier factor
LRD	local reachability density
MRD	minimal residual disease
PBMCs	peripheral mononuclear cells
PPV	positive predictive value
RD	reachability distance
ROC	receiver operating operating characteristic
scRNA-seq	single-cell RNA-sequencing
SOM	self-organizing map

SW	switched memory (B cells)
TP/TN	true positive/negative
t-SNE	t-distributed stochastic neighbor embedding
UMAP	uniform manifold approximation and projection
US	unswitched memory (B cells)

Acknowledgments

Firstly, I would like to thank Professor Giuseppe Pantaleo for giving me the great opportunity to do my PhD in his laboratory and for providing me all the tools necessary to accomplish it successfully.
I would also like to thank my thesis supervisor, Dr. Mauro Delorenzi who was my mentor from the beginning of my Master studies. Under his guidance, instruction and advices, I have learned a lot about bioinformatics, statistics and developped a critical thinking about data analysis and results interpretation.
I would like to thank my physics Professor Tran Mihn Tan who supported me with precious advices and words of wisdom during my whole accademic path, from the Bachelor until today.
Also, I would like to thank Dr. Alessandra Noto, Dr. Denis Comte and Dr. Craig Fenwick for all the time and passion they had in helping me to understand biology, immunology and medicine. I would also like to thank other members of the lab and friends such as Morgane, Victor, Patricia and Erica for the help and support during these years.
The biggest thank is to my entire family, without them I could not have the possibility to do all my studies. My parents Amina and Loris and my brother Eios supported and helped me with love and care during this long journey.

Chapter 1

Introduction

1.1 An overview of flow and mass cytometry applications

Technological advances in single-cell measurements such as flow or mass cytometry (CyTOF, cytometry by time of flight mass spectrometry) have allowed researchers to achieve a deeper understanding of the cellular heterogeneity within populations that were once assumed to be homogeneous [Spitzer and Nolan, 2016]. These technologies are a staple of biological research with applications in immunology and cell biology, and widely used in clinical diagnostics.

Flow cytometry uses fluorescence-labeled antibodies for the rapid and simultaneous analysis of multiple parameters (> 20). Fluorescent molecules with different excitation and emission properties are used in combinations to detect proteins on the surface or within cells. When cells labeled with fluorescent molecules pass through the laser beam, the peak of photon emission is recorded and the resulting fluorescence intensity is used to determine to the protein abundance [Bashashati and Brinkman, 2009]. The passage through the laser beam also provide information about cell size and shape. Moreover, this technology allows the viable separation (or sorting) of purified cell populations [Perfetto et al., 2004]. Although flow cytometry is in continue evolution permitting an increasing number of parameters to be analyzed, fluorophore emission spectral overlap (cross-talk between channels) remains a limitation to the number of parameters that can be recorded [Palit et al., 2019].

Mass cytometry represents a next generation flow cytometry platform that uses unique stable heavy-metal isotopes (*i.e.* rare earth metals) labeled antibodies, not naturally found in cells. The use of these isotope-labeled antibodies results in very little cross-talk between channels due to the different atomic weights that can be discriminate with high accuracy and enabling the quantification of over 50 parameters at the single-cell level [Bruggner et al., 2014a] (Figure 1). This high parameterization allows to identify lineage or maturation states of cells of interests (by measuring levels of transmembrane proteins expressed on the cell surface, called surface markers) and to study the cellular behavior (*e.g.* cell-signaling receptors, phosphorylation of signaling proteins, receptor ligands, adhesion molecules, transcription factors, cytokines production etc.) with several

additional parameters. Consequently, this technology allows to significantly increase our ability of characterize complex cellular populations. In addition, mass cytometry does not restrict the investigation to a single level of cellular metabolism. Indeed, proteins levels, posttranslational modifications and proteolysis products can be evaluated in a single experiment [Spitzer and Nolan, 2016].

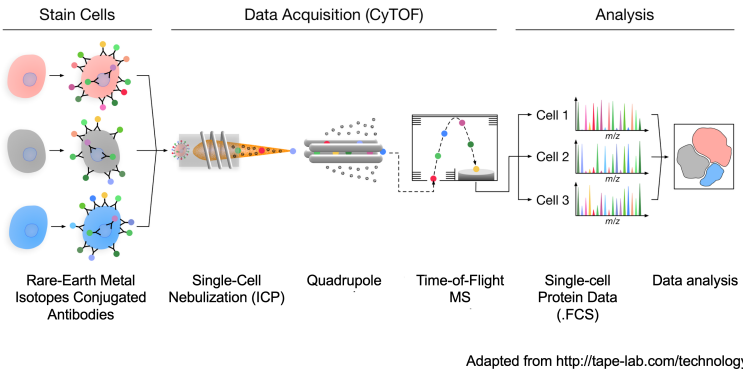


Figure 1: Mass cytometry uses antibodies labeled with isotopically enriched chelated metals to detect and quantify surface and intracellular proteins. Labeled cells are diluted in water, dispersed in droplets through a nebulizer and sent into an inductively-coupled argon plasma (ICP) to be vaporized, atomized and ionized. This process leads to the formation of ion clouds that contain the ions derived from single-cells metal-conjugated probes. Ions in ion clouds are filtered by a quadrupole and sent to the time-of-flight (TOF) mass spectrometer (MS) that separates ions on the basis of mass-to-charge ratio. Ions' profile from each single cell are recorded and compiled into flow cytometry standard (FCS) files [Bendall et al., 2011].

More recently, have been developed methods that control batch effects such as fluctuations in instrument sensitivity or other experimental variation [Lai et al., 2015; Finck et al., 2013]. In the first case, calibration beads are used as internal standards and spiked in samples to correct for temporal variations in signal intensity: both for the variations associated to different mass cytometer instruments and for the signal drift during acquisition [Abdelrahman et al., 2010; Finck et al., 2013; Newell and Cheng, 2016]. In the second case, mass-tag 'barcoding' (generally through CD45 for its ubiquitous expression on cell membrane) is used to simultaneously stain and acquire a large number of cellular samples (up to 20) at the same time and eliminate technical variability in both sample preparation and acquisition [Lai et al., 2015].

The main limitations of mass cytometry include the lower rate of cell acquisition (about 500 cells/sec) compared to flow cytometry (tens of thousands cells/sec), the high operating costs, the lower sensitivity for features expressed at very low levels and the inability to sort cells, since they are destroyed during the acquisition process [Newell and Cheng, 2016]. Despite that, CyTOF combines a high level of throughput, high resolution (at the single-cell level) and high parametrization which provides the advantage of capturing the complexity of cellular samples and the possibility of detecting and/or discovering rare cell populations that would not necessarily be analyzed using other single-cell technologies (*e.g.* single-cell RNA-sequencing).

A representative example highlighting the unique features of mass cytometry is presented by Sachs et al. [2014]. Application of the method led to the detection of rare cell populations in the context of an acute myeloid leukemia (AML) study. The main

challenge in finding stem cell abnormalities causing relapse in more than 60% of AML patients is the identification and classification of rare and heterogeneous cell populations. Using mass cytometry, Sachs et al. identified NRAS-dependent signaling changes, driven by mutations found in a large portion of those AML patients that present rare and aggressive subtypes of leukemic cells. They showed that NRAS had cell-type-specific effects, meaning that the same NRAS mutation could have different effects depending on the leukemic subpopulation. The marker combinations required to identify these rare stem cell populations and the examination of NRAS-dependent signaling needed more parameters than flow cytometry machines were able to support [Insights powered by mass cytometry, Fluidigm, 2020].

1.2 Relevance of rare cell populations identification

As previously mentioned, the comprehensive understanding of the samples' cell heterogeneity via mass cytometry can reveal previously uncharacterized immune cell types, help to understand their differentiation and function, and possibly provide new diagnostic biomarkers or novel therapeutic targets [De Biasi et al., 2017]. In particular, the identification and characterization of rare cell populations is of paramount importance for understanding the onset, progression and pathogenesis of diseases such as autoimmunity, immune deficiencies or cancer [Schreier et al., 2018]. Often, the health and disease status of the patients depend on minor group(s) of cells with frequencies largely below 1% of the total cell population [Proserpio and Lönnberg, 2016]. Consequently, the identification and quantification of rare cells can provide valuable information on the status of the patient, thus improve medical diagnostics. The subsequent characterization of these populations can improve not only the understanding of disease mechanisms, but also the definition of novel therapeutic targets [De Biasi et al., 2017]. The next paragraphs briefly discuss few examples.

Circulating endothelial cells (CECs, 0.01-0.001%) are mature endothelial cells found in the blood stream. The physical barrier between blood and tissue that the endothelium constitutes have important functional roles in trafficking regulation, coagulation and regulation of blood pressure. The detection of elevated CEC levels has been associated to different cardiovascular diseases, offering a non-invasive option to diagnose and eventually prevent cardiovascular diseases [Farinacci et al., 2019].

Endothelial progenitor cells (EPCs, 0.01-0.001%) are a rare cell population that moves from the bone marrow to the peripheral circulation to join sites of vessel injury in order to contribute to vasculogenesis [Li et al., 2012]. EPCs have been associated to neovascularization, thus defined as a biomarker of tumor angiogenesis [Farinacci et al., 2019; Mancuso et al., 2003].

Circulating tumor cells (CTCs, 0.001%) are used as an outstanding tool to evaluate the biology of metastatic cancers, monitor progression and are used as a relapse indicator [Castro-Giner and Aceto, 2020]. For example, the detection of measurable ("minimal")

residual disease (MRD) in AML patients posttherapy serves as a strong prognostic marker for the increased risk of relapse and shorter survival. It can also be used to define risk-stratification and assess treatment response [Ravandi et al., 2018].

Despite their low frequency, *antigen-specific T and B cells* are critical for the functional immune response as they are at the basis of cellular and humoral immunological responses, respectively [Sweedler and Arriaga, 2007]. In particular, the humoral immune response starts when antigen-reactive B cells encounter antigens in secondary lymphoid tissues (*i.e.* lymph nodes, spleen) where the antigen is delivered to the B cell surface and presented to T cells via the MHC class II [Cyster and Allen, 2019]. In the specific case of a T-dependent immune response, rare antigen-engaged B cells have to come in contact with rare cognate antigen-specific T cells. The interaction between B and CD4 helper T cells induces B cell proliferation, differentiation and the formation of germinal centers (GC) in the lymph nodes. In GCs the interaction between B cells and T follicular helper (Tfh) cells, promote B cell proliferation, affinity maturation, isotype class-switching and generation of long-lasting memory B cells and plasma cells. The importance of understanding B cells affinity maturation lies the capacity of responding to infections, vaccine antigens and generation of autoantigens. Consequently, the quantification of the resulting antigen-specific B cells allows to evaluate B cell compartment response when challenged by infection or vaccination and can provide important information about past exposures.

Invariant natural killer T (iNKT) cells are a population of specialized T cells representing less than 0.1% of peripheral blood mononuclear cells (PBMCs). iNKT cells are "innate-like" adaptive lymphocytes that have a crucial role in responding early during an infection (within hours compared to adaptive lymphocytes responses that takes days) [Krovi and Gapin, 2018]. These cells express both T lymphocytes (CD3, CD4, CD8) and NK (CD16, CD56) markers. Moreover, they are mainly characterized by the expression of invariant TCR ($V\alpha 24 - J\alpha 18$), which does not interact with peptides presented by MHC molecules but recognize self and foreign lipids as cognate antigen presented by CD1d [Krovi and Gapin, 2018]. Despite their very low frequency in peripheral blood, these cells have been shown to balance immune activation and tolerance through proliferation and release of pro-/anti-inflammatory cytokines. Thus, iNKT cells are involved in the pathogenesis of many diseases, such as autoimmune and allergic diseases and cancer [Brennan et al., 2013]. Defects in iNKT cells have been shown to predispose to autoimmune diseases due failure of immune regulation [Chen et al., 2015]. In particular, the reduced frequency and impaired function of iNKT cells in patients with systemic lupus erythematosus suggest that they play a protective role in autoimmunity [Hofmann et al., 2013]. Moreover, a subset of iNKT cells expressing CD4 and HIV-1 co-receptors CCR5 and CXCR6 has been shown to be susceptible to infection by HIV, thus constituting a reservoir for HIV [Motsinger et al., 2002]. In addition, a reduction in the number of iNKT in patients with AML has been linked to a poor prognosis with lower overall survival [Najera Chuc et al., 2012].

1.3 State-of-the-art methods in mass cytometry

1.3.1 The curse of dimensionality in manual gating

Traditional analysis of flow or mass cytometry data consists in manual gating, that involves the inspection of a series of 2-dimensions (*i.e.* markers) at a time to reveal information about cellular hierarchy and identify known cell populations. The major limitations of this technique are the variations in populations definition, which are associated with the level of expertise of the investigator, the difficulties in detecting unknown cell populations and the time required to inspect $p(p-1)/2$ (p, nb. markers) bi-plots. Moreover, by considering one or two markers at a time, we can lose a lot of information. Importantly, the mentioned limitations increase with the increasing of the dimensionality of the data and has been referred as the "curse of dimensionality". These terms refer to the difficulties associated with the exponential increase of space volume as the number of features of a dataset increase [Newell and Cheng, 2016]. Even if each dimension is considered only as a binary variable (*i.e.* cells considered positive or negative for each marker, and not considering intermediate states), the number of potential states in terms of markers combinations increases exponentially and rises to a number beyond a trillion when 40 dimensions are considered [Newell and Cheng, 2016].

Automatic methods

During the past decade, many efforts have been made to develop methods for the automatic and unbiased detection of cell populations in order to extract relevant information from the increasing complexity of the datasets. Most of these methods aimed at detecting already well characterized cell populations or discover unknown or not well-characterized cell populations (*i.e.* non-canonical cell subsets) that could explain or describe differences between control and disease samples.

The main reason for the development of automatic methods is the need of circumvent the subjectivity of manual gating, which introduces variability in the data as well as a lack of reproducibility. This is particularly important in the context of multi-center clinical trials, where data analysis has to be standardized, reproducible and comparable over time and centers [Finak et al., 2014] to be efficiently usable in all subsequent statistical data analyses.

In response to these needs, the **FlowCAP consortium** (Flow Cytometry: critical assessment of population identification methods) took the initiative to promote the development of computational methods for the automatic identification of cell populations. The first computational tools were designed for the *automated gating* of cell populations that aimed at support the application of similar operations to a collection of samples. These methods include flowCore [Hahne et al., 2009] and flowViz [Sarkar et al., 2008] (an others) that have been included in the Opencyto framework (a Bioconductor infrastructure), which extends these flow cytometry packages. Briefly, Opencyto implements a hierarchical automated gating pipeline for data-driven automated gating using a so-called gating template. These templates are panel-specific and allow to standardize the analyses of experiments with the same panel. However, it has been shown that such automated analysis has limitations in accurately identifying and quantifying rare cell subsets

[Hunter-Schlichting et al., 2020]. Although the increased automation of the gating analysis process, these algorithms-assisted methods require some kind of user information such as a pre-defined gating hierarchy (*e.g.* the gating template in Opencyto).

The efforts on automatic gating approaches were followed by a rapid growth in the number of supervised and unsupervised clustering and other computational methods for flow cytometry data analysis, most of which are currently applied to mass cytometry datasets as well. The methods described and discussed in the following sections include those that do not require prior information about the data, but identify cellular groups (*i.e.* cluster) in an agnostic manner in the high-dimensional space. The advantage of these methods is their unbiased and exhaustive way of analyzing the data and identify novel cellular phenotypes in a data-driven manner.

1.3.2 Unsupervised methods

The unsupervised clustering methods aim at organize single cells into consistent groups, called clusters, based on features similarities (*i.e.* markers expression profiles). Compared to single-cell RNA-sequencing (scRNA-seq) methods that processes between tens of thousands to hundreds of thousands of cells and deals with an extremely high dimensionality of the data ($p >> N$, p = features, N = number of observations), the algorithms applied to cytometry data have to manage up to millions of cells but with reduced dimensionality. Consequently, most of the methods for cytometry data do not require a dimension reduction step and generally can perform calculations on the original high-dimensional space. However, calculations on such a large number of observations require greater efficiency (in terms of runtime) compared to scRNA-seq data analysis methods. Contrary to scRNA-seq, cytometry data analysis is not faced to technical variation such as dropout, strong batch effect or to the large number of different data generation protocols available, which have big effects on the data characteristics [Duò et al., 2018]. As expected, the different nature of these single-cell measurements has an impact on the data analysis. Indeed, in Duò et al. [2018] the authors showed that unsupervised clustering methods specifically designed for scRNA-seq methods (*i.e.* SC3, Seurat) have overall best performance compared to methods developed for other single-cell data (*i.e.* FlowSOM) and more general approaches (*i.e.* hierarchical clustering, k-means).

In the context of mass cytometry, there have been a few studies [Weber and Robinson, 2016; Melchiotti et al., 2017; Liu et al., 2019; Saeys et al., 2016] comparing the performance of clustering methods in terms of accuracy, efficiency and stability, and/or summarizing their features. In the next chapters, we recapped and discussed the most used methods in mass cytometry and their applicability for rare cell populations detection based on literature or direct experience, with a major focus on unsupervised techniques since they allow for the discovery of not-previously described cell populations [Weber and Robinson, 2016]. Indeed, the separation of cells into clusters is exclusively based on cells' features profile, meaning that no prior knowledges are required to identify aberrations of the immune system.

As our interest is focused on methods aiming at detecting rare cell populations rather than major cell populations, we selected and discussed some of the best performing algorithms available [Weber and Robinson, 2016] for flow and mass cytometry and their

ranking in terms of performance (mean F1 score and runtime) (see Table 1). The F1 score is generally used to evaluate classification systems by considering both precision (number of true positive divided by the number of predicted positive results, *i.e.* $TP/(TP+FP)$, or positive predictive value) and recall (number of true positive divided by the number of relevant items, *i.e.* $TP/(TP+FN)$). The traditional F1 score is the harmonic mean of precision and recall:

$$F1 = \frac{2 \cdot precision \cdot recall}{precision + recall}, \tag{1.1}$$

Note, the dataset [Naim et al., 2014; Mosmann et al., 2014] used for the evaluation of the different methods in Weber and Robinson [2016] is a simple case analysis. Indeed, only one healthy sample was analyzed, which is not representative of experimental design complexity and the number of samples measured in disease studies. Moreover, the target rare cells are mostly localized in a distinct part of the t-SNE plot (Figure 2), suggesting a unique expression profile of the rare population compared to the rest of the sample.

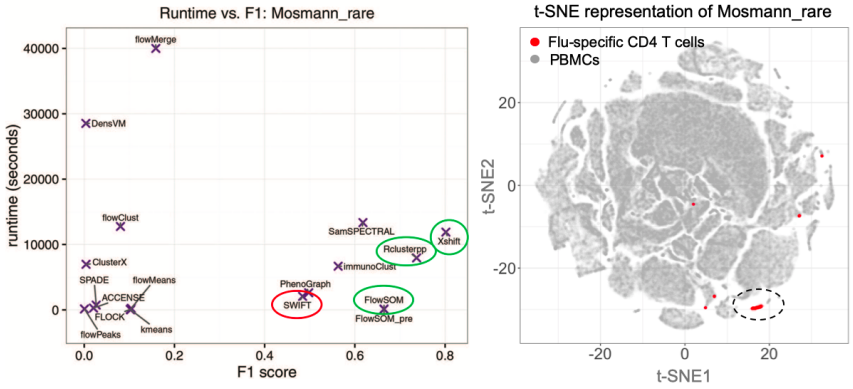


Figure 2: Left-plot: performance (F1 score) vs. runtime of unsupervised clustering methods for the detection of rare cell populations (adapted from Weber and Robinson [2016]). Green circles show the top-three methods for rare cell detection. Red circle shows a method designed for rare cells identification. Right-plot: t-SNE representation of Mosmann dataset. Grey dots represent 396'460 PBMCs and red dots represent 109 manually gated Flu-specific CD4 T cells (0.03%).

Consequently, the main limitations of the recent publications dedicated to the detection of rare cell populations in high-throughput single-cell data is that they were tested on datasets that do not accurately reflect the complexity of physiological cellular samples in terms of variable cell phenotypes, heterogeneous population compositions and the lower frequencies (<1%) of rare cells to be detected [Weber and Robinson, 2016; Jindal et al., 2018b].

Unsupervised clustering methods evaluation in Weber and Robinson [2016] showed that methods such as FlowSOM, X-shift [Samusik et al., 2016] or Rclusterpp [Linderman et al., 2013], which were not specifically designed for the detection of rare cell subsets, have the best performance [Weber and Robinson, 2016] (Figure 2). However, both X-shift and Rclusterpp require much higher computer power to complete an analysis. Typically, they can only be used with machines equipped with multiple processor cores and Rclusterpp in addition usually requires also subsampling to reduce the number of cells actually used in one run. Such a sub-sampling seems suited for detection of major populations but

Method	Description	Ranking	
		Major	Rare
FlowSOM^{M}	Self-organizing map & consensus clustering	**1**	**3**
flowMeans^{M}	k-means based	2	7
Xshift^{**,M}	weighted kNN density estimation	**3**	**1**
PhenoGraph^{M}	kNN graph	6	4
ClusterX^{*,M}	density-based clustering	4	6
Rclusterpp^{**,M}	large-scale hierarchical clustering	**5**	**2**
SWIFT^{*,**,R}	gaussian mixture models by expectation maximization	7	5

Table 1: Ranking of best performing clustering methods (accordingly to [Weber and Robinson, 2016] for both main and rare cell populations identification. Performance was established in terms of mean F1 score and runtime. Some methods require subsampling* and/or a multicore processor**. Methods designed for main^{M} or rare^{R} populations detection

is problematic when one is searching for rare cells as their identification is uncertain even in the whole dataset. Surprisingly, SWIFT (scalable Weighted Iterative Flow-clustering Technique) [Naim et al., 2014; Mosmann et al., 2014] that was designed for rare cell populations detection was found to be much less performant in terms of F1 score (despite its very short runtime) compared to the three previously mentioned methods in the data used by the authors.

FlowSOM

FlowSOM [Van Gassen et al., 2015] is a powerful tool originally designed for flow cytometry data analysis but now widely used in mass cytometry data analysis. It has been integrated into different analysis pipelines [Nowicka et al., 2017; Chen et al., 2016; Weber et al., 2019] given its accuracy and rapidity in identifying both major and rare cell populations [Weber and Robinson, 2016; Liu et al., 2019]. The FlowSOM algorithm consists in the construction of a self-organizing map (SOM) [Kohonen, 1990], where the points (*i.e.* single cells) are mapped to nodes (also called codes) of a 2-dimensional SOM-grid (typically a grid of $10x10$ units, each interpreted later as a cluster of similar cells thus calling "cell types") [Van Gassen et al., 2015].

The SOM aims at reproducing the topology of the data in the high-dimensional space by maintaining the same neighbors, and this starting from a distance matrix [Wehrens, 2007]. However, the concept of distances in SOM is different and cannot be interpreted as estimates of the true distances between points but as relatedness when mapped to the same node in the grid. Indeed, SOM focuses on the largest similarities between points rather than dissimilarities. This is of great advantage in the context of large single cell data sets because it ensures the grouping of cells with very similar profiles.

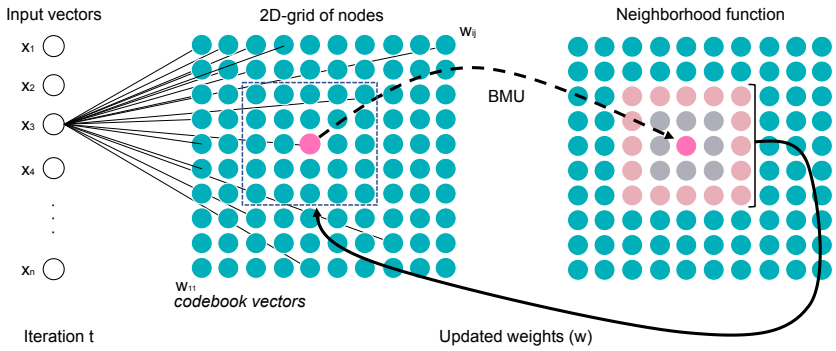


Figure 3: Scheme of SOM learning process: The *2*-dimensional grid of nodes (representing neurons of an artificial neural network) is connected to input data (x) by local coordinates (ij) and weights (w, codebook vectors). The closest point to the randomly selected input point is called best matching unit (BMU). Nearest points of BMU are identified by a neighborhood function and weights for these nodes are updated at each iteration t [Friedel and Iwashita, 2013].

The mapping of cells to the SOM nodes is an iterative process that starts by a training step, which consists in building the map using input information (Figure 3). The training begins by the random selection of k points in the dataset that are assigned to the k nodes (corresponding to the number of clusters, by default 100). This represents the assignment of the so-called codebook vector (vector of weights, which are the positions in the input space) to every node that defines the pattern or prototype of the node. At each iteration, a cell is randomly selected and the nearest node (*i.e.* best matching unit or BMU) to that cell is identified as well as nearest nodes using the Chebyshev distance (called neighborhood function). At this point, the selected node (and the corresponding neighbors with less extent) are updated to become more similar to the selected cell according to the learning rate, which decreases at each iteration until convergence of the map. The weight (w) for node j are updated at each iteration as follow (2):

$$\omega_{t+1}(j) = \omega_t(j) + \varepsilon_t \times h_t(i, j_{BMU}) \times (x - \omega_t(j)), \tag{1.2}$$

where ε corresponds to the learning rate (typically with a value of 0.05), h represents the neighborhood function around the BMU node (best matching unit) and x is the target input vector.

At the end of the process, each cell in the dataset is assigned to one of the hundred nodes corresponding to the cluster label of the final clustering. In this way, we identify a much larger number of clusters than the expected number of cell types. By adding a step of consensus clustering (CC) on the hundred nodes, we obtain a smaller number of (meta-)clusters. The FlowSOM package provides as clustering method the consensus hierarchical clustering (metaClustering_consensus function), which is iteratively applied to a subset of nodes. The process by which the subsampled nodes are partitioned into one of the numbers of meta-clusters is repeated hundred times in order to calculate a consensus rate between all pairs of nodes [Şenbabaoğlu et al., 2014]. The resampling

procedure simulates perturbations in the data and is used to assesses the stability of the meta-clusters. Consensus rate is defined as the proportion of pairs of nodes that are grouped together in several subsamples, which represents the agreement among the multiple iterations.

diffcyt

The diffcyt [Weber et al., 2019] framework incorporates FlowSOM for the differential discovery analyses in high-dimensional cytometry data. The authors proposed the combination of high-resolution clustering, which consists in the definition of an extremely high number of clusters (*i.e.* from 400 to 1600 clusters tested) by increasing the total size of the 2d-grid of nodes in the hope that one of these clusters coincides with the cells to be discovered, and the application of statistical approaches, originally developed for transcriptomic applications (*e.g.* edgeR [Robinson et al., 2010], voom [Law et al., 2014]) to find those clusters that are differentially abundant between study groups. To test the capacity of the diffcyt approach in detecting rare cell populations the authors used in silico spiked-in semi-simulated datasets. They used mass cytometry bone marrow mononuclear cells (BMMCs) from healthy donors that they split into two parts; one half was used to define the healthy group of donors and the second half to simulate minimal residual disease (MRD) in acute myeloid leukemia (AML) patients. For the MRD group they added real experimental AML blast cells at different frequencies (5%, 1% and 0.1%), which represented the ground truth signal to be detected. In order to test the capacity of the approach in detecting the target rare cells they defined and tested 400 clusters for differential abundance between the two groups and used ROC curves to assess the quality of this detection method. They showed that the diffcyt method succeeded in detecting the blast cells as a cell population specific for the MRD group, and also that their detection is more difficult at low frequency. In particular, the authors reported very low precision (PPV = 0.25) in detecting AML blast cells at 0.1%, meaning that the relevant cluster was strongly "contaminated" by non-blast cells, which hinder the identification of their biological role. At the higher tested frequencies of 5% and 1%, the AML blast cells were split into different clusters, meaning that the use of FlowSOM does not guarantees the identification of a unique cluster containing the target cells. An additional difficulty of the approach is that one has to adapt the number of clusters used to the data, and there is no simple method to know how to do so.

cydar

Cydar [Samusik et al., 2016] represents a computational approach that aimed at detecting differentially abundant cell populations by assigning cells to overlapping hyperspheres (spheres in the high-dimensional space, p-dim hyperspheres where p is the number of features) that are then tested for significant differences between study groups and controlled for the so-called spatial false discovery rate. Hyperspheres are centered on an existing cell and its radius is $0.5\sqrt{(p)}$ to balance the increasing sparsity of the data as the number of features increases. Then, cells are assigned to the hyperspheres (or clusters for analogy to the other methods) within the specified radius and then the hyperspheres are tested for significant differences between study groups. The testing is performed using nega-

tive binomial models (implemented in the edgeR package from Bioconductor originally designed for bulk transcriptomic data) that account for the discrete nature of the cell counts per hypersphere. The resulting p-values per hypersphere are then used to control the FDR.

Importantly, edgeR assumes that input counts are filtered from low average counts hyperspheres. This because these hyperspheres do not provide enough evidence to reject the null hypothesis (H_0: there is no change in the average counts between study groups within each hypersphere) even if they contain consistent changes in abundance. Therefore, the removal of low average counts hyperspheres reduces the total number of tests and the severity of multiple test correcting. The discard of such low average counts hyperspheres can potentially have negative effect on the capacity of detecting rare cell populations. Since there is no power to detect group imbalances in hyperspheres with low average counts, these are not considered in the tests, thus reducing the severity of the necessary multiple testing correction. The discard of such low average counts hyperspheres can potentially preclude the detection of some rare cell populations.

1.3.3 Supervised methods

The second groups of algorithms include supervised machine learning clustering methods, which are based on biological or clinical information (*i.e.* external variables) describing the study groups of a dataset. These methods focus on the identification of cellular correlates of an independent biological variable (*e.g.* biological condition or clinical outcome) [Aghaeepour et al., 2016]. With that intent, these methods rely on annotated training sets as input in order to identify patterns associated with groups of cells that at best predict a sample's group (or end-point of interest). Generally, supervised methods require different steps of analysis; firstly, markers' measurements for each cell events (*i.e.* dependent variable) are used to assign cells to a cell type, then each cell type is tested – based on its features – for an association with the external variable. In this way, markers are used to train a model on the basis of a test dataset, in which it learns and identifies patterns in the data that are used to define groups of cells, which are subsequently tested for association. The resulting model provides information about the data (*e.g.* the cell types that are associated with an external variable) that can be further investigated. These types of analysis approach seek at stratifying cell sub-populations whose abundance or behavior is correlated with an end-point of interest. For instance, the identification of cell populations whose frequency predicts a disease status, the effectiveness of a treatment or can predict patient outcome and survival. Currently, supervised methods are a minority compared unsupervised techniques. The most famous methods include Citrus, which is not adapted for rare cell detection (discussed below) and CellCNN, which is the only supervised method designed for the detection of rare disease-associated cell subsets.

Citrus

Citrus [Bruggner et al., 2014b] is an automated method for the detection of stratifying cellular sub-populations, which combines unsupervised clustering and machine learning supervised association testing. Clusters of cells are identified using agglomerative hierarchical clustering on down-sampled data (fixed number of cells per sample). Data

down-sampling is required (by default 5'000 cells per sample) due to the runtime limitation associated to performing hierarchical clustering, which represent a major limitation for the identification of rare cell populations. After that, clusters are selected based on a minimum frequency criterion and examined for subsequent analysis. To identify clusters with stratifying signals Citrus uses regularized regression or classification models to selected features that are predictive and thus associated to an outcome of interest, these features include cluster abundance (frequency) or median marker expression. In principle, the use of agglomerative hierarchical clustering could be a valid option to identify rare cell. Populations at the lower levels of the hierarchy. However, the necessity of down-sampling the data due to the computational limitation of the algorithm considerably reduces the chances that the randomly selected down-samples contain (enough) rare cell events from the same population to be identified.

CellCNN

CellCNN [Arvaniti and Claassen, 2017] is a supervised method specifically designed for the detection (and/or discovery) of rare cell populations that are associated with disease status. It uses convolutional neural networks to identify cell subsets that differ in terms of frequency between study groups without considering all the input features but using filters that correspond to molecular profiles (*i.e.* combinations of markers expression that do not necessarily correspond to known cell populations). CellCNN has been shown to have overall comparable performance with the *diffcyt* approach in the detection of differentially abundant rare cell populations.

1.4 Anomaly detection algorithms

1.4.1 General introduction

Probably the first definition of anomaly was given by Grubbs (1969): "An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs". Anomaly detection algorithms are used within diverse research areas and have different application domains for the identification of observations that deviate from the norm, so they are by definition "rare" compared to "normal instances". Many of the algorithms were specifically developed for certain applications, while other are more generic. Historically, one the main reason for identifying anomalies (or outliers) was their removal because of algorithms' sensitivity to these observations in the data [Goldstein and Uchida, 2016a]. Generally, anomalies are divided into global anomalies (formally called collective anomalies) and local anomalies (formally called contextual anomalies); Figure 4 gives a schematic representation of the difference between the two concepts [Goldstein and Uchida, 2016a]. Observations that are very different from all other observations in the dataset are considered global anomalies. These could for example be due to a defective measurement or an unexpected contamination of the objects under study. Therefore they are frequently relatively easy to identify, at least in principle. While local anomalies are considered as such only when a close-by neighborhood of the data is considered.

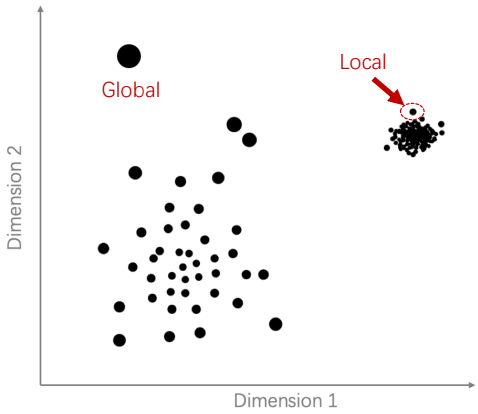


Figure 4: Simple schematic 2-dimensional example showing the difference between global and local anomalies.

The main challenge of anomaly detection includes the definition of normal regions, since boundaries between normality and anomaly is often not well-defined. An additional challange is the distinction of subgroups caused by measurement noise in the data from authentic subgroups with bona-fide reproducible anomalous characteristics. The applications of anomaly detection algorithms are wide and typically include intrusion detection (e.g. network traffic and server applications), fraud detection (e.g. financial transactions, credit card payments, insurance claim, insider trading detection), data leakage prevention (e.g. sensitive information protection) or industrial damage detection (e.g. faults in mechanical units or structural detects) [Goldstein and Uchida, 2016a]. In the medical

context it comprises patient monitoring (*e.g.* ECG or EEG signals) or image analysis (e.g. computed tomography), where the cost of wrongly classifying an anomaly as normal can be very high.

1.4.2 Classification of anomaly detection methods

Anomalies detection methods are divided into two major groups: supervised and unsupervised Methods (see Figure 5). The first group consists of methods that use a set of labeled observations to identifiy the features that can distinguish normal from anomalous observations. This group also include semi-supervised methods, where the labels of only normal instances are provided. The identification of such features includes the training of a model (i.e. classifier) that allows then to classify new unlabeled observations. The second group of methods includes unsupervised anomaly detection algorithms are used to detect and/or discover rare observations of unknown nature. In this case, the input of the algorithm consists in unlabeled data. Unsupervised anomaly detection techniques generally compute an anomaly score (or degree of anomaly, that is a continuous score) for each observation in the dataset.

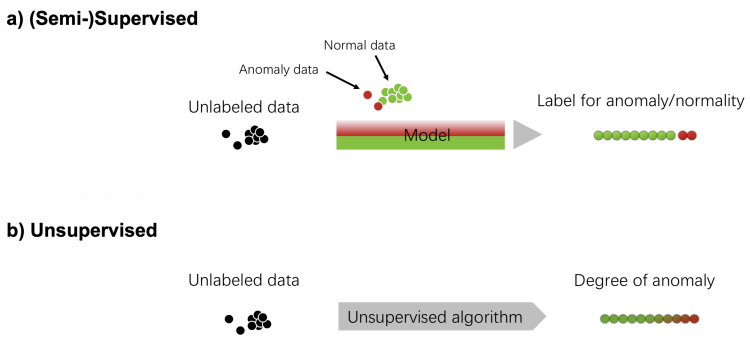


Figure 5: Anomaly detection categories: a) supervised (and semi-supervised) techniques uses both normal and anomalous labeled observations (fully labeled) or only nomal labeled observation to train the model, b) unsupervised techniques does not need any labels, indeed these methods are uniquely based on the intrinsic properties of the data.

Since this study is focusing on the detection and eventually discovery of rare observations in cytometry data, we focused our attention uniquely on unsupervised anomaly detection algorithms. Figure 6 summarizes the three main groups of unsupervised anomaly detection principles: nearest-neighbors based, clustering-based and statistical methods, more closely described in the following paragraphs.

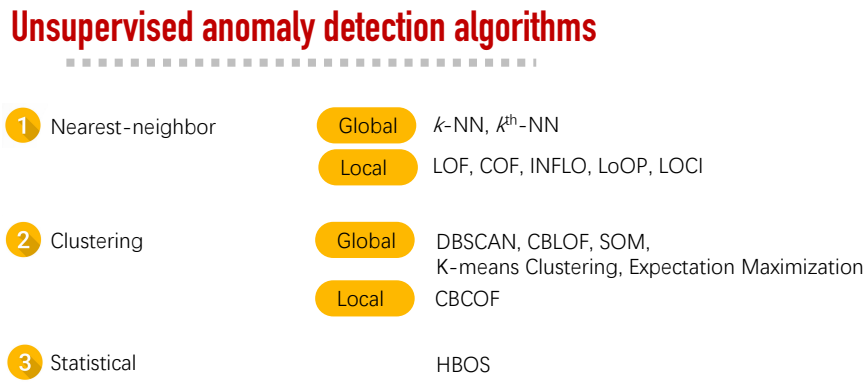


Figure 6: Overview of unsupervised anomaly detection algorithms. Local methods from nearest-neighbor based algorithms include LOF (local outlier factor, from dbscan package allows entering arguments of kNN algorithm), COF (connectivity-based outlier factor), INFLO (influenced outlierness), LoOP (local outlier probability), LOCI (local correlation integral). The additional kNN methods from DDoutlier package in R include the KNN-AGG (aggregated k-nearest neighbors distance over different ks), KNN-IN (n-degree for observations in a k-nearest neighbors graph) and KNN-SUM (Sum of distance to k-nearest neighbors).

Nearest-neighbor (NN) methods are again divided into two groups depending on the targeted anomalies: global or local. NN methods that target global anomalies include those using a distance (or similarity) measure to identify the k^{th}-NN (a single observation) or average distance to all k-NNs [Angiulli and Pizzuti, 2002] . NN methods targeting local anomalies use measures of relative density instead of a distance measure. Density-based anomaly detection techniques estimate the density around each observation by taking in consideration a neighborhood of the data points. An observation in a low-density region is considered anomalous, while an observation in a high-density region is considered as normal. The distinction between normal and anomalous observation is determined based on a score that describes the density in its neighborhood. The neighborhood is composed by the k nearest-neighbors respectively to the hypersphere centered at the given observation and with radius given by the distance to the k^{th} nearest-neighbor. This distance represents an estimate of the inverse density in this space; the higher the distance between the given observation and its k nearest-neighbors the lower the density around the given point is. Then, the estimated densities are used to compute the anomaly score, that consists in the ratio of the density of a given observation and its neighborhood.
The second group of unsupervised anomaly detection techniques consists of **clustering-based methods**. These methods are divided into different categories depending on the type of assumption that the algorithm is based on. Clustering algorithms that do not force all observation to belong to a cluster (*e.g.* DBSCAN [Ester, 1996]), consider observations not assigned to a cluster as anomalous. However, these methods are designed to optimize the quality of the clusters and not to find anomalies. Other methods calculate distances to clusters' centroids to obtain an anomaly score, which assumption is based on the fact that anomalies are far away from the closest cluster centroid. Finally, there are clustering methods that consider as anomalous any small cluster.
The third main group of unsupervised anomaly detection algorithms include **statistical based methods**. This group of methods is probably the less adapted to biological

applications given the requirement of assumptions on data distribution. In few words, statistical techniques fit a model that is based on normal behavior of the data and then preform statistical inference to evaluate if an observation belongs or not to the defined model.

In [Goldstein and Uchida 2016], the authors have shown that NN-based algorithm have overall better performance compared to clustering-based algorithms, especially for local anomalies identification. These methods were tested on ten datasets of varying dimensions and frequencies of rare instances. The tested datasets include breast cancer images (separate cancer to healthy patients, $N_{normal} = 367$, $N_{rare} = 10$, $p = 30$), pen-based recognition of handwritten text (digits from different writers, few digits kept at low frequencies, $N_{normal} = 6724$, $N_{rare} = 10$, $p = 16$), speech accent data (normal consists in American accent and anomalies other accents, $N_{normal} = 3686$, $N_{rare} = 61$, $p = 400$), satellite images ($N_{normal} = 5100$, $N_{rare} = 75$, $p = 36$), thyroid disease features (distinguish healthy non-thyroid and thyroid patients, $N_{normal} = 6916$, $N_{rare} = 250$, $p = 21$), detection of abnormal radiator flow in a NASA space shuttle ($N_{normal} = 46464$, $N_{rare} = 878$, $p = 9$), object images taken under different light conditions and viewing angles shuttle ($N_{normal} = 50000$, $N_{rare} = 1508$, $p = 27$), and intrusion detection of a computer network environment shuttle ($N_{normal} = 620098$, $N_{rare} = 1052$, $p = 38$). In general, Goldstein and Uchida 2016 showed that the performance of NN-based methods was better in most of the tested datasets compared to clustering-basted algorithms. They also observed a much higher stability of the results even when a not-perfect k is selected, probably due to the non-deterministic nature of the tested clustering-based methods. The advantage of clustering-based methods is the lower computation time. Consequently, the authors recommend the use of NN-based methods if precision is a more important issue than computation time. As mentioned, NN-based algorithms have higher computation time. In fact, the computation complexity of finding the nearest neighbors is $O(n2)$, while the remaining computations can be neglected ($< 1\%$ of runtime) [Goldstein and Uchida 2016]. According to their analysis, the authors concluded that among the NN-based methods, the best performing algorithm for local anomaly detection was the *local outlier factor* (LOF) when the target rare observations involve local anomalies. At present, to the best of our knowledge, there is no publicly available benchmark study testing local anomaly detection algorithms on single-cell data such as cytometry data, consequently part of this study involves evaluation of anomaly detection methods on these data. Based on the remarkable paper reported above [Goldstein and Uchida 2016], we decided to limit the study to NN-based algorithms, and, for feasibility, we only considered methods that were readily available in R software packages. The advantage of NN-based over statistical-based methods is the absence of any assumption about the data (e.g. distribution, independency of variables) that statistical methods require. The only assumption of NN-based methods is that normal observations are found in dense regions, while anomalies occur in low density regions, as previously described. In particular, density-based approaches have the advantage of considering the different densities of the neighborhoods in the data.

1.4.3 Local outlier factor

Local outlier factor (or LOF) [Breunig et al., 2000] is the most popular anomaly detection algorithm and the first introducing the concept of *local* anomaly.
Formal definition of local outliers - the basic idea of local outlier factor is that being outlier is not a binary property, instead each observation has an outlier factor that has a degree of being outlier.
LOF is based on the concept of local density that is defined by the number of nearest neighbors (k). Comparing local density of a data point x to the local densities of its k-nearest neighbors (k-NN) allows identifying regions of similar and lower density, thus outliers.
The distance of a data point x to the k-NN ($k - distance(x)$) is the distance $d(x, o)$ and includes all data points that are at the same distance, consequently could include more data points than the k value (Figure 7). This set of points at the same distance from x are notated as $N_k(x)$. The $k - distance(x)$ is used to define the reachability distance (RD) (Equation 1.3). The RD of x from o is the true distance ($d(x, o)$) of the object o but at least the $k - distance(o)$. In other words, if the data point o is far away from x, then the RD is the actual distance between the two data points, if o is "sufficiently" close to x, then RD is replaced by the $k - distance(o)$. This smoothing effect allows to reduce fluctuations in results and is controlled by the number of designated k_s; the higher the number of k value the more similar the RD for points within the same neighborhood.

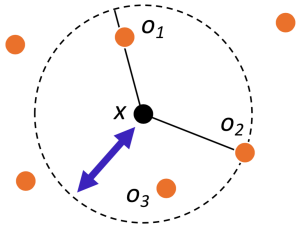


Figure 7: Reachability distance according to the number of k-nearest neighbors.

$$RD_k(x, o) = max\{k - distance(o), d(x, o)\} \tag{1.3}$$

The local density of a data point is then estimated by the distance at which it can be "reached" from its neighbors. The local reachability density (LRD) is defined in Equation 1.4.

$$LRD_k(x) = 1/(\frac{\sum_{o \in N_k(x)} RD_k(x,o)}{N_k(x)}) \tag{1.4}$$

In order to detect density-based outliers, local reachability densities (LRD) are then compared with those of the neighbors by means of Equation 1.5. LOF for a point x is the average local reachability density of the k neighbors ($LDR(o)$) divided by its own

LDR.

$$LOF_k(x) = \frac{\sum_{o \in N_k(x)} \frac{LDR(o)}{LDR(x)}}{N_k(x)} \tag{1.5}$$

LOF values of 1 corresponds to a data point with density comparable to its neighbors, while increasing values (> 1) indicate regions of lower density.

1.4.4 Other NN-based methods

The aim of this study was not to provide a comprehensive evaluation of all anomaly (or outlier) detection algorithms available, consequently we focused on well-documented methods and on those available as an R package. We based our selection on literature search with major emphasis on a study proposing a comparative evaluation of unsupervised anomaly detection algorithms for multivariate data [Goldstein and Uchida, 2016a]. In this publication the authors reported nearest neighbors-based methods to be better performant compared to clustering-based methods. We excluded statistical-based methods for the required assumptions on the data, because we might not have enough knowledge or information about the analyzed data and their underlying distribution. The dbscan and DDoutlier packages propose different nearest-neighbors outlier detection algorithms for multi-dimensional (multivariate) datasets, most of those are distance or density based. We will briefly describe the methods that were tested in order to be able to select candidate methods that we tested on single-cell mass cytometry datasets.

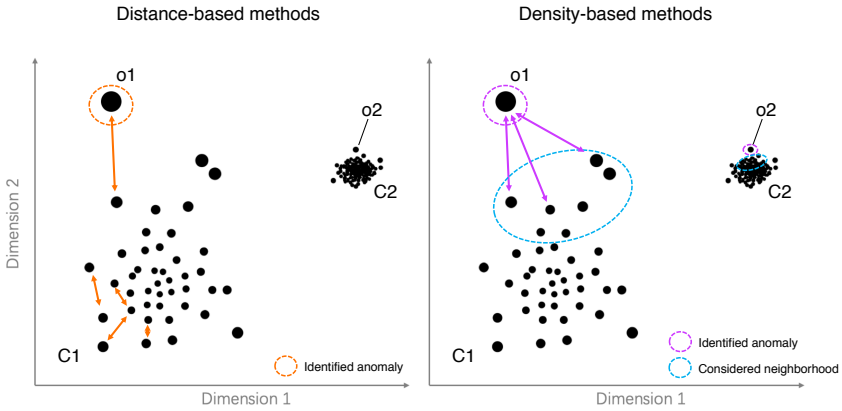


Figure 8: NN-based methods schematic representation of anomaly identification. Distance-based methods (left representation) uses distances calculated between each pair of points as anomaly score. The distances calculated between point o1 and points in cluster C1 or the distances between points within cluster C1 are larger compared to the distance between o2 and points in cluster C2. Meaning that o1 and points in C1 will have high score compared to point o2 and points in C2. Density-based methods (right representation) uses a close-by neighborhood of the points to estimate densities around each point, thus able to deal with neighborhoods of different densities (*i.e.* different densities of C1 and C2). Consequently, density-based techniques identify o2 as an anomaly as well.

Distance-based methods - these methods are based on the k-NN algorithm (from dbascan package) in order to find the k-nearest neighbor observations of all the points in the dataset. Since these methods do not deal with neighborhoods of different densities, they are considered global anomaly detection techniques (see Figure 8, left). The distance measure for an observation to the k-NNs is interpreted as follow: observations with low distance have a dense neighborhood, while observations with high distance have sparse neighborhoods, thus deviating characteristics from dense regions and considered as anomalies (or outliers). We tested three variants of these methods called; k-NN sum, k-NN aggregation and k-NN in-degree. The k-NN sum method [Jacob H. Madsen] sums the distance to the k-NNs and use the resulting values as anomaly or outlier score. The k-NN aggregation was thought by Angiulli and Pizzuti [2002] to reduce the subjectivity of selecting the appropriate number of neighbors (k) and consists in aggregating the distances (*i.e.* sum of the distances) to the k-NNs over a range of k values and use the resulting value as outlier score. Finally, the k-NN in-degree method compute the in-degree for each observation, that represent the number of reverse neighbors (retrieves all points that have the given observation as their nearest neighbor) [Hautamaki and Ismo, 2004]. The smaller the in-degree, the greater the degree of outlierness of the observation.

Density-based methods - these methods are also based on the k-NN algorithm but estimates the density around each observation based on the distances of the k neighbors. These methods overcome the drawback of distance-based methods that consists in do not dealing with neighborhoods of different densities (see Figure 8, right). In addition of the

LOF method (available in both dbscan and DDoutlier packages), the DDoutlier package provides functions for the following (tested) methods: INFLO, COF, LOCI, LoOP and KDEOS (see Figure 6). The connectivity-based outlier factor (COF) [Tang, Chen, Fu and Cheung, 2002] is similar to LOF, but the density estimation for each observation is performed using a shortest-path approach (*i.e.* chaining distance, is the minimum sum of all the distances that connect an observation and its k nearest-neighbors). Then, the chaining distances are compared between observations as a ratio of distances, which corresponds to the anomaly score. The influenced outlierness (INFLO) method uses in addition of the k-NN method a reverse nearest-neighborhood set, that are then combined. Then, the local density is calculated in the same way as LOF [Goldstein and Uchida, 2016a]. The local outlier probability (LoOP) produces as output an anomaly probability instead of a score as the previously described methods. The advantage of this method is the possibility of comparing anomalous observations across different datasets [Goldstein and Uchida, 2016a]. The assumption behind this method is that distances to the nearest-neighbors follow a normal distribution, and given the positivity of distances LoOP assumes half-normal distribution and calculates standard deviations as probabilistic set of distance, that are used as local density estimation. The LoOP anomaly score consists of a ratio between each observation and its k-nearest neighbors. The local correlation integral (LOCI) [Papadimitriou, Gibbons and Faloutsos, 2003] address the inconvenience of defining the number of neighbors to use. It uses a maximization approach in order to include all possible k values by defining the radius-neighborhood (that is expanded over time) and then use the maximum score. As LoOP, it estimates local density using the half-normal distribution of the amount of observations in the neighborhood instead of distances [Goldstein and Uchida, 2016a]. However, the local estimation is calculated by comparing two neighborhoods of different size instead of the ratio of local densities.
The previous methods that use the k-NN algorithm to find the k-nearest neighbors have typically a computational complexity of $O(n2)$, and the additional operations are negligible. On the other hand, LOCI includes an additional step of expanding the radius, which increases the computational complexity to $O(n3)$. Finally, the kernel density estimation outlier score (KDEOS) [Schuber, Zimek and Kriegle, 2014] calculates a kernel density estimation over a range of k values. The resulting anomaly score is normalized between 1 (lowest density estimation) and 0 (higher density estimation).

1.4.5 Finder of Rare Entities

In Jindal et al. [2018], the authors proposed a new method for the identification of rare cells in single-cell RNA-sequencing data called FiRE (Finder of Rare Entities). FiRE is an unsupervised anomaly detection technique that computes a rareness score to individual cells without learning from a set of labeled examples. The rareness score is an estimate of the density around each point in the multi-dimensional space using the Sketching technique, which is a ranked-based analysis for similarity search. The Sketching technique consists in compacting data structures into features vectors by encoding each data points into bit vectors (values 0, 1). These vectors are constructed by applying a threshold to the expression profile (values between the minimum and maximum expression value). In addition, when variables are very numerous, vectors can be shortened by randomly

selecting only a subset of the variables (defined by the parameter M). Then, distances between features vectors are calculated using the Hamming distance, which consists in the number of bit positions that are different (Figure 9).

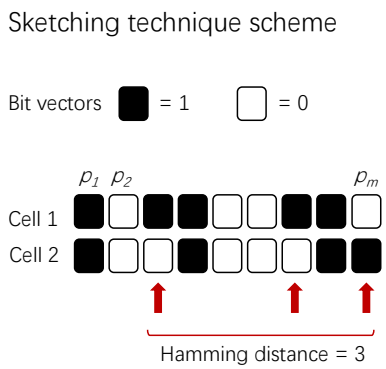


Figure 9: Scheme describing the principle of the Sketching technique. In the scheme are shown bit vectors (taking values in the set $\{0, 1\}$) for two cells. Bit vectors are defined based on a randomly selected threshold of markers expression (p_n), if the cell express higher levels a 1 is assigned, else a 0. Then, the distances (*i.e.* similarities or dissimilarities) between bit vectors consists in the number of bit positions in which the two bit vectors are different (shonw by the red arrows). The Sketching technique is used for compacting data structures and approximate distances between high-dimensional points (*i.e.* cells).

This can be done for all pairs with very fast computer code. When variables subsets are used the process is iteratively repeated (parameter L) for different variables subsets. This generates hash codes (also called buckets of cells, which contains cells that are close in the high-dimensional space). Then, bit vectors are mapped to the hash codes. The density is estimated for each bucket as the number of cells in it divided by the total number of cells. The density estimate corresponds to the probability that a randomly picked cell is assigned to the bucket containing that cell. Buckets containing large amounts of cells correspond to large clusters while buckets containing few cells correspond to rare cells. The FiRE score is calculated as follow: Figure 10 shows the performance assessed for FiRE, LOF and two other methods for single-cell RNA-sequencing in Jindal et al. [2018], for the identification of rare events. The authors showed that two methods, GiniClust and RaceID, are computationally expensive and failed in detecting the majority of the labeled rare events (as shown in Figure 10, left). While FiRE and LOF performed well for increasing frequencies of "rare" target cells in this dataset. However, the application of LOF to transcriptome datasets is not adapted to the size of dimensionality as previously reported by [Goldstein and Uchida, 2016a]. One of the limitations of the testing dataset used by authors is the artificial nature of the data. Indeed, the simulated datasets used consist of human Jurkat T cells and human 293T embryonic kidney cells, which are very distinct cell populations between each other and very homogeneous among them. Moreover, these populations are cell lines, thus not representing a real biological sample. The nature of the data used allow the cell populations separation quite straightforward even without the use of such a specialized method (as shown in Figure 10, right).

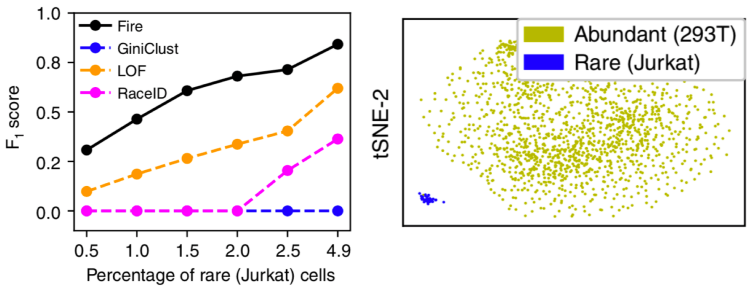


Figure 10: Figures from [Jindal et al., 2018]. Left-side plot shows the performance (F1 score) as a function of the frequency of rare Jurkat cells (human T cell line) among 293T cells (human cell line derived from human embryonic kidney cells) of FiRE and other scRNA-sequencing methods for rare cell detection, plus LOF. Right-side t-SNE plot shows the abundant 293T cells and rare cells (Jurkat).

1.5 Motivation and aims of the study

In the last decade, substantial efforts have been done with the objective of developing methods for the automatic detection and eventually discovery of unknown (or not well-characterized) cell populations. Currently, the analytical tools for the detection of major cell populations and subsets with mass cytometry applications are widely present and continue to increase. However, only few data analysis methods specifically designed for the efficient identification of rare cell populations are available. In particular, mass cytometry data present significant challenges due to the high number of data points (single-cell resolution and high-throughput) in the fairly high and growing dimensionality given by multiplexed panels. Because of this scarcity, this study aimed at implementing a straightforward approach that efficiently supports a data analyst to identify disease-associated rare cell populations in large and complex biological samples and within reasonable limits of time and computational infrastructure. In this study, we proposed a novel computational framework called D-AREdevil (disease-associated rare cells detection) for cytometry datasets (see section x for details). This new computational approach combines the use of two unsupervised methods; an anomaly detection algorithm called LOF (local outlier factor) [Breunig et al., 2000] and a clustering algorithm called FlowSOM [Van Gassen et al., 2015] (self-organizing map). In our approach, the LOF score serves to select a set of candidate cells belonging to one or more groups of similar rare cell populations. Then, the clustering method is used to identify subgroups within the selected set of cells and this allows to identify any subgroup that is associated with a patient group, disease type, clinical outcome or other characteristic of interest. The major objectives of this study included:

- Selection of candidate anomaly detection and unsupervised clustering algorithms based on the study of the current literature in the field
- Testing of the selected anomaly detection algorithms with the aim of determining the best performing one in the context of single-cell mass cytometry data
- Definition of performance criteria (ROC curves and AUC values, runtime)

- Testing on different datasets containing abundant and target rare observations in real experimental mass cytometry datasets
- Overall performance evaluation of the combined approaches in a classification setting, where manually gated target rare cells are used as reference-standard for testing
- The comparison and evaluation of the performance with similar method available for single-cell RNA-sequencing data (see Section 1.4.5).
- Propose a straightforward analysis pipeline (as a "way-to-proceed") for the approach application
- Apply and validate the approach on different datasets containing abundant and target rare observations in real experimental mass cytometry datasets

The main objective and challenge of this study was to find methods for the identification of rare cells among a specific cell population. For instance, identify B cells specific for a given known antigen among total B cells. Importantly, we aimed at targeting rare cells differ from abundant cell types by a restricted number of markers. In the language used by specialists in the field of "anomaly detection", such rare cell populations are referred to as local anomalies, because their peculiarity can be recognized only when a close-by neighborhood of data points is considered (see section 1.4.4 for details). In this study, we focused on local anomalies detection since their identification by state-of-the-art dimensionality reduction techniques or unsupervised clustering methodologies were unable to isolate these cells from abundant cell types. Indeed, using dimensionality reduction techniques, these cells showed to be poorly resolved with other data points in a 2-dim representation. On the other hand, rare cell types that correspond to global anomalies are very distinct both in terms of frequency and features profile compared to abundant cell types that can be easily identified (*i.e.* isolated) using clustering algorithms or visualized in a 2-dim plot. Currently, to the best of our knowledge, there are no published studies that integrate the LOF (or any other anomaly detection algorithms) for this task, within the context of mass cytometry. In this study, we reported the properties and implementation of the D-AREdevil computational framework and presented an evaluation of its performances and applications on three different spiked-in datasets based on real mass cytometry (experimental). We took advantage of real experimental mass cytometry datasets (two generated in our lab and one publicly available [Weber and Soneson, 2019]) to generate spiked-in datasets with one or more known rare cell populations at varying frequencies (all below 1%) and tested the ability of our approach to identify the labeled target cells in order to bring them to the attention of the data analyst. This is a key step in the process of finding cell subgroups that are associated with a disease or outcome of interest, when their existence and identification is not previously known and has yet to be discovered.

1.6 D-AREdevil (disease-associated rare cell populations detection) framework

1.6.1 Additional considerations on previous methods

The main limitations of the recent publications dedicated to the detection of rare cell populations in high-throughput single-cell data is that they were tested on datasets that do not accurately reflect the complexity of physiological cellular samples in terms of variable cell phenotypes, heterogeneous population compositions and lower frequencies (<1%) of rare cells to be detected. Indeed, most of the studies use (semi-)simulated datasets in which the separation of rare from the abundant cell types could already be visualized in plots obtained by standard dimensionality reduction techniques such as t-stochastic neighbor embedding (t-SNE) or tested detecting of cell populations with frequencies above 1%. One of the major aims of this study is to test and establish methods that can perform the detection of rare cells in cell populations where these are extremely underrepresented and which profile deviates from abundant cell types by only a few features; such as antigen-specific B cells among total non-naïve B cells. In these examples, even the best state-of-the-art dimensionality reduction techniques appear so far to be unsuited for detecting and isolate these rare cell events.

1.6.2 Overview of D-AREdevil

The D-AREdevil framework takes as input single cell expression data from samples representing different experimental or medical conditions (*i.e.* including samples from different study groups, time-points or datasets) that include major cell populations and potentially rare cell types, associated with disease status or other relevant clinical or biological information, to be identified. D-AREdevil proceeds in several steps (Figure 11). Firstly, (1) we apply an anomaly detection scoring algorithm (*i.e.* LOF, FiRE) to rank cells. Afterward, we take the set of top-scoring cells above a selected cutoff. By varying the cutoff, we can adjust the number of selected cells. Then, (2) we apply dimensionality reduction to visualize the selected rare cells and watch out for differences between conditions. On the reduced dimensions (3) we apply unsupervised clustering (*i.e.* FlowSOM) to group the selected cells into clusters with a similar profile. Finally (4) we test each of these clusters for differential abundance between conditions.

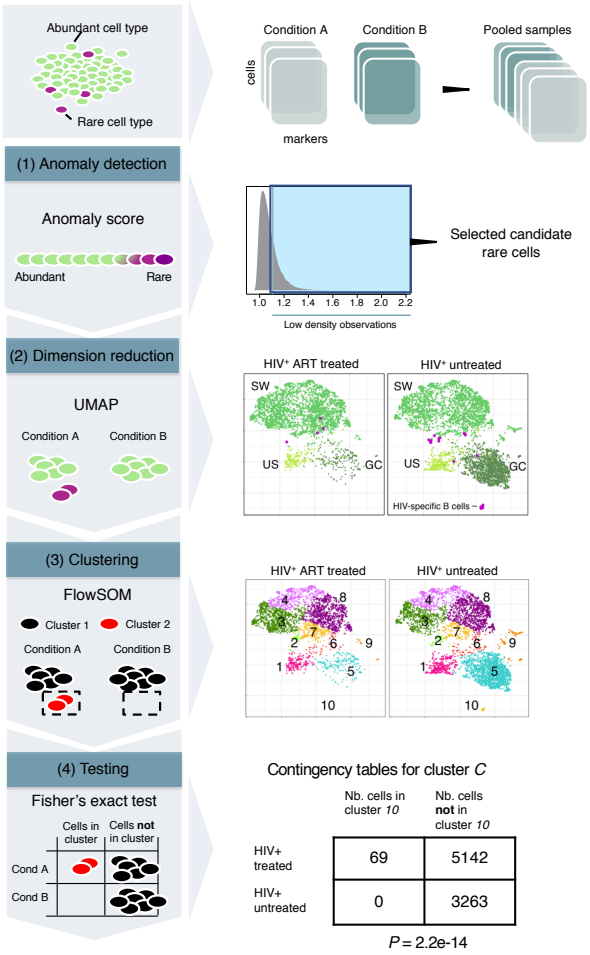


Figure 11: Schematic overview of analysis steps of the D-AREdevil methodology. The analysis starts with pooling and normalization of cells from all samples (i.e. including samples of different study groups, time-points or datasets). (1) Next, we apply an anomaly detection scoring (LOF) to rank cells. LOF computes the local density of each cell with respect of its k-nearest neighbors (the LOF score). Afterward, we take the set of cells above a selected LOF score cutoff. By varying the cut-off, we can adjust the number of selected cells. (2) Then, we apply dimensionality reduction to visualize the selected rare cells and watch out for differences between conditions. (3) On the reduced dimensions we apply unsupervised clustering (*i.e.* FlowSOM) to identify subgroups of cells with a similar profile and (4) test each of them for differential abundance between conditions.

Chapter 2

Results

2.1 Anomaly detection algorithms selection

This section of results shows how we selected the local outlier factor (LOF) algorithm among the several methods available in R software packages.

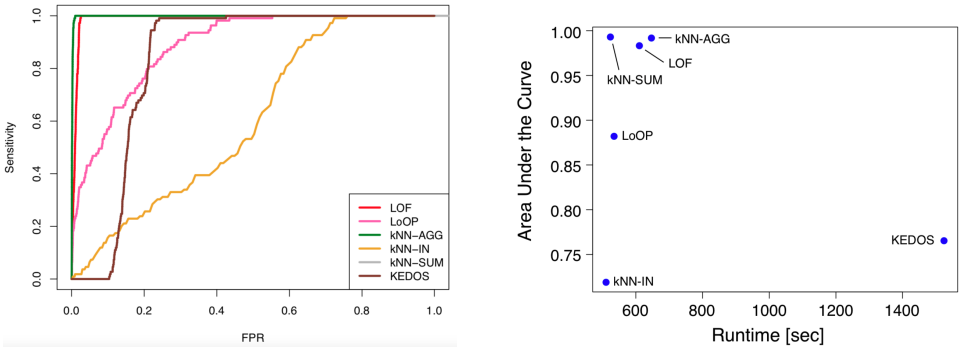
We tested the performance of different nearest neighbors-based anomaly detection algorithms on a two simple-case datasets created for benchmarking purposes to select the most interesting method.

The first datasest is a publicly available flow cytometry (*Mosmann dataset* [2]) dataset that has been previously used in Weber and Robinson [2016] to evaluate the performance of different unsupervised clustering methods for the detection of rare cell populations. It consists of PBMCs from healthy human donors stimulated *ex vivo* with a peptides pool of H1N1 strains of influenza A. The PBMCs were antibody-labeled using 14 surface and intracellular markers. The data we used here were already pre-processed Weber and Robinson [2016] and consists of 396'460 PBMCs from one healthy donor and 109 influenza-specific memory CD4 T cells (representing 0.03% of the total dataset).

The second dataset has been described in the manuscript at the end of the thesis (*HIV-specific B cells dataset*). Briefly, it consists of LNMCs from ART treated and untreated HIV^+ patients that were analyzed by mass cytometry (CyTOF) using a B cell panel of 32 surface and intracellular markers. It contains 79'886 non-naive B cells and 80 HIV-specific B cells (representing 0.1% of the total dataset).

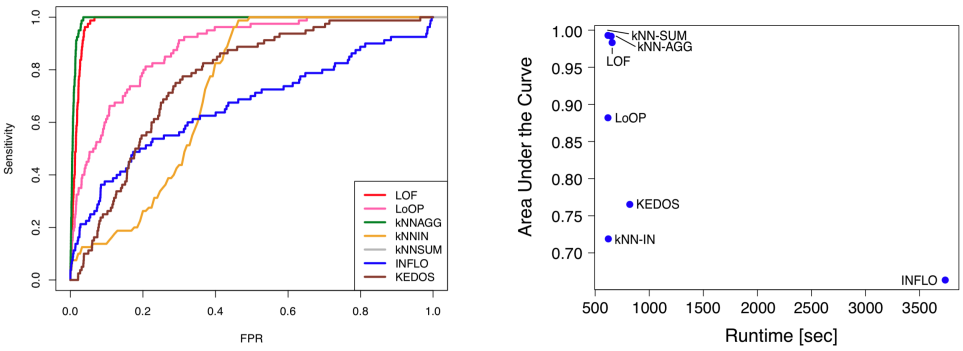
Figures 12-13 show ROC curves (a) and AUC values as a function of runtime (b) for the tested algorithms. Note, results for COF and LOCI were not available due to the too demanding memory requirements and we discarted INFLO due the too high computation time observed when ran on *Mosmann* and *HIV-specific B cells* datasets.

[2]Downloaded from: https://flowrepository.org/id/FR-FCM-ZZPH



(a) ROC curves (b) AUC values vs. runtime

Figure 12: Performance of different nearest neighbors-based anomaly detection algorithms: LOF (local outlier factor), INFLO (influenced outlierness), KDEOS (kernel density estimation outlier score), LOOP (local outlier probability), KNNAGG (aggregated k-nearest neighbors distance over different k's), KNNIN (in-degree for observations in a k-nearest neighbors graph), KNNSUM (sum of distance to k-nearest neighbors) on *Mosmann* dataset.



(a) ROC curves (b) AUC values vs. runtime

Figure 13: Performance of different nearest neighbors-based anomaly detection algorithms: LOF (local outlier factor), INFLO (influenced outlierness), KDEOS (kernel density estimation outlier score), LOOP (local outlier probability), KNNAGG (aggregated k-nearest neighbors distance over different k's), KNNIN (in-degree for observations in a k-nearest neighbors graph), KNNSUM (sum of distance to k-nearest neighbors) on *HIV-specific B cells dataset.*

Overall, the testing on the three datasets show that the best performing algorithms were LOF (in red), a density-based local anomaly detection method, k-NN sum (in grey) and k-NN aggreagation (in green), two distance-based global anomaly detection methods. Indeed, the three methods showed the highest AUC values and shortest runtime. Based on these results and previous studies that reported the higher performance of LOF to accurately identify local anomalies, we focus further evaluations on LOF algorithm rather than global anomaly detection techniques.

2.2 Paper I : A novel computational approach for discovering disease-associated rare cell populations in cytometry data (unpublished)

In this manuscript we proposed and described the application of a novel multi-step computational framework called D-AREdevil. This methodology was developed to identify and discover rare cell populations that are associated with a disease or condition of interest. To this end, our strategy was to take advantage of anomaly detection algorithms (*i.e.* LOF and FiRE, described in the Introduction) to select a set of candidate rare cells for further investigation. The identification of rare cells subsets is then performed by clustering the set of candidate rare cells to find sub-groups that are then tested for assiciation.

The design of the methodology is implemented to deal with the high throughput of cytometry dataset, where the detection of minor cell populations ($< 1\%$) remains challenging. We presented step-by-step the methodology application on three different test datasets. The test datasets were generated by keeping low frequencies of manually gated rare cell populations (0.1, 0.05 and 0.01%) amongst abundant cell types.

The performance of D-AREdevil framework was evaluated in terms of specificity and precision in detecting the target rare cells. We observed overall good performance of our approach, especially in the dataset 1 (AML blast cells among BMMCs), where the expression profile of rare cells was more distinct compared to abundant cell types. The detection of rare cells in dataset 2 (iNKT cells among CD3+ T cells) and 3 (HIV-specific B cells among non-naive B cells) was more challenging due to the restricted number of features distinguishing the rare cells from the rest of the data collection. In particular, for frequencies below 0.05% we obtained reduced precision (*i.e.* contamination of cell types that were not manually gated as rare cells). Despited that, we investigated the expression profile of clusters and observed that the "contaminations" consisted of cells with very similar profile to the target rare cells. Meaning that our approach identified homogenously similar rare cell types that the subjective nature of manual gating did not take into account.

2.3 Paper II : The Deficiency in Th2-like Tfh Cells Affects the Maturation and Quality of HIV-Specific B Cell Response in Viremic Infection (unpublished, format for submission to Nature Communication)

T follicular helper CD4 T cells (Tfhs) promote the development of germinal centers and maturation of B cells. Germinal center (GC) Tfhs are a very heterogeneous population that is characterized phenotypically by the co-expression of CXCR5 and PD-1 and by the expression of B-cell lymphoma 6 protein (BCL-6) transcriptor factor. HIV infection is characterized by the expansion of Tfh cell in viremic individuals and the viremia levels in these patients correlates with increased frequencies of GC B cells. Despite the higher frequency of Tfh cells in HIV-infected individuals, they are less effective at providing adequate B-cell help and even if they are capable of responding to HIV antigens the response is affected.

In this study, **I mainly contributed in the dissection of the phenotypic heterogeneity of Tfh cells** in lymph nodes of HIV infected viremic, long-term ART treated and healthy HIV negative individuals (**Result section: *Characterization of Tfh cells*, Figure 1**). We used unsupervised clustering on pooled Tfh cells from the three study groups, *i.e.* FlowSOM, in combination with consensus clustering in order to define 20 different Tfh clusters. Among the 20 defined clusters we identified those that were significantly differentially abundant between the three study groups.

We observed that clusters with higher frequencies in viremic individuals co-expressed CXCR3 and CD38 with varying levels of CD57, CXCR4, HLA-DR and CD25, thus identified CXCR3 and CD38 as the markers driving the Tfh heterogeneity in lymph nodes. The increased co-expression of these two markers by Tfh cells from viremic compared to ART treated and healthy individuals was then confirmed by manual gating. We further investigated the markers characterizing CD38+CXCR3+ Tfh cells and found that an un-regulation of markers of T cell activation (Ki-67, CD25 and HLA-DR), classic Tfh markers (BCL-6, ICOS, CD40L) and the HIV co-receptor CCR5, suggesting that this subset might be more susceptible to HIV infection.

Chapter 3

Discussion

In this thesis, we presented a novel computational framework that we called D-AREdevil. This approach targets disease-associated rare cell populations in a multi-step, flexible and straighforward manner in high-dimensional mass cytometry datasets. D-AREdevil has been designed to help bioinformaticians to identify rare unknown cell populations (cell type, cell state, aberrant cancer cell or the like) of potential high importance in high dimensional mass cytometry data. This is often the case when a cell population is challenging to identify because it represents a rare minority of the full cell collection and thus typically remains hidden among the other cells. In the case of a cell population associated with a disease or condition under study, a test of association between the number of cells in a cluster and conditions can highlight cluster(s) for further analyses. This problem has so far received only little attention in the literature and no standard leading approach exists.

The method is easily implemented and relies on an optimal combination of existing and well-documented unsupervised methods; unsupervised anomaly detection techniques (*i.e.* LOF or FiRE) and unsupervised clustering (*i.e.* FlowSOM). The selection of FlowSOM, was mainly based on previous studies [Weber and Robinson, 2016; Nowicka et al., 2017; Weber and Soneson, 2019] that demonstrated, for applications in the context of mass cytometry, a high performance and an extremely fast runtime compared to other populations-finding methods. On the other hand, the selection of anomaly detection techniques was based on testing and comparisons performed in **here** (Results section 2.1).

Heretofore, no study (to the best of our knowledge) showed the applicability and practical employment of anomaly detection algorithms in the context of single-cell analysis. The applicability of the approach developed in this thesis was demonstrated through different datasets types, in particular in cytometry data.

3.1 Discussion of the different steps

3.1.1 Anomaly detection

In Results section 2.1 we showed the performance of anomaly detection techniques in terms of AUC values and runtime. We selected and tested the state-of-the-art methods

based on two criteria: the methods that are available in R software and focused on those that provide a continuous anomaly score for individual data points. In particular, our selection was based on Goldstein and Uchida [2016a], which presented nearest neighbors-based techniques as the best performant for *local* anomaly detection despite the higher computation time. Among those, we preferred methods providing continuous anomaly score over binary output in order to have more flexibility when selecting rare cell populations. Since the absolute value of the score strongly depends on the dataset, number of variables and normalization, a binary output "hides" important information about the cell ranking and restrincts the investigator to a defined subset of cells.

The R packages available for anomaly (or in the case of the following packages are called "outlier") detection include HighDimOut (2015), DDoutlier (2018) and OutlierDetection (2019), all available from CRAN. The methods available from OutlierDetection package were excluded because only providing a binary output (*i.e.* list of anomalous observations) without any information about the cell ranking. The HighDimOut package contains angle-based (ABOD) and feature bagging (FBOD) outlier detection algorithms, methods suggested to be more appropriate for dataset with increasing dimensionality [Lazarevic and Kumar, 2005; Breunig et al., 2000]. However, in Domingues et al. [2018] the author showed poor performance of ABOD compared to LOF in terms of scalability, memory consumption and precision in their test datasets. We rapidly evaluated both ABOD and FBOD but did not included them in further evaluation since they did not provide significant improvents over the methods tested from DDoutlier package. From the DDoutlier package we observed the higher performance, in terms of AUC values and runtime, for LOF, k-NN sum and k-NN aggregation methods (Result section 2.1). Due to the very similar results obtained for the three methods, we based our decision of selecting LOF for further evaluation on previous assessments [Goldstein and Uchida, 2016a] that showed and recommended LOF for local anomalies detection tasks over the k-NN methods that targets global anomalies more efficiently (Introduction section 1.4.4). One of the major drawbacks of the selected anomaly detection technique remains the runtime complexity associated with the k-nearest neighbors search. Generally, the k-NN search have a time complexity of $O(n^2)$. The kNN function used in LOF (from dbscan package) uses a space-partitioning data structure called kd-tree to identify neighbors in the high-dimensional space. Even if this approach allows to reduce the number of distance computations, it remains costly in large datasets [Wu and Jermaine, 2006]. In addition, this technique does not scale well as the number of variables increases [Chandola et al., 2009; Bentley, 1975]. The other computations of LOF, that is density estimation and score calculations (Introduction section 1.4.3, Equations 1.3-1.5), can be neglected since representing less than 1% of runtime [Goldstein and Uchida, 2016b].

Another method that was selected for examination is called FiRE (finder of rare entities) published in [Jindal et al., 2018a]. FiRE has been shown to be a very fast algorithms ($O(n)$, where n is the number of cells), which takes advantage of the Sketching technique to estimate similarities between data points. We compared the two anomaly detection methods LOF and FiRE in the context our approach, and discussed the results in the following paragraphs.

In summary, we used benchmark mass cytometry datasets to compare the detection performance of the two selected methods within our D-AREdevil framework: the *local outlier factor* (LOF) and *finder of rare entities* (FiRE). We included these anomaly detection techniques in our approach to serve as selectors for candidate rare cells. In this way, we were able to filter-out the majority of abundant cells and focus on potentially relevant disease-associated rare cell subsets. Importantly, the rare cell populations that we are targeting using this approach are not necessarily the rarest cells in the datasets, we rather want to focus on those cell types that are present in different abundance between conditions. With that purpose, we tested different cut-offs in order to provide an indication to the users concerning the most appropriate ones and to address decision-making in the context of mass cytometry data analysis. We showed that the use of a permissive cut-off (*i.e.* that selects a large number of candidate rare cells) such as *q75*, maximises the capacity of identifying the majority of rare disease-associated cell types. However, when the starting number of cells from pooled samples was very large, taking 25% of the cells is not sufficient to filter-out enough abundant cells and thus to focus on rare cell populations. An overall observation is that less permissive cut-offs (*i.e.* that selects a reduced number of candidate rare cells compared *q75*) such as *q95* and IQR-*q95* provided an enogh filtered set of candidate rare cells to allow good sensitivity and precision in the identification of disease-associated rare cell types. The most stringent cut-off (*i.e.* that selects a very limited number of candidate rare cells), IQR-*q95*, was tested because proposed in the FiRE publication and because it corresponds to the "standard" way of identifying outliers. We observed that this cut-off was not adapted for FiRE application in the context of mass cytometry datasets. Indeed, in each tested dataset this cut-off did not select any cell when applied to FiRE score. However, the use of IQR-*q95* on LOF score provides good results in terms of PPV but has the disadvantage to lose some of the already rare target cells due to its high stringency.

Overall, our results demonstrated the applicability of FiRE in our framework by its high sensitivity in detecting the target rare cell populations. On the other hand, LOF showed longer computation time. Despite that, LOF showed not only high sensitivity but overall higher precision when using the different cut-offs for identifying the target rare cells compared to FiRE. Although the lower precision of FiRE, the two methods can be used in conjunction to validate the results when both methods identify the same rare cell populations. Although any discovery using this approach requires to be validated using targeted confirmatory experiments, the agreement between the two methods provides an indication about the probable existence of identified rare cell populations.

3.1.2 Dimension reduction and Clustering

Once a set of candidate rare cells is selected on the basis of their anomaly score and a defined cut-off, we used dimensionality reduction techniques such as UMAP to visualize the selected cells in a 2-dimensional plot per condition. We suggest the use of UMAP for its scalability for rapidly increasing sample sizes (conversely to ISOMAP, Diffusion Map or t-SNE) and preservation of global structures of the data. The visualization of the selected rare cell allows to have a first snapshot of the possible differences in these cells distribution between conditions.

Dimensionality reduction is not only used for visualization purposes but also to improve the clustering of these cells and identify homogeneous sub-groups that are specifically associated with a disease or outcome of interest. Indeed, when applying FlowSOM using the first two UMAP coordinates in addition to the whole markers set, we observed an improved separation of the target rare cells. The use of all markers in the dataset and the first two UMAP coordinates allows to maximise our capacity of identifying the target rare cells while maintaining the unsupervised and unbiased way of proceeding. We also tested whether adding additional UMAP coordinates would improve our capacity of identifying the target rare cells, but it was not the case for the tested datasets.

The use of FlowSOM as unsupervised clustering technique was based on previous studies reporting this method as one of the best performing and the one showing the fastest runtime. We defined ten different clusters independently of the dataset type, anomaly score technique applied and anomaly score cut-off used. We did not focus on the optimization of the clustering step since the previous step (*i.e.* anomaly detection) should have filter-out enough abundant cell types to reveal, in principle, the interesting rare cell subsets at the dimensionality reduction and visualization step of the analysis. Indeed, we observed that the methodology was mostly effective when the target rare cells were distinctly separated when the selected candidate rare cells were visualized in lower dimension plot. Among the ten different clusters that we defined, we identified the cluster containing the higher number of target rare cells and evaluated sensitivity and precision on that cluster.

In general, our results demonstrated that with this approach one can easily identify multiple rare cell subsets (*i.e.* dataset 1, AML blast cell subsets) by inspecting the expression profile of cells once these are conveniently enriched and grouped in one cluster. This is the case even when the PPV of the cluster is low because the target cells represent a minority in the selected cluster. Of note, in all the three datasets we observed that by plotting the expression profile of the cells (those contained in the cluster with the target cells) the LOF score is homogeneously higher compared to the other cell types. This indicates that these cells are found together (in close proximity to each other) in low density regions in the high-dimensional space.

3.1.3 Testing association

The final step of the analysis consists in testing all the clusters represeting less than 1% of the total sample size for association with the conditions' variables. Since the target cells correspond to few cell counts, we used a simple Fisher's exact test on a 2x2 table, with the numbers of cells in the cluster per condition, is used to help identify the rare cell populations of interest for further investigations. This test provides a p-value for the difference in cell counts between conditions and the odds ratio value that provides indication of the strength of association. One aspect that could be studied further is the association test, where regression methodologies that could include single sample or patient information and additional covariates would be a logic extension.

3.2 Conclusions and future persepectives

Advances in single-cell technologies increase progressively the potentialities of cellular heterogeneity understanding, including the discovery of previously undetectable cell types and in particular rare cell populations. Because the data generated have increasing resolution, both in terms of throughput and dimensionality, we have now the means to search for and discover cell populations that are very under-represented (<1%) in the total population. Despite the many efforts that have been done in the last past years, methods enabling the automatic, accurate and rapid detection of rare cell populations are nearly non-existent. In particular for mass cytometry, where the throughput is considerably higher compared to scRNA-sequencing, which identification represent a real "needle in a haystack" scenario.

Globally, this study showed very good results for the identification of rare cell populations that are present at frequencies below 0.1% in the total cell population under investigation. Our computational framework uses anomaly detection (*i.e.* LOF or FiRE) in combination with unsupervised clustering to maximize and improve the detection capacity of rare cell populations with very good sensitivity and precision compared to previous work. In particular, we showed that this approach performed well in detecting single (HIV-specific B cells or invariant natural killer T cells) or multiple rare cell populations (two different subsets of AML blast cells).

Of note, our approach performed particularly well if we consider the complexity of the testing datasets we used to demonstrate the performance of the D-AREdevil approach. The complexity of the datasets is intended not only in terms of markers diversity; containing a large majority of phenotypic markers (cell surface), transcription factors or phospho proteins, but also in terms of cellular homogeneity if compared to total PMBCs. We believe that proposing an approach involving multiple analysis steps is of advantage to adapt the analyses to the biological question(s). Indeed, it makes the approach very flexible and allows the user to make decisions according to the results obtained at the different steps of analysis. Given the demonstrated adaptability of the approach to different dataset types, D-AREdevil could potentially be adapted and applied to data beyond flow and mass cytometry. A potential application is scRNA-seq data, which requires adaptations of the approach to the different type of dimensionality. ScRNA-seq data requires a feature selection step when dealing with increasing dimensionality ($p >> N$) or the use of an expression matrix with reduced dimensionality (*e.g.* using UMAP, t-SNE or other techniques). Alternatively, the selection of an anomaly detection method able to deal with many variables (*e.g.* FBOD).

In conclusion, this approach has the potentiality of being used as a pipeline for direct data analysts in the field and/or representing a bridge towards more sophisticated solutions for detecting rare cell population of biological relevance. In addition, it deserves to be validate on additional testing datasets and eventually updated with new developed state-of-the-art methods at the different steps of analysis.

Bibliography

A. I. Abdelrahman, O. Ornatsky, D. Bandura, V. Baranov, R. Kinach, S. Dai, S. C. Thickett, S. Tanner, and M. A. Winnik. Metal-containing polystyrene beads as standards for mass cytometry. *Journal of Analytical Atomic Spectrometry*, 25(3):260–268, 2010. ISSN 02679477. doi: 10.1039/b921770c.

N. Aghaeepour, P. Chattopadhyay, M. Chikina, T. Dhaene, S. Van Gassen, M. Kursa, B. N. Lambrecht, M. Malek, G. McLachlan, Y. Qian, et al. A benchmark for evaluation of algorithms for identification of cellular correlates of clinical outcomes. *Cytometry Part A*, 89(1):16–21, 2016.

E. Arvaniti and M. Claassen. Sensitive detection of rare disease-associated cell subsets via representation learning. *Nature communications*, 8(1):1–10, 2017.

A. Bashashati and R. R. Brinkman. A Survey of Flow Cytometry Data Analysis Methods. *Advances in Bioinformatics*, 2009(December):1–19, 2009. ISSN 1687-8027. doi: 10.1155/2009/584603.

S. C. Bendall, E. F. Simonds, P. Qiu, E.-a. D. Amir, P. O. Krutzik, R. Finck, R. V. Bruggner, R. Melamed, A. Trejo, O. I. Ornatsky, R. S. Balderas, S. K. Plevritis, K. Sachs, D. Pe, S. D. Tanner, and G. P. Nolan. Single-Cell Mass Cytometry of Differential. *Science*, 332(May):687–695, 2011. ISSN 0036-8075. doi: 10.1126/science.1198704.

J. L. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

P. J. Brennan, M. Brigl, and M. B. Brenner. Invariant natural killer T cells: An innate activation scheme linked to diverse effector functions. *Nature Reviews Immunology*, 13 (2):101–117, 2013. ISSN 14741733. doi: 10.1038/nri3369.

M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

R. V. Bruggner, B. Bodenmiller, D. L. Dill, R. J. Tibshirani, and G. P. Nolan. Automated identification of stratifying signatures in cellular subpopulations. *Proceedings of the National Academy of Sciences*, 111(26):E2770–E2777, 2014a. ISSN 0027-8424. doi: 10.1073/pnas.1408792111.

R. V. Bruggner, B. Bodenmiller, D. L. Dill, R. J. Tibshirani, and G. P. Nolan. Automated identification of stratifying signatures in cellular subpopulations. *Proceedings of the National Academy of Sciences*, 111(26):E2770–E2777, 2014b.

F. Castro-Giner and N. Aceto. Tracking cancer progression: From circulating tumor cells to metastasis. *Genome Medicine*, 12(1):1–12, 2020. ISSN 1756994X. doi: 10.1186/s13073-020-00728-3.

V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

H. Chen, M. C. Lau, M. T. Wong, E. W. Newell, M. Poidinger, and J. Chen. Cytofkit: a bioconductor package for an integrated mass cytometry data analysis pipeline. *PLoS computational biology*, 12(9):e1005112, 2016.

J. Chen, M. Wu, J. Wang, and X. Li. Immunoregulation of nkt cells in systemic lupus erythematosus. *Journal of immunology research*, 2015, 2015.

J. G. Cyster and C. D. Allen. B cell responses: cell interaction dynamics and decisions. *Cell*, 177(3):524–540, 2019.

S. De Biasi, L. Gibellini, M. Nasi, M. Pinti, and A. Cossarizza. Rare cells: focus on detection and clinical relevance. In *Single Cell Analysis*, pages 39–58. Springer, 2017.

R. Domingues, M. Filippone, P. Michiardi, and J. Zouaoui. A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern Recognition*, 74: 406–421, 2018.

A. Duò, M. D. Robinson, and C. Soneson. A systematic performance evaluation of clustering methods for single-cell rna-seq data. *F1000Research*, 7, 2018.

M. Farinacci, T. Krahn, W. Dinh, H.-D. Volk, H.-D. Düngen, J. Wagner, T. Konen, and O. von Ahsen. Circulating endothelial cells as biomarker for cardiovascular diseases. *Research and Practice in Thrombosis and Haemostasis*, 3(1):49–58, 2019. ISSN 2475-0379. doi: 10.1002/rth2.12158.

G. Finak, J. Frelinger, W. Jiang, E. W. Newell, J. Ramey, M. M. Davis, S. A. Kalams, S. C. De Rosa, and R. Gottardo. OpenCyto: An Open Source Infrastructure for Scalable, Robust, Reproducible, and Automated, End-to-End Flow Cytometry Data Analysis. *PLoS Computational Biology*, 10(8), 2014. ISSN 15537358. doi: 10.1371/journal.pcbi.1003806.

R. Finck, E. F. Simonds, A. Jager, S. Krishnaswamy, K. Sachs, W. Fantl, D. Pe'er, G. P. Nolan, and S. C. Bendall. Normalization of mass cytometry data with bead standards. *Cytometry Part A*, 83 A(5):483–494, 2013. ISSN 15524922. doi: 10.1002/cyto.a.22271.

M. Goldstein and S. Uchida. A comparative study on outlier removal from a large-scale dataset using unsupervised anomaly detection. *ICPRAM 2016 - Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods*, pages 263–269, 2016a. doi: 10.5220/0005701302630269.

M. Goldstein and S. Uchida. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one*, 11(4):e0152173, 2016b.

F. Hahne, N. LeMeur, R. R. Brinkman, B. Ellis, P. Haaland, D. Sarkar, J. Spidlen, E. Strain, and R. Gentleman. flowCore: A Bioconductor package for high throughput flow cytometry. *BMC Bioinformatics*, 10, 2009. ISSN 14712105. doi: 10.1186/1471-2105-10-106.

S. C. Hofmann, A. Bosma, L. Bruckner-Tuderman, M. Vukmanovic-Stejic, E. C. Jury, D. A. Isenberg, and C. Mauri. Invariant natural killer t cells are enriched at the site of cutaneous inflammation in lupus erythematosus. *Journal of Dermatological Science*, 71(1):22–28, 2013.

D. Hunter-Schlichting, J. Lane, B. Cole, Z. Flaten, H. Barcelo, R. Ramasubramanian, E. Cassidy, J. Faul, E. Crimmins, N. Pankratz, et al. Validation of a hybrid approach to standardize immunophenotyping analysis in large population studies: the health and retirement study. *Scientific Reports*, 10(1):1–9, 2020.

A. Jindal, P. Gupta, Jayadeva, and D. Sengupta. Discovery of rare cells from voluminous single cell expression data. *Nature Communications*, 9(1):1–11, 2018a. ISSN 20411723. doi: 10.1038/s41467-018-07234-6.

A. Jindal, P. Gupta, Jayadeva, and D. Sengupta. Discovery of rare cells from voluminous single cell expression data. *Nature Communications*, 9 (1), 2018b. ISSN 20411723. doi: 10.1038/s41467-018-07234-6. URL `http://dx.doi.org/10.1038/s41467-018-07234-6`.

T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

S. H. Krovi and L. Gapin. Invariant natural killer t cell subsets—more than just developmental intermediates. *Frontiers in immunology*, 9:1393, 2018.

L. Lai, R. Ong, J. Li, and S. Albani. A CD45-based barcoding approach to multiplex mass-cytometry (CyTOF). *Cytometry Part A*, 87(4):369–374, 2015. ISSN 15524930. doi: 10.1002/cyto.a.22640.

C. W. Law, Y. Chen, W. Shi, and G. K. Smyth. voom: Precision weights unlock linear model analysis tools for rna-seq read counts. *Genome biology*, 15(2):R29, 2014.

A. Lazarevic and V. Kumar. Feature bagging for outlier detection. (June 2014):157, 2005. doi: 10.1145/1081870.1081891.

D. W. Li, Z. Q. Liu, J. Wei, Y. Liu, and L. S. Hu. Contribution of endothelial progenitor cells to neovascularization (review). *International Journal of Molecular Medicine*, 30 (5):1000–1006, 2012. ISSN 11073756. doi: 10.3892/ijmm.2012.1108.

M. Linderman, R. Bruggner, and M. R. Bruggner. Package 'rclusterpp'. 2013.

X. Liu, W. Song, B. Y. Wong, T. Zhang, S. Yu, G. N. Lin, and X. Ding. A comparison framework and guideline of clustering methods for mass cytometry data. *Genome biology*, 20(1):297, 2019.

P. Mancuso, A. Calleri, C. Cassi, A. Gobbi, M. Capillo, G. Pruneri, G. Martinelli, and F. Bertolini. Circulating endothelial cells as a novel marker of angiogenesis. In *Novel Angiogenic Mechanisms*, pages 83–97. Springer, 2003.

R. Melchiotti, F. Gracio, S. Kordasti, A. K. Todd, and E. de Rinaldis. Cluster stability in the analysis of mass cytometry data. *Cytometry Part A*, 91(1):73–84, 2017. ISSN 15524930. doi: 10.1002/cyto.a.23001.

T. R. Mosmann, I. Naim, J. Rebhahn, S. Datta, J. S. Cavenaugh, J. M. Weaver, and G. Sharma. Swift—scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, part 2: biological evaluation. *Cytometry Part A*, 85(5):422–433, 2014.

A. Motsinger, D. W. Haas, A. K. Stanic, L. Van Kaer, S. Joyce, and D. Unutmaz. CD1d-restricted human natural killer T cells are highly susceptible to human immunodeficiency virus 1 infection. *Journal of Experimental Medicine*, 195(7):869–879, 2002. ISSN 00221007. doi: 10.1084/jem.20011712.

I. Naim, S. Datta, J. Rebhahn, J. S. Cavenaugh, T. R. Mosmann, and G. Sharma. Swift—scalable clustering for automated identification of rare cell populations in large, high-dimensional flow cytometry datasets, part 1: Algorithm design. *Cytometry Part A*, 85(5):408–421, 2014.

A. E. Najera Chuc, L. A. Cervantes, F. P. Retiguin, J. V. Ojeda, and E. R. Maldonado. Low number of invariant NKT cells is associated with poor survival in acute myeloid leukemia. *Journal of Cancer Research and Clinical Oncology*, 138(8):1427–1432, 2012. ISSN 01715216. doi: 10.1007/s00432-012-1251-x.

E. W. Newell and Y. Cheng. Mass cytometry: Blessed with the curse of dimensionality. *Nature Immunology*, 17(8):890–895, 2016. ISSN 15292916. doi: 10.1038/ni.3485.

M. Nowicka, C. Krieg, L. M. Weber, F. J. Hartmann, S. Guglietta, B. Becher, M. P. Levesque, and M. D. Robinson. Cytof workflow: differential discovery in high-throughput high-dimensional cytometry datasets. *F1000Research*, 6, 2017.

S. Palit, C. Heuser, G. P. De Almeida, F. J. Theis, and C. E. Zielinski. Meeting the challenges of high-dimensional single-cell data analysis in immunology. *Frontiers in Immunology*, 10(JUL):1–12, 2019. ISSN 16643224. doi: 10.3389/fimmu.2019.01515.

S. P. Perfetto, P. K. Chattopadhyay, and M. Roederer. Seventeen-colour flow cytometry: Unravelling the immune system. *Nature Reviews Immunology*, 4(8):648–655, 2004. ISSN 14741733. doi: 10.1038/nri1416.

V. Proserpio and T. Lönnberg. Single-cell technologies are revolutionizing the approach to rare cells. *Immunology and Cell Biology*, 94(3):225–229, 2016. ISSN 14401711. doi: 10.1038/icb.2015.106.

F. Ravandi, R. B. Walter, and S. D. Freeman. Evaluating measurable residual disease in acute myeloid leukemia. *Blood Advances*, 2(11):1356–1366, 2018. ISSN 24739537. doi: 10.1182/bloodadvances.2018016378.

M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1): 139–140, 2010.

Z. Sachs, R. S. LaRue, H. T. Nguyen, K. Sachs, K. E. Noble, N. A. M. Hassan, E. Diaz-Flores, S. K. Rathe, A. L. Sarver, S. C. Bendall, N. A. Ha, M. D. Diers, G. P. Nolan, K. M. Shannon, and D. A. Largaespada. NRASG12V oncogene facilitates self-renewal in a murine model of acute myelogenous leukemia. *Blood*, 124(22):3274–3283, 2014. ISSN 15280020. doi: 10.1182/blood-2013-08-521708.

Y. Saeys, S. Van Gassen, and B. N. Lambrecht. Computational flow cytometry: Helping to make sense of high-dimensional immunology data. *Nature Reviews Immunology*, 16 (7):449–462, 2016. ISSN 14741741. doi: 10.1038/nri.2016.56.

N. Samusik, Z. Good, M. H. Spitzer, K. L. Davis, and G. P. Nolan. Automated mapping of phenotype space with single-cell data. *Nature Methods*, 13(6):493–496, 2016. ISSN 15487105. doi: 10.1038/nmeth.3863.

D. Sarkar, N. Le Meur, and R. Gentleman. Using flowviz to visualize flow cytometry data. *Bioinformatics*, 24(6):878–879, 2008.

S. Schreier, S. Borwornpinyo, R. Udomsangpetch, and W. Triampo. An update of circulating rare cell types in healthy adult peripheral blood: findings of immature erythroid precursors. *Annals of Translational Medicine*, 6(20):406–406, 2018. ISSN 23055839. doi: 10.21037/atm.2018.10.04.

M. H. Spitzer and G. P. Nolan. Mass Cytometry: Single Cells, Many Features. *Cell*, 165(4):780–791, 2016. ISSN 10974172. doi: 10.1016/j.cell.2016.04.019. URL `http://dx.doi.org/10.1016/j.cell.2016.04.019`.

J. V. Sweedler and E. A. Arriaga. Single cell analysis. *Analytical and Bioanalytical Chemistry*, 387(1):1–2, 2007. ISSN 16182642. doi: 10.1007/s00216-006-0921-4.

S. Van Gassen, B. Callebaut, M. J. Van Helden, B. N. Lambrecht, P. Demeester, T. Dhaene, and Y. Saeys. Flowsom: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry Part A*, 87(7):636–645, 2015.

L. M. Weber and M. D. Robinson. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry Part A*, 89(12):1084–1096, 2016. ISSN 15524930. doi: 10.1002/cyto.a.23030.

L. M. Weber and C. Soneson. HDCytoData: Collection of high-dimensional cytometry benchmark datasets in Bioconductor object formats. *F1000Research*, 8:1459, 2019. ISSN 20461402. doi: 10.12688/f1000research.20210.2.

L. M. Weber, M. Nowicka, C. Soneson, and M. D. Robinson. diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering. *Communications Biology*, 2(1), 2019. ISSN 23993642. doi: 10.1038/s42003-019-0415-5. URL `http://dx.doi.org/10.1038/s42003-019-0415-5`.

R. Wehrens. <Kohonen-Manual.Pdf>. *JSS Journal of Statistical Software*, 21(5), 2007. URL `http://www.jstatsoft.org/`.

M. Wu and C. Jermaine. Outlier detection by sampling with accuracy guarantees. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 767–772. ACM, 2006.

Y. Şenbabaoğlu, G. Michailidis, and J. Z. Li. Critical limitations of consensus clustering in class discovery. *Scientific reports*, 4(1):1–13, 2014.

Articles

A novel computational approach for discovering disease-associated rare cell populations in cytometry data

Madeleine Suffiotti,[1,4] Alessandra Noto,[1] Morgane Humbel[1], Victor Joo[1], Denis Comte[1], Craig Fenwick[1], Giuseppe Pantaleo[1,2], Mauro Delorenzi[3,4*]

[1]Service Immunology and Allergy, Lausanne University Hospital, Lausanne, Switzerland

[2]Swiss Vaccine Research Institute, Lausanne University Hospital, Lausanne, Switzerland
[3]Translational Bioinformatics and Statistics, Department of Oncology, Swiss Cancer Center Leman, University of Lausanne, Lausanne, Switzerland
[4]Bioinformatics Core Facility, SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland

*Corresponding author; E-mail: mauro.delorenzi@unil.ch

Abstract

Advances in single-cell technologies such as mass cytometry provide increasing resolution of the complexity of cellular samples, allowing researchers to deeply investigate cellular heterogeneity and potentially discover previously undetected rare cell populations. The identification of rare cell populations is of paramount importance for understanding the onset, progression and pathogenesis of many diseases while remaining a challenge due to the always increasing dimensionality and throughput of the data generated. We demonstrate in three different mass cytometry datasets the efficacy of a novel computational framework called D-AREdevil (disease-associated rare cells detection) that efficiently supports a data analyst to identify low frequencies (< 0.1%) disease-associated rare cell populations in large and complex biological samples. Our approach takes advantage of anomaly detection techniques in combination with modern dimension reduction and unsupervised clustering method.

Introduction (926)

Technological advances in single-cell technologies such as flow or mass cytometry allow researchers to achieve a deeper understanding of the cellular heterogeneity within populations that were once considered to be homogeneous [Spitzer and Nolan, 2016]. In particular, mass cytometry (or CyTOF, “cytometry by time-of-flight mass spectrometry”) enables the quantification of over 50 parameters at the single-cell level and profiling of millions of cells from a single sample [Bruggner et al., 2014].

The capacity of accurately characterizing rare cell populations is crucial for understanding the onset, progression and pathogenesis in diseases such as autoimmune diseases, immunodeficiencies, or cancer [Schreier et al., 2018]. The comprehensive understanding of the samples’ heterogeneity can reveal previously uncharacterized immune cell types, help to understand their differentiation and function, and possibly provide new diagnostic biomarkers or novel therapeutic targets [De Biasi et al., 2017]. Indeed, health and disease status of patients often depend on minor groups of cells with frequencies largely below 1%. Examples of these rare cell populations include circulating endothelial cells, which have been associated with cardiovascular damages [Farinacci et al., 2019] or endothelial progenitor cells, which are biomarkers of tumor angiogenesis [Li et al., 2012]. In the case of hematologic malignances, such as acute myeloid leukemia (AML), the detection of minimal residual disease (MRD) serves as a strong prognostic marker for a high risk of relapse post-therapy [Ravandi et al., 2018]. However, the identification of these rare cell populations remains challenging in mass cytometry data due to the high number of data points (single-cell resolution and high-throughput) in the growing dimensionality given by multiplexed panels.

In the last decade, significant efforts have been made to develop methods for the automatic detection and discovery of unknown or not well-characterized cell populations, but few methods were designed and demonstrated to be efficient for rare cell populations detection. In Weber and Robinson [2016], several unsupervised clustering methods for cytometry data have been evaluated for the detection of rare cell populations, both in terms of performance and runtime, on a simple case dataset.

Xshift [Samusik et al., 2016], Rclusterpp [Linderman et al., 2013] and FlowSOM [Van Gassen et al., 2015] were reported as the best performing methods. In Weber et al. [2019] the authors proposed high-resolution clustering (*i.e.* definition of an extremely high number of clusters) using FlowSOM for the discovery of differentially abundant rare cell types. However, the enrichment for the rare cells of interest was not optimal when the cells were present at frequencies below 1%.

In machine learning, the use of anomaly detection algorithms to find rare observations has a long tradition for tasks such as intrusion detection (including network traffic and server applications), fraud detection (financial transactions), patient monitoring (ECG and EEG signals) or image analysis (computed tomography), much less for molecular applications [Goldstein and Uchida, 2016].

The anomaly detection method called Local Outlier Factor (LOF) [Breunig et al., 2000] has been shown to be the best performing algorithm among the nearest neighbor-based methods to detect local anomalies, even though it has a relatively high computation time [Goldstein and Uchida, 2016]. The advantage of nearest neighbors-based methods is their simplicity; the only assumption is that normal observations are found in dense regions of data points while anomalies occur in low density regions [Chandola et al., 2009]. In particular, LOF assigns to each observation a continuous anomaly score, which represents an estimate of the local density obtained based on the number of nearest neighbors (k).

A more recent method was proposed by Jindal et al. [2018], called FiRE (finder of rare entities) assigns a score of anomaly to individual cells as well but in a very rapid timeframe. The computation of the FiRE score combines the use of the Sketching technique that compacts the expression profile of cells into bit vectors and the creation of low-dimensional bit signatures (hash codes) to group cells of similar profile. Then, the populousness of hash codes is used as estimate of cell density.

The purpose of this study is to present an approach that uses unsupervised anomaly detection algorithms to support a data analyst to finding extremely rare cell populations (with frequencies < 0.1%) in large and complex biological samples within reasonable limits of time and computational infrastructure. The approach aims to find rare cells

that have homogeneously consistent expression profile with only few deviating features, for example antigen-specific B cells among total B cells. We propose a novel computational framework called D-AREdevil (disease-associated rare cells detection) for cytometry datasets. The main characteristic of our computational framework is the combination of an anomaly detection algorithm (*i.e.* LOF or FiRE) with a fast unsupervised clustering method (*i.e.* FlowSOM). In our approach, the anomaly score serves to select a set of candidate rare cells to be further investigated, thus filter-out the vast majority of abundant cell types. Then, the selected set of cells is analyzed to identify sub-groups of similar cells. This is a key step in the process of finding cell sub-groups that are associated with a disease (or condition of interest) when their existence and identification is not previously known and has to be discovered.

Currently, to the best of our knowledge, there are no standard analysis workflows that integrate anomaly detection algorithms with the end of detecting disease-associated rare cell populations within the context of mass cytometry. Here, we report the properties and implementation of D-AREdevil and present applications on three different test datasets. We generated the test datasets by mixing one or more known rare cell populations at varying frequencies and tested the ability of our approach to find the target rare cells.

Results

Overview of D-AREdevil

D-AREdevil is a novel multi-step computational framework designed for the detection of disease-associated rare cell populations in molecular profile data. It consists of four major steps (**Figure 1a**): (1) anomaly detection, (2) dimension reduction, (3) unsupervised clustering and (4) statistical testing for association. The analysis starts with single cell expression data from pooled samples (**Figure 1b**), *i.e.* samples representing different "conditions", for example healthy and disease. The input data might contain one or several unknown rare cell populations potentially associated with the disease status or with other relevant clinical-biological features, which are difficult to detect because of the overwhelming number of cells. The steps 1-3 are useful for exploring or discovering rare cell populations in absence of different conditions.

The first major step aims at eliminating the vast majority of abundant cells and to enrich for the target rare cells we want to further investigate. A set of candidate rare cells is obtained by taking the top-ranking cells according to the anomaly-detection method. It is not straightforward to find an optimal cut-off value for the anomaly score, here we tested four different cut-off values (**Figure 1c**): two permissive cut-offs, the 75th and the 95th percentile (called *q75* and *q95*, selecting 25% respectively 5% of cells with the highest score) and two more stringent cut-offs, where the interquartile range IQR times 1.5 is added to *q75* or *q95* (called IQR-*q75* and IQR-*q95*), thus selecting a small number of cells with very high anomaly score. We tested two different anomaly scoring algorithm, LOF and FiRE (described Introduction and Methods). The following two steps are done in parallel and both on the selected candidate rare cells. Firstly, the dimensionality of these cells is reduced and then visualized using a modern technique called UMAP (uniform manifold approximation and projection [Becht et al., 2019]) (**Figure 1d**). Then, unsupervised clustering is applied to sub-group similar cells within the selected set. UMAP allows to efficiently visualize the spatial distribution of the individual cells, their sub-groups and eventually to identify differences between conditions. We identify sub-group using the FlowSOM and generate a fixed number of 10 clusters. In some cases, the UMAP representation can suggest that a higher

number of clusters should be defined. The clusters are shown using different colors in the UMAP representation (**Figure 1e**). Finally, to identify clusters significantly associated with a condition (**Figure 1f**) we perform Fisher's exact tests on clusters representing <1% of the total dataset. In principle, all clusters with statistically significant p-value and relevant odds ratio value should be investigated.

In order to test the proposed procedure, we applied it to three test datasets containing known rare cells, called target cells. As a measure of the ability to identify the target cells, we report sensitivity and precision values after the cut-off application (step 1) as well as in the cluster containing the majority of target cells (step 3). Moreover, we present examples of the UMAP pattern and the results of the association test aimed at their identification.

Test-dataset 1: Detection of rare AML blast cells subsets in simulated minimal residual disease

The first test dataset we used consists of data previously used for a similar purpose [Weber et al., 2019]). The dataset consists in bone marrow mononuclear cells (BMMCs) from healthy donors in which two subsets of AML blast cells ($CD34^{+}CD45^{mid}$) were introduced at low frequency (0.04%) to simulate minimal residual disease (MRD) and to test the ability to identify these cancer cells (see Methods, **Table 1**). The target AML blast cells were either cytogenetically normal (CN subset) or had one of the classic rearrangements t(8;21) and inv(16) that affect the core-binding transcription factor translocation (CBF subset). We applied LOF and FiRE on this dataset and show 2-dimensional UMAP representations of the selected candidate rare cells based on four different cut-offs on LOF score (**Figure 2a**). The UMAP representations show that AML blast cells were clearly isolated from the other cells when applying the *q95*, IQR-*q75* and IQR-*q95* cut-offs. While for *q75* cut-off, the AML blast cells are melted within the other cells, likely due to the high number of cells this cut-off selects. Nonetheless, FlowSOM defined one cluster that was sufficiently enriched of the target cells (Cluster 2 in blue, **Figure 2a**). The hierarchical clustering and heat-map representation of the expression profile of this cluster (**Figure 2b,** including selected markers) show that the two AML blast cells subsets form separated

sub-groups and have higher LOF score compared to the other cells included in cluster 2. The CBF AML blasts are $CD38^-$ and $CD7^-$ representing an immature phenotype of blast cells, while the CN AML blasts express both markers. This cluster is also prominent in a test of association with the AML condition (**Table 2,** *q75* cluster 2). Note, few target CN AML blast cells are found among other cell types (UMAP representations for *q95*, IQR-*q75* and IQR-*q95*) due to the absence of expression of CD7, which profile is more similar to other BMMCs (**Figure 2b**, arrows in the heatmap).

The results demonstrate that we can identify multiple rare cell subsets by inspecting the expression profile of cells contained in clusters that result associated with a condition. As shown here, rare subsets are identified even when the target cells are only a minority in the selected cluster, thus resulting in low precision of the cluster (**Figure 2c**, black square). In this dataset, the target cells are relatively easy to find due to their differences in markers profile compared to the other cell types. Indeed, the LOF-FlowSOM and FiRE-FlowSOM methodologies have both good sensitivity and positive predictive value, except that sensitivity is reduced at the most stringent cut-off (IQR-*q95)* and precision tends to be lower at permissive cut-offs. This emphasized the potential importance of an optimal cut-off, stringent enough for high precision but without losing high sensitivity. In the following, we tested the methodologies in more challenging situations and less artificially designed test datasets.

Test-dataset 2: Detection of rare invariant natural killer T cells associated with potential reduced protection in autoimmune diseases

Invariant natural killer T cells (iNKT) cells represent less than 0.1% of peripheral blood mononuclear cells (PBMCs). These cells are interesting due to their possible protective role in autoimmunity [Hofmann et al., 2013] as they are present at reduced frequencies in autoimmune diseases such as systemic lupus erythematosus (SLE) compared to healthy subjects. We created the second test datasets by selecting $CD3^+$ T cells from SLE patients and healthy individuals that represent the abundant cell types (see Methods). We used manually gated iNKT cells (characterized by TCR-Vα14–Jα18 expression) as target rare cells (**Supplementary Figure 2**). We tested

three different datasets with decreased frequencies of target cells (0.1, 0.05 and 0.01%). We applied the four different cut-offs on LOF and FiRE anomaly scores on these three datasets. Similarly to dataset 1, the 2-dimensional UMAP representations (**Figure 3a**) show the iNKT cells distribution among $CD3^+$ T cells in the dataset with frequency of 0.01%. The target iNKT cells (violet dots) are not all clustered together (*q95*, IQR-*q75* and IQR-*q95*) but some of them are scattered amongst other cell types. This is likely due to the fact that, with the exception of TCR-Vα14–Jα18 expression, these iNKT cells have an expression profile very similar to $CD3^+$ T cells. Consequently, solely iNKT with homogeneous phenotype are grouped together after clustering. We show the cells' expression profile of the cluster containing the majority of target iNKT cells for the cut-off *q95* (**Figure 3b**, heatmap of cluster 3), which have the lowest precision value (**Figure 3c** bottom, triangle in black box). The heat-map shows that the target cells represent a sub-population of $CD4^-CD8^-$ (double negative) T cells, which is itself a rare population in PMBCs.

An objective and rapid identification of interesting sub-groups of cells can be obtained by testing the FlowSOM clusters for association with the healthy condition (**Table 3**). The interesting clusters can be further investigated as previously demonstrated despite the low PPV (**Figure 3c**), as is the case at frequency 0.01%. We obtained multiple clusters with a statistically significant association with the healthy condition and the highest log(OR) for cluster 3 (2.09, CI95% [1.51, 2.75]) (**Table 3**). The p-value of the association test tends to improve with increasing total number of cells in a cluster. Therefore, we recommend prioritizing the highest log(OR) values for ranking clusters in order of interest.

The FiRE-FlowSOM methodology was less effective: there were no target cells selected at the two stringent cutoffs. At the permissive cut-offs the performance was good for frequencies 0.1 and 0.05%. A low PPV shows that no cluster specific enough for target cells was obtained at 0.01% (**Figure 3c**).

Test-dataset 3: Number of nearest-neighbors evaluation and runtime comparison between LOF and FiRE

The third test dataset is a mass cytometry dataset acquired from lymph node mononuclear cells (LNMCs) isolated from HIV$^+$ untreated and HIV$^+$ ART treated patients (see Methods). We used HIV-specific B cells as target cells that were manually gated on IgG$^+$ memory B cells specific for the HIV envelope glycoprotein gp140 and used memory B cells as abundant cell types (**Supplementary Figure 3**).
We took advantage of this dataset to explore the performance variation of LOF accordingly to the number of nearest-neighbor (*k*) points selected by the user.
We generated 100 sub-samples test datasets by randomly selecting 85 HIV-specific B cells and mixing them with a fixed set of memory B cells to a frequency of 0.05%. We report and compare sensitivity and precision (or PPV) for the cluster containing the majority of target cells after applying the whole procedure with LOF or FiRE (step 1-3 of the methodology) (**Figure 4a, Supplementary Table 1**).
The results show that the LOF-FlowSOM and FiRE-FlowSOM methodologies typically have good sensitivity and precision for *q95*, IQR-*q75* and IQR-*q95* cut-offs. Concerning LOF-FlowSOM, higher *k* values and in particular $k \geq 300$ performs well in sensitivity and in precision for all tested cut-offs, while the lowest value ($k = 50$) performs poorly. As expected, when using stringent cut-offs, the sensitivity tends to be reduced due to the smaller number of target cells that are selected. For both anomaly scores, the *q95* cutoff is a good choice for this test dataset, that is when 5% of the cells with highest score is selected. The precision has high variation for *q75*, likely due to the fact that selecting 25% of the most anomalous cells (about 430k cells) is a too large number to isolate the 85 target cells in one distinct cluster out of the 10 defined.
Overall, by selecting a large k value (≥ 300) for LOF we obtain better precision compared to FiRE for *q95*, IQR-*q75* and IQR-*q95* cut-offs.
Concerning FiRE-FlowSOM, its performance is similar to that of LOF-FlowSOM but is more variable depending on the selected cut-off and sub-sample of target cells.

Afterwards, we evaluated how the *k* value influences the computation time. We tracked the runtime of LOF (with varying *k*) and FiRE on the same data, while increasing the input data size (**Figure 4b**). FiRE shows a faster runtime (about 20 seconds for 320k cells) compared to LOF (on average across the different *k* values, about 4h for 320k cells) with an Intel Core i7 processor and clock speed of 2.9 GHz, and 16GB RAM. The runtime of LOF increased of 1.7-fold from k = 40 to k = 1280.
Overall, the results show that LOF tends to give slightly better performance than FiRE on this dataset, when k ≥ 300. FiRE also showed a good performance, and was substantially faster in execution, thus it is suited for larger datasets than those used here.

Test-dataset 3: Detection of rare HIV-specific B cells associated with increased viremia

HIV-specific B cells are more frequent in HIV$^+$ untreated compared to the HIV$^+$ ART treated patients, as their expansion is driven by HIV replication and the presence of antigen [Cubas et al., 2013]. We generated three datasets by randomly selecting HIV-specific B cells at frequencies of 0.1%, 0.05% and 0.01% of the total dataset.

We decided to compare more in detail how well cells' ranking based on LOF and FiRE scores allow to select the target cells (step 1 of the methodology). We generated receiver operating characteristic (ROC) curves and calculated the area under the curves (AUC) values (**Figure 5a, Supplementary Figure 4**). The curves confirm that overall both methods are excellent at attributing a high anomaly score to the target cells and thus distinguishing them from the other cell types. The 2-dimensional UMAP representations (**Figure 5b**) show the selected candidate rare cells based on the different cut-offs and FlowSOM results on the dataset containing 0.01% of HIV-specific B cells. The *q75* cut-off is not shown because not suited (too permissive) given the input data size of this dataset (**Table 1**). We inspected the expression profile of the cluster containing the majority of target cells for IQR-*q75* cut-off (**Figure 5c**, cluster 2), which has low precision (**Figure 5d**). Interestingly, the low precision of this cluster is due to the presence of additional HIV-specific B cells that were not considered in the

gating strategy since IgG^-. Thus, with our approach we identified additional HIV-specific B cells (with less affinity to gp140 trimer) (**Supplementary Figure 3**).

We tested the 10 clusters defined using FlowSOM for association with HIV^+ untreated patients (**Table 4**). Among them, several clusters resulted to be significantly associated with viremia and with strong association log(OR) that could be further investigated. In particular, cluster 2 (**Figure 5b and c**, IQR-*q75*) has with $p < 10E\text{-}18$ and log(OR) = 2.49[1.74, 3.41]. Cluster 2 is one of two showing strong positive association with viremia.

Discussion

The computational strategy we present here has been designed to help bioinformaticians that aim at finding unknown cell populations (cell types or cell states, for example aberrant cancer cells or rare immune cells) of potential high importance in high throughput mass cytometry data. There is a lack of computational strategies designed to find cell populations that represents a small minority (< 1%) of the full cell collection. Indeed, such rare populations do not appear as a separate group among the full cell collection when using clustering algorithms (*e.g.* FlowSOM) or dimensionality reduction technique (*e.g.* UMAP) (**Supplementary Figures 1-3**). This is particularly the case when the distinction between rare and abundant cell types is defined by one or few features. Importantly, in our methodology we focus our attention on rare cell populations that are associated with a disease or condition of interest. In this case, identification is guided by an association test. The strategy of filtering-out the vast majority of abundant cell types using anomaly detection algorithms, leads to clusters sufficiently enriched in rare cells to be revealed by a classic test of association.

In the context of mass cytometry, the problem has so far received little attention in the literature and no standard leading approach exists. Apart the novelty of the proposed approach, D-AREdevil framework has the merits of simplicity and rapidity. It does not require much code-writing, as it is fully based on existing robust methods implemented in packages for the R environment.

We used three test mass cytometry datasets to explain and exemplify the methodology. Compared to existing approaches, we provided a strategy that improves the detection performance of rare cell populations in complex datasets. The datasets used reflect the complexity of physiological cellular samples in terms of population composition and low frequencies of rare cells to be detected. Moreover, this approach allows to better cover the entirety of phenotype-associated rare cell populations due to the flexibility of the continuous LOF or FiRE score compared to unsupervised clustering that assigns a cell label. In particular, the visualization of the LOF score in concomitance with the expression profile of the selected rare cells gives information about the level of rareness and profile deviation compared to abundant cells.

The three test datasets were also used to compare performance and runtime of the two anomaly detection methods (LOF and FiRE) in the context of our approach. Overall, in the test datasets studied, LOF is more precise and stable in the selection of points situated in low density regions. FiRE also performs well and has the advantage of being very fast with execution time complexity O(n), linear in the number of cells. Therefore, for very large cell collections or for a first-step analysis, FiRE may be preferable to LOF when analysis time is a constraint. On the other hand, LOF has computational complexity $O(n^2)$, due to the k-NN algorithm for calculating distances between data points. Moreover, FiRE is designed to deal with high dimensionality of the data (*e.g*. single-cell RNA-sequencing), while LOF is not scalable for such dimensionality of the data and requires the use of dimension reduction (*e.g.* UMAP) before application. The two methods could be used in conjunction as they might find different cell clusters of interest or increase the confidence in a discovery when both identify the same cell subsets.

One limitation of our framework resides in the exploratory nature of the approach. Any discovery will require confirmation in new and independent data sets, additional experimental wet lab investigations or follow-up clinical studies. The current approach requires human intervention and the empirical choice of parameters: the cut-off value on the anomaly score and the number of nearest-neighbor (*k* values), if LOF is used. A related issue is the presence systematic background noise (*s* antibody labeling) that could be identified as rare observations and which exclusion requires biological knowledges.

Concerning the cut-offs selection, we made the following conclusions. By using a permissive cut-off such as *q75*, we maximize our capacity of identifying the majority of rare disease-associated cell types. In situations where the starting number of cells from pooled samples is very large, this restriction to 25% of the cells is not sufficient to filter-out enough abundant cell types. Overall, we observe that less permissive cut-offs provide a suitably filtered set of candidate rare cells with good sensitivity and precision. This was especially true for populations that have only subtle distinguishing characteristics. In practice, the data analyst might want to try a few different

parameters, in particular when no priori knowledges about the frequency of rare cells to be detected.

We did not address the question of a data-driven optimization and automation of these steps, nor of rigorous statistical control, as these seem hard to achieve. In our view, the proposed flexible exploratory approach is very valuable by itself for the analysis of many current projects. One aspect that could be improved is the association test, by considering samples' information. For instance, using regression methodologies that include single sample labels for cells and evaluate statistical significance with resampling or label-permutation methods, in order to achieve a higher degree of statistical robustness.

Methods

Datasets

We use three mass cytometry datasets (**Table 1**), which are available upon request by the authors.

The dataset 1 (AML blast cells) consists of bone marrow mononuclear cells (BMMCs) that was treated as follow based on Weber et al., 2019. BMMCs comes from healthy donors (N = 5) that were split into two parts; the first half represents the healthy subjects, the second represents the simulated minimal residual disease (MRD) subjects, in which were introduced AML blast cells (two subsets) at low proportions (**Supplementary Figure 1**).

We constructed the datasets 2 and 3 with data generated at the University Hospital of Lausanne (**Supplementary Methods**). The principle used to create test datasets was to select a rare cell type in real experimental datasets. Then we combined a number of these target rare cells with a large number of "abundant" cell types in order to obtain the desired relative frequencies of target cells. The cells from all samples were analyzed as a unique pool, in order to track conditions of the cells [Nowicka et al., 2017].

Dataset 2 (*iNKT cells*): Peripheral blood mononuclear cells (PBMCs) from 6 systemic lupus erythematosus (SLE) patients and 6 healthy donors were profiled by CyTOF using 34 extracellular markers recognizing and characterizing $CD3^+$ T cells. Cells were manual gated by an expert as $CD3^+$ invariant natural killer T cells (iNKT cells) positive for the TCR-Vα24-Jα18 protein (**Supplementary Figure 2a**). iNKT cells were used as target rare cells among $CD3^+$ T cell sub-populations.

Dataset 3 (*HIV-specific B cells*): Lymph node mononuclear cells (LNMCs) from 24 HIV+ patients (14 untreated viremic and 10 ART treatead aviremic) were profiled by CyTOF with 33 B cells surface and intracellular markers. HIV-specific B cells were manually gated by an expert on the basis of IgG and gp140 trimers (HIV envelope glycoprotein that is recognized by the IgG receptor on B cells). We used non-naïve B

cells as abundant cell types, that were gated on the basis of IgD and CD27 (**Supplementary Figure 3a**).

D-AREdevil Procedure

We provide here some information to complement the description of our method given in the section results (**Figure 1**).

Step 1: Anomaly detection.

The D-AREdevil framework is intended for the identification of so-called local anomalies, points occurring in local low-density regions in the high-dimensional feature space. Local anomalies are found in the neighborhood of a large cluster of data points and difficult to identify by other means. We use two methods for assigning anomaly score to cells, Local Outlier Factor (LOF) [Breunig et al., 2000] and Finder of Rare Entities (FiRE) [Jindal et al., 2018], implemented in the R packages DDoutlier and FiRE respectively. LOF generates a score that has larger values for points located in regions of low density. The score depends on the parameter *k*, which is the (minimum) number of nearest neighbors' points used to estimate the local point density. As indication, a good choice of *k* is generally larger than the (unknown) number of rare cells to detect. FiRE uses a fast algorithm to estimate the local densities and to derive an anomaly score for each cell. It uses the Sketching technique to map each data point into a bit vector hash code obtained by binarizing the features (here the marker expression values) by ensuring that points mapped to the same hash code are nearby in the space of all features. To define similarities, it subsamples a number M of features from the measurements at the same time and repeats the operation L times with new subsamples of features in order to deal with the high dimensionality of transcriptomic datasets. Then, the L density estimates are combined in the final score. FiRE does not have parameters that need to be tuned given the reduced dimensionality of cytometry datasets, so we used the parameters suggested by the authors.

Steps 2 and 3: Clustering and visualization.

Once a set of candidate rare cells is selected on the basis of their anomaly score, the aim is to identify any homogeneous sub-group of these cells that is specifically associated with a disease or outcome of interest. Concomitant visualization of the selected candidate rare cells in the two conditions is obtained by applying the UMAP technique [Becht et al., 2019] from the package uwot to all selected candidate rare cells and plotting them in two side-by-side graphs per condition. We performed clustering with the FlowSOM package [Van Gassen et al., 2015]. The clustering was performed on the selected candidate rare cells using all the markers and the first two UMAP coordinates. We used the SOM function with default options, which uses a two-dimensional self-organizing map grid of size 10x10 leading to 100 clusters. Successively, we reduced the number of clusters to 10 with the consensus clustering method implemented in the metaClusteringconsensus function of the same package.

Step 4: Association testing.

To identify clusters of cells with different abundance between conditions, we apply the Fisher's exact test to each of the 10 clusters to a 2-by-2 contingency table per cluster. The contingency table consists in the number of cells for each condition belonging or not to the cluster. We report p-values adjusted with a Bonferroni correction (with factor 10, for the 10 clusters) and report estimated log odds ratios and their 95% confidence intervals. Note, in our analyses we focused on clusters representing <1% of the total dataset, thus rare cell populations.

Runtime analysis

The runtime analysis was tracked using peakRAM package.

Figure legends

Figure 1: Schematic overview of analysis steps of D-AREdevil methodology. **(a)** Schematic representation of the four major steps in the D-AREdevil workflow: (1) anomaly detection, (2) dimension reduction, (3) clustering and (4) association testing. **(b)** Input data are provided as an expression matrix of pooled samples (*i.e.* cells from different conditions analyzed together). **(c)** Anomaly detection applied on the input matrix produces a score of anomaly (or rareness) for individual cells. Left plot shows a typical right-skewed distribution of LOF score, the blue region (score values > 1) represents cells with high score value. The middle and right plots show the LOF and FiRE score distributions, respectively. The vertical lines show the tested cut-offs to selected candidate rare cells (red: *q75* = 75th percentile, green: *q95* = 95th percentile, the blue/black lines show IQR based cut-offs). **(d)** 2D-UMAP plots of the selected candidate rare cells for LOF (left) and FiRE (right), where an example of target rare cells is shown be present in one condition (minimal residual disease (MRD) in acute myeloid leukemia patients) but not in the other (healthy subjects) **(e)** Same 2D-UMAP (as in d) but with clusters produced by FlowSOM labeled by different colors. Dashed circles show the cluster containing target rare cells **(f)** Contingency tables for clusters representing < 1% of the total dataset and tested for association using a Fisher's exact test. Results for LOF-FlowSOM are on the left and FiRE-FlowSOM on the right.

Figure 2: Benchmark results for the identification of rare AML blast cell subsets (0.04%) in simulated minimal residual disease of AML patients. **(a)** UMAP representations of candidate rare cells selected by applying: (from left to right) *q75* cut-off (N = 39,389 cells) on LOF score and subsequent application of FlowSOM (in blue is shown cluster 2 that contains the target AML blast cells, CBF subset in green and CN subset in red), *q95* (N = 7,878 cells), IQR-*q75* (N = 8,020 cells) and IQR-*q95* (N = 1,866 cells). **(b)** Heat-map showing the expression profile of cells in cluster 2. Blue-to-yellow color-code shows low-to-high markers expression. LOF score is shown in the first column (green shades). Cell labels are shown in the second column. Arrows and black boxes indicate similar profile between some CN blasts and some BMMCs **(c)** Detection performance for LOF (purple) and FiRE (blue). For each score cut-off is

shown sensitivity after anomaly selection (step 1) respectively after selection of a FlowSOM cluster (step 3) along with the positive predictive value (PPV).

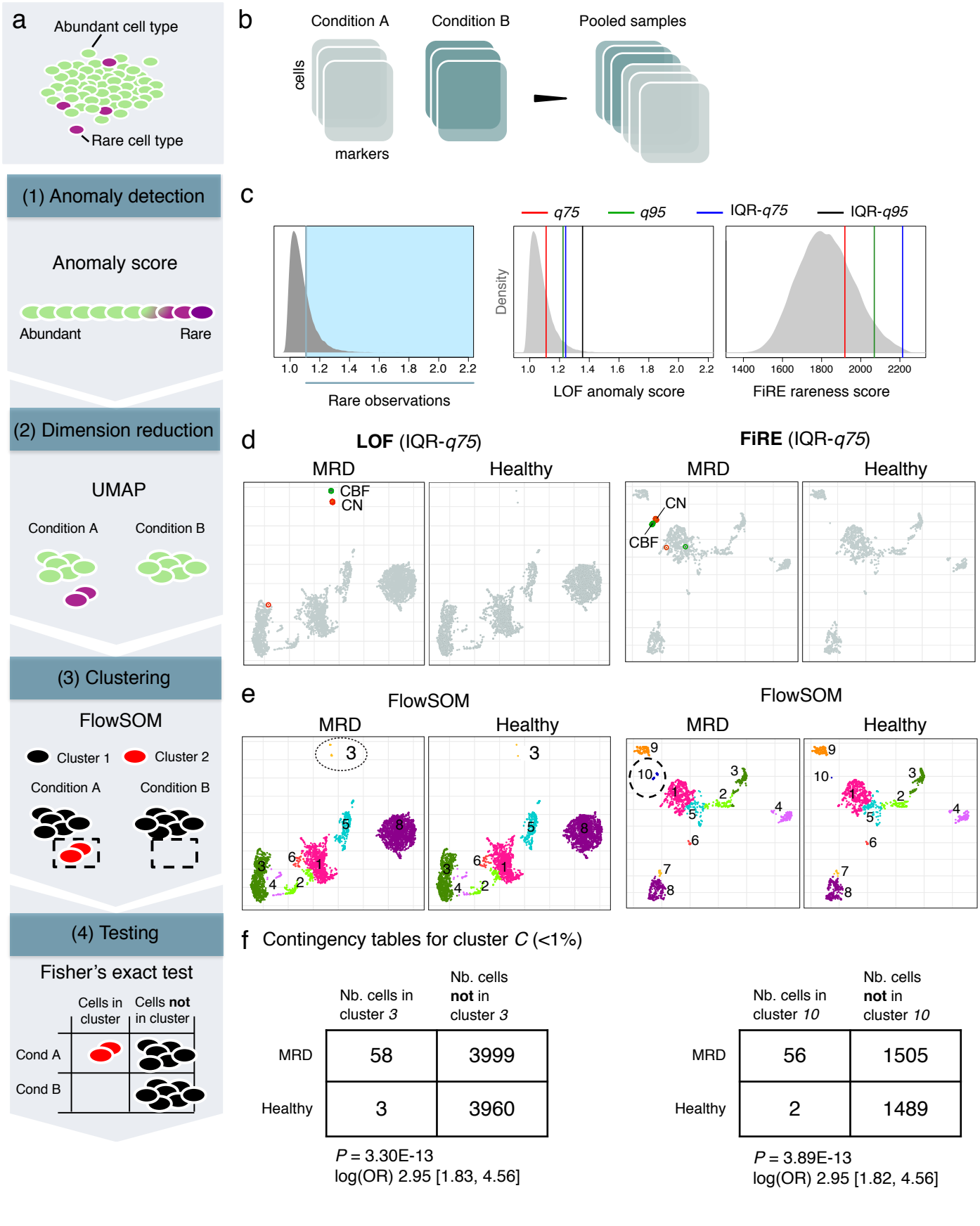
Figure 3: Benchmark results for the identification of rare iNKT cells (0.01%) in systemic lupus erythematosus patients and healthy donors (dataset 2). **(a)** UMAP representations of candidate rare cells selected by applying different cut-offs to the LOF score: *q75* (N = 89,013 cells), *q95* (N = 17,803), IQR-*q75* (N = 11,558 cells) and IQR-*q95* (N = 2,166 cells). The clusters defined by FlowSOM are shown by different colors and black squares show the cluster containing the majority of iNKT cells (violet dots). **(b)** Heat-map showing the expression profile of cells in cluster 3 (*q95*). Blue-to-yellow color-code shows low-to-high markers expression. LOF score is shown in the first column (green shades). Cell labels are shown in the second column. **(c)** Detection performance for LOF-FlowSOM (purple) and FiRE-FlowSOM (blue) for datasets containing 0.1, 0.05 and 0.01 % of iNKT cells. For each score cut-off (circle = *q75*, triangle = *q95*, square = IQR-*q75* and prism = IQR-*q95*) the plots show sensitivity (top) and positive predictive value (bottom) after the whole procedure (step 1-3). Black boxes indicate the results for clusters in black boxes in (a).

Figure 4: Performance results for the identification of HIV-specific B cells among non-naïve B cells in the lymph nodes of HIV$^+$ ART treated and untreated patients (dataset 3). **(a)** Boxplots of sensitivity and positive predictive value for 100 sub-samples of HIV-specific B cells when using LOF-FlowSOM (purple colors distinguishing the different number of nearest-neighbors' values (*k*)) and FiRE-FlowSOM (blue) applied on the dataset containing 0.05% of target cells. The results are reported for cut-off. **(b)** Execution time recorded for FiRE (blue) and LOF (using different *k* values, purple colors) with increasing number of cells from 10k to 320k.

Figure 5: Benchmark results for the identification of rare HIV-specific B cells among non-naïve B cells in the lymph nodes of HIV$^+$ ART treated and untreated patients (dataset 3). **(a)** ROC curves and corresponding AUC values on datasets containing 0.1%, 0.05% or 0.01% of rare target cells (LOF at the top, FiRE at the bottom). Marks on the curves indicate the position of the four tested cut-offs. There are no cells that reach the most stringent cutoff on the FiRE score. The color scale of the curve

indicates the anomaly score values. **(b)** UMAP representations for dataset containing 0.01% of target cells selected by applying different cut-offs to the LOF score: *q75* (N = 107,548 cells), *q95* (N = 21,510), IQR-*q75* (N = 18,577cells) and IQR-*q95* (N = 4,114 cells). The clusters defined by FlowSOM are shown by different colors and black squares show the cluster containing the majority of iNKT cells (violet dots). **(b)** Heat-map showing the expression profiles of cells in cluster 2 (IQR-*q75*). Blue-to-yellow color-code shows low-to-high markers expression. LOF score is shown in the first column (green shades). Cell labels are shown in the second column. **(c)** Detection performance for LOF-FlowSOM (purple) and FiRE-FlowSOM (blue) for datasets containing 0.1, 0.05 and 0.01 % of target cells. For each score cut-off (circle = *q75*, triangle = *q95*, square = IQR-*q75* and prism = IQR-*q95*) the plots show sensitivity (top) and positive predictive value (bottom) for a selected cluster (step 1-3). Black boxes indicate the results for clusters shown in (a).

Figure 1



f Contingency tables for cluster *C* (<1%)

	Nb. cells in cluster *3*	Nb. cells **not** in cluster *3*
MRD	58	3999
Healthy	3	3960

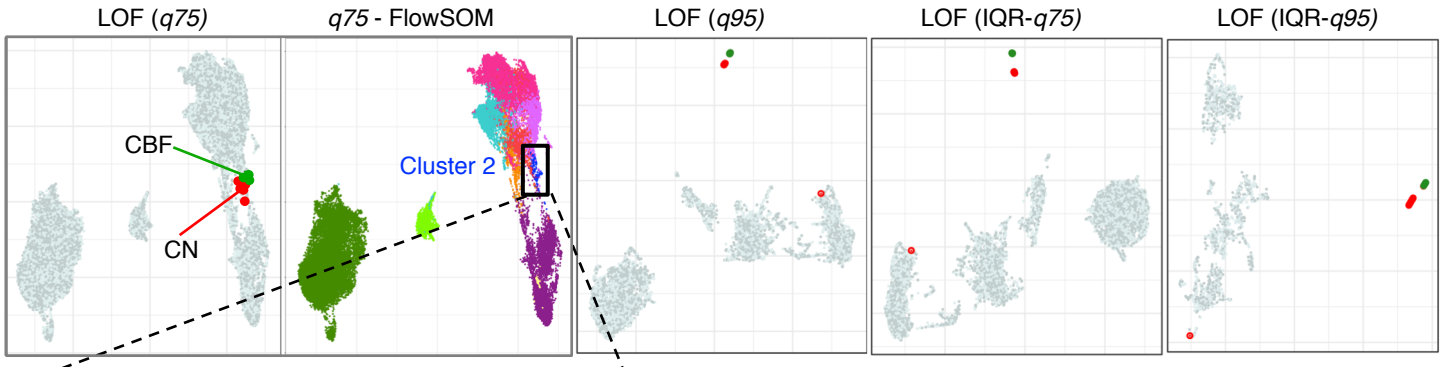
P = 3.30E-13
log(OR) 2.95 [1.83, 4.56]

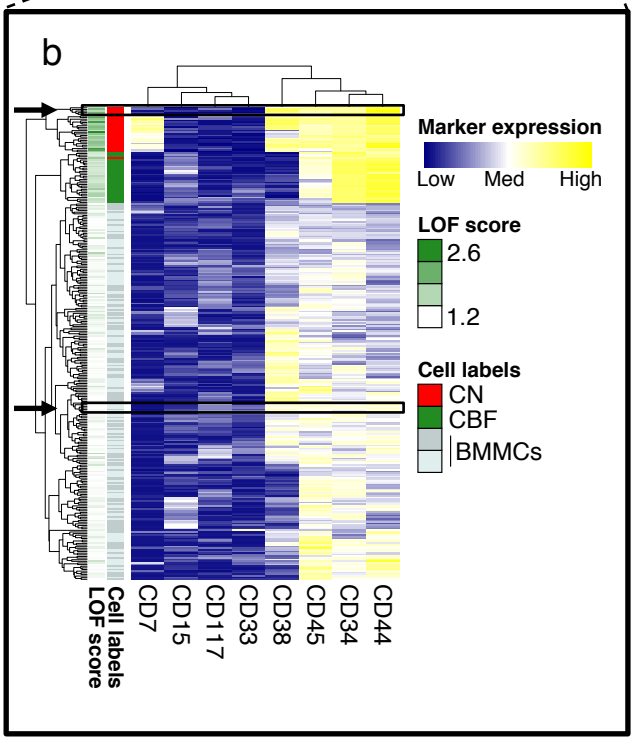
	Nb. cells in cluster *10*	Nb. cells **not** in cluster *10*
MRD	56	1505
Healthy	2	1489

P = 3.89E-13
log(OR) 2.95 [1.82, 4.56]

Figure 2

a


LOF (q75)
q75 - FlowSOM
LOF (q95)
LOF (IQR-q75)
LOF (IQR-q95)
CBF
CN
Cluster 2


b
Marker expression
Low
Med
High
LOF score
2.6
1.2
Cell labels
CN
CBF
BMMCs
Cell labels
LOF score
CD7
CD15
CD117
CD33
CD38
CD45
CD34
CD44

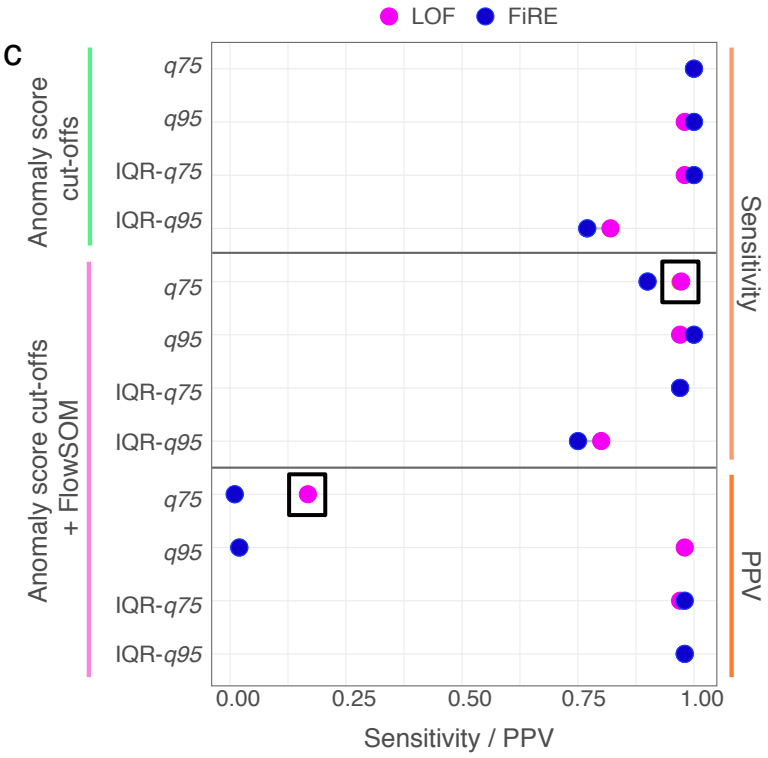

c
LOF
FiRE
Anomaly score cut-offs
Anomaly score cut-offs + FlowSOM
q75
q95
IQR-q75
IQR-q95
Sensitivity
PPV
0.00
0.25
0.50
0.75
1.00
Sensitivity / PPV

Figure 3

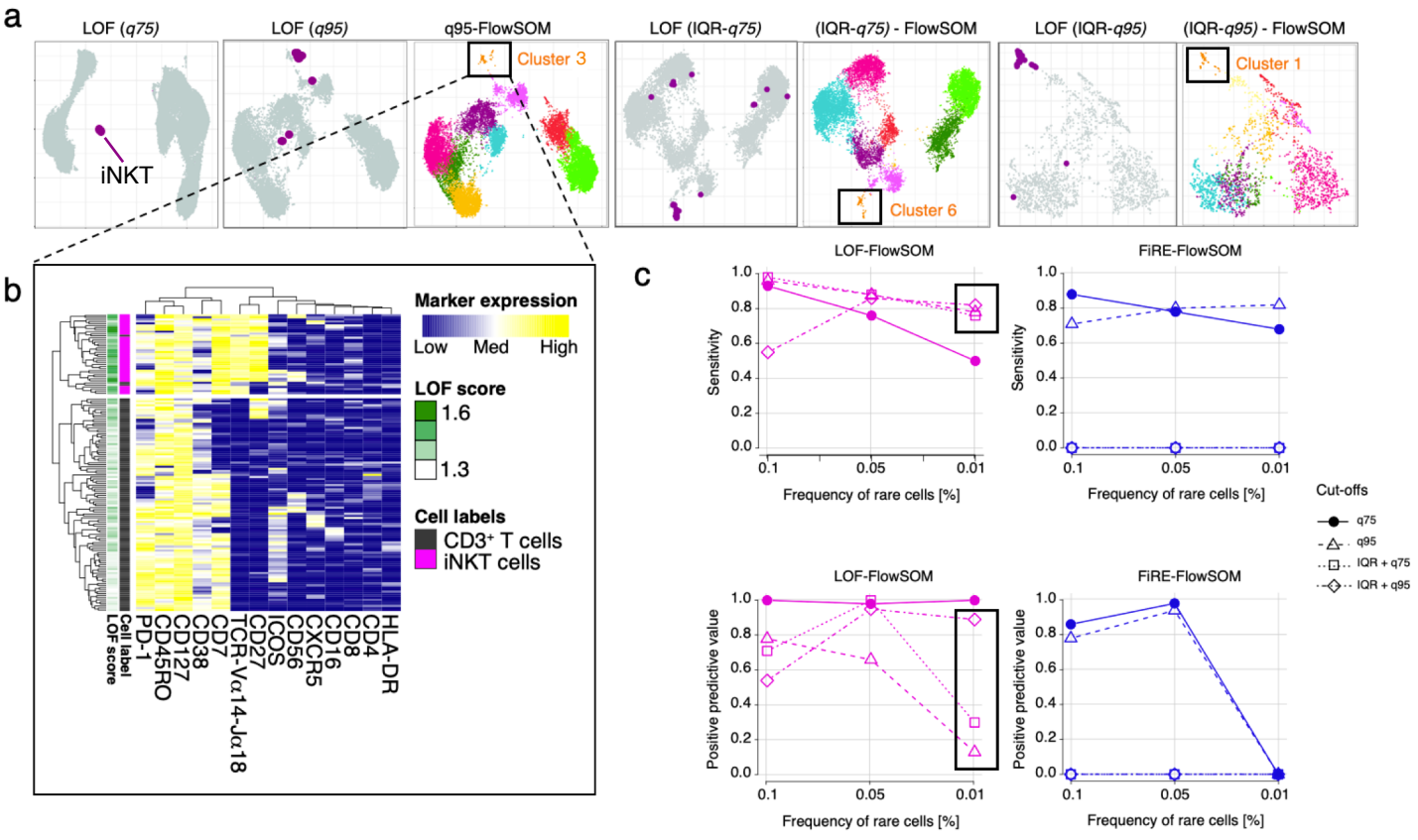

a
LOF (q75)
LOF (q95)
q95-FlowSOM
Cluster 3
iNKT
LOF (IQR-q75)
(IQR-q75) - FlowSOM
Cluster 6
LOF (IQR-q95)
(IQR-q95) - FlowSOM
Cluster 1
b
Marker expression
Low
Med
High
LOF score
1.6
1.3
Cell labels
CD3+ T cells
iNKT cells
Cell label
LOF score
PD-1
CD45RO
CD127
CD38
CD7
TCR-Vα14-Jα18
CD27
ICOS
CD56
CXCR5
CD16
CD8
CD4
HLA-DR
c
LOF-FlowSOM
FiRE-FlowSOM
Sensitivity
Positive predictive value
Frequency of rare cells [%]
0.1
0.05
0.01
Cut-offs
q75
q95
IQR + q75
IQR + q95

Figure 4

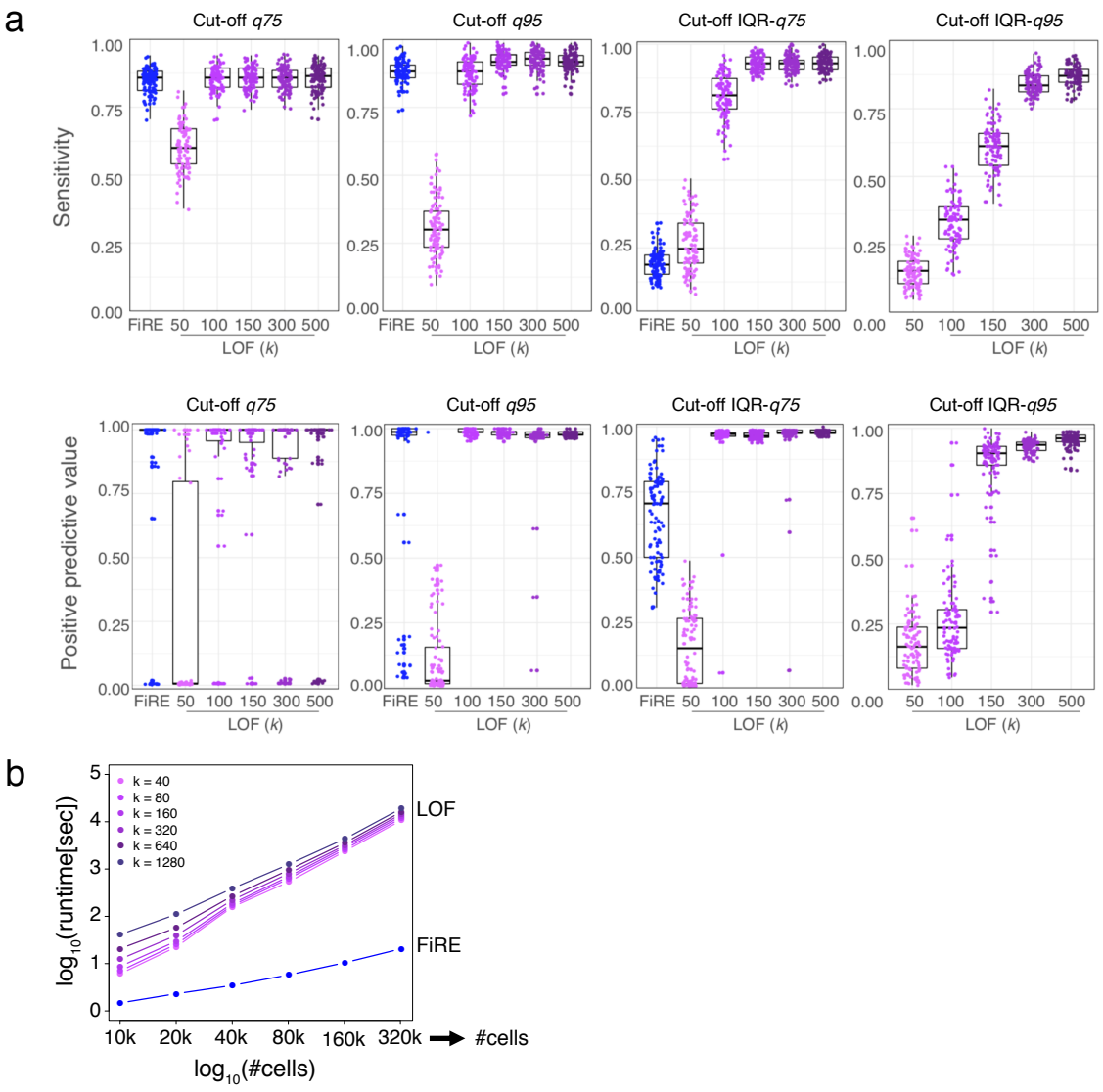
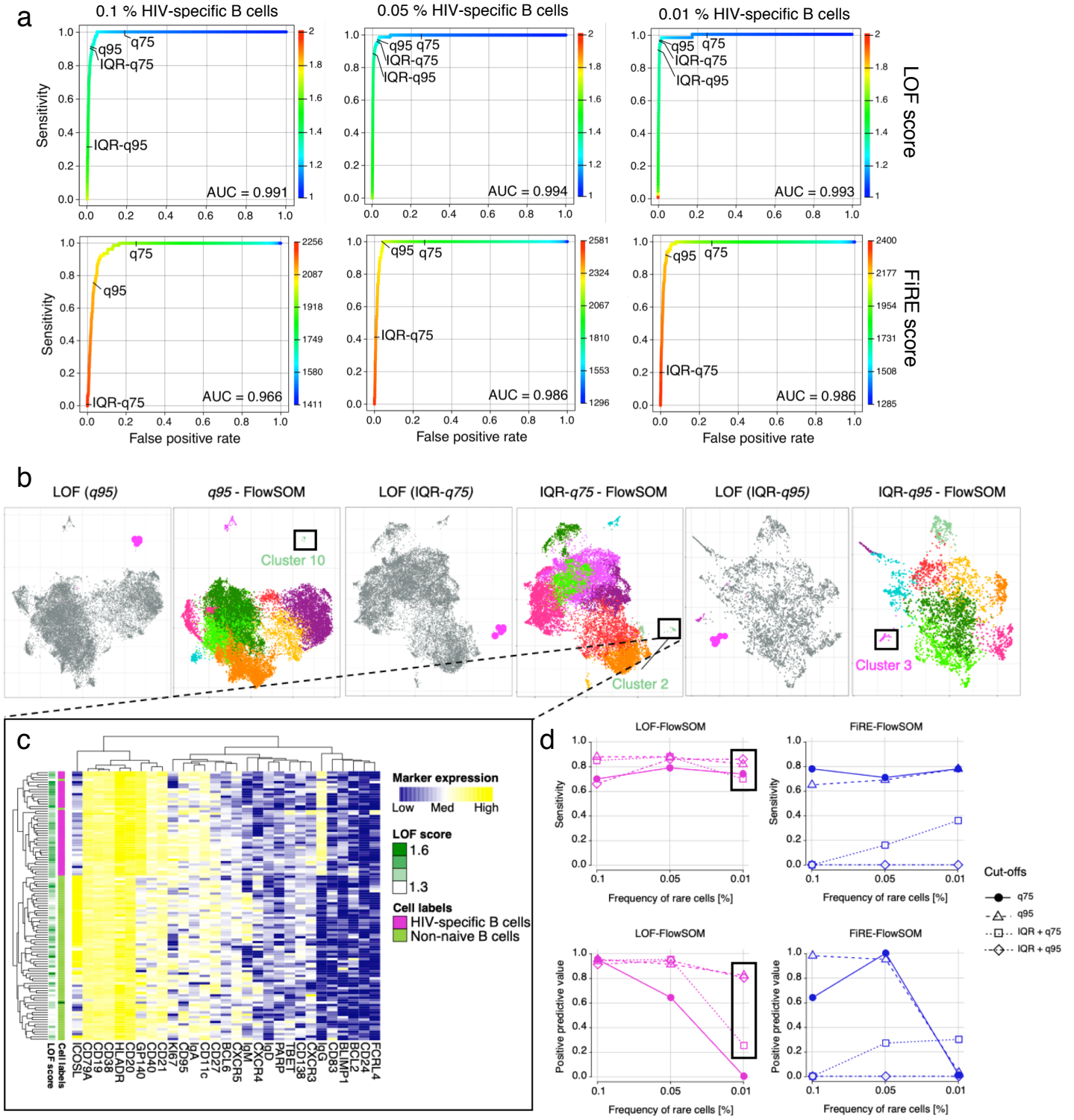

a
Cut-off q75
Cut-off q95
Cut-off IQR-q75
Cut-off IQR-q95
Sensitivity
Positive predictive value
1.00
0.75
0.50
0.25
0.00
FiRE 50 100 150 300 500
50 100 150 300 500
LOF (k)
b
log10(runtime[sec])
k = 40
k = 80
k = 160
k = 320
k = 640
k = 1280
LOF
FiRE
10k 20k 40k 80k 160k 320k
#cells
log10(#cells)

Figure 5


a
0.1 % HIV-specific B cells
0.05 % HIV-specific B cells
0.01 % HIV-specific B cells
Sensitivity
False positive rate
q75
q95
IQR-q75
IQR-q95
AUC = 0.991
AUC = 0.994
AUC = 0.993
AUC = 0.966
AUC = 0.986
AUC = 0.986
LOF score
FiRE score
b
LOF (q95)
q95 - FlowSOM
LOF (IQR-q75)
IQR-q75 - FlowSOM
LOF (IQR-q95)
IQR-q95 - FlowSOM
Cluster 10
Cluster 2
Cluster 3
c
Marker expression
Low Med High
LOF score
1.6
1.3
Cell labels
HIV-specific B cells
Non-naive B cells
LOF score
Cell labels
iCOSL
CD79A
CD19
CD38
HLADR
CD20
GP140
CD40
CD21
Ki67
CD95
IgA
CD11c
CD27
BCL6
CXCR5
IgM
CXCR4
IgD
PARP
TBET
CD138
CXCR3
IgG
CD83
BLIMP1
BCL2
CD24
FCRL4
d
LOF-FlowSOM
FiRE-FlowSOM
Sensitivity
Positive predictive value
Frequency of rare cells [%]
Cut-offs
q75
q95
IQR + q75
IQR + q95

Tables

Table 1: Summary of the datasets used to evaluate the performance of the D-AREdevil approach.

Dataset	Nb. cells	Nb. variables	Rare cells	Abundant cells	% rare cells
Dataset 1: AML blast cells	157,553	31	Two AML blast cells subsets	BMMCs	0.04
Dataset 2: iNKT cells	80,080	34	iNKT cells	$CD3^+$ T cells	0.1
	160,080	34	iNKT cells	$CD3^+$ T cells	0.05
	356,050	34	iNKT cells	$CD3^+$ T cells	0.01
Dataset 3: HIV-specific B cells	79,966	33	HIV-specific B cells	Non-naive B cells	0.1
	168,081	33	HIV-specific B cells	Non-naive B cells	0.05
	430,189	33	HIV-specific B cells	Non-naive B cells	0.01

Table 2: Fisher's exact test results for dataset 1 (AML blast cells) using the LOF-FlowSOM methodology with parameter k = 100. We reported results in log for the odds ratio (OR) and its 95% confidence interval. In bold are shown the clusters containing the target rare cells. A positive log(OR) corresponds to a positive association with the simulated MRD condition.

Cut-off	Cluster	*P*-value	Log(OR)	CI95% Log(OR)
q75	**2**	5.81E-02	0.30	[0.05, 0.55]
	5	1.00E+00	0.06	[-0.16, 0.29]
	7	1.00E+00	0.08	[-0.08, 0.23]
	10	2.77E-01	0.22	[-0.02, 0.45]
q95	2	1.00E+00	0.01	[-0.17, 0.19]
	3	1.00E+00	-0.31	[-0.89, 0.25]
	4	1.00E+00	-0.02	[-0.15, 0.12]
	5	1.00E+00	-0.08	[-0.23, 0.07]
	7	1.71E-15	4.05	[2.29, 7.73]
	8	1.00E+00	0.30	[-0.26, 0.87]
	9	6.20E-01	-0.22	[-0.47, 0.03]
	10	1.00E+00	0.14	[-0.21, 0.49]
IQR-q75	2	1.00E+00	[0.33	[-0.18, 0.84]
	3	2.65E-13	2.95	[1.83, 4.56]
	4	1.00E+00	-0.03	[-0.22, 0.16]
	5	1.00E+00	-0.03	[-0.16, 0.10]
	6	1.00E+00	-0.11	[-0.26, 0.05]
	7	1.00E+00	0.11	[-0.18, 0.39]
	8	1.00E+00	-0.17	[-0.47, 0.13]
	10	1.00E+00	-0.04	[-0.25, 0.17]
IQR-q95	1	2.70E-01	-0.22	[-0.43, -0.02]
	2	1.00E+00	-0.16	[-0.52, 0.19]
	4	5.71E-12	3.83	[2.06, 7.52]
	5	1.00E+00	0.24	[-0.18, 0.68]
	6	1.00E+00	-0.19	[-0.43, 0.06]
	7	1.00E+00	0.06	[-0.15, 0.28]
	8	1.00E+00	0.15	[-0.37, 0.67]
	9	1.00E+00	0.24	[-0.27, 0.76]
	10	1.00E+00	-0.01	[-0.52, 0.51]

Table 3: Fisher's exact test results for dataset 2 (iNKT cells) using the LOF-FlowSOM methodology with parameter k = 100. We reported results in log for the odds ratio (OR) and its 95% confidence interval. In bold are shown the clusters containing the target rare cells. A positive log(OR) corresponds to a positive association with the healthy condition.

Dataset 2: iNKT cells

0.1%				
Cut-off	Cluster	*P*-value	Log(OR)	CI95% Log(OR)
q75	3	1.00E+00	-0.60	[-1.95, 0.65]
	6	4.15E-04	0.86	[0.39, 1.37]
	1	4.20E-08	-1.11	[-1.53, -0.71]
	3	1.00E+00	0.11	[-0.07, 0.30]
q95	4	2.65E-52	-2.53	[-3.00, -2.10]
	5	3.94E-16	2.33	[1.61, 3.20]
	6	2.17E-04	0.68	[0.34, 1.03]
	7	3.41E-04	-0.39	[-0.59, -0.20]
	8	2.12E-08	0.55	[0.36, 0.74]
	9	1.16E-01	-0.91	[-1.73, -0.15]
IQR-q75	1	1.00E+00	-0.08	[-0.25, 0.09]
	2	3.00E-07	0.61	[0.38, 0.83]
	3	2.17E-05	-0.77	[-1.11, -0.44]
	4	2.16E-01	0.30	[0.04, 0.55]
	5	2.74E-03	0.68	[0.29, 1.08]
	6	1.60E-01	-1.17	[-2.31, -0.18]
	7	1.63E-02	0.31	[0.12, 0.51]
	8	9.83E-06	-0.50	[-0.71, -0.30]
	9	2.41E-15	2.22	[1.53, 3.03]
	10	1.83E-31	-2.59	[-3.25, -2.02]
IQR-q95	1	4.34E-09	2.96	[1.61, 5.10]
	2	1.00E+00	-0.50	[-1.34, 0.34]
	3	1.00E+00	-0.09	[-0.64, 0.47]
	4	1.07E-02	-2.71	[-6.49, -0.72]
	5	4.02E-10	-3.88	[-7.59, -2.07]
	6	1.89E-01	-0.48	[-0.89, -0.07]
	7	1.00E+00	0.43	[-0.30, 1.22]
	8	1.00E+00	-0.12	[-0.83, 0.61]
	9	5.40E-01	0.40	[-0.01, 0.81]
	10	1.00E+00	0.33	[-0.18, 0.87]

0.05%				
Cut-off	Cluster	*P*-value	Log(OR)	CI95% Log(OR)
q75	**5**	2.06E-13	3.25	[1.92, 5.38]
	10	4.30E-04	-0.74	[-1.15, -0.35]
q95	2	1.71E-59	-1.94	[-2.23, -1.66]
	4	9.07E-04	-0.25	[-0.37, -0.12]
	5	7.14E-08	0.66	[0.43, 0.90]
	6	2.05E-14	0.79	[0.58, 1.00]
	7	2.21E-10	0.41	[0.29, 0.53]
	8	9.56E-34	-1.42	[-1.68, -1.17]
	9	2.04E-15	2.13	[1.47, 2.89]
	10	1.25E-05	1.02	[0.56, 1.50]
IQR-q75	1	1.71E-17	-0.79	[-0.97, -0.61]
	2	1.91E-14	0.48	[0.36, 0.60]
	3	1.00E+00	-0.10	[-0.30, 0.10]
	4	2.52E-52	-2.23	[-2.61, -1.87]
	5	2.23E-07	1.65	[0.97, 2.43]
	6	6.96E-15	0.85	[0.63, 1.07]
	7	5.76E-07	-1.54	[-2.24, -0.91]
	8	1.00E+00	1.43	[-0.76, 5.29]
	9	1.30E-01	-0.15	[-0.27, -0.03]
	10	1.91E-04	0.32	[0.17, 0.47]
IQR-q95	1	1.27E-06	0.90	[0.54, 1.27]
	2	1.03E-01	-0.54	[-0.97, -0.11]
	3	7.68E-02	-0.67	[-1.19, -0.16]
	4	2.51E-02	-0.49	[-0.81, -0.16]
	5	1.64E-01	0.42	[0.06, 0.78]
	6	1.00E+00	0.10	[-0.44, 0.66]
	7	1.61E-03	1.09	[0.47, 1.79]
	8	6.62E-03	-0.71	[-1.14, -0.28]
	9	6.22E-02	0.76	[0.19, 1.38]
	10	7.99E-03	-0.69	[-1.11, -0.27]

0.01%				
Cut-off	Cluster	*P*-value	Log(OR)	CI95% Log(OR)
q75	**7**	1.26E-05	3.01	[1.20, 6.73]
q95	1	1.55E-187	-2.16	[-2.35, -1.99]
	2	4.14E-16	-1.16	[-1.46, -0.87]
	3	2.44E-19	2.09	[1.52, 2.75]
	4	1.00E+00	-0.02	[-0.12, 0.08]
	6	6.40E-34	-0.61	[-0.71, -0.51]
	7	3.74E-17	0.60	[0.46, 0.74]
	9	2.79E-27	0.46	[0.37, 0.54]
IQR-q75	1	1.00E+00	-0.06	[-0.15, 0.03]
	2	3.70E-12	0.37	[0.27, 0.47]
	3	2.45E-18	0.39	[0.31, 0.48]
	4	1.00E+00	-0.25	[-2.26, 1.76]
	5	6.78E-23	1.27	[0.99, 1.56]
	6	1.48E-04	-0.32	[-0.46, -0.17]
	7	3.63E-21	-0.58	[-0.70, -0.46]
	8	2.81E-14	1.49	[1.06, 1.96]
	9	1.29E-12	0.64	[0.47, 0.82]
	10	4.69E-134	-2.10	[-2.31, -1.90]
IQR-q95	**1**	4.77E-03	0.82	[0.33, 1.33]
	2	9.33E-20	-1.50	[-1.86, -1.16]
	3	4.25E-02	-0.48	[-0.82, -0.14]
	4	1.28E-01	0.53	[0.10, 0.98]
	5	1.00E+00	0.17	[-0.33, 0.70]
	6	1.00E+00	0.23	[-0.50, 1.00]
	7	1.00E+00	-0.08	[-0.27, 0.12]
	8	4.78E-02	0.38	[0.11, 0.65]
	9	1.00E+00	1.17	[-0.41, 3.42]
	10	7.94E-02	0.26	[0.06, 0.46]

Table 4: Fisher's exact test results for dataset 3 (HIV-specific B cells) using the LOF-FlowSOM methodology with parameter k = 300 (based on results obtained in Figure 4a). We reported results in log for the odds ratio (OR) and its 95% confidence interval. In bold are shown the clusters containing the target rare cells. A positive log(OR) corresponds to a positive association with the untreated condition.

Dataset 3: HIV-specific B cells

0.1%				
Cut-off	Cluster	*P*-value	Log(OR)	CI95% Log(OR)
q75	2	1.08E-08	0.65	[0.42, 0.88]
	5	1.00E+00	0.04	[0.19, 0.27]
	10	5.08E-09	Inf	[1.89, Inf]
q95	1	3.79E-12	Inf	[2.22, Inf]
	3	7.54E-07	2.27	[1.15, 3.87]
	4	5.63E-13	0.76	[0.55, 0.98]
	5	1.00E+00	0.94	[-0.55, 3.16]
	6	3.47E-28	-1.00	[-1.18, -0.83]
	7	1.03E-02	1.18	[0.39, 2.14]
	8	3.01E-02	-0.57	[-0.94, -0.18]
	9	3.50E-04	-0.37	[-0.55, -0.19]
	10	1.22E-73	-1.66	[-1.84, -1.47]
IQR-q75	1	1.01E-35	3.24	[2.38, 4.38]
	2	8.13E-38	1.52	[1.25, 1.79]
	3	2.27E-06	2.24	[1.12, 3.84]
	4	9.27E-09	-0.65	[-0.85, -0.44]
	5	1.31E-13	1.25	[0.89, 1.65]
	6	1.25E-02	0.66	[0.24, 1.11]
	7	1.78E-22	-1.35	[-1.63, -1.08]
	8	2.50E-03	-0.38	[-0.59, -0.18]
	9	1.16E-24	-1.24	[-1.47, -1.00]
	10	2.18E-15	-1.48	[-1.86, -1.11]
IQR-q95	1	2.13E-06	-1.68	[-2.41, -0.99]
	2	1.00E+00	0.14	[-0.43, 0.73]
	3	6.04E-01	-0.67	[-1.41, 0.08]
	4	2.02E-04	1.68	[0.76, 2.86]
	5	9.33E-04	-1.41	[-2.20, -0.66]
	6	1.00E+00	0.05	[-0.58, 0.73]
	7	7.71E-09	2.05	[1.21, 3.09]
	8	3.56E-05	2.97	[1.18, 6.67]
	9	3.46E-07	-0.93	[-1.27, -0.59]
	10	1.51E-01	0.47	[0.08, 0.87]

0.05%				
Cut-off	Cluster	*P*-value	Log(OR)	CI95% Log(OR)
q75	**10**	1.76E-22	4.30	[2.56, 7.97]
q95	1	1.79E-13	-0.43	[-0.54, -0.32]
	2	1.94E-07	-0.49	[-0.66, -0.32]
	3	2.73E-57	-0.91	[-1.03, -0.80]
	4	4.28E-41	-0.77	[-0.89, -0.66]
	5	1.20E-12	0.52	[0.37, 0.66]
	6	8.61E-09	2.02	[1.19, 3.06]
	7	1.18E-177	2.04	[1.86, 2.21]
	8	9.99E-16	2.24	[1.53, 3.11]
	9	6.17E-13	2.84	[1.73, 4.44]
	10	8.80E-32	Inf	[3.26, Inf]
IQR-q75	1	9.18E-05	0.36	[0.20, 0.52]
	2	5.78E-13	3.80	[2.05, 7.49]
	3	1.76E-19	-0.99	[-1.21, -0.77]
	4	3.80E-43	-0.90	[-1.03, -0.77]
	5	2.44E-29	-0.83	[-0.98, -0.69]
	6	6.31E-12	1.36	[0.93, 1.83]
	7	1.84E-166	2.31	[2.10, 2.53]
	8	2.81E-07	-0.43	[-0.58, -0.28]
	9	1.00E+00	-0.10	[-0.24, 0.05]
	10	3.22E-13	3.19	[1.87, 5.32]
IQR-q95	1	4.42E-01	0.50	[0.00, 1.02]
	2	6.77E-05	1.23	[0.63, 1.92]
	3	9.40E-11	0.97	[0.67, 1.28]
	4	7.58E-08	1.96	[1.12, 3.00]
	5	2.36E-01	-0.32	[-0.61, -0.03]
	6	3.04E-07	-0.77	[-1.04, -0.49]
	7	6.81E-01	-0.41	[-0.86, 0.05]
	8	2.25E-12	3.78	[2.02, 7.46]
	9	3.76E-07	-1.18	[-1.63, -0.73]
	19	4.78E-08	-0.94	[-1.27, -0.61]

0.01%				
Cut-off	Cluster	*P*-value	Log(OR)	CI95% Log(OR)
q75	5	7.56E-36	-0.85	[-0.99, -0.71]
	8	1.03E-56	2.49	[2.05, 2.98]
	9	1.09E-13	2.63	[1.65, 3.96]
q95	1	1.38E-84	-1.43	[-1.58, -1.27]
	2	7.92E-05	-0.51	[-0.75, -0.28]
	5	1.00E+00	-0.09	[-0.20, 0.03]
	6	1.38E-25	3.76	[2.45, 5.88]
	7	6.67E-33	-0.43	[-0.50, -0.36]
	9	3.48E-30	0.85	[0.70, 1.01]
	10	9.34E-08	Inf	[1.82, Inf]
IQR-q75	1	0.00E+00	2.14	[2.02, 2.25]
	2	1.26E-19	2.49	[1.74, 3.41]
	3	1.21E-11	0.63	[0.44, 0.81]
	4	6.76E-02	0.23	[0.06, 0.40]
	5	3.83E-24	3.35	[2.25, 4.94]
	7	9.03E-03	-0.13	[-0.21, -0.05]
	8	9.57E-41	-0.64	[-0.73, -0.54]
	9	9.52E-83	-1.44	[-1.60, -1.28]
	10	3.31E-26	-0.57	[-0.67, -0.46]
IQR-q95	1	1.95E-12	Inf	[2.32, Inf]
	2	4.50E-01	0.28	[0.01, 0.57]
	3	1.03E-10	Inf	[2.16, Inf]
	4	2.11E-46	1.10	[0.94, 1.27]
	5	1.56E-03	0.53	[0.24, 0.83]
	6	2.96E-19	-0.72	[-0.88, -0.57]
	7	1.00E+00	-0.07	[-0.25, 0.10]
	8	8.65E-26	-1.36	[-1.63, -1.10]
	9	3.59E-03	-0.48	[-0.75, -0.21]
	10	1.63E-07	-0.70	[-0.94, -0.45]

References

E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell. Dimensionality reduction for visualizing single-cell data using umap. *Na- ture biotechnology*, 37(1):38–44, 2019.

M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. Lof: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data*, pages 93–104, 2000.

R. V. Bruggner, B. Bodenmiller, D. L. Dill, R. J. Tibshirani, and G. P. Nolan. Automated identification of stratifying signatures in cellular subpopulations. *Proceedings of the National Academy of Sciences*, 111(26):E2770–E2777, 2014.

F. Castro-Giner and N. Aceto. Tracking cancer progression: From circulating tumor cells to metastasis. *Genome Medicine*, 12(1):1–12, 2020. ISSN 1756994X. doi: 10.1186/s13073- 020-00728-3.

V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009.

R. A. Cubas, J. C. Mudd, A.-L. Savoye, M. Perreau, J. Van Grevenynghe, T. Metcalf, E. Con- nick, A. Meditz, G. J. Freeman, G. Abesada-Terk, et al. Inadequate t follicular cell help impairs b cell immunity during hiv infection. *Nature medicine*, 19(4):494–499, 2013.

S. De Biasi, L. Gibellini, M. Nasi, M. Pinti, and A. Cossarizza. Rare cells: focus on detection and clinical relevance. In *Single Cell Analysis*, pages 39–58. Springer, 2017.

M. Farinacci, T. Krahn, W. Dinh, H.-D. Volk, H.-D. Düngen, J. Wagner, T. Konen, and O. von Ahsen. Circulating endothelial cells as biomarker for cardiovascular diseases. *Research and Practice in Thrombosis and Haemostasis*, 3(1):49–58, 2019. ISSN 2475-0379. doi: 10.1002/rth2.12158.

G. Finak, W. Jiang, and R. Gottardo. Cytoml for cross-platform cytometry data sharing. *Cy- tometry Part A*, 93(12):1189–1196, 2018.

M. Goldstein and S. Uchida. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PloS one*, 11(4):e0152173, 2016.

D. Grün, A. Lyubimova, L. Kester, K. Wiebrands, O. Basak, N. Sasaki, H. Clevers, and A. Van Oudenaarden. Single-cell messenger rna sequencing reveals rare intestinal cell types. *Nature*, 525(7568):251–255, 2015.

S. C. Hofmann, A. Bosma, L. Bruckner-Tuderman, M. Vukmanovic-Stejic, E. C. Jury, D. A. Isenberg, and C. Mauri. Invariant natural killer t cells are enriched at the site of cutaneous in- flammation in lupus erythematosus. *Journal of Dermatological Science*, 71(1):22–28, 2013.

L. Jiang, H. Chen, L. Pinello, and G.-C. Yuan. Giniclust: detecting rare cell types from single- cell gene expression data with gini index. *Genome biology*, 17(1):144, 2016.

A. Jindal, P. Gupta, Jayadeva, and D. Sengupta. Discovery of rare cells from voluminous single cell expression data. *Nature Communications*, 9(1):1–11, 2018. ISSN 20411723. doi: 10.1038/s41467-018-07234-6.

J. H. Levine, E. F. Simonds, S. C. Bendall, K. L. Davis, E. A. D. Amir, M. D. Tadmor, O. Litvin, H. G. Fienberg, A. Jager, E. R. Zunder, R. Finck, A. L. Gedman, I. Radtke, J. R. Downing, D. Pe'er, and G. P. Nolan. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*, 162(1):184–197, 2015. ISSN 10974172. doi: 10.1016/j.cell.2015.05.047. URL http://dx.doi.org/10.1016/j.cell.2015.05.047.

D. W. Li, Z. Q. Liu, J. Wei, Y. Liu, and L. S. Hu. Contribution of endothelial progenitor cells to neovascularization (review). *International Journal of Molecular Medicine*, 30(5):1000–1006, 2012. ISSN 11073756. doi: 10.3892/ijmm.2012.1108.

M. Linderman, R. Bruggner, and M. R. Bruggner. Package ‘rclusterpp’. 2013.
L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning*

research, 9(Nov):2579–2605, 2008.

M. Nowicka, C. Krieg, L. M. Weber, F. J. Hartmann, S. Guglietta, B. Becher, M. P. Levesque, and M. D. Robinson. Cytof workflow: differential discovery in high-throughput high- dimensional cytometry datasets. *F1000Research*, 6, 2017.

F. Ravandi, R. B. Walter, and S. D. Freeman. Evaluating measurable residual disease in acute myeloid leukemia. *Blood Advances*, 2(11):1356–1366, 2018. ISSN 24739537. doi: 10.1182/bloodadvances.2018016378.

N. Samusik, Z. Good, M. H. Spitzer, K. L. Davis, and G. P. Nolan. Automated mapping of phe- notype space with single-cell data. *Nature Methods*, 13(6):493–496, 2016. ISSN 15487105. doi: 10.1038/nmeth.3863.

S. Schreier, S. Borwornpinyo, R. Udomsangpetch, and W. Triampo. An update of circulat- ing rare cell types in healthy adult peripheral blood: findings of immature erythroid pre- cursors. *Annals of Translational Medicine*, 6(20):406–406, 2018. ISSN 23055839. doi: 10.21037/atm.2018.10.04.

M. H. Spitzer and G. P. Nolan. Mass Cytometry: Single Cells, Many Features. *Cell*, 165(4):780–791, 2016. ISSN 10974172. doi: 10.1016/j.cell.2016.04.019. URL http://dx.doi.org/10.1016/j.cell.2016.04.019.

S. Van Gassen, B. Callebaut, M. J. Van Helden, B. N. Lambrecht, P. Demeester, T. Dhaene, and Y. Saeys. Flowsom: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry Part A*, 87(7):636–645, 2015.

L. M. Weber and M. D. Robinson. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry Part A*, 89(12):1084–1096, 2016. ISSN 15524930. doi: 10.1002/cyto.a.23030.

L. M. Weber, M. Nowicka, C. Soneson, and M. D. Robinson. diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering. *Communications biology*, 2(1): 1–11, 2019.

H. Zhang, V. Vakil, M. Braunstein, E. L. Smith, J. Maroney, L. Chen, K. Dai, J. R. Berenson, M. M. Hussain, U. Klueppelberg, et al. Circulating endothelial progenitor cells in multiple myeloma: implications and significance. *Blood*, 105(8):3286–3294, 2005.

Supplementary Information

Supplementary Methods

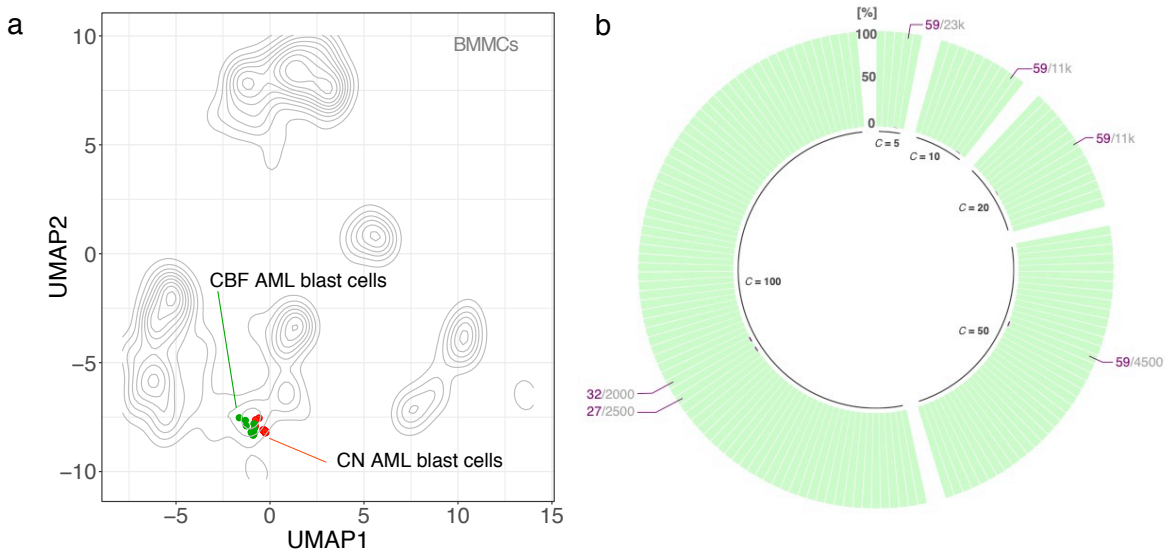
FCS files (dataset 1 and 3) were normalized to the EQ Four Element Calibration Beads using the CyTOF software. FCS files (dataset 3) were de-barcoded using Cytobank software. Cell populations used as reference-standard for the analyses (dataset 1 and 3) were manually gated for major and rare cell populations in CytoBank Data Analysis Software (dataset 1) or FlowJo v10.4.2 (Treestar, Inc., Ashland, CR) (dataset 3). FCS files as well as major and rare cell populations' labels were imported in R software v3.6 using CytoML package (from Bioconductor) [Finak et al., 2018]. Input data consist in a single protein-expression matrix for each dataset, where rows correspond to cells and columns to markers and meta-data (*i.e.* sample IDs, study groups, cell label). Markers intensity values (ion counts) were transformed using an inverse hyperbolic sine (arcsinh) with cofactor 5 [Nowicka et al., 2017]. The arcsinh transformation allows to reduce skewness of markers distribution, similarly to log transformation it de-emphasizes high values but can handle zeros or negative values (quasi-linear close to zero values). The cofactor of the function controls the width of the quasi-linear region; cofactor 5 correspond to standard value for mass cytometry, while 150 is used for flow cytometry [Weber et al., 2019].

Supplementary Tables

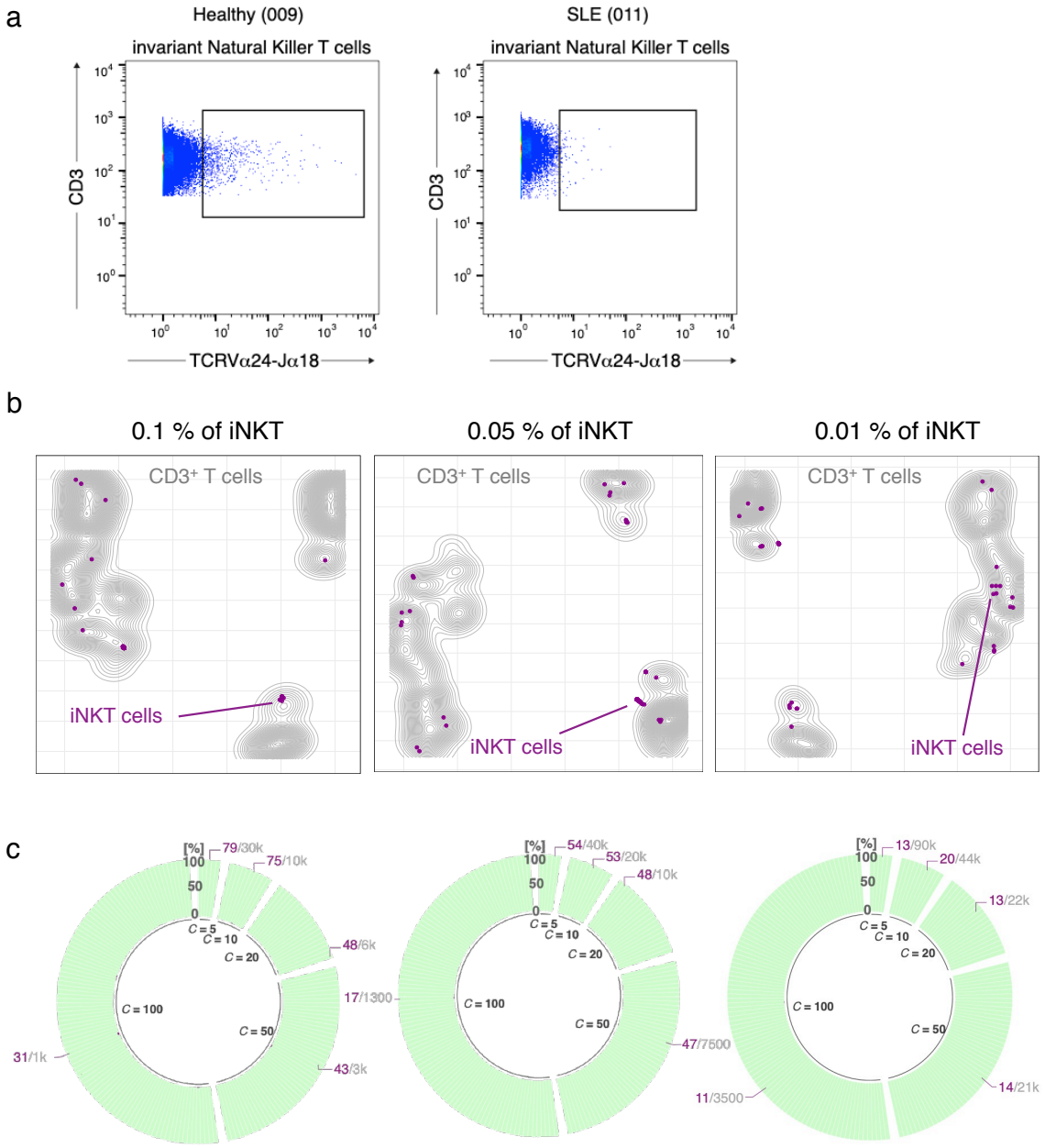
Supplementary Table 1: Sensitivity (true positive rate, TPR) and precision (or positive predictive value, PPV) of 100 sub-sampling of HIV-specific B cells (target rare cells) in dataset 3. We report 95% confidence intervals for TRP and PPV and the percentage of sub-samples with TPR/PPV higher than 0.5 when applying *q75*, *q95*, IQR-*q75* and IQR-*q95* cut-offs. We show the results for FiRE and LOF with different number of nearest-neighbors (*k*).

Method (*q75*)	TPR (95%CI)	PPV (95%CI)	%TPR > 0.5	%PPV > 0.5
FiRE	0.85 ± 0.008	0.87 ± 0.064	100	88
LOF ($k = 50$)	0.60 ± 0.017	0.26 ± 0.084	88	26
LOF ($k = 100$)	0.86 ± 0.008	0.81 ± 0.074	100	83
LOF ($k = 150$)	0.86 ± 0.008	0.85 ± 0.066	100	87
LOF ($k = 300$)	0.86 ± 0.008	0.83 ± 0.071	100	84
LOF ($k = 500$)	0.86 ± 0.008	0.88 ± 0.061	100	89
Method (*q95*)	TPR (95%CI)	PPV (95%CI)	%TPR > 0.5	%PPV > 0.5
FiRE	0.88 ± 0.008	0.86 ± 0.006	100	87
LOF ($k = 50$)	0.30 ± 0.020	0.11 ± 0.030	5	0
LOF ($k = 100$)	0.87 ± 0.010	0.99 ± 0.002	100	100
LOF ($k = 150$)	0.92 ± 0.007	0.98 ± 0.002	100	100
LOF ($k = 300$)	0.92 ± 0.007	0.96 ± 0.022	100	98
LOF ($k = 500$)	0.92 ± 0.006	0.98 ± 0.002	100	100
Method (IQR-*q75*)	TPR (95%CI)	PPV (95%CI)	%TPR > 0.5	%PPV > 0.5
FiRE	0.19 ± 0.009	0.826 ± 0.66	0	73
LOF ($k = 50$)	0.26 ± 0.020	0.16 ± 0.077	0	1
LOF ($k = 100$)	0.81 ± 0.016	0.96 ± 0.067	100	99
LOF ($k = 150$)	0.93 ± 0.005	0.97 ± 0.066	100	100
LOF ($k = 300$)	0.93 ± 0.006	0.96 ± 0.070	100	99
LOF ($k = 500$)	0.93 ± 0.006	0.98 ± 0.070	100	100
Method (IQR-*q95*)	TPR (95%CI)	PPV (95%CI)	%TPR > 0.5	%PPV > 0.5
LOF ($k = 50$)	0.15 ± 0.011	0.66 ± 0.034	0	2
LOF ($k = 100$)	0.33 ± 0.017	0.16 ± 0.028	3	5
LOF ($k = 150$)	0.60 ± 0.017	0.96 ± 0.020	89	96
LOF ($k = 300$)	0.84 ± 0.008	0.97 ± 0.003	100	100
LOF ($k = 500$)	0.87 ± 0.007	0.96 ± 0.020	100	100

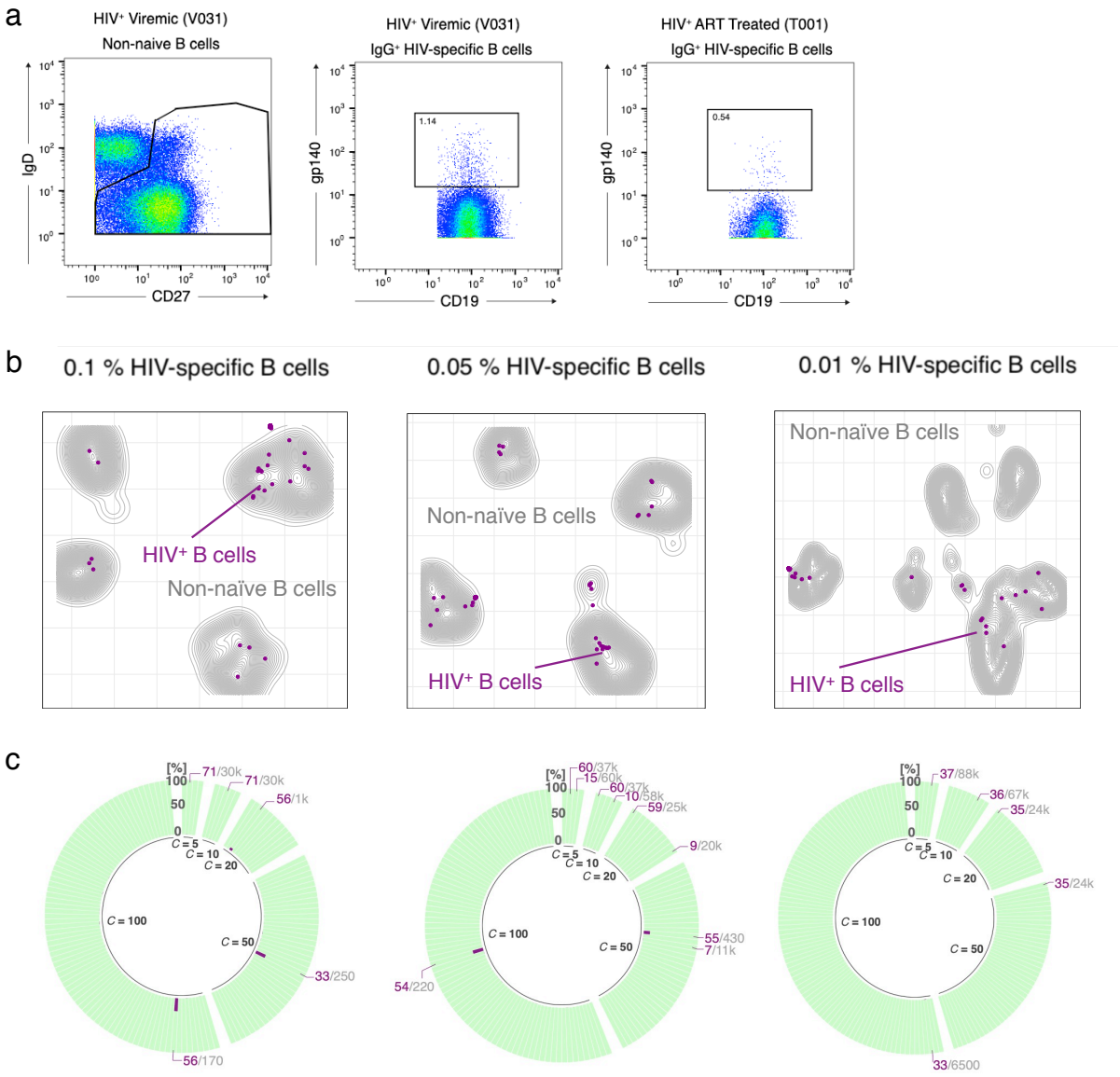
Supplementary Figures



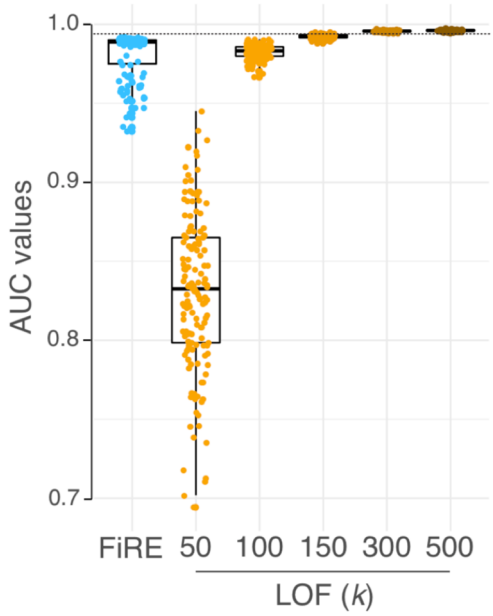
Supplementary Figure 1: AML blast cells dataset **(a)** UMAP representation of the whole dataset. 0.02% of AML blast cells CBF subset (green dots) and 0.02% of CN subset (red dots) among BMMCs (grey density lines) **(b)** FlowSOM results on the whole dataset for increasing number of total clusters (C = 5 clusters, C = 10, C = 20, C = 50 and C = 100). The green bars represent non-target cells (*i.e.* BMMCs) and in violet are shown the target AML blast cells. We indicated the number of target cells (in violet) among the BMMCs (approx. value) for the most enriched cluster.



Supplementary Figure 2: iNKT cells dataset **(a)** Gating strategy of iNKT cells in one healthy subject and one SLE patient. **(b)** UMAP representations of the whole datasets. iNKT are present at different frequencies 0.1%, 0.05% and 0.01% (violet dots) among CD3⁺ T cells (grey density lines). **(c)** FlowSOM results on the whole dataset for increasing number of total clusters (C = 5 clusters, C = 10, C = 20, C = 50 and C = 100). The green bars represent non-target cells (*i.e.* CD3⁺ T cells) and in violet are shown the target iNKT cells. We indicated the number of target cells (in violet) among the CD3⁺ T cells (approx. value) for the most enriched cluster.



Supplementary Figure 3: HIV-specific B cells dataset (a) Gating strategy of non-naïve B cells and HIV-specific B cells in one HIV+ viremic subject and one HIV+ ART treated subject. (b) UMAP representations of the whole datasets. HIV-specific B cells are present at different frequencies 0.1%, 0.05% and 0.01% (violet dots) among non-naïve B cells (grey density lines). (c) FlowSOM results on the whole dataset for increasing number of total clusters (C = 5 clusters, C = 10, C = 20, C = 50 and C = 100). The green bars represent non-target cells (*i.e.* non-naive B cells) and in violet are shown the target HIV+ B cells. We indicated the number of target cells (in violet) among the non-naïve B cells (approx. value) for the most enriched cluster.



Supplementary Figure 4: Boxplots of area under the ROC curve (ROC AUC) calculated on test samples for 100 sub-samples of HIV-specific B cells (at frequency 0.05%) and fixed major B cell types in lymph nodes of HIV+ ART treated and untreated patients. In blue are shown results for FiRE and in orange colors are shown LOF results for increasing number of nearest-neighbors (k).

The Deficiency in Th2-like Tfh Cells Affects the Maturation and Quality of HIV-Specific B Cell Response in Viremic Infection

Alessandra Noto[1], Madeleine Suffiotti[1], Line Esteves-Leuenberger[1], Francesco A. Procopio[1], Victor Joo[1], Guy Cavet[2], Yvonne Leung[2], Jean-Marc Corpataux[3], Matthias Cavassini[4], Agostino Riva[5], Leonidas Stamatatos[6,], Raphael Gottardo[7], Adrian McDermott[8], Richard A. Koup[8], Craig Fenwick[1], Giuseppe Pantaleo[1,9]*

[1]Service Immunology and Allergy, Lausanne University Hospital, University of Lausanne, Lausanne, Switzerland.

[2]Atreca, Redwood City, CA USA

[3]Service of Vascular Surgery, Lausanne University Hospital, University of Lausanne, Lausanne, Switzerland.

[4]Service of Infectious Diseases, Lausanne University Hospital, University of Lausanne, Lausanne, Switzerland.

[5]Division of Infectious Diseases, Luigi Sacco Hospital, University of Milan, Italy.

[6]Department of Global Health, Seattle University of Washington, Seattle, United States of America.

[7]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA

[8]Vaccine Research Center, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, USA

[9]Swiss Vaccine Research Institute, Lausanne University Hospital, University of Lausanne, Lausanne, Switzerland

*Corresponding author. Email: giuseppe.pantaleo@chuv.ch

Abstract

T follicular helper cells (Tfhs) promote the development of germinal centers and maturation of B cells. We have found that the expression of CXCR3 defines distinct single IFN-γ and dual IL-21/IFN-γ Th1-like Tfh ($CXCR3^+$) and single IL-4 and dual IL-21/IL-4 Th2-like Tfh ($CXCR3^-$) cytokine secreting cells. $CXCR3^-$ Th2-like Tfhs are significantly reduced during ongoing HIV replication. While the percentage of Th2-like Tfhs correlates with that of total and cycling HIV-specific B cells, the percentage of $CXCR3^+$ Th1-like Tfhs correlates with HIV-specific B cells expressing T-bet and CXCR3. Of note, only IL-4 and IL-21 cytokines boosted efficient maturation of HIV-specific B cells while IFN-γ induced expression of T-bet and CXCR3 in B cells and suppressed immunoglobulin production. Interestingly, total and HIV-specific $CXCR3^+$ B cells showed lower rate of somatic hypermutation, as compared to $CXCR3^-$ B cells. Therefore, the imbalance in Th2/Th1-like Tfhs is a major cause of B cell abnormalities in viremic HIV infection.

Introduction

T follicular helper CD4 T cells (Tfhs) are essential for germinal center formation, B cell proliferation, affinity maturation, isotype class switching, and generation of long lasting memory B cells and plasma cells[1-5]. Germinal center (GC) Tfhs are a very heterogenous population, phenotypically defined by coexpression of CXCR5 and PD-1 and by expression of the B-cell lymphoma 6 protein (BCL-6) transcription factor, and functionally characterized by the production of IL-21 and IL-4 cytokines that together optimally drive B cell maturation[6-11].

HIV infection is characterized by the expansion of Tfhs in viremic individuals and viremia levels correlate with increased frequency of GC B cells and plasma cells and decreased frequencies of naive, unswitched memory and switched memory B cell populations[12]. Despite an increase in their frequency, Tfhs from HIV-infected individuals are less effective at providing adequate B-cell help[13]. They remain capable of responding to HIV antigens but become functionally skewed[14,15]. The increased expression of Programmed death-ligand 1 (PD-L1) by GC B cells and the delivery of the inhibitory signal through the PD-1/PDL-1 interaction[13] has been suggested to explain the dysfunction of Tfhs.

Lymph node Tfhs from simian immunodeficiency virus (SIV) infected animals and peripheral Tfhs from HIV-1 infected individuals have been shown to have a polarized T helper (Th)1-like phenotype and to express increased levels of T-bet and CXCR3[16,17]. Futhermore, Th1-like Tfhs have increased production of Interferon (IFN)-γ and contain more copies of SIV DNA as compared to $CXCR3^-$ Tfhs in chronic SIV infection[16].

Of note, studies performed in Non-Human Primates (NHPs) and in humans have demonstrated that the frequency and the quality of Tfhs specific to Env drives the magnitude and the quality of the Env-specific B cell response[18] and that they are required for the development of

HIV broadly neutralizing antibodies[19]. In this regard, an elegant study performed in mice has shown that the progressive differentiation of Tfhs secreting IL-21 and IL-4 regulates the GC response[11].

Two recent studies performed in tonsils and lymph nodes obtained from healthy HIV negative and/or HIV-infected viremic individuals have started to uncover the phenotypic heterogeneity of Tfhs and highlighted certain phenotypic and functional (cytokines) differences between healthy and viremic individuals[15,20]

Previous studies have demonstrated that PD-1^{+}/Tfh CD4 T cells are the major cell reservoir in lymph nodes in both viremic and long term treated individuals[12,21]. Whether the role of Tfhs as a major cell reservoir for HIV has an impact on the generation of the antibody response remains unclear[21, 22, 23, 12].

In the present study, we have dissected the phenotypic and functional heterogeneity of Tfhs in lymph nodes of HIV infected viremic, long-term ART treated and healthy HIV negative individuals, and determined their impact on the development of the Env-specific B cell response. We show that the expression of CXCR3 defines a Th1-like Tfh cell population functionally enriched in single IFN-γ and dual IFN-γ/IL-21 cells while CXCR3^{-} Th2-like Tfh cells are enriched in single IL-4 and dual IL-4/IL-21 producing cells. Of note, the Th2-like Tfh cell population was significantly reduced in viremic individuals. We provide evidence that the reduction in this Tfh cell population and the imbalance in the Th2 and Th1-like Tfh cell populations are a major cause of the phenotypic and functional abnormalities in the B cell responses of HIV viremic individuals. These results will be also instrumental in developing strategies to optimize the induction of the magnitude and quality of antibody responses following vaccination.

Results

Characterization of Tfh cells

In order to dissect the heterogeneity of Tfhs, lymph node biopsies were obtained from 8 healthy HIV negative, 12 long-term ART treated and 9 viremic individuals (Table 1). Lymph node mononuclear cells (LNMCs) were then characterized with a unique panel of 30 markers of T cell activation, memory differentiation, chemokine receptors and HIV coreceptors (Supplementary Table 1a). Tfhs were defined by gating on memory CD4 T cells on the basis of the expression of PD-1 and CXCR5 and high levels of BCL-6 (Figure 1a and 1b). Unsupervised clustering was performed on pooled Tfh cells from the three study groups by a data-driven unsupervised clustering method, i.e., FlowSOM, in combination with consensus clustering. This analysis defined 20 different populations (i.e. clusters) within $CXCR5^{high}PD\text{-}1^{high}$ Tfhs. T-distributed stochastic neighbor embedding (t-SNE) was used to perform dimensions reduction and visualize our data in a two-dimensional plot that placed cells with similar phenotypic characteristics (in high dimensional space) in close proximity (Fig. 1c, Supplementary Fig. 1a and b). After clusters definition, we used the heat map (Fig. 1d) to show the phenotypic profile of each cluster in terms of median marker intensity for an individual marker. The 20 defined clusters showed variations in the expression of CD38, CXCR3, CD57, HLA-DR, CD127, CXCR4, and to less extend of CCR5, CD25, CCR7 and CD32 while Tfhs were homogeneous for the expression of CD27, ICOS and CD40L (Fig. 1d).

Next, we found that only 8 out of the 20 Tfh clusters were significantly differentially distributed in the three study groups (Fig. 1e and Supplementary Fig. 1c). Clusters 1, 8 and 11 were significantly decreased in viremics as compared to ART treated and healthy individuals.

These three clusters accounted for the 29.3% and 26.84% of the total Tfh cell population in ART treated and healthy individuals, respectively, and in viremics only 5.9% ($P < 0.0001$) (Fig. 1f).

Clusters 5, 10, 13 and 15 were over represented in viremics and accounted for 49.17% of the total Tfh cell population (cluster 5: 24.4%; cluster 10: 15.53%; cluster 13: 0.89%; cluster 15: 8.35%), while in ART treated and healthy individuals they accounted for 22.14% and 18.66%, respectively ($P < 0.0001$) (Fig. 1f).

Only cluster 12 was significantly increased in healthy as compared to ART treated ($P = 0.048$) and viremics ($P = 0.0014$) (on average healthy: 5.12%; ART treated: 2.63%; viremics: 0.89%) (Fig. 1e and f).

Clusters with higher frequencies in viremics co-expressed CXCR3 and CD38 with varying levels of CD57, CXCR4, HLA-DR and CD25 (Fig. 1d) while those abundant in ART treated and healthy individuals expressed lower levels of CD38 and CCR5 and were heterogeneous in CXCR3 expression (Fig. 1d). This was further confirmed by unsupervised principal component analysis (PCA) which showed that the markers that contributed the most to Tfh heterogeneity were CXCR3 and CD38 (Fig. 1g).

Alltogether these data highlight the phenotypic heterogeneity of Tfhs and define distinct phenotypic subsets of Tfhs that are differentially distributed in healthy, ART treated and viremic individuals.

Relationship between CD38 and CXCR3 Tfh cell populations

Having delineated the phenotypic markers defining the Tfh cell populations differentially distributed in the three study groups, we sought to investigate further the relationship between CD38 and CXCR3 and to confirm the findings generated by FlowSOM in combination with

consensus clustering by manual gating. The additional analyses confirmed that Tfhs co-expressing CD38 and CXCR3 were consistently increased in viremics as compared to healthy and ART treated individuals (Fig. 2a, b). SPICE analyses demonstrated that the frequency of the $CD38^{+}CXCR3^{+}$ Tfh cell population was 44.1% in viremics, 17.4% in ART treated individuals ($P < 0.0001$) and 6.7% in healthy ($P < 0.0001$). $CD38^{-}CXCR3^{+}$ cells were significantly reduced in viremics as compared to ART treated individuals (10.5% vs 21.19%, $P = 0.001$) and to healthy (20%, $P = 0.017$) (Fig. 2a, b). $CD38^{+}CXCR3^{-}$ cells were significantly increased in viremics as compared to healthy individuals (28.9% vs 14.11%, $P < 0.0001$) but not as compared to ART treated subjects (22.9%, $P = 0.14$) (Fig. 2a, b).

We next determined the expression of markers of memory cell differentiation, cell activation, and cell trafficking within the Tfh cell populations defined by the expression of CD38 and CXCR3 in the three study groups (Fig. 2c). The heat map shows that the large majority of markers defining T cell activation (Ki-67, CD25, HLA-DR) as well as classic Tfh markers (BCL-6, ICOS, CD40L) were significantly increased in $CD38^{+}CXCR3^{+}$ Tfhs as compared to the other Tfh cell populations. Of note, the expression of the HIV coreceptor CCR5 was greatly enriched within $CD38^{+}CXCR3^{+}$ Tfhs in the three study groups (67% in healthy, 55.8% ART treated and 44.5 % in HIV viremics) (Fig. 2c), suggesting that this subset might potentially be more susceptible to HIV infection. Interestingly, we found higher levels of cell-associated HIV RNA in $CXCR3^{+}$ as compared to $CXCR3^{-}$ Tfhs (Fig. 3). Therefore, these findings have identified a cell subset serving as HIV reservoir within the total Tfh cell population. Previous studies[16,24], have shown an enrichment of inducible replication competent HIV in blood $CXCR3^{+}$ CD4 T cells and an enrichment of SIV DNA in the same cell population.

The Th1 cell lineage T-box transcription factor (T-bet) was significantly increased in $CD38^+CXCR3^+$ Tfhs as compared to $CD38^-CXCR3^+$, $CD38^+CXCR3^-$ and $CD38^-CXCR3^-$ Tfhs ($P < 0.001$), while the Th2 transcription factor GATA-3 was significantly increased in $CD38^-CXCR3^-$ Tfhs (Fig. 2c) ($P < 0.01$). Of note, GATA-3 expression within the $CD38^-CXCR3^-$ Tfhs was significantly decreased in HIV infected viremic individuals as compared to ART treated subjects (0.81% in viremics vs 30.8% in ART treated, $P < 0.0001$) and HIV uninfected individuals (12.04%, $P = 0.0006$). Similarly, the Th2 specific chemokine, CCR4, was strongly reduced in all Tfhs of viremic individuals compared to both healthy and ART treated donors (1.6 to 2 fold, $P < 0.01$).

Taken together, these results indicate that the co-expression of CD38 and CXCR3 identifies a population of activated and Th1 polarized Tfhs, while Th2 polarized Tfhs were $CD38^-CXCR3^-$ and significantly reduced in viremic individuals.

Functional characterization of $CD38^+CXCR3^+$ Tfh cells

Tfhs produce high levels of IL-21, a cytokine that is critical for GC formation and B cell maturation[25,26]. Although IL-21 is the signature cytokine of Tfhs , studies have shown that Tfhs are also able to produce cytokines typical of other cell lineages of helper CD4 T cells[12]. Therefore, we determined the cytokine profile of total Tfhs and of the four Tfh cell populations defined by CXCR3 and CD38 expression. LNMCs from HIV uninfected, HIV-1 infected ART treated and viremic individuals were stimulated for 5 hours in the presence of PMA and ionomycin and the cytokine profile was evaluated by mass cytometry. The production of IFN-γ, IL-21, IL-4, IL-2 and TNF-α by total Tfhs within the three study groups using SPICE analysis is shown in Supplementary Figure 2a. Tfhs from healthy subjects were enriched in polyfunctional Tfhs producing five ($IFN\text{-}\gamma^+IL\text{-}21^+IL\text{-}4^+IL\text{-}2^+TNF\text{-}\alpha^+$) and four ($IFN\text{-}\gamma^-IL\text{-}21^+IL\text{-}4^+IL\text{-}2^+TNF\text{-}\alpha^+$)

cytokines as compared to ART treated and viremic subjects, while Tfhs from viremics were generally enriched in IL-21^{+}IL-4^{-}IL-2^{-} regardless of their ability to produce IFN-γ and TNF-α (Supplementary Fig. 2a).

Next, we assessed the cytokine profile in the four Tfh cell populations defined by the expression of CD38 and CXCR3 (Fig. 4a, b). t-SNE analysis showed a dichotomy in the distribution of IFN-γ and IL-4 between CXCR3^{+} and CXCR3^{-} Tfh cell populations whereas the two cytokines were spread within CD38^{+} and CD38^{-} Tfhs (Fig. 4a). The dichotomy in the distribution of IFN-γ and IL-4 between CXCR3^{+} and CXCR3^{-} Tfh cell populations was further confirmed in the heat map (Fig. 4b) analysing the distribution of all cytokines within the four Tfh cell populations. CD38^{-}CXCR3^{-} and CD38^{+}CXCR3^{-} Tfhs on one side and CD38^{-}CXCR3^{+} and CD38^{+}CXCR3^{+} on the other side share similar functional cytokine profiles and are closely related to each other. CXCR3^{-}CD38^{-} and CXCR3^{-}CD38^{+} Tfhs were predominantly enriched in IL-4 whereas CXCR3^{+}CD38^{-} and CXCR3^{+}CD38^{+} Tfhs were significantly enriched in IFN-γ and also in IL-21, IL-2 and TNF-α. Of note, in viremics there was a significant trend in the decrease of Tfhs secreting IL-2 across the four Tfh cell populations and in the percentage of Tfhs secreting IL-4 (about 2.7 fold reduction) as compared to healthy individuals (Fig. 4b and Supplementary Fig. 2b). The cytokine profile observed in ART treated individuals was generally intermediate beween healthy and viremic individuals apart from IL-4, which was strongly reduced in all CXCR3^{+} and/or CD38^{+} cell populations (Fig. 4b and Supplementary Fig. 2b).

Taken together these results demonstrate that the expression and/or the lack of CXCR3 defines functionally distinct Tfh cell populations, i.e. CXCR3^{+} Tfh/Th1-like IFN-γ^{+} and CXCR3^{-} Tfh/Th2-like IL-4^{+}.

A recent study performed in mice has demonstrated that IL-21 promotes efficient B cell proliferation and that both IL-21 and IL-4 are required for full maturation of B cells[11]. We therefore analyzed the distribution of IL-21, IL-4, and IFN-γ producing Tfhs within $CXCR3^+$ and $CXCR3^-$ cell populations in the three study groups. Interestingly, we identified Tfhs producing only IL-21 (single IL-21), only IFN-γ (single IFN-γ), IL-21 and IFN-γ (dual IL-21/IFN-γ), only IL-4 (single IL-4), or IL-21 and IL-4 (dual IL-21/IL-4). The expression and/or lack of CXCR3 segregated the single IL-4 and dual IL-21/IL-4 cytokine producing Tfhs within the $CXCR3^-$ Th2-like Tfhs and single IFN-γ and the dual IL-21/IFN-γ within the $CXCR3^+$ Th1-like Tfh cell populations (Fig. 4c). Similar to what was observed for the total IL-4 secreting Tfhs, the percentage of dual IL-21/IL-4, and to a lesser extent of single IL-4 Tfh, was reduced in viremic as compared to healthy individuals (Fig. 4c). Of note, the Th1-like/Th2-like Tfhs ratio calculated by the ratio between the frequency of total IFN-γ^+ (i.e. single IFN-γ^+ plus dual IFN-γ^+/IL-21) Tfhs and the frequency of total IL-4 producing Tfhs (Supplementary Fig. 2c) was significantly increased in viremic as compared to HIV negative individuals (5.1 vs 1.2 , $P = 0.003$) (Fig. 4d). Furthermore, treatment with ART showed a trend towards the recovery of Th2-like Tfhs (Fig. 4d and Supplementary Fig. 2c).

Therefore, these results indicate that the $CXCR3^-$ Th2-like Tfh cell populations potentially important for promoting efficient B cell maturation are quantitatively reduced in viremic individuals.

Frequency and distribution of gp140-specific B cells in ART treated and viremic individuals

We have previously shown that during the viremic phase of HIV infection the expansion of Tfh cells positively correlates with the frequency of total GC B cells[12]. Here, we investigated the relationship between Tfhs and antigen-specific B cells. We have taken advantage of the use of

a biotinylated gp140 trimers linked to metal conjugated streptavidin in order to identify HIV-specific B cells from lymph nodes by mass cytometry[27]. As comparator, we have analyzed Flu-specific B cells within the same lymph nodes of healthy, viremic and ART treated individuals. After gating on $CD19^+$ cells to identify B cells and on IgG^+ cells to enrich in memory B cells, gp140 trimers and HA protein were used to identify HIV-specific and Flu-specific B cells, respectively (Fig. 5a). The frequency of LN gp140-specific B cells was significantly higher in viremics versus long-term ART treated individuals (2.7% vs 1.2%, $P = 0.0005$) (Fig. 5a, b) and in LNs as compared to peripheral blood (viremics: 2.7% in LNs vs 1% in blood, $P = 0.006$; ART Treated: 1.2% in LNs vs 0.5% in blood, $P = 0.002$) (Supplementary Fig.3a, b). Flu-specific B cells were detected at similar frequencies in a subset of the LNs from the three study groups (about 0.3% of IgG^+ B cells) but at significantly lower frequencies as compared to the percentage of gp140-specific B cells in viremic and ART treated LNs ($P = 0.002$ and $P = 0.001$, respectively) (Fig. 5a, b). Next, we determined the distribution of gp140- and Flu-specific B cells within memory B cell subsets using mass cytometry. In LNs, IgD and CD38 identify four populations of non-naive B cells: unswitched memory IgD^+CD38^-, switched memory IgD^-CD38^-, GC B cells IgD^-CD38^+ and plasma cells IgD^-CD38^{hi} [12]. After gating on non-naive B cells (naive IgD^+CD27^- (<1%) were excluded) the phenotypic analysis showed that gp140-specific B cells in LNs from viremic individuals were primarily contained within GC B cells (52.6%) and switched memory (42.3%), while gp140-specific B cells from ART treated individuals were mostly contained within the switched memory B cell population (78.3%) (Supplementary Fig. 4a and Fig. 5c). Flu-specific B cells were mostly contained within switched memory B cells in the three study groups (around 80%) (Fig. 5d). The frequency of gp140-specific B cells was positively correlated with the total percentage of GC B cells ($r = 0.82$, $P < 0.0001$) and negatively correlated with the frequency of

switched memory B cells (r = -0.79, P < 0.0001) (Supplementary Fig. 4b). Moreover, plasma viral load was positively correlated with the percentage of gp140-specific B cells within the total GC B cells (r = 0.61, P = 0.037) and negatively correlated with the percentage of gp140-specific B cells within switched memory B cells (r = -0.74, P = 0.007) (Supplementary Fig. 4b).

To further confirm the differences in the distribution of gp140- and Flu-specific B cells in the different manually gated memory B cell populations (Fig. 5c, d and Supplementary Fig. 4a) we performed a multivariate analysis with 31 markers (listed in Table S1b) using multi-dimensional scaling (MDS). MDS provides a graphical summary of the cell subsets similarities in the expression patterns of the 31 markers. The 3D MDS plots (Fig. 5e) were used to directly compare the phenotypic relationship between gp140- and Flu-specific B cells and the LN memory B cell subsets. In support of the distribution of gp140-specific B cells in the different manually gated memory B cell populations (Fig. 5c, d), the MDS plot clearly showed that gp140-specific B cells clustered closer to GC B cells in HIV viremic individuals while in ART treated subjects these cells clustered closer to switched memory B cells (Fig. 5e). Consistent with the data generated by manual gating, the MDS analysis showed that Flu-specific B cells clustered with switched memory B cells in the three study groups (Fig. 5e).

Taken together these results strongly suggest that the enrichment of gp140-specific B cells within the total GC B cell population is driven by HIV replication.

Next, we analyzed the expression of the same panel of B cell markers (Supplementary Table 1b) in gp140-specific B cells of both HIV^+ viremic and ART treated individuals. We observed that among all markers, 14 were significantly differentially expressed (FDR < 0.05) between the two groups. Increased levels of BCL-6 (26.8% vs 11.7%, P = 0.0002), Ki-67 (22.8% vs 5.5%, P = 0.0002), and CD38 (31.2% vs 7.5%, P = 0.0002) were found in gp140-specific B

cells from viremic individuals as compared to gp140-specific B cells from ART treated individuals (Fig. 6a). Similarly, gp140-specific B cells expressed higher levels of T-bet (19% vs 6.5%, P = 0.0005) and of the inhibitory molecule Fc-receptor-like 4 (FCRL4) (28.7% vs 9.4%, P = 0.003) (Fig. 6a). Interestingly, gp140-specific B cells from viremic individuals had an apoptotic phenotype as indicated by the decreased expression of BCL-2 (28.7% vs 75.1%, P = 0.0009) and increased CD95 expression (21.4% vs 7.2%, P = 0.0002) (Fig. 6a). Loss of CD21 expression in the blood of HIV infected individuals has been described as a marker of active HIV infection and disease progression[28]. LN gp140-specific B cells from viremic individuals showed a trend towards reduced expression of CD21 and decreased expression of CD27 (CD21: 75.8% vs 86.8%, P = 0.015; CD27: 81.3% vs 90.9%, P = 0.005) as compared to gp140-specific B cells of ART treated individuals (Fig. 6a). Loss of CXCR5 and CXCR3 expression (CXCR5: 50.8% vs 86.8%, P < 0.0001; CXCR3: 55.4 vs 78.2%; P = 0.005) and increased CD11c (13.7% vs 9.7%, P = 0.03) on gp140-specific B cells suggests that these changes in the expression of trafficking receptors may result in abnormal trafficking of gp140-specific B cells between the dark and the light zone and also in an accelerated exit of B cells from the GC.

We then compared the phenotypic profiles of gp140 versus Flu-specific B cells in viremic individuals presenting both responses. Expression of CD21, CD40, CXCR4 and CXCR5 was significantly downregulated in gp140-specific B cells as compared to Flu-specific B cells (CD21: 75.8% vs 87%, P = 0.005; CD40: 63.1% vs 78.1%, P = 0.01; CXCR4: 66.6% vs 71.8%, P = 0.03; CXCR5: 50.8% vs 85.3%, P < 0.0001) (Fig. 6b). Furthermore, the expression of markers of cell activation and maturation such as BCL-6, CD38 and Ki-67 was significantly increased in gp140-specific B cells (BCL-6: 26.8% vs 17%, P = 0.04; CD38: 31.2% vs 9%, P = 0.002; Ki-67: 22.8%

vs 10.8%, $P = 0.02$). The differences in the phenotypic profile between gp140- and Flu-specific B cells were less important in ART treated individuals (Supplementary Fig. 5).

These results further indicate that the ongoing HIV replication shapes the phenotypic profile of HIV-specific B cells.

In order to determine the impact of HIV replication on the phenotypic profile of Flu-specific B cells, the profile was compared between LNs of healthy versus viremic individuals (Fig. 6c). Interestingly, expression of the anti-apoptosis marker BCL-2 was downregulated in LN Flu-specific B cells of viremic as compared to healthy individuals (36.7% vs 70.8%, $P < 0.0001$) and pro-apotosis CD95 marker was upregulated in viremic individuals (24% vs 11.9%, $P = 0.015$). Expression of CXCR5 and ICOS-L was also downregulated in LNs of viremic individuals (CXCR5: 85.3% vs 93%, $P = 0.03$; ICOS-L: 21.6% vs 39%, $P = 0.03$) while markers of activation and maturation such as CD38 and Ki-67 were upregulated (CD38: 9% vs 4.2%, $P = 0.001$; Ki-67: : 10.8% vs 0.9%, $P < 0.0001$). Finally, T-bet expression was significantly upregulated in Flu-specific B cells of viremic individuals (24.2% vs 9.9%, $P = 0.02$). Therefore, the phenotypic profile observed in Flu-specific B cells of viremic as compared to healthy individuals is similar to that of gp140-specific B cells in viremic individuals.

These results indicate that ongoing HIV replication in LNs causes changes in the phenotypic profile of Ag-specific B cells that are not restricted exclusively to HIV-specific B cells.

Relationship between Tfh cell populations and gp140-specific B cells

Having defined phenotypically and functionally distinct Tfh cell populations and the phenotype of LN gp140-specific B cells associated with ongoing HIV replication, we next analyzed the relationship between the functional profile of Tfh cells and the phenotype and

function of gp140-specific B cells. As shown above, the $CXCR3^-$ Tfh cell population consists of single IL-21, dual IL-21/IL-4 and single IL-4 producing Tfh cells thus defining a Th2-like Tfh cell population while the $CXCR3^+$ Tfh cell population defines a Th1-like Tfh cell population containing dual IL-21/IFN-γ and single IFN-γ Tfh cells (Fig. 4c). The correlation heat map in Figure 7 shows that the frequency of $gp140^+$ B cells in viremics was positively correlated with the proportion of dual Tfh IL-21/IL-4 $CXCR3^-$ Th2-like Tfh cells (r = 0.82, P = 0.001). Similarly, the proportion of dual IL-21/IL-4 $CXCR3^-$ Tfhs was negatively correlated with the frequency of $CD95^+gp140^+$ B cells (r = -0.61, P = 0.04) and was positively correlated with the frequency of $Ki\text{-}67^+gp140^+$ B cells (r = 0.87, P = 0.0004) suggesting that IL-4 in combination with IL-21 drives the expansion of HIV specific B cells. However, the frequency of single IFN-γ $CXCR3^+$ Th1-like Tfh cells in viremics did not correlate with the percentage of gp140-specific B cells and was positively correlated with increased expression of T-bet (r = 0.71, P = 0.014), CXCR3 (r = 0.7, P = 0.015) and FCRL4 (r = 0.71, P = 0.013) on $gp140^+$ B cells. A similar correlation heat map from ART treated individuals is shown in Supplementary Figure 6.

Having identified distinct signatures in both the Tfh and B cell compartments in HIV-1 infected individuals, we sought to provide formal demonstration of the influence of the Th1-like Tfh and Th2-like Tfh cells on modulating the function and phenotype of B cells.

Firstly, we evaluated the effects of cytokine treatment with IL-21, IL-4 and IFN-γ on the induction of the expression of CXCR3 and T-bet in different memory B cell populations from tonsils of HIV negative pediatric donors and on the immunoglobulin production in T-B cell co-cultures in vitro. Tonsil cells were used due to the large number of tissue B cells necessary to perform these experiments. IFN-γ consistently induced expression of CXCR3 and T-bet in unswitched, switched and GC B cell populations (Fig. 8a). IL-21 and IL-4 showed no effects, with

the exception of IL-4 inducing T-bet expression in unswitched memory B cells. R848, a strong agonist of TLR7/TLR8, was used as positive control and showed an effect similar to IFN-γ.

Secondly, we determined the effects of IFN-γ treatment on immunoglobulin (Ig) production in Tfh-GC B cell co-culture. Interestingly, IFN-γ treatment caused significant reduction in the production of IgG1, IgG2, IgG3 and IgM and a trend towards reduction in IgG4 and IgA (Fig. 8b).

Thirdly, we investigated the impact of IL-4 and IFN-γ treatment on the maturation of HIV-specific B cell responses from 9 HIV viremic individuals after 4 days of stimulation in the presence of gp140 protein, IL-21 and a suboptimal doses of R848. As shown in Figure 8c, HIV-specific B cell responses from unstimulated LNMC, as measured by ELISPOT, were low but detectable and their frequencies significantly increased in all the conditions when cells were stimulated with gp140+R848 ($P < 0.01$). Interestingly, IL-4 stimulation led to a significant increase in the proportion of gp140 ASC cells when compared to gp140+R848 stimulated cells and R848+IFN-γ stimulation, 1.5 fold, ($P = 0.0039$) and 1.8 fold ($P = 0.01$), respectively. Of note, the frequencies of HIV-specific B cells after IFN-γ stimulation were not significantly different from those observed with gp140+R848 stimulation alone.

Fourthly, we determined whether there were differences in the quality of the antibodies produced by $CXCR3^+$ versus $CXCR3^-$ GC B cell populations. The rationale for these experiments is supported by our observation that Th1-like Tfhs drive the expansion of the $CXCR3^+T\text{-}bet^+$ GC B cells and substantially affect Igs production. Furthermore, consistent with previous studies[29,30] we observed an increased frequency of $T\text{-}bet^+CXCR3^+$ B cells in viremics as compared to HIV uninfected and ART treated individuals ($P < 0.0001$) (Supplementary Fig. 7a), and all $T\text{-}bet^+$ B cells were contained within the $CXCR3^+$ B cell population ($P < 0.0001$) (Supplementary Fig. 7b).

Therefore, we isolated CXCR3$^+$ and CXCR3$^-$ IgG$^+$ B cells from five viremic individuals and assessed levels of somatic hypermutation (SHM) by carrying out error-corrected sequence analysis of natively paired heavy and light chain genes for LN B cells. We found that the level of SHM differed by B cell CXCR3 status (Fig. 8d). On average, the CXCR3$^+$ phenotype was associated with significantly lower levels of SHM than the CXCR3$^-$ phenotype ($P = 0.0003$ for the heavy chain and $P < 0.0001$ for the light chain). Along the same line, the level of SHM was significantly lower in gp140-specific CXCR3$^+$ as compared to CXCR3$^-$ B cells ($P = 0.0025$ for the heavy chain and $P = 0.0088$ for the light chain) (Fig. 8e).

Taken together these results indicate that the skewed Th1-like Tfh versus Th2-like Tfh cells cytokine profiles associated with active HIV replication influence the phenotype, the maturation, the magnitude and the quality of HIV-specific B cell responses.

Discussion

In the present study we have used mass cytometry to dissect the heterogeneity of Tfh cells. Phenotypic and functional diversity has previously been shown within different Tfh cell populations from tonsils using mass cytometry[20]. Another study performed using lymph nodes of HIV viremic individuals showed that persistent antigen stimulation likely causes the selection of an oligoclonal HIV-specific Tfh cell population with a dominant IL-21 functional profile[15].

We have studied lymph nodes from three study groups including healthy HIV negative, HIV infected ART treated and untreated viremic individuals to dissect the differences in phenotypic and functional profiles and to determine whether the different profiles influence the development of B cell responses. The unsupervised approach used to analyse distributions of thirty markers of T cell activation, memory differentiation, chemokine receptors and HIV coreceptors has allowed us to identify a number of phenotypic markers defining differences between the three study groups. Four markers including CXCR3, CD38, HLA-DR and CD57 contributed most to the diversity of Tfhs between the three study groups. In particular, the frequency of Tfhs expressing CD38, HLA-DR and CD57 was significantly lower in the Tfh clusters differentiating HIV^- and ART treated from HIV viremic individuals. A recent study performed in blood and tonsils of healthy individuals has shown that the co-expression of CD57 within $PD\text{-}1^{hi}$ Tfhs defines Tfhs with reduced cytokine (IL-21 and IL-10) production and increased cytotoxic potential thus suggesting a role in the regulation of GC responses[31]. The significant decrease observed in the $CD57^+$ Tfhs in healthy and ART treated individuals may suggest that in the absence of antigen-specific activation of GCs the proposed regulatory role of $CD57^+PD\text{-}1^{hi}$ Tfhs in terminating the GC reaction is no longer needed. The frequencies of $CXCR3^+$ Tfhs were higher in the clusters of HIV viremic individuals. The Tfh cell population greatly expanded in viremic individuals was

characterized by the co-expression of CXCR3 and CD38 while a Tfh cell population lacking both markers was largely (about 60%) represented in HIV negative individuals. Interestingly, $CD38^{+}CXCR3^{+}$ Tfhs expressed lower levels of BCL-2 in viremic individuals, suggesting that they were more prone to apoptosis. Furthermore, the expression or the lack of CXCR3 distinguishes between Th1-like Tfhs expressing T-bet and Th-2-like Tfhs expressing GATA-3, respectively.

These results are consistent with the conclusions from other studies pointing out an increased activation of Tfhs and a higher proportion of $CXCR3^{+}$ Tfhs in viremic infection[12,15,29,32]. However, we demonstrate that CXCR3 is the phenotypic marker distinguishing between Th1-like and Th-2 like Tfhs both in healthy and HIV infected individuals.

Of note, we also demonstrate that the HIV co-receptor CCR5 is greatly expressed on $CXCR3^{+}$ Tfhs and defines the population of Tfhs with highest levels of HIV RNA transcription. The identification of the major HIV cell reservoir within Tfhs may have implications for monitoring the efficacy of virus suppression in lymphoid tissue following ART or other intervention strategies being developed in the arena of HIV cure.

An additional important observation of our study is the functional dichotomy between $CXCR3^{+}$ and $CXCR3^{-}$ Tfhs in the secretion of Th1 (IFN-γ) and Th2 (IL-4) cytokines. Previous studies performed in mice have shown the critical role of IL-21 in driving the expansion of GC B cells[33-40]. However, IL-21 is not sufficient for the optimal maturation of the GC response, which also requires Tfhs producing IL-4, which seem to further regulate the migration of GC B cells between the dark and light zones[11]. $CXCR3^{+}$ Tfhs contain single IL-2, dual IL-21/IFN-γ and single IFN-γ cells producing cytokines while $CXCR3^{-}$ Tfhs contain single IL-21, dual IL-21/IL-4 and single IL-4 cells. We have not observed a defect in IL-21 producing Tfhs, which is consistent with the significant expansion of GC B cells (both total and HIV-specific) associated with active HIV

replication. However, we have observed a selective defect, about a 4 fold reduction, in the total percentage of Tfhs producing IL-4 (dual IL-21/IL-4 + single IL-4) between healthy and viremic individuals, thus suggesting a potential defect in the maturation of the B cell response in viremic individuals. Of note, ART was not able to significantly recover the Th2-like Tfh population.

We then characterized in depth the phenotype and the maturation of HIV-specific B cells and used as comparator B cells specific to Flu isolated from lymph nodes of the same individuals. HIV-specific B cells were selectively enriched in lymph nodes as compared to blood while Flu-specific B cells show similar frequencies in the two compartments. HIV-specifc B cells were differently distributed as compared to Flu-specific, being the former mostly contained within GC B cells and the latter in the switch memory B cells. HIV-specific B cells showed significant higher percentage of cycling cells and an activated phenotype and also a pro-apoptotic profile as indicated by the reduced expression of BCL-2 and increased expression of CD95. These observations further support previous studies[12,14] indicating that B cell expansion during actively replicating HIV infection is driven by HIV. However, when lymph node Flu-specific B cells from viremic were compared to those of healthy individuals, they also showed significant increase in markers of activation and maturation, an apoptotic phenotypic profile and increase expression of T-bet. Therefore, despite the absence of Flu-specific stimulation these results suggest that the Th1 cytokine microenviroment associated with viremic HIV infection may be responsible for the changes in maturation, activation and phenotype of Flu-specific B cells from viremic as compared to healthy individuals.

We then determined the influence of the cytokine microenvironment and, in particular, of the Th1/Th2 cytokine imbalance on the phenotype and the optimal maturation of the B cell response in individuals with active replicating HIV infection.

Our results indicate that IFN-γ and not IL-4 induces the phenotype, i.e. expression of T-bet and CXCR3, observed in GC B cells of viremic individuals. The additional changes in the phenotype of GC B cells such as the increased in the CD11c and FCRL-4 expression and the decrease in CD21 and CD27 are in line with previous studies indicating that these phenotypic abnormalities are associated with chronic stimulation[30,41-44]. Furthermore, the positive correlation of the percentage of single IFN-γ Th1-like Tfhs with T-bet$^+$, CXCR3$^+$, CD95$^+$ and FCRL4$^+$ HIV-specific B cells further support the results obtained *in vitro* and also indicates that HIV-specific B cells induced by Th1-like Tfhs are potentially prone to apoptosis and limited proliferation. In contrast, dual IL-21/IL-4 Th2-like Tfhs cells were positively correlated with the percentage of total and dividing (Ki-67$^+$) HIV-specific B cells and negatively correlated with the expression of CD95.

Therefore, the dominant Th1-like Tfhs cytokine profile resulting from the reduction of Th2-like Tfhs is the driving force of the phenotypic abnormalities observed.

We provide several lines of evidence that the imbalance in the Th-l versus Th-2-like Tfh cytokine profile affects the maturation of B cell response. Firstly, IFN-γ suppresses Igs production in T-B cell co-cultures. Secondly, only IL-4 and not IFN-γ in combination with IL-21 increases the maturation of gp140-specifc B cells. Thirdly, HIV-specific B cells have significantly reduced expression of CXCR4 and CXCR5 indicating that the migration between the dark and light zone of the GC may be impaired in the presence of a defect of Th2-like Tfh cells with a negative influence on the affinity maturation of B cells. Fourthly, total and gp140-specific lymph node CXCR3$^+$ B cells, which contain almost all T-bet$^+$ cells, and which are induced by IFN-γ, show significantly lower levels of somatic hypermutation as compared to CXCR3$^-$ cells.

Our results therefore support the model that the defect in Th2-like/Tfh cells in favor of Th1-like Tfh cells is an important mechanism to explain the unique phenotypic profile of B cells

and the impaired maturation of B cell response in viremic individuals. These results also indicate that HIV immunization strategies aimed at the development of quantitatively and qualitatively effective antibody responses should target the development of optimal Th2-like Tfh cell responses.

Materials and Methods

Experimental Design

Lymph node biopsies were performed in 24 HIV-1 infected viremic individuals naive to antiretroviral therapy and 29 ART treated subjects (Table 1). With regard to HIV negative subjects, lymph node biopsies (inguinal lymph nodes) were performed in 18 subjects who underwent vascular (varicose vein stripping) and general (uncomplicated bilateral inguinal herniorrhaphy) surgery. Tonsils were obtained from young patients who underwent tonsillectomy. These studies were approved by the Institutional Review Board of the Centre Hospitalier Universitaire Vaudois, and all subjects gave written informed consent. For all the experiments, participant ID were randomized, and the samples were randomly numbered to perform the experiments. No outliers were excluded from the analyses.

Isolation of lymph node and tonsil mononuclear cells

Lymph node and tonsil mononuclear cells were isolated by mechanical disruption as previously described[45] and cells were cryopreserved in liquid nitrogen.

CyTOF marker labeling and detection

Cryopreserved lymph node mononuclear cells (LNMCs) were thawed and resuspended in complete RPMI medium (Gibco; Life Technologies; 10% heat-inactivated FBS [Institut de Biotechnologies Jacques Boy], 100 IU/ml penicillin, and 100 µg/ml streptomycin [BioConcept]). For the T cell panel 2^X10^6 cells/ml were stimulated or not with 100 ng/ml PMA (Sigma-Aldrich) and 1 µg/ml ionomycin (Sigma-Aldrich) in the presence of golgi plug (BD) for 5 hours at 37°C.

For the B cell panel cells were blocked using unlabeled anti-CD4 pure (clone SK3, BD Bioscience) antibody as previously described[46]. Cells were washed twice and then incubated for 30 minutes at 4°C with gp140 (Consensus B) biotinylated bound to a streptavidin PE. Two biotinylated flu probes bound to APC were used as previously shown[47]: one from H1 strain CA09 (for samples collected during or after the 2009-2010 season) and one from NC-99 (for samples collected prior to the 2009-2010 season).

Viability of cells in 500 µl of PBS was identified by incubation with 50 µM cisplatin (Sigma-Aldrich) for 5 min at RT and quenched with 500 µl fetal bovine serum. Next, cells were incubated for 30 min at 4°C with a 50 µl cocktail of cell surface metal conjugated antibodies (Fluidigm/DVS Science). Cells were washed and fixed for 10 min at RT with 2.4% PFA. Next, cells were permeabilized for 45 min at 4°C with Foxp3 Fixation/Permeabilization kit (eBioscience), washed and stained at 4°C for 30 min with a 50 µl cocktail of transcription factor and cytokine metal conjugated antibodies. Cells were washed and fixed for 10 min at RT with 2.4% PFA. Total cells were identified by DNA intercalation (1 µM Cell-ID Intercalator, Fluidigm/DVS Science) in 2% PFA at 4°C overnight. The list of metal isotopes antibodies used are listed in Table S1a, b. Labeled samples were assessed by the CyTOF1 instrument that was upgraded to CyTOF2 (Fluidigm) using a flow rate of 0.045 ml/min.

CyTOF data analysis

FCS files were normalized to the EQ Four Element Calibration Beads using the CyTOF software. For conventional cytometric analysis of B and Tfh cell populations, FCS files were imported into Cytobank Data Analysis Software or FlowJo v10.4.2 (Treestar, Inc., Ashland, CR) and SPICE v5.3 (developed by Mario Roederer, National Institute of Health)[48]. Gated Tfh cells were imported

into R software using the flowWorkspace framework[49]. Marker intensity values were arcsinh (hyperbolic inverse sine) with cofactor 5 transformed. Unsupervised clustering was conducted using FlowSOM[50] (BuiltSOM function in FlowSOM package) on pooled Tfhs from all samples (92'116 cells) (in combination with hierarchical consensus meta-clustering (metaClustering_consensus function in FlowSOM package). Dimensionality reduction was performed using the Barnes-Hut implementation of t-distributed stochastic neighbor embedding (Rtsne function in Rtsne package).

Principal component analysis was performed on single cell data and the absolute values of the marker loadings of the first two principal components were averaged and reported in Figure 1F.

Sorting of Tfh cell populations

Cryopreserved lymph node mononuclear cells were thawed and stained with the violet LIVE/DEAD stain kit and with anti-CD3 APC-H7 (BD), anti-CD4 Alexa700 (Biolegend), anti-CD45RA ECD (Beckman Coulter), anti-CXCR5 FITC (BD), and anti-PD-1 Pe-Cy7 (BD), anti-CXCR3 PE (Biolegend) and anti-CD38 V450 (BD) at 4°C for 20 min, and the $CD38^{-}CXCR3^{-}$, $CD38^{+}CXCR3^{-}$, $CD38^{-}CXCR3^{+}$ and $CD38^{+}CXCR3^{+}$ Tfh populations were sorted using FACSAria (BD). In all sorting experiments, the grade of purity of the sorted populations was > 95%.

Quantification of cell-associated RNA

Cell-associated HIV-1 RNA (unspliced HIV RNA LTR-gag region) from individual samples was extracted from Tfh cell populations sorted on the basis of CD38 and CXCR3 expression ($CD38^{-}CXCR3^{-}$, $CD38^{+}CXCR3^{-}$, $CD38^{-}CXCR3^{+}$, $CD38^{+}CXCR3^{+}$) and subjected to DNase treatment

(RNAqueous-4PCR Kit, Ambion)[51]. RNA standard curves were generated after isolation and quantification of viral RNA from supernatant of ACH2 culture as previously described[51]. One-step cDNA synthesis and pre-amplification were performed as previously described[52].

B cell/T cell co-culture assay

B cells from tonsil mononuclear cells were enriched using CD19-positive selection (STEMCELL Technologies), and dead cells were excluded using the violet LIVE/DEAD stain kit at 4°C for 15 min and stained at 4°C for 25 min with the anti-CD19 APC–Cy7, anti-IgD PE, anti-CD27 Pe-Cy7 and anti-CD38 ECD mAbs. GC B cells ($CD19^+IgD^-CD38^+$) were sorted from enriched B cell fraction (CD19 positive) using a FACSAria. The CD19-negative fraction was stained with the violet LIVE/DEAD stain kit at 4°C for 15 min and stained at 4°C for 25 min with the anti-CD3 APC-H7, anti-CD4 FITC, anti-CD45RA ECD, anti-CXCR5 APC, and anti–PD-1 PeCy7. Sorted Tfh cells populations (10^5 cells) were co-cultured with sorted autologous GC B cells (10^5 cells) in the presence of 250 ng/ml SEB (Sigma-Aldrich) and in presence or absence of 100 ng/ml of recombinant IFN-γ (R&D) in 96-well U-bottom plates as previously described[12]. As positive controls, GC B cells were cultured alone in presence of 5×10^4 pfu/ml of inactivated Staphylococcus aureus and 25 μg/ml CpG. Secretion of IgM, IgG1, IgG2, IgG3, IgG4 and IgA was assessed at day 5 by Luminex (Affimetrix).

Elispot assay

LNMCs were stimulated or not for 4 days with 0.1 ug/ml of gp140 and 1 ug/ml of R848 (InvivoGen), 10 ng/ml of IL-2 (Miltenyi Biotec) and 100 ng/ml of IL-21 (Miltenyi Biotec). In the stimulated conditions cells were treated or not with 100 ng/ml of IL-4 (Miltenyi Biotec) or IFN-

γ (Miltenyi Biotec). ELISPOT plates (BD) were coated with 15 ug/ml of anti-Ig antibodies (Mabtech) at 4°C overnight. Next, plates were washed and cells were added for 24 hours at 37°C followed by addition of biotinylated antibody against IgG or biotinylated proteins gp140 or the control protein keyhole limpet hemocyanin (KLH), and finally addition of a streptavidin-HRP (Mabtech). Frequencies of gp140-specific antibody secreting cells (ASC) were calculated from triplicate wells plated with 100.000 LNMCs per well. Specificity was verified with PBMCs of HIV-uninfected individuals.

Paired Chain Antibody Sequencing

Single $CD19^{+}CD20^{+}CD3^{-}CD14^{-}IgA^{-}IgM^{-}IgD^{-}$ IgG^{+} cells and $gp140^{+}$ B cell were sorted on the basis of CXCR3 expression into wells of 384-well plates by FACS. Generation of barcoded cDNA, PCR amplification, and sequencing of IgG genes were performed as described in Tan et al. 2014[53], with the following modifications: biotinylated Oligo(dT) and RT maxima H- (Fisher Scientific Company) were used for reverse transcription, cDNA was extracted using Streptavidin C1 beads (Life Technologies), and DNA concentrations were determined using qPCR (KAPA SYBR® FAST qPCR Kit for Titanium, Kapabiosystems). V(D)J assignment and mutation identification was performed using a variant of SoDA[54].

Statistical analysis

GraphPad PRISM and R softwares were used to perform statistical analyses.

Linear regressions were performed to compare frequencies (log10 transformed) of Tfh clusters or antigen specific B cells among the different study groups (Figure 1e, f and and 6) and the resulting

P values were adjusted for multiple testing using the Benjamini-Hochberg FDR method (with significance cutoff set at 0.05).

Statistical analyses comparing cell surface markers, transcription factors and cytokine production (log10 transformed) (Fig. 2c and 4b) in Tfh subsets defined by CD38 and CXCR3 were assessed by linear mixed-effect models accounting for differences between patient groups (healthy, ART-treated and viremic individuals) with patient-level random intercepts. *P* values were adjusted using the Benjamini-Hochberg FDR method (with significance cutoff set at 0.05).

Two-tailed Mann-Whitney unpaired test was used to compare frequencies of B cell subsets between the three different groups (HIV-, ART treated and viremic individuals).

Correlative analyses on Figure 7 were performed on log10 transformed frequencies using Pearson's test. The Wilcoxon signed-rank paired test was used to detect differences between variables from the same sample.

References and Notes:

1 Kim, C. H. *et al.* Subspecialization of CXCR5+ T cells: B helper activity is focused in a germinal center-localized subset of CXCR5+ T cells. *The Journal of experimental medicine* 193, 1373-1381 (2001).

2 Breitfeld, D. *et al.* Follicular B helper T cells express CXC chemokine receptor 5, localize to B cell follicles, and support immunoglobulin production. *The Journal of experimental medicine* 192, 1545-1552 (2000).

3 Schaerli, P. *et al.* CXC chemokine receptor 5 expression defines follicular homing T cells with B cell helper function. *The Journal of experimental medicine* 192, 1553-1562 (2000).

4 Fazilleau, N., Mark, L., McHeyzer-Williams, L. J. & McHeyzer-Williams, M. G. Follicular helper T cells: lineage and location. *Immunity* 30, 324-335, doi:10.1016/j.immuni.2009.03.003 S1074-7613(09)00114-9 [pii] (2009).

5 Fazilleau, N., McHeyzer-Williams, L. J., Rosen, H. & McHeyzer-Williams, M. G. The function of follicular helper T cells is regulated by the strength of T cell antigen receptor binding. *Nature immunology* 10, 375-384, doi:10.1038/ni.1704 ni.1704 [pii] (2009).

6 Crotty, S. T Follicular Helper Cell Biology: A Decade of Discovery and Diseases. *Immunity* 50, 1132-1148, doi:10.1016/j.immuni.2019.04.011 (2019).

7 Haynes, N. M. *et al.* Role of CXCR5 and CCR7 in follicular Th cell positioning and appearance of a programmed cell death gene-1high germinal center-associated subpopulation. *Journal of immunology (Baltimore, Md. : 1950)* 179, 5099-5108, doi:10.4049/jimmunol.179.8.5099 (2007).

8 Nurieva, R. I. *et al.* Bcl6 mediates the development of T follicular helper cells. *Science* 325, 1001-1005, doi:10.1126/science.1176676 (2009).

9 Yu, D. *et al.* The transcriptional repressor Bcl-6 directs T follicular helper cell lineage commitment. *Immunity* 31, 457-468, doi:10.1016/j.immuni.2009.07.002 (2009).

10 Reinhardt, R. L., Liang, H. E. & Locksley, R. M. Cytokine-secreting follicular T cells shape the antibody repertoire. *Nature immunology* 10, 385-393, doi:10.1038/ni.1715 (2009).

11 Weinstein, J. S. *et al.* TFH cells progressively differentiate to regulate the germinal center response. *Nature immunology* 17, 1197-1205, doi:10.1038/ni.3554 (2016).

12 Perreau, M. *et al.* Follicular helper T cells serve as the major CD4 T cell compartment for HIV-1 infection, replication, and production. *The Journal of experimental medicine* 210, 143-156, doi:10.1084/jem.20121932
jem.20121932 [pii] (2013).

13 Cubas, R. A. *et al.* Inadequate T follicular cell help impairs B cell immunity during HIV infection. *Nat Med* 19, 494-499, doi:10.1038/nm.3109 (2013).

14 Noto, A. & Pantaleo, G. B-cell abnormalities and impact on antibody response in HIV infection. *Current opinion in HIV and AIDS* 12, 203-208, doi:10.1097/coh.0000000000000359 (2017).

15 Wendel, B. S. *et al.* The receptor repertoire and functional profile of follicular T cells in HIV-infected lymph nodes. *Science immunology* 3, doi:10.1126/sciimmunol.aan8884 (2018).

16 Velu, V. *et al.* Induction of Th1-Biased T Follicular Helper (Tfh) Cells in Lymphoid Tissues during Chronic Simian Immunodeficiency Virus Infection Defines Functionally Distinct Germinal Center Tfh Cells. *Journal of immunology (Baltimore, Md. : 1950)* 197, 1832-1842, doi:10.4049/jimmunol.1600143 (2016).

17 Cubas, R. *et al.* Reversible Reprogramming of Circulating Memory T Follicular Helper Cell Function during Chronic HIV Infection. *Journal of immunology (Baltimore, Md. : 1950)* 195, 5625-5636, doi:10.4049/jimmunol.1501524 (2015).

18 Yamamoto, T. *et al.* Quality and quantity of TFH cells are critical for broad antibody development in SHIVAD8 infection. *Science translational medicine* 7, 298ra120, doi:10.1126/scitranslmed.aab3964 (2015).

19 Havenar-Daughton, C. *et al.* Direct Probing of Germinal Center Responses Reveals Immunological Features and Bottlenecks for Neutralizing Antibody Responses to HIV Env Trimer. *Cell reports* 17, 2195-2209, doi:10.1016/j.celrep.2016.10.085 (2016).

20 Wong, M. T. *et al.* Mapping the Diversity of Follicular Helper T Cells in Human Blood and Tonsils Using High-Dimensional Mass Cytometry Analysis. *Cell reports* 11, 1822-1833, doi:10.1016/j.celrep.2015.05.022 (2015).

21 Banga, R. *et al.* PD-1(+) and follicular helper T cells are responsible for persistent HIV-1 transcription in treated aviremic individuals. *Nat Med* 22, 754-761, doi:10.1038/nm.4113 (2016).

22 Boritz, E. A. *et al.* Multiple Origins of Virus Persistence during Natural Control of HIV Infection. *Cell* 166, 1004-1015, doi:10.1016/j.cell.2016.06.039 (2016).

23 Fukazawa, Y. *et al.* B cell follicle sanctuary permits persistent productive simian immunodeficiency virus infection in elite controllers. *Nat Med* 21, 132-139, doi:10.1038/nm.3781 (2015).

24 Banga, R. *et al.* Blood CXCR3(+) CD4 T Cells Are Enriched in Inducible Replication Competent HIV in Aviremic Antiretroviral Therapy-Treated Individuals. *Frontiers in immunology* 9, 144, doi:10.3389/fimmu.2018.00144 (2018).

25 Chtanova, T. *et al.* T follicular helper cells express a distinctive transcriptional profile, reflecting their role as non-Th1/Th2 effector cells that provide help for B cells. *Journal of immunology (Baltimore, Md. : 1950)* 173, 68-78 (2004).

26 Rasheed, A. U., Rahn, H. P., Sallusto, F., Lipp, M. & Muller, G. Follicular B helper T cell activity is confined to CXCR5(hi)ICOS(hi) CD4 T cells and is independent of CD57 expression. *European journal of immunology* 36, 1892-1903, doi:10.1002/eji.200636136 (2006).

27 Kardava, L. *et al.* Abnormal B cell memory subsets dominate HIV-specific responses in infected individuals. *J Clin Invest* 124, 3252-3262, doi:10.1172/jci74351 (2014).

28 Moir, S. & Fauci, A. S. B cells in HIV infection and disease. *Nat Rev Immunol* 9, 235-245, doi:10.1038/nri2524 (2009).

29 Del Alcazar, D. *et al.* Mapping the Lineage Relationship between CXCR5(+) and CXCR5(-) CD4(+) T Cells in HIV-Infected Human Lymph Nodes. *Cell reports* 28, 3047-3060.e3047, doi:10.1016/j.celrep.2019.08.037 (2019).

30 Austin, J. W. *et al.* Overexpression of T-bet in HIV infection is associated with accumulation of B cells outside germinal centers and poor affinity maturation. *Science translational medicine* 11, eaax0904, doi:10.1126/scitranslmed.aax0904 (2019).

31 Alshekaili, J. *et al.* STAT3 regulates cytotoxicity of human CD57+ CD4+ T cells in blood and lymphoid follicles. *Scientific reports* 8, 3529, doi:10.1038/s41598-018-21389-8 (2018).

32 Noto, A. *et al.* CD32(+) and PD-1(+) Lymph Node CD4 T Cells Support Persistent HIV-1 Transcription in Treated Aviremic Individuals. *Journal of virology* 92, doi:10.1128/jvi.00901-18 (2018).

33 Linterman, M. A. *et al.* IL-21 acts directly on B cells to regulate Bcl-6 expression and germinal center responses. *The Journal of experimental medicine* 207, 353-363, doi:10.1084/jem.20091738 (2010).

34 Kasaian, M. T. *et al.* IL-21 limits NK cell responses and promotes antigen-specific T cell activation: a mediator of the transition from innate to adaptive immunity. *Immunity* 16, 559-569, doi:10.1016/s1074-7613(02)00295-9 (2002).

35 Ozaki, K. *et al.* A Critical Role for IL-21 in Regulating Immunoglobulin Production. *Science* 298, 1630-1634, doi:10.1126/science.1077002 (2002).

36 Pène, J. *et al.* Cutting Edge: IL-21 Is a Switch Factor for the Production of IgG₁ and IgG₃ by Human B Cells. *The Journal of Immunology* 172, 5154-5157, doi:10.4049/jimmunol.172.9.5154 (2004).

37 Ettinger, R. *et al.* IL-21 Induces Differentiation of Human Naive and Memory B Cells into Antibody-Secreting Plasma Cells. *The Journal of Immunology* 175, 7867-7879, doi:10.4049/jimmunol.175.12.7867 (2005).

38 Kuchen, S. *et al.* Essential Role of IL-21 in B Cell Activation, Expansion, and Plasma Cell Generation during CD4⁺ T Cell-B Cell Collaboration. *The Journal of Immunology* 179, 5886-5896, doi:10.4049/jimmunol.179.9.5886 (2007).

39 Nurieva, R. I. *et al.* Generation of T follicular helper cells is mediated by interleukin-21 but independent of T helper 1, 2, or 17 cell lineages. *Immunity* 29, 138-149, doi:10.1016/j.immuni.2008.05.009 (2008).

40 Vogelzang, A. *et al.* A fundamental role for interleukin-21 in the generation of T follicular helper cells. *Immunity* 29, 127-137, doi:10.1016/j.immuni.2008.06.001 (2008).

41 Moir, S. *et al.* Evidence for HIV-associated B cell exhaustion in a dysfunctional memory B cell compartment in HIV-infected viremic individuals. *The Journal of experimental medicine* 205, 1797-1805, doi:10.1084/jem.20072683 (2008).

42 Kardava, L. *et al.* Abnormal B cell memory subsets dominate HIV-specific responses in infected individuals. *J Clin Invest* 124, 3252-3262, doi:10.1172/JCI74351 (2014).

43 Karnell, J. L. *et al.* Role of CD11c(+) T-bet(+) B cells in human health and disease. *Cell Immunol* 321, 40-45, doi:10.1016/j.cellimm.2017.05.008 (2017).

44 Naradikian, M. S. *et al.* Cutting Edge: IL-4, IL-21, and IFN-γ Interact To Govern T-bet and CD11c Expression in TLR-Activated B Cells. *Journal of immunology (Baltimore, Md. : 1950)* 197, 1023-1028, doi:10.4049/jimmunol.1600522 (2016).

45 Pantaleo, G. *et al.* Lymphoid organs function as major reservoirs for human immunodeficiency virus. *Proceedings of the National Academy of Sciences of the United States of America* 88, 9838-9842 (1991).

46 Doria-Rose, N. A. *et al.* Frequency and phenotype of human immunodeficiency virus envelope-specific B cells from patients with broadly cross-neutralizing antibodies. *Journal of virology* 83, 188-199, doi:10.1128/jvi.01583-08 (2009).

47 Whittle, J. R. *et al.* Flow cytometry reveals that H5N1 vaccination elicits cross-reactive stem-directed antibodies from multiple Ig heavy-chain lineages. *Journal of virology* 88, 4047-4057, doi:10.1128/jvi.03422-13 (2014).

48 Roederer, M., Nozzi, J. L. & Nason, M. C. SPICE: exploration and analysis of post-cytometric complex multivariate datasets. *Cytometry. Part A : the journal of the International Society for Analytical Cytology* 79, 167-174, doi:10.1002/cyto.a.21015 (2011).

49 Finak, G. *et al.* OpenCyto: an open source infrastructure for scalable, robust, reproducible, and automated, end-to-end flow cytometry data analysis. *PLoS Comput Biol* 10, e1003806, doi:10.1371/journal.pcbi.1003806 (2014).

50 Van Gassen, S. *et al.* FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry. Part A : the journal of the International Society for Analytical Cytology* 87, 636-645, doi:10.1002/cyto.a.22625 (2015).

51 Vandergeeten, C. *et al.* Cross-clade ultrasensitive PCR-based assays to measure HIV persistence in large-cohort studies. *Journal of virology* 88, 12385-12396, doi:10.1128/jvi.00609-14 (2014).

52 Procopio, F. A. *et al.* A Novel Assay to Measure the Magnitude of the Inducible Viral Reservoir in HIV-infected Individuals. *EBioMedicine* 2, 874-883, doi:10.1016/j.ebiom.2015.06.019 (2015).

53 Tan, Y. C. *et al.* High-throughput sequencing of natively paired antibody chains provides evidence for original antigenic sin shaping the antibody response to influenza vaccination. *Clinical immunology (Orlando, Fla.)* 151, 55-65, doi:10.1016/j.clim.2013.12.008 (2014).

54 Volpe, J. M., Cowell, L. G. & Kepler, T. B. SoDA: implementation of a 3D alignment algorithm for inference of antigen receptor recombinations. *Bioinformatics (Oxford, England)* 22, 438-444, doi:10.1093/bioinformatics/btk004 (2006).

Acknowledgments

We are grateful to Alex Farina, Thibaut Decaillon, Manon Geiser and Michael Moulin for technical assistance. This work was supported by grants of Bill and Melinda Gates Foundation and the Swiss National Fund. The funders had no role in the study design, data collection and interpretation, or the decision to submit the work for publication.

Author contributions

A.N. designed and performed the experiments, analyzed the data, and wrote the manuscript. M.S. performed data analyses. L.E.-L. and V.J. performed mass cytometry stainings. F.A.P. performed the HIV RNA quantification. G.C. and Y.L. performed the Paired Chain Antibody Sequencing. J-M.C. performed lymph node biopsy. M.C. and A.R. recruited participants. L.S., A.M. and R.K. provided gp140 and HA probes and helped with experimental design. R.G. and C.F. provided help in mass cytometry data analysis. G.P. desined the overall study, provided conceptual advice and wrote the manuscript.

Competing interests

All the authors declare no competing financial interests. **Data and materials availability**: All relevant data are within the paper.

Figure legend:

Fig. 1. *High dimensional analysis of Tfh cells in HIV infected and uninfected individuals.* Mass cytometry staining was performed on LN mononuclear cells isolated from 8 HIV uninfected, 12 HIV infected ART treated, and 9 viremic individuals. (**a**) Representative Tfh mass cytometry gating strategy on the basis of CXCR5 and PD-1 expression in $CD45RA^{-}CD4^{+}$ cells from representative viremic HIV infected individual and (**b**) BCL-6 expression on Tfh and non-Tfh cells. (**c**) t-SNE was performed after pooling the three study groups and gating on Tfh cells. The numbered and colored clusters Tfh clusters were obtained using FlowSOM. (**d**) Heat map showing median marker expression (arcsinh-transformed) of cell surface markers of the indicated clusters identified in (c). Median marker expression values are color-coded from blue (low) to yellow (high). (**e**) Frequency of Tfh clusters that are significantly different between the three groups (HIV uninfected (green), ART treated (red) and viremic (blue) individuals). (**f**) Pie charts representing frequencies of Tfh clusters in HIV uninfected, ART treated and viremics. Arcs show frequencies of clusters that are significantly different between HIV uninfected, ART treated and viremics. (**g**) Bar plot showing the relative contribution of markers to Tfh clusters heterogeneity. (Y-axis: average marker loadings in the first two principal components of a PCA). *P* values were obtained by linear regressions and corrected using FDR method with a cutoff of 0.05. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

Fig. 2. *$CD38^{+}CXCR3^{+}$ Tfh cells are increased in viremics and represent a distinct Tfh population.* Mass cytometry staining was performed on LN mononuclear cells isolated from 8 HIV uninfected, 20 HIV infected ART treated and 18 untreated HIV-infected viremic subjects. Cells were stained with antibodies against PD-1, CXCR5, CD38, CXCR3. (**a**)

Representative profile by mass cytometry of Tfh cells gated on the basis of CD38 and CXCR3 from one representative HIV uninfected, one ART treated and one viremic individual. (**b**) All the possible combinations of CD38 and CXCR3 expression are shown on the x axis, whereas the frequencies of the Tfh-cell populations are shown on the y axis. Pie charts represent all the possible combinations of the two markers. Arcs show the total proportion of the expression of the specified marker. Statistical analyses of the global CD38 and CXCR3 expression (pie charts) were performed by partial permutation tests using the SPICE software. (**c**) Heat map of scaled mean marker expression (percentage of positive cells) in Tfh cells defined on the basis of CD38 and CXCR3 expression of 8 HIV uninfected, 12 ART treated and 9 viremic individuals. The bottom panel shows significant differences between subsets of cells for all possible comparisons. Differences between subsets were calculated on all cohort samples using linear mixed-effect models. In (c) *P* values were obtained by linear regressions and corrected using FDR method with a cutoff of 0.05. Stars indicate statistical significance * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

Fig. 3. *Levels of cell-associated HIV RNA in Tfh populations*. Levels of cell-associated unspliced HIV RNA (copies/million cells) in sorted $CD38^-CXCR3^-$, $CD38^+CXCR3^-$, $CD38^+CXCR3^+$ and $CD38^-CXCR3^+$ Tfh cells isolated from 6 viremic HIV infected individuals. *P* values were obtained by Wilcoxon signed-rank test. *$P < 0.05$, **$P < 0.01$. Error bars denote mean ± S.E.M.

Fig. 4. *Functional analyses of Tfhs on the basis of $CD38^+$ and $CXCR3^+$ expression*. LNMCs were isolated from HIV uninfected (N = 12), HIV-1 infected ART treated individuals (N = 12) and viremics (N = 14) and stimulated with PMA-ionomycin for 5 hours and stained with antibodies against PD-1, CXCR5, CD38, CXCR3, IFN-γ, IL-21, IL-4, IL-2 and TNF-α.

(**a**) t-SNE plot of polled Tfh cells from the three study groups (51'171 cells). (**b**) Heat map of scaled mean marker expression (percentage of producing cytokines) in Tfhs gated on the basis of CD38 and CXCR3. The bottom panel shows significant differences between subsets of cells for all possible comparisons. (**c**) Simultaneous analysis of the functional profile of $CXCR3^-$ and $CXCR3^+$ Tfh cells on the basis of IL-21, IL-4 and IFN-γ production. (**d**) Th1/Th2 ratio in Tfh cells from HIV uninfected, HIV infected viremic and ART treated individuals. Ratio was calculated by dividing the frequencies of IFN-γ producing Tfh cells and the frequencies of IL-4 producing Tfh cells in HIV- and HIV viremics. In (b) differences between subsets were calculated on all cohort samples using linear mixed-effect models and *P* values were corrected using FDR method with a cutoff of 0.05. In (c and d) *P* values were obtained by a Mann-Whitney test to compare the three study groups and a Wilcoxon signed-rank test to compare frequencies between $CXCR3^-$ and $CXCR3^+$ populations. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Error bars denote mean ± S.E.M.

Fig. 5. *Comparison between lymph node gp140 and influenza specific B cells.* LNMCs from HIV infected ART treated (N = 11) and HIV infected viremic individuals (N = 12) were stained with a panel of 37 markers (Table S1b). (**a**) Representative mass cytometry profiles of LN $CD19^+$ IgG^+ B cell populations binding to gp140 or Flu (H1-CA09) probes in representative HIV^-, ART treated and viremic subjects. (**b**) Cumulative data on the frequencies of Flu and gp140-specific B cells in LNs of HIV^- (green), ART treated (red) and viremic (blue) individuals. Cumulative data on the distribution of (**c**) gp140-specific B cells and (**d**) Flu-specific B cells within the unswitched memory (IgD^+CD38^-), switched memory (IgD^-CD38^-), GC (IgD^-CD38^+) and plasma cells (IgD^-CD38^{hi}) B cell populations. (**e**) Multi-dimentional scaling (MDS) of Flu^+ B cells, $gp140^+$ B cells and memory B cell

subsets from HIV$^-$, ART-treated and viremic HIV-infected subjects. In (b-c-d) *P* values were obtained by Mann-Whitney test to compare frequencies between the three study groups and by Wilcoxon signed-rank test to compare frequencies within the same study groups. Error bars denote mean ± S.E.M. Stars indicate statistical significance * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

Fig. 6. *Phenotype of lymph node gp140 and Flu-specific B cells.* LNMCs from HIV infected ART treated (N = 11) and HIV infected viremic individuals (N = 12) were stained with a panel of 37 markers (Table S1b). Heat map of scaled mean marker expression (% of positive cell gated) in (**a**) gp140-specific B cells from ART treated vs viremic individuals, (**b**) Flu-specific vs gp140-specific B cells from the same viremic individuals and (**c**) Flu-specific B cells from HIV uninfected vs Flu-specific B cells from viremics. All markers shown are significantly different between groups (FDR < 0.05). Differences were calculated using linear regressions.

Fig. 7. *Correlations between cytokines producing Tfhs and frequency and phenotype of gp140 specific B cells*. Correlative heat maps between the frequency of cytokines produced by Tfh cells from HIV infected viremic individuals (Fig. 3c) and the proportion of gp140 specific B cells and their phenotype (Fig. 4b and Fig. 5a) (N = 11). Correlative analyses were performed on log10 transformed frequencies using Pearson's test.

Fig. 8. *Effect of in vitro cytokine stimulation of B cells on antibody production and SHM*. Tonsil (TN) mononuclear cells were cultured for 3 days in the presence or absence of IL-21 (100 ng/ml), IL-4 (100 ng/ml), IFN-γ (100 ng/ml) and R848 (1 μg/ml). Mass cytometry staining was performed using anti-CD19, anti-CD27, anti-CD38, anti-IgD, anti-CXCR3, and anti-T-bet antibodies (N = 3). (**a**) Percentage of T-bet$^+$ and CXCR3$^+$ B cells after 3 days of

culture. TN Tfh cells were cultured with autologous GC B cells in the presence of SEB and in the presence or absence of recombinant IFN-γ (100 ng/ml) (N = 6). (**b**) Immunoglobulin production was assessed at day 5 by Luminex. Lymph node (LN) mononuclear cells from 9 HIV^+ viremic individuals were stimulated or not for 4 days presence of gp140 (0.1 ug/ml) and R848 and in prensence or absence of IL-4 (100 ng/ml) and IFN-γ (100 ng/ml). (**c**) Frequencies of gp140-specific antibody secreting cells (ASC) were measured by ELISPOT. The left panel shows representative counting of spot-forming cells (SFC) and the right panel shows the frequency of gp140 ASC calculated from triplicate wells plated with 100.000 LNMCs per well. Total somatic mutations per antibody were determined by paired chain sequencing of $CXCR3^+$ and $CXCR3^-$ IgG B cells (d) or $CXCR3^+$ and $CXCR3^-$ gp140-specific B cells (e) from five donors. Error bars correspond to mean ± SEM and statistical significance was evaluated using Wilcoxon signed-rank test and in (d and e) by Mann-Whitney test. P values * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$

863 ***Table 1.*** *Study cohort: clinical data.*

Patient ID	Age	Sex	Duration of HIV Infection (years)	Viral Load (copies/ml)	CD4 count (cells/ul)	ART status	Assays
T008	53	M	7.78	<20	NA	ART Treated	2
T010	54	F	15.76	<20	703	ART Treated	2
T011	57	M	20.1	<20	854	ART Treated	3-4
T006	53	M	9	<20	648	ART Treated	3-4
T001	49	M	25.23	<20	549	ART Treated	3-4
T042	58	M	11.1	<20	635	ART Treated	1-2
T058	46	F	10.86	<20	666	ART Treated	2-3-4
T060	42	M	21	<20	786	ART Treated	1-2
T061	43	M	23	<20	698	ART Treated	2-3-4
T062	47	M	5.12	<20	487	ART Treated	3-4
T067	41	M	14.01	<20	609	ART Treated	3
T068	45	M	3.33	<20	928	ART Treated	1-2
T069	47	M	5.12	<20	487	ART Treated	1-2
T070	55	M	7.06	<20	615	ART Treated	3-4
T072	47	M	18	<20	614	ART Treated	1-2
T078	38	M	6.27	<20	728	ART Treated	2
T081	51	M	22.45	<20	1236	ART Treated	1-2
T082	52	M	23.49	<20	1093	ART Treated	2
T083	54	F	23.27	<20	941	ART Treated	2
T093	41	F	3	<20	336	ART Treated	3-4
T094	47	M	8	<20	451	ART Treated	1-2
T096	54	F	18	<20	1253	ART Treated	1-2
T100	36	F	13.06	<20	643	ART Treated	3-4
T103	41	M	1.62	<20	217	ART Treated	1-2
T104	41	F	10	<20	455	ART Treated	3-4
T108	50	M	11	<20	490	ART Treated	2-3-4
T123	61	M	9.25	<20	1270	ART Treated	1-2
T136	44	M	7.5	<20	732	ART Treated	1-2
T137	39	M	3.59	<20	700	ART Treated	1-2
V005	38	M	8	5186	651	Treatment Naive	3-4-5-6
V102	36	M	NA	83000	602	Treatment Naive	5-7
V037	37	M	0.3	25000	704	Treatment Naive	3-4-7
V150	53	M	10	6900	578	Treatment Naive	1-2
V017	34	M	0.2	32000	602	Treatment Naive	2-3-4-6
V018	48	M	2	260000	475	Treatment Naive	2-3-4-5-6
V019	NA	F	NA	73400	634	Treatment Naive	2-3-4
V035	34	M	6.1	3000	517	Treatment Naive	2-3-4-6
V031	32	F	0.2	90000	819	Treatment Naive	2-3-4-5
V034	24	M	0.3	5000	549	Treatment Naive	2-3-4-5
V143	31	M	0.1	10000000	174	Treatment Naive	2-3-5
V038	38	M	0.1	1500000	550	Treatment Naive	2-3-4
V049	45	M	0.1	20000	704	Treatment Naive	2-3-4-5-6
V056	35	F	13	830	670	Treatment Naive	1-2-3-4
V106	37	M	0.1	160000	501	Treatment Naive	1-2
V117	51	M	5.21	54000	504	Treatment Naive	1-2-4
V119	23	M	1.3	13000	498	Treatment Naive	1-2-3
V118	29	M	0.06	14000	538	Treatment Naive	1-2-3
V124	46	F	0.06	510000	468	Treatment Naive	1-2-7
V125	35	M	0.16	17000	511	Treatment Naive	1-2
V140	23	M	0.08	360000	427	Treatment Naive	1-2
V149	39	M	0.1	18000	280	Treatment Naive	5-7
V015	25	M	0.14	17000	416	Treatment Naive	5-7
V148	38	M	0.06	20000	717	Treatment Naive	7

864

865 Assay 1: FlowSom clustering (Fig.1); Assay 2:Tfh phenotype (Fig.2b); Assay 3: PMA/ionomicyn stimulation
866 (Fig.4); Assay 4: Ag-specific B cell phenotype (Fig.5); Assay 5:B cell Elispot (Fig. 8c); Assay 6: SHM (Fig. 8d);
867 Assay 7: cell associated HIV-RNA (Fig. 3). NA=not applicable
868

Supplementary materials

Supplementary Fig. 1. *High-dimensional analysis of Tfh cells by mass cytometry.* **a**) Individual t-SNE plots of each healthy (N = 8), HIV^+ ART treated (N = 12) and HIV^+ viremic individual (N = 9). **b**) Signal intensity of individual markers on t-SNE plots. Tfh cells were pooled (N = 92'162 cells) and colored according to the scaled expression level of indicated markers. **c**) Tfh clusters not significantly different between the three study groups. Linear regressions were performed to compare frequencies of Tfh clusters across the three groups (HIV^- (green), HIV^+ ART treated (red) and HIV^+ viremic (blue) individuals.

Supplementary Fig. 2. *Cytokine profile of Tfh cells from the three study groups.* LNMCs were isolated from HIV^- (N = 12), HIV^+ ART treated individuals (N = 12) and viremics (N = 14) and stimulated with PMA-ionomycin for 5 hours. **a**) Simultaneous analysis of the functional profile of Tfh cells on the basis of IL-21, IL-4, IFN-γ, IL-2 or TNF-α production. All the possible combinations of the various functions are shown on the x axis, whereas the percentages of the distinct cytokine-producing Tfhs are shown on the y axis. The pie charts summarize the data, and each slice corresponds to the proportion of Tfh cells positive for a certain combination of functions. **b)** Cytokine's production by Tfh cells gated on the basis of CD38 and CXCR3 expression. Tfh cells were gated on $CD38^-CXCR3^-$, $CD38^+CXCR3^-$, $CD38^-CXCR3^+$ and $CD38^+CXCR3^+$ and the percentage of IFN-γ, IL-21, IL-4, IL-2 and TNF-α positive cells was analyzed by mass cytometry. **c)** Frequencies of IFN-γ, IL-21 and IL-4 producing cells in the total Tfh cells the three study groups. Statistical significance (*P* values) in (b) were calculated using one-way ANOVA followed by test for multiple comparison and in (c) using Mann-Whitney. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Error bars denote mean ± S.E.M.

Supplementary Fig. 3. *Frequency of gp140 specific B cells from HIV infected ART treated and viremics in blood versus Lymph nodes.* **a**) Representative mass cytometry profiles of blood and LN $CD19^+$ IgG^+ B cell populations binding to gp140 probes in representative ART treated and viremic subjects. **b**) Cumulative data on the frequencies of gp140-specific B cells in blood and LN mononuclear cells of ART treated (red) and viremic (blue) individuals. Statistical significance (*P* values) were calculated using Mann-Whitney test to compare the two groups and a Wilcoxon signed-rank test to compare frequencies between blood and LNs. * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Error bars denote mean ± S.E.M.

Supplementary Fig. 4. a) *Representative mass cytometry profile of gp140 and flu specific B cells in LN B cell populations defined by IgD and CD38 expression.* $gp140^+$ B and Flu^+ B cells (black dots) from representative HIV^-, ART treated and viremic individuals by mass cytometry. Gates were set using total memory B cells (red dots) **b**) *Correlative analyses between the percentage of GC $gp140^+$ B cells with the percentage of GC B, the percentage of switched memory B cells and plasma viral load*. A Spearman rank test was used for correlations.

Supplementary Fig. 5. *Phenotype of Flu and gp140 specific B cells from ART treated individuals.* Heat map of scaled mean marker expression (% of positive cell gated) in Flu and gp140 specific B cells from ART treated individuals. All markers shown are significantly different between ART treated and viremic individuals (FDR < 0.05). Differences were calculated using linear regressions.

Supplementary Fig. 6. *Correlative analyses between the frequency of $gp140^+$ B cells and their phenotype and Tfh cytokine's production from ART treated individuals.* Correlative analyses were performed on log10 transformed frequencies using Pearson's test.

Supplementary Fig. 7. *Frequency of T-bet$^+$ CXCR3$^+$ B cells subsets from HIV uninfected, HIV infected ART treated and viremics.* Non naive B cells were gated on unswitched memory (IgD$^+$CD38$^-$), switched memory (IgD$^-$CD38$^-$), GC (IgD$^-$CD38$^+$) and plasma cells (IgD$^-$ CD38^{hi}) B cell populations and (**a**) the percentage of T-bet$^+$ CXCR3$^+$ B cells was analyzed by mass cytometry. (**b**) Percentage of T-bet$^+$ B cells within the CXCR3$^+$ and CXCR3$^-$ B cells. In (a) statistical significance (*P* values) were calculated using Mann-Whitney test while in (b) by Wilcoxon signed-rank test * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Error bars denote mean ± S.E.M.

Supplementary Table 1. *Mass cytometry panels.* **a**) Mass cytometry T cell panel. **b**) Mass cytometry B cell panel.

Figure 1

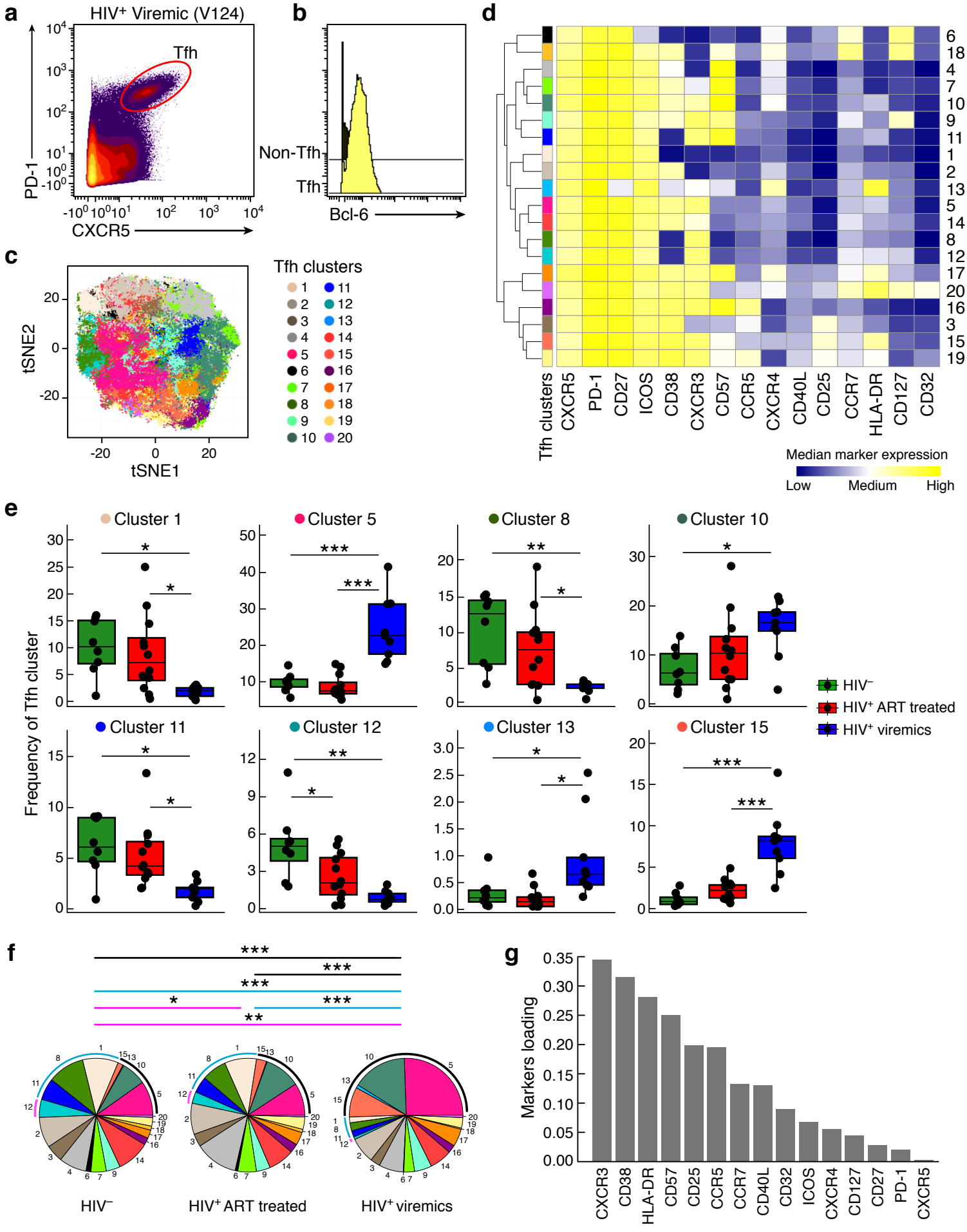

a
HIV+ Viremic (V124)
Tfh
PD-1
CXCR5
b
Non-Tfh
Tfh
Bcl-6
c
Tfh clusters
tSNE2
tSNE1
d
Tfh clusters
CXCR5
PD-1
CD27
ICOS
CD38
CXCR3
CD57
CCR5
CXCR4
CD40L
CD25
CCR7
HLA-DR
CD127
CD32
Median marker expression
Low
Medium
High
e
Cluster 1
Cluster 5
Cluster 8
Cluster 10
Cluster 11
Cluster 12
Cluster 13
Cluster 15
Frequency of Tfh cluster
HIV⁻
HIV+ ART treated
HIV+ viremics
f
HIV⁻
HIV+ ART treated
HIV+ viremics
g
Markers loading
CXCR3
CD38
HLA-DR
CD57
CD25
CCR5
CCR7
CD40L
CD32
ICOS
CXCR4
CD127
CD27
PD-1
CXCR5

Figure 2

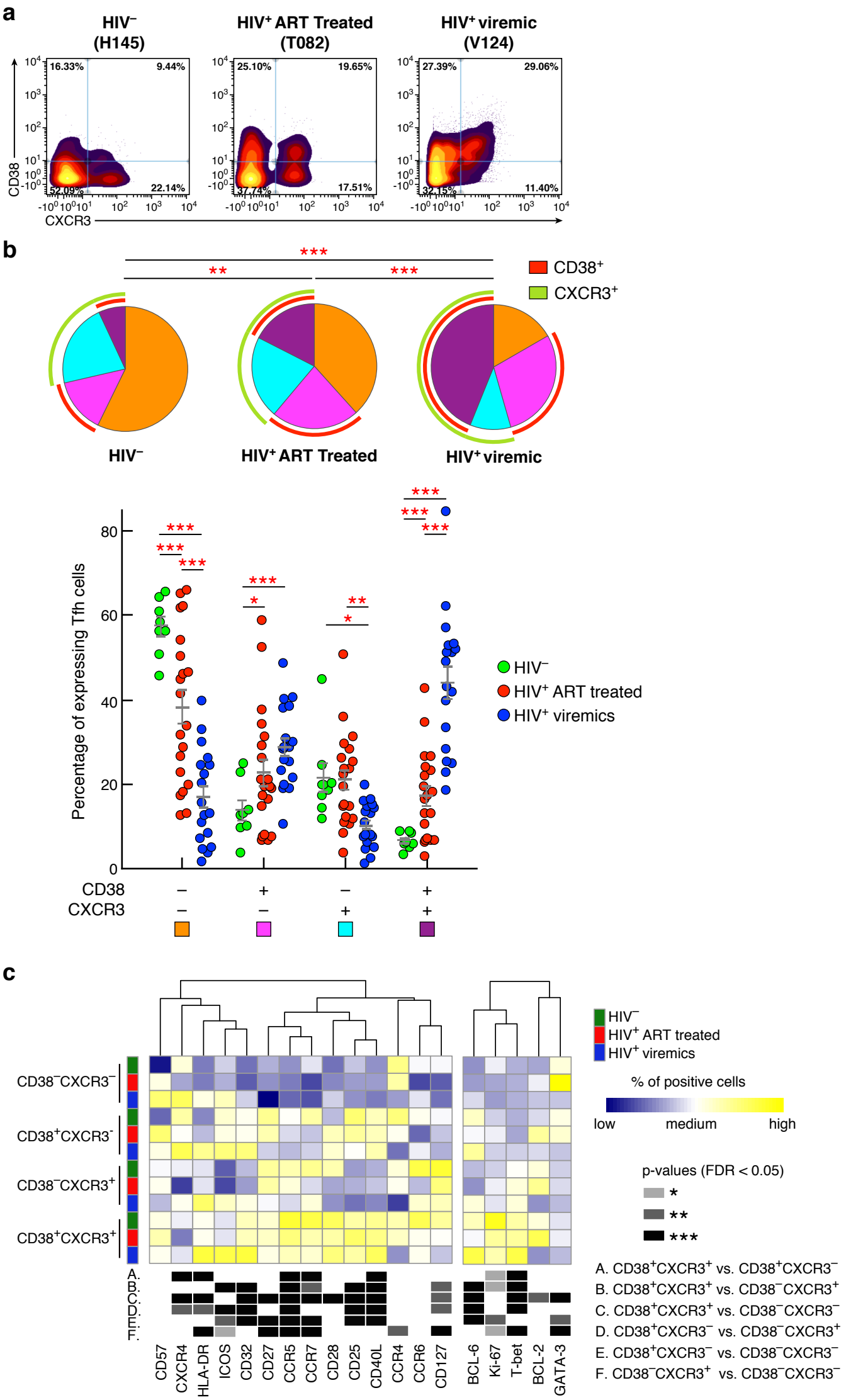

a
HIV⁻ (H145)
HIV⁺ ART Treated (T082)
HIV⁺ viremic (V124)
16.33%
9.44%
52.09%
22.14%
25.10%
19.65%
37.74%
17.51%
27.39%
29.06%
32.15%
11.40%
CD38
CXCR3
b
CD38⁺
CXCR3⁺
HIV⁻
HIV⁺ ART Treated
HIV⁺ viremic
Percentage of expressing Tfh cells
HIV⁻
HIV⁺ ART treated
HIV⁺ viremics
CD38 − + − +
CXCR3 − − + +
c
CD38⁻CXCR3⁻
CD38⁺CXCR3⁻
CD38⁻CXCR3⁺
CD38⁺CXCR3⁺
CD57 CXCR4 HLA-DR ICOS CD32 CD27 CCR5 CCR7 CD28 CD25 CD40L CCR4 CCR6 CD127 BCL-6 Ki-67 T-bet BCL-2 GATA-3
HIV⁻
HIV⁺ ART treated
HIV⁺ viremics
% of positive cells
low medium high
p-values (FDR < 0.05)
A. CD38⁺CXCR3⁺ vs. CD38⁺CXCR3⁻
B. CD38⁺CXCR3⁺ vs. CD38⁻CXCR3⁺
C. CD38⁺CXCR3⁺ vs. CD38⁻CXCR3⁻
D. CD38⁺CXCR3⁻ vs. CD38⁻CXCR3⁺
E. CD38⁺CXCR3⁻ vs. CD38⁻CXCR3⁻
F. CD38⁻CXCR3⁺ vs. CD38⁻CXCR3⁻

Figure 3

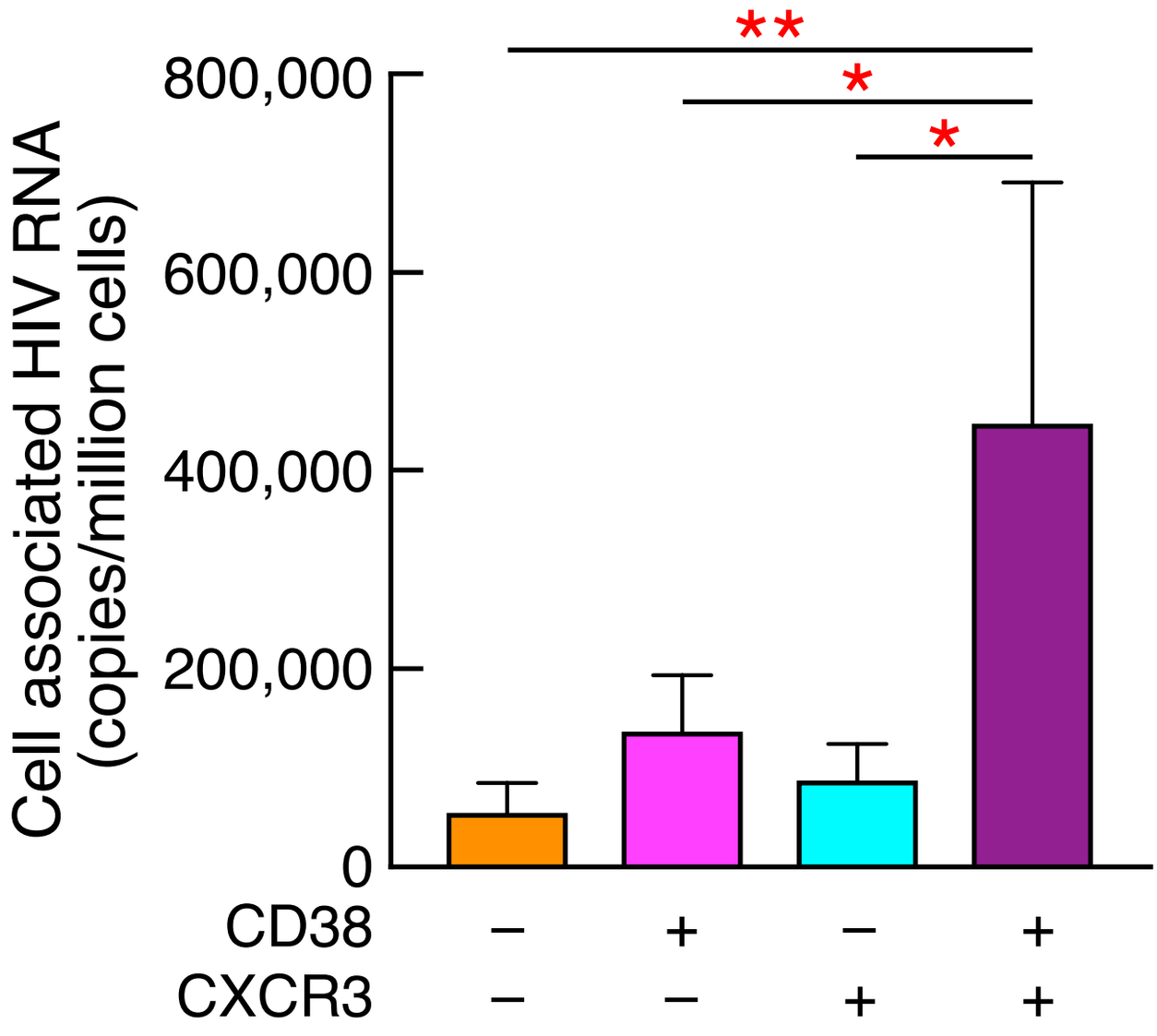

**
*
*
800,000
600,000
400,000
200,000
0
Cell associated HIV RNA
(copies/million cells)
CD38
CXCR3
–
+
–
+
–
–
+
+

Figure 4

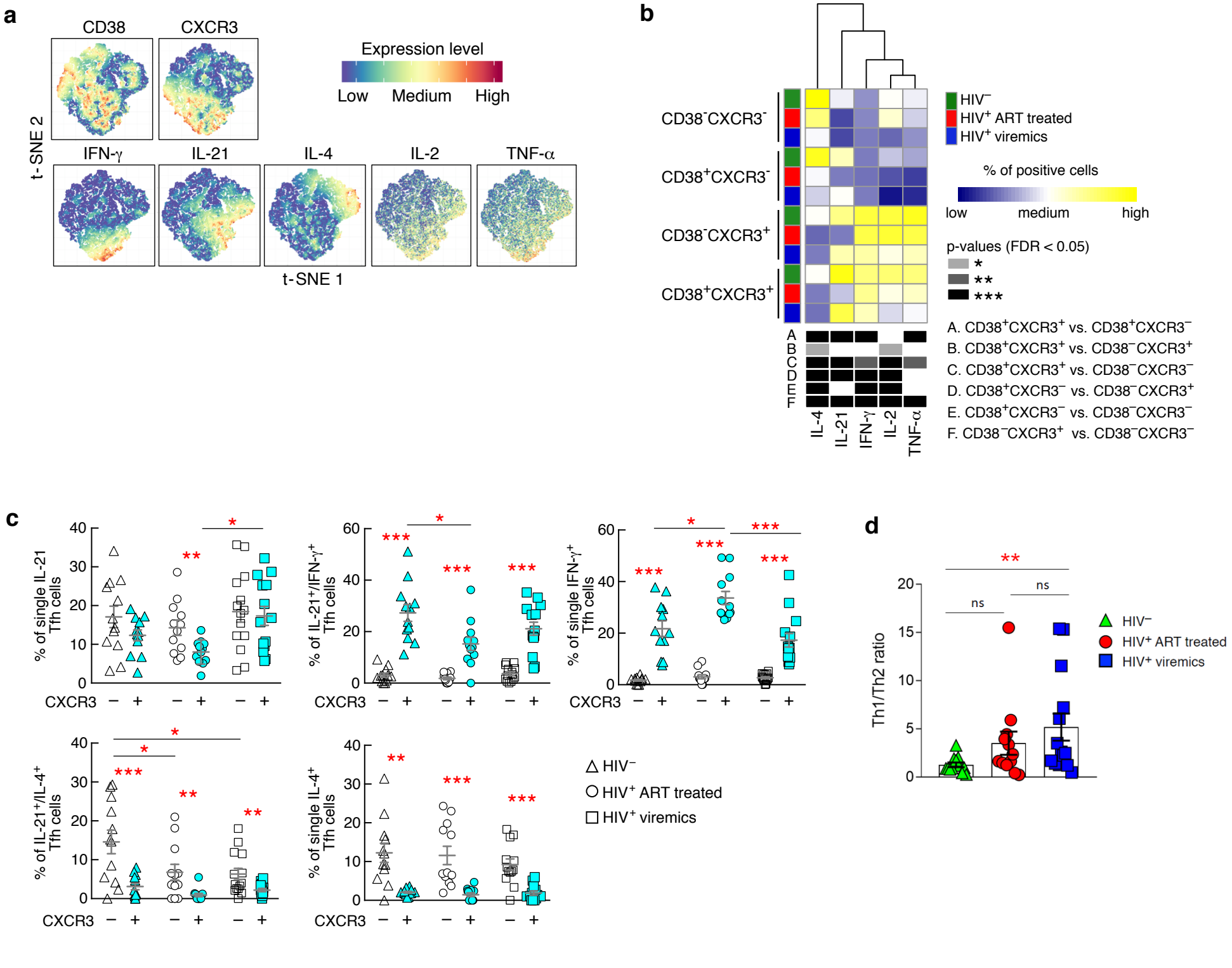

a
CD38
CXCR3
Expression level
Low
Medium
High
IFN-γ
IL-21
IL-4
IL-2
TNF-α
t-SNE 2
t-SNE 1
b
CD38⁻CXCR3⁻
CD38⁺CXCR3⁻
CD38⁻CXCR3⁺
CD38⁺CXCR3⁺
HIV⁻
HIV⁺ ART treated
HIV⁺ viremics
% of positive cells
low
medium
high
p-values (FDR < 0.05)
*
**
***
A. CD38⁺CXCR3⁺ vs. CD38⁺CXCR3⁻
B. CD38⁺CXCR3⁺ vs. CD38⁻CXCR3⁺
C. CD38⁺CXCR3⁺ vs. CD38⁻CXCR3⁻
D. CD38⁺CXCR3⁻ vs. CD38⁻CXCR3⁺
E. CD38⁺CXCR3⁻ vs. CD38⁻CXCR3⁻
F. CD38⁻CXCR3⁺ vs. CD38⁻CXCR3⁻
IL-4
IL-21
IFN-γ
IL-2
TNF-α
c
% of single IL-21 Tfh cells
% of IL-21⁺/IFN-γ⁺ Tfh cells
% of single IFN-γ⁺ Tfh cells
% of IL-21⁺/IL-4⁺ Tfh cells
% of single IL-4⁺ Tfh cells
CXCR3
HIV⁻
HIV⁺ ART treated
HIV⁺ viremics
d
Th1/Th2 ratio
ns
HIV⁻
HIV⁺ ART treated
HIV⁺ viremics

Figure 5

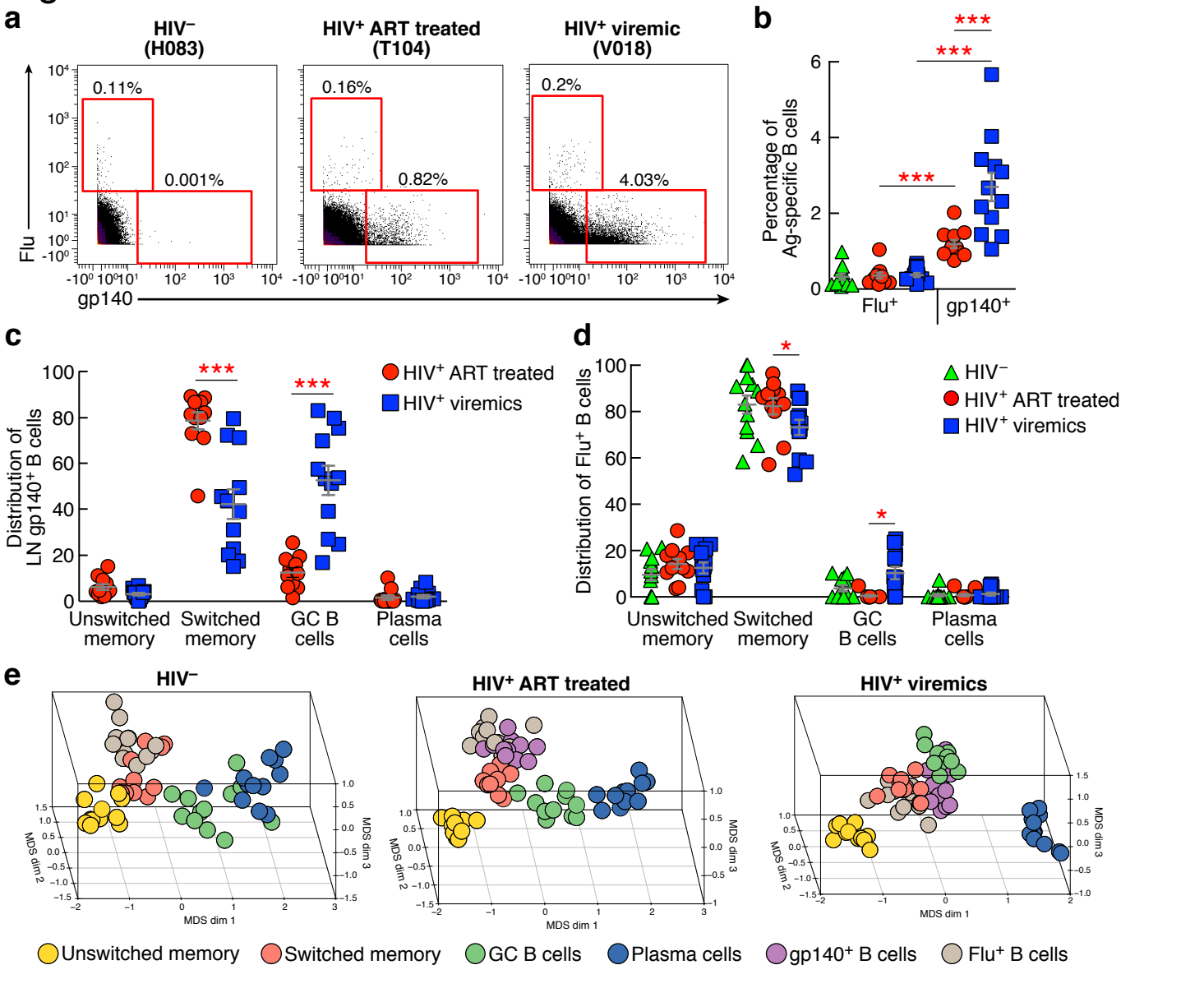

a
HIV⁻ (H083)
HIV⁺ ART treated (T104)
HIV⁺ viremic (V018)
0.11%
0.001%
0.16%
0.82%
0.2%
4.03%
Flu
gp140
b
Percentage of Ag-specific B cells
***
***
***
Flu⁺
gp140⁺
c
Distribution of LN gp140⁺ B cells
***
***
HIV⁺ ART treated
HIV⁺ viremics
Unswitched memory
Switched memory
GC B cells
Plasma cells
d
Distribution of Flu⁺ B cells
*
*
HIV⁻
HIV⁺ ART treated
HIV⁺ viremics
Unswitched memory
Switched memory
GC B cells
Plasma cells
e
HIV⁻
HIV⁺ ART treated
HIV⁺ viremics
MDS dim 1
MDS dim 2
MDS dim 3
Unswitched memory
Switched memory
GC B cells
Plasma cells
gp140⁺ B cells
Flu⁺ B cells

Figure 6

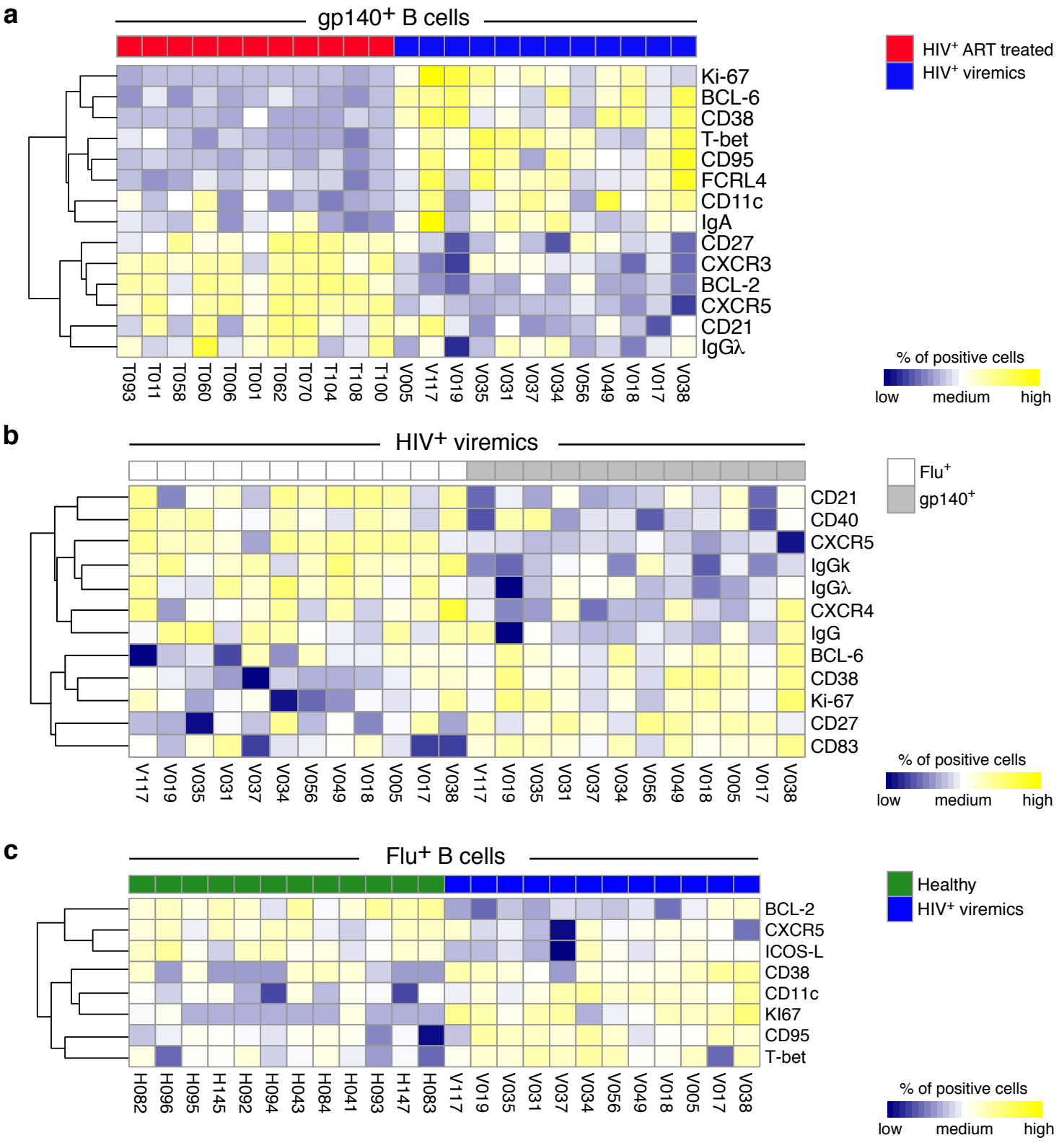

a
gp140+ B cells
HIV+ ART treated
HIV+ viremics
Ki-67
BCL-6
CD38
T-bet
CD95
FCRL4
CD11c
IgA
CD27
CXCR3
BCL-2
CXCR5
CD21
IgGλ
T093 T011 T058 T060 T006 T001 T062 T070 T104 T108 T100 V005 V117 V019 V035 V031 V037 V034 V056 V049 V018 V017 V038
% of positive cells
low medium high
b
HIV+ viremics
Flu+
gp140+
CD21
CD40
CXCR5
IgGk
IgGλ
CXCR4
IgG
BCL-6
CD38
Ki-67
CD27
CD83
V117 V019 V035 V031 V037 V034 V056 V049 V018 V005 V017 V038 V117 V019 V035 V031 V037 V034 V056 V049 V018 V005 V017 V038
% of positive cells
low medium high
c
Flu+ B cells
Healthy
HIV+ viremics
BCL-2
CXCR5
ICOS-L
CD38
CD11c
KI67
CD95
T-bet
H082 H096 H095 H145 H092 H094 H043 H084 H041 H093 H147 H083 V117 V019 V035 V031 V037 V034 V056 V049 V018 V005 V017 V038
% of positive cells
low medium high

Figure 7

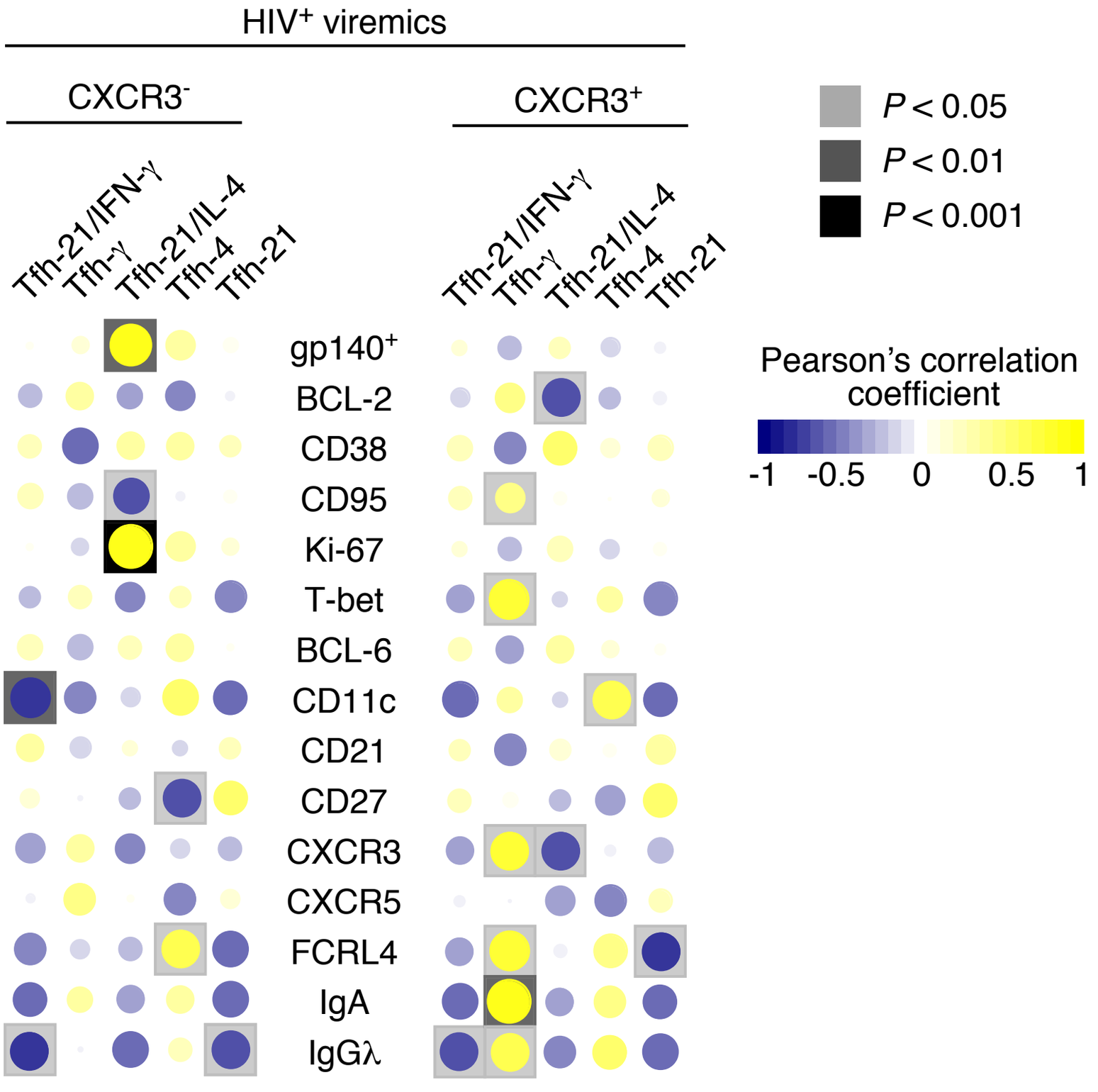
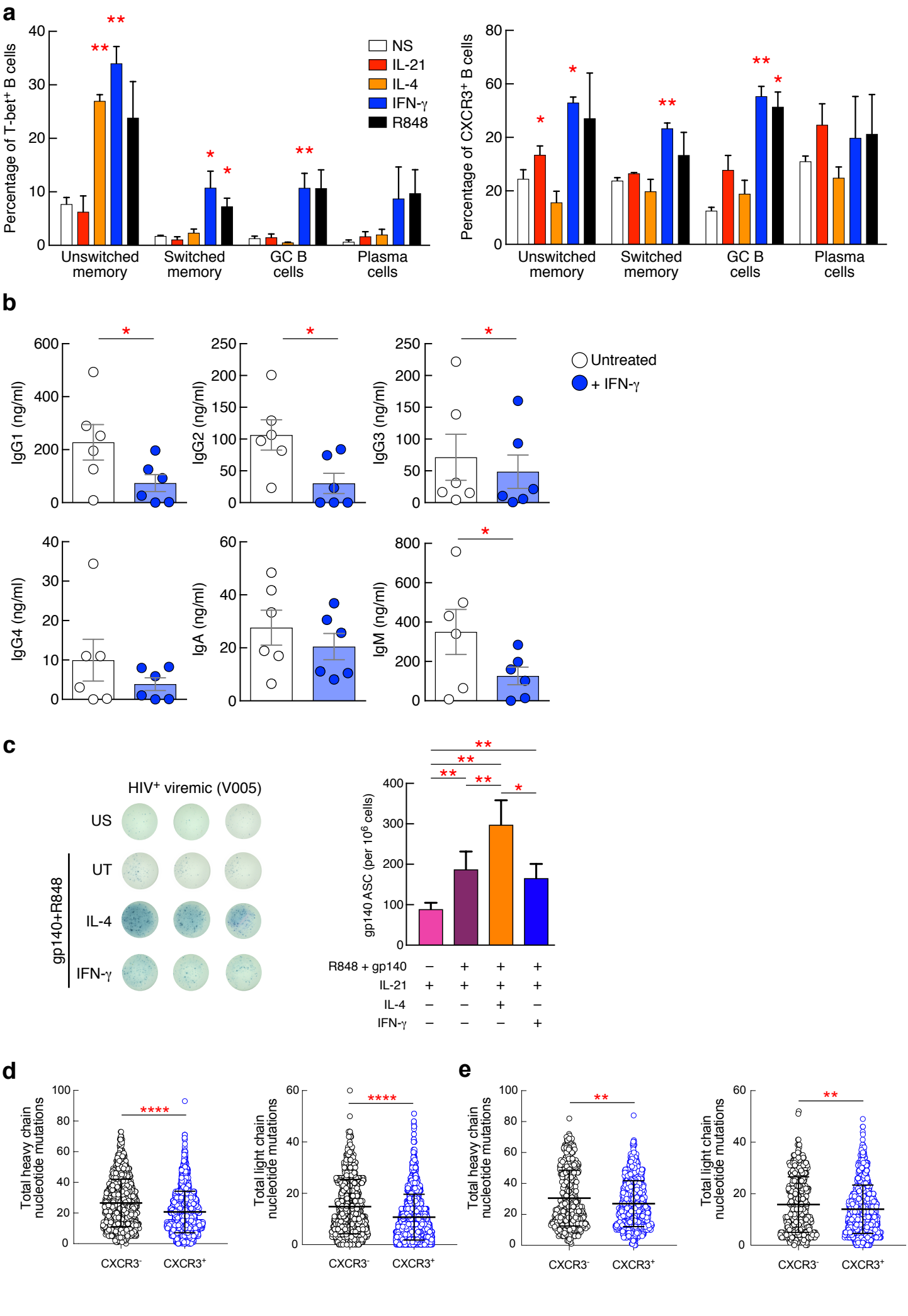
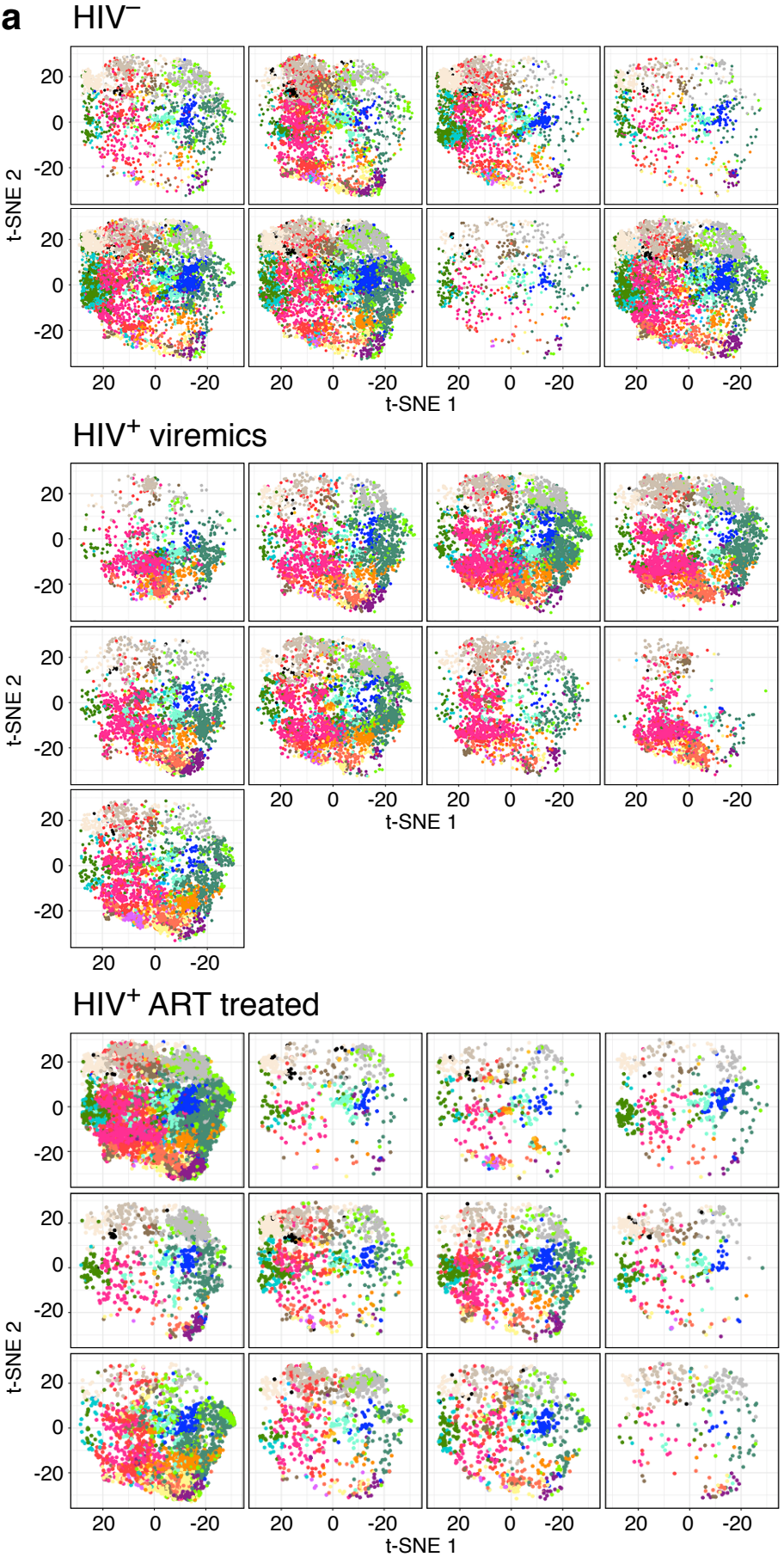

HIV+ viremics
CXCR3-
CXCR3+
Tfh-21/IFN-γ
Tfh-γ
Tfh-21/IL-4
Tfh-4
Tfh-21
gp140+
BCL-2
CD38
CD95
Ki-67
T-bet
BCL-6
CD11c
CD21
CD27
CXCR3
CXCR5
FCRL4
IgA
IgGλ
P < 0.05
P < 0.01
P < 0.001
Pearson's correlation coefficient
-1
-0.5
0
0.5
1

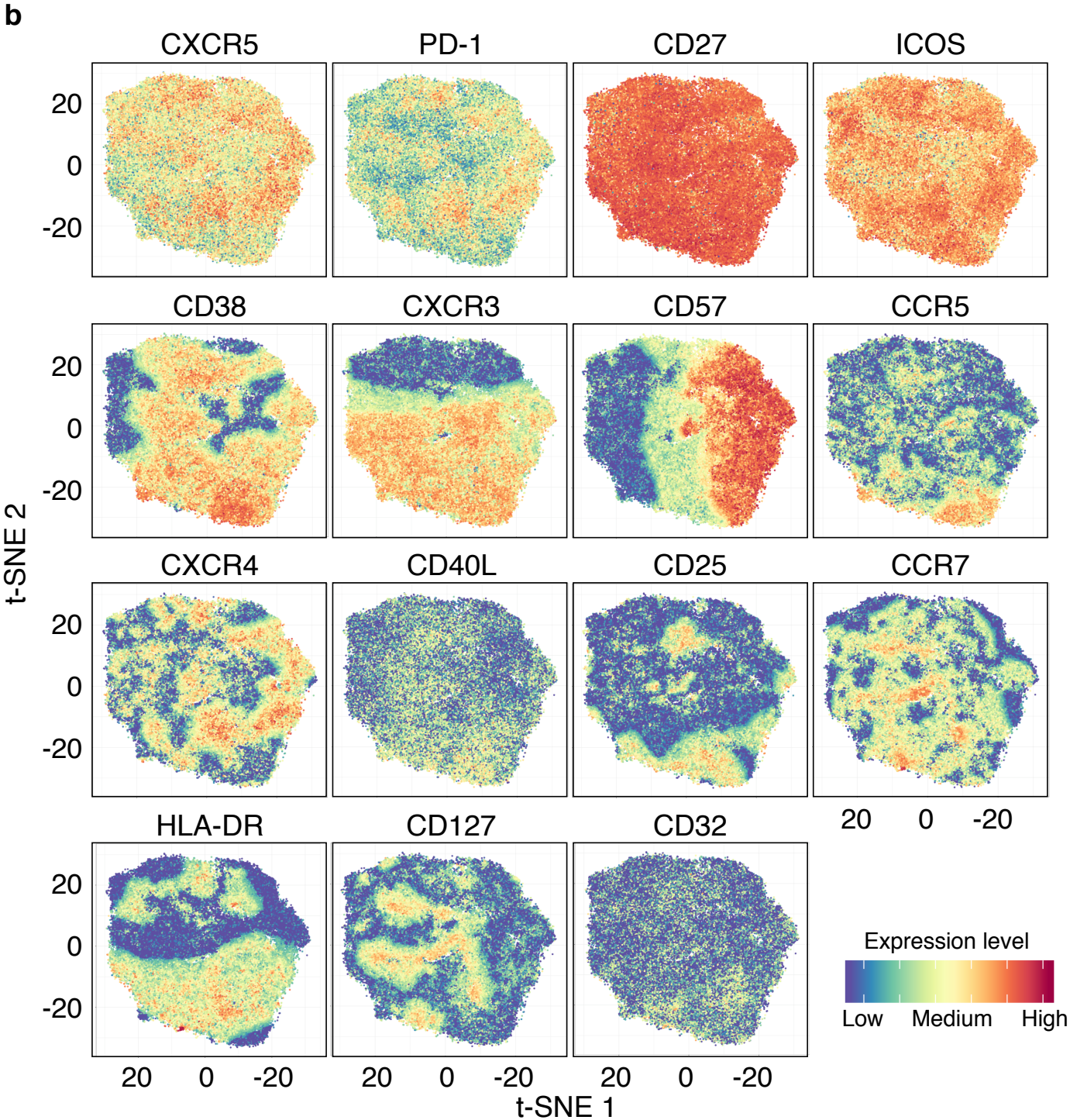
Figure 8

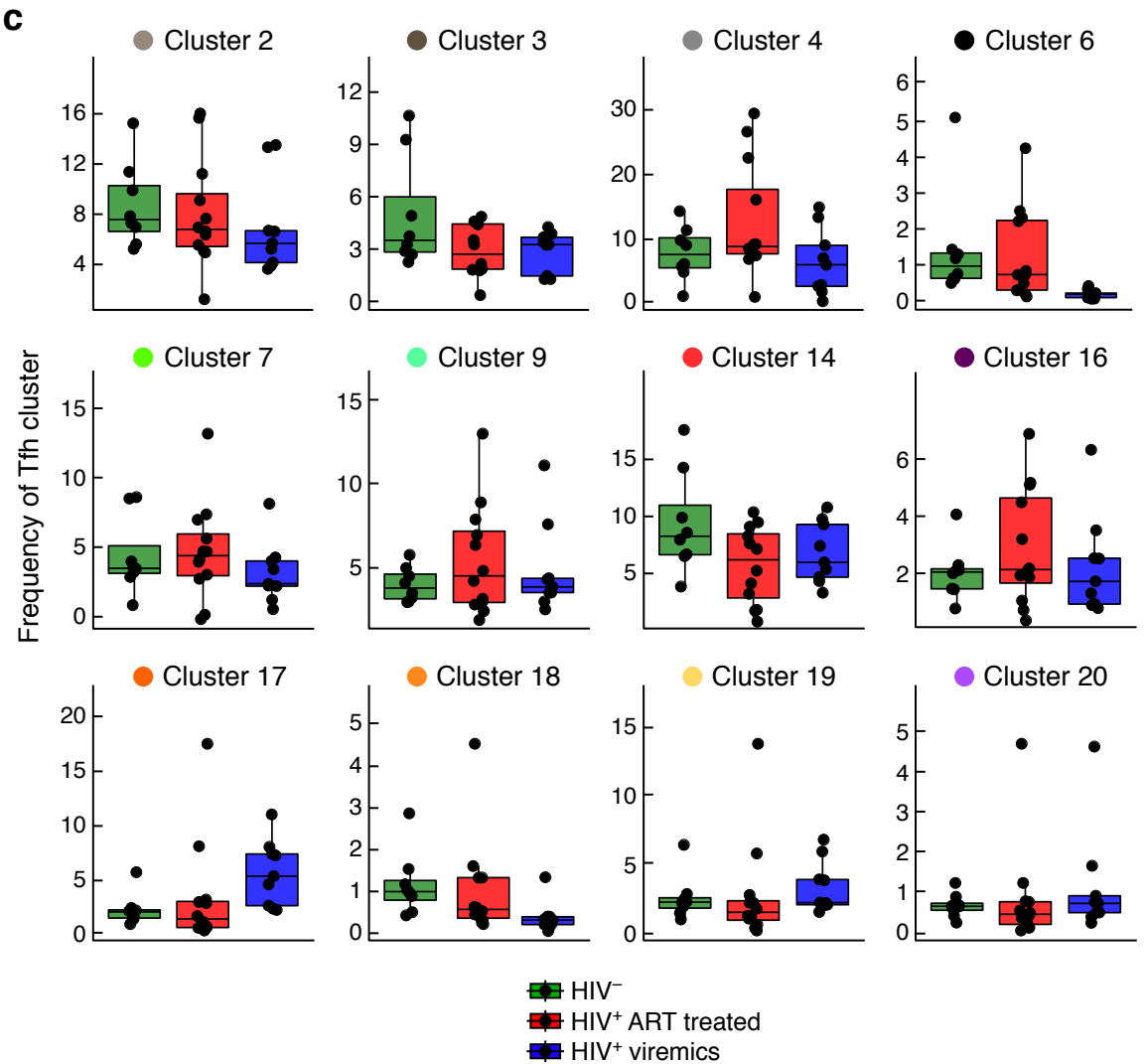

a
Percentage of T-bet+ B cells
NS
IL-21
IL-4
IFN-γ
R848
Unswitched memory
Switched memory
GC B cells
Plasma cells
Percentage of CXCR3+ B cells
b
IgG1 (ng/ml)
IgG2 (ng/ml)
IgG3 (ng/ml)
IgG4 (ng/ml)
IgA (ng/ml)
IgM (ng/ml)
Untreated
+ IFN-γ
c
HIV+ viremic (V005)
US
UT
IL-4
IFN-γ
gp140+R848
gp140 ASC (per 10^6 cells)
R848 + gp140
IL-21
IL-4
IFN-γ
d
Total heavy chain nucleotide mutations
Total light chain nucleotide mutations
CXCR3-
CXCR3+
e
Total heavy chain nucleotide mutations
Total light chain nucleotide mutations
CXCR3-
CXCR3+

Supplementary Figure 1

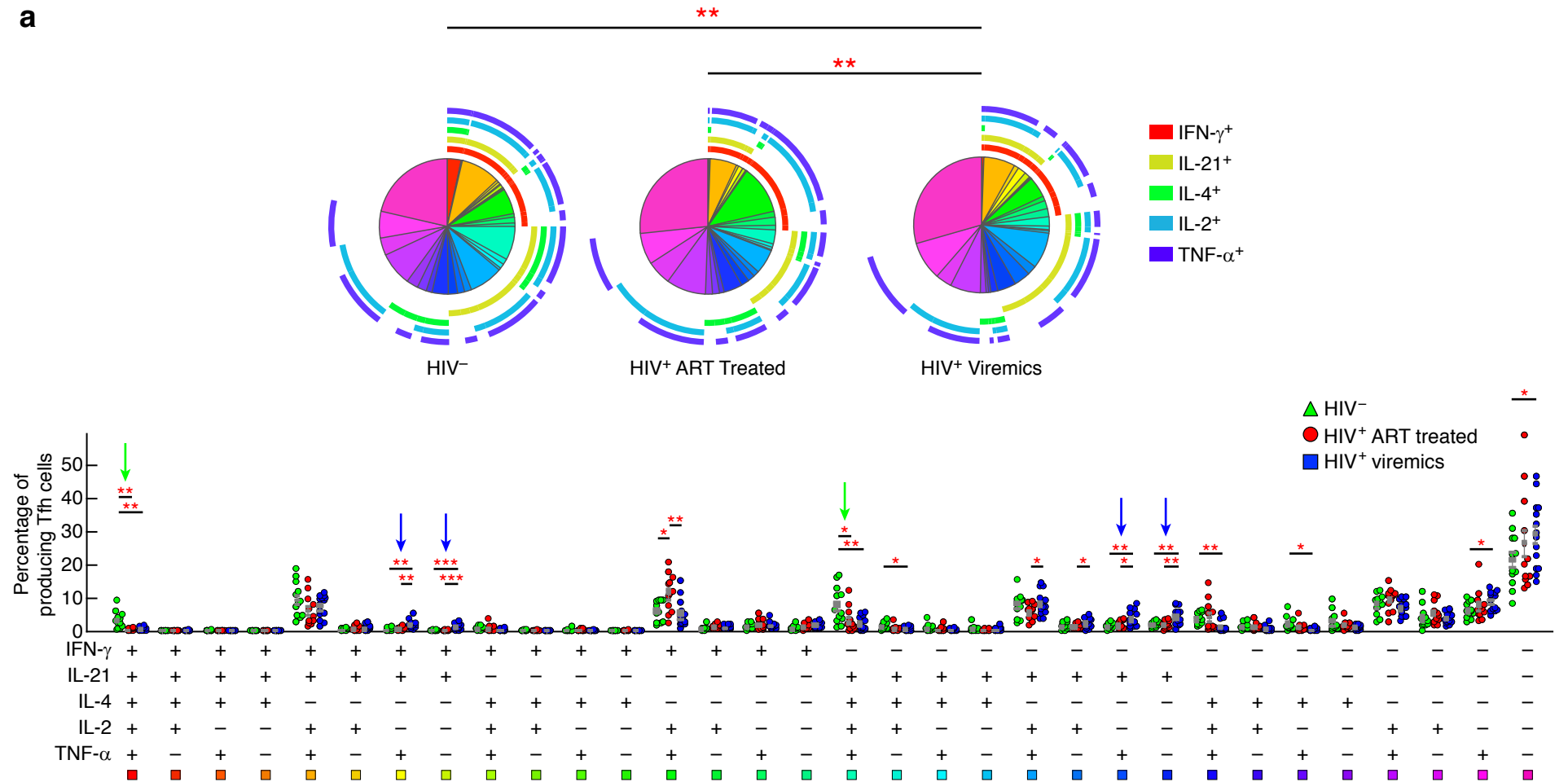
a

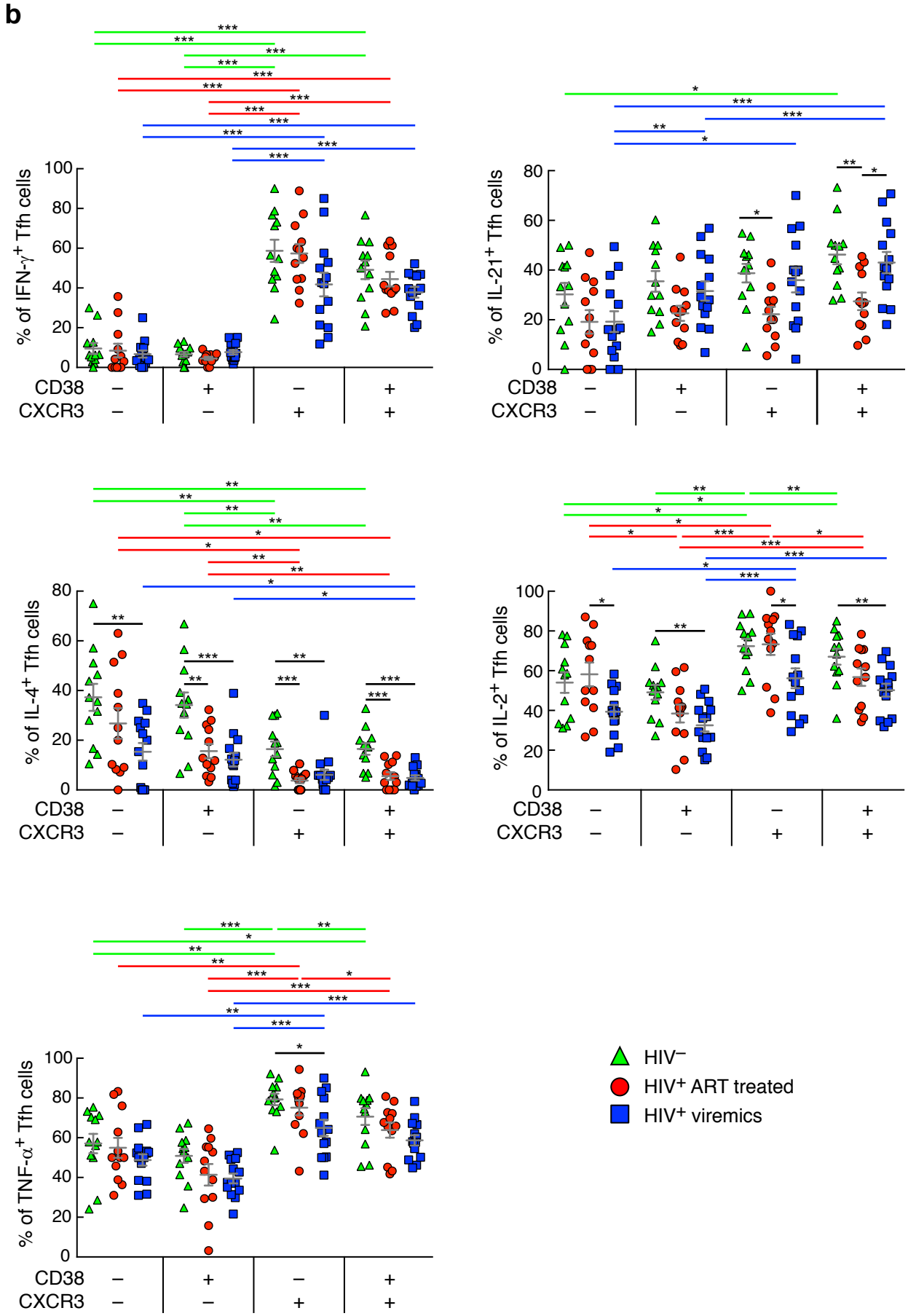


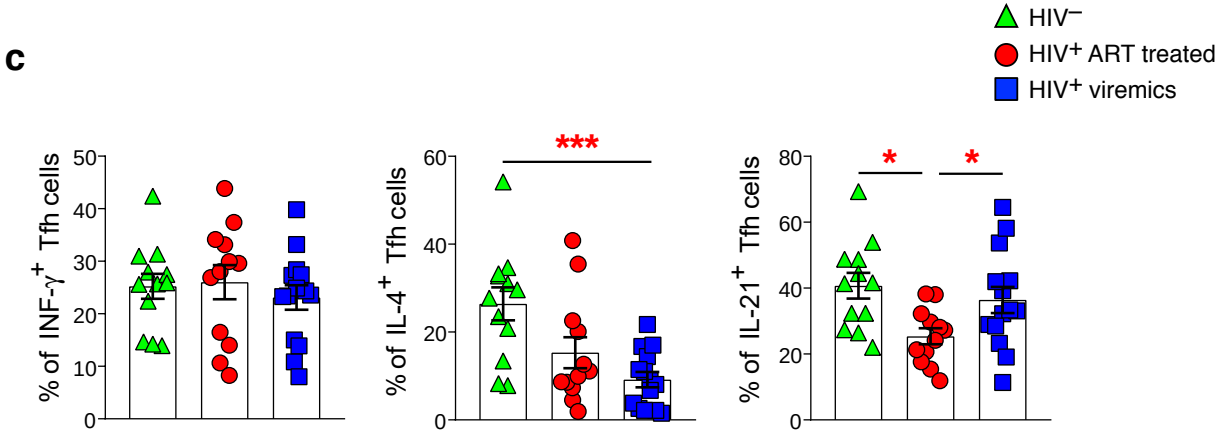

b
CXCR5
PD-1
CD27
ICOS
CD38
CXCR3
CD57
CCR5
CXCR4
CD40L
CD25
CCR7
HLA-DR
CD127
CD32
20
0
-20
t-SNE 2
t-SNE 1
Expression level
Low
Medium
High


c
Cluster 2
Cluster 3
Cluster 4
Cluster 6
Cluster 7
Cluster 9
Cluster 14
Cluster 16
Cluster 17
Cluster 18
Cluster 19
Cluster 20
Frequency of Tfh cluster
HIV⁻
HIV⁺ ART treated
HIV⁺ viremics

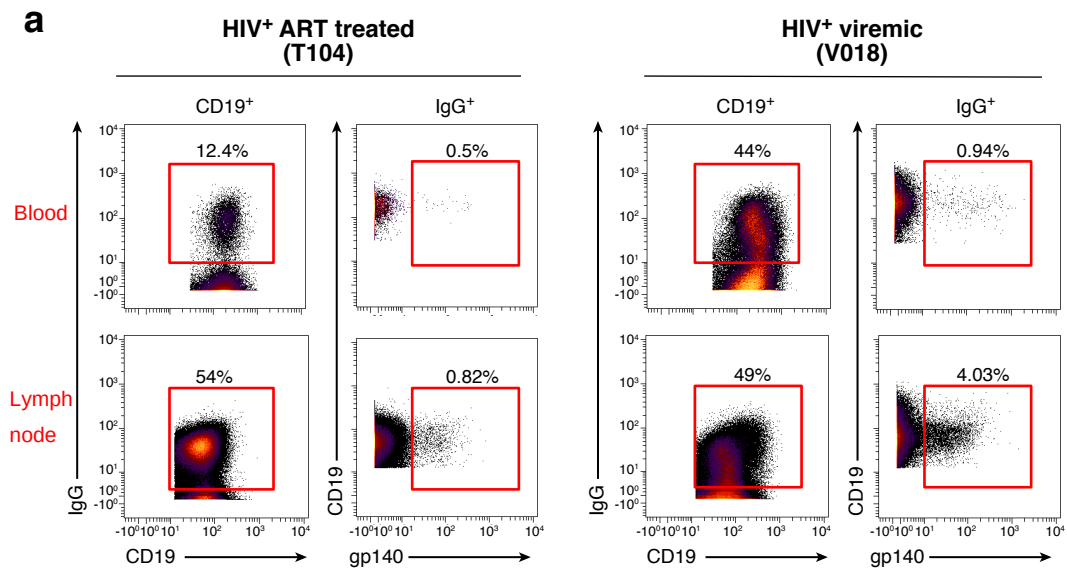
Supplementary Figure 2

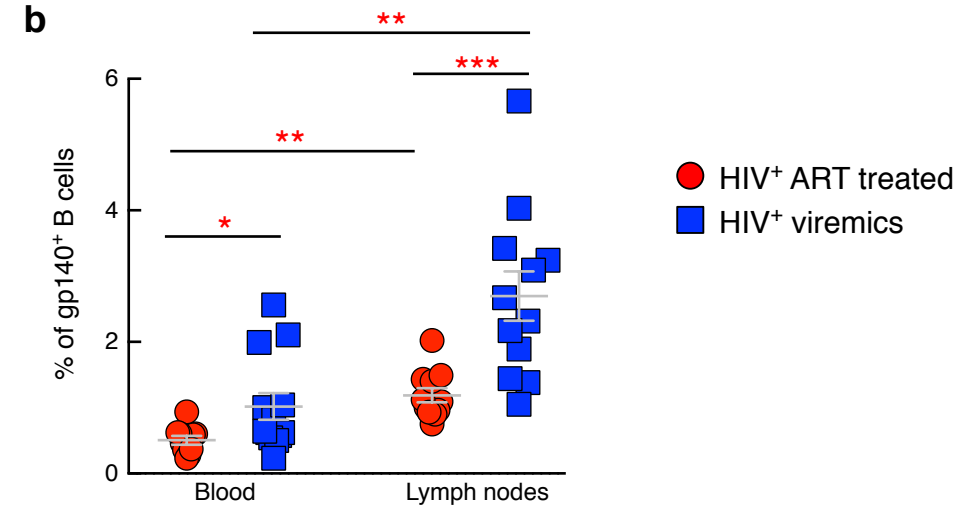



b
% of IFN-γ+ Tfh cells
% of IL-21+ Tfh cells
% of IL-4+ Tfh cells
% of IL-2+ Tfh cells
% of TNF-α+ Tfh cells
CD38
CXCR3
HIV−
HIV+ ART treated
HIV+ viremics

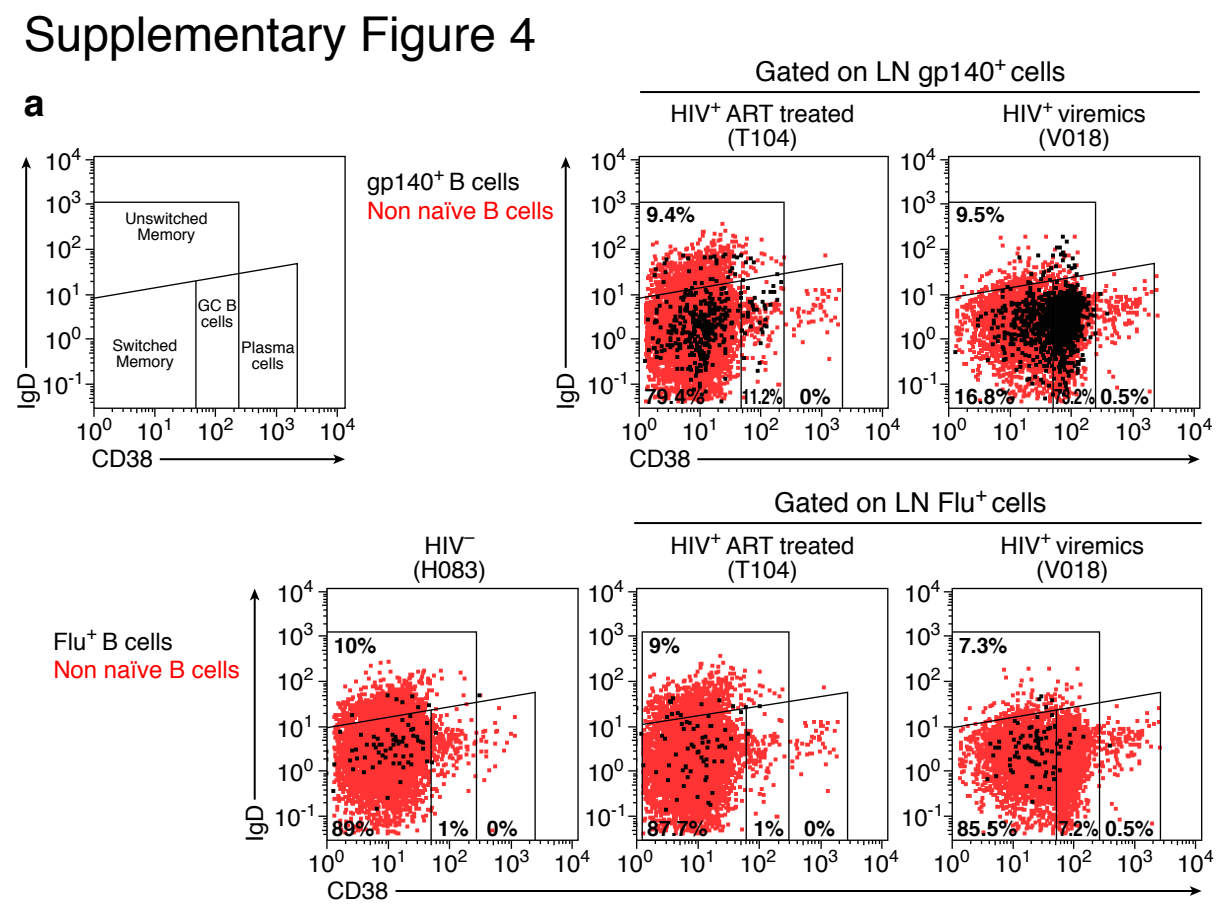

c
HIV−
HIV+ ART treated
HIV+ viremics
% of INF-γ+ Tfh cells
50
40
30
20
10
0
% of IL-4+ Tfh cells
60
40
20
0
***
% of IL-21+ Tfh cells
80
60
40
20
0
*
*

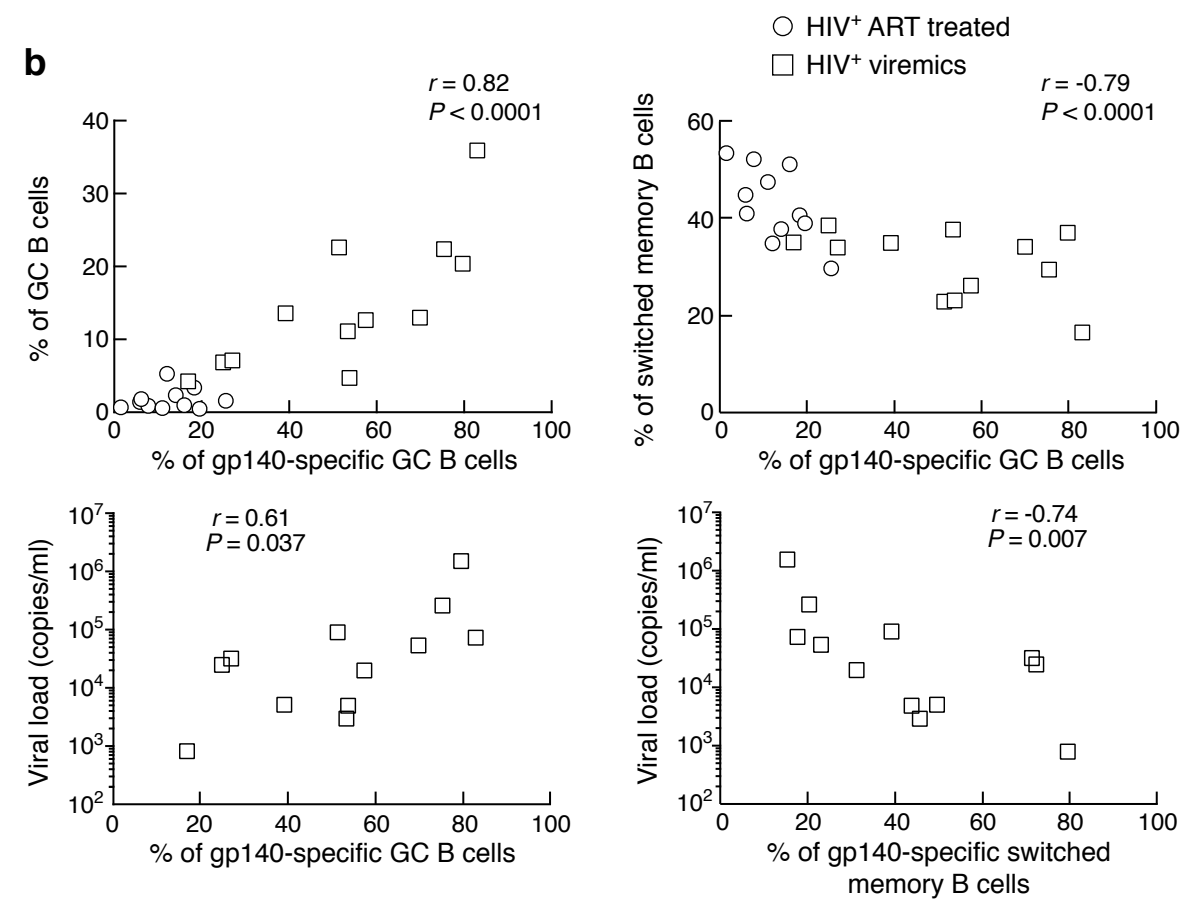
Supplementary Figure 3



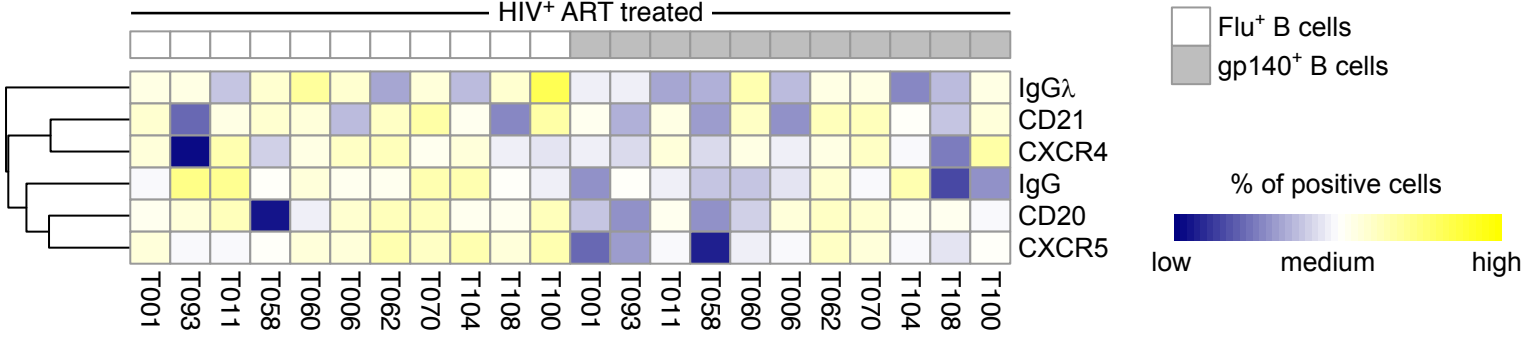


Supplementary Figure 4

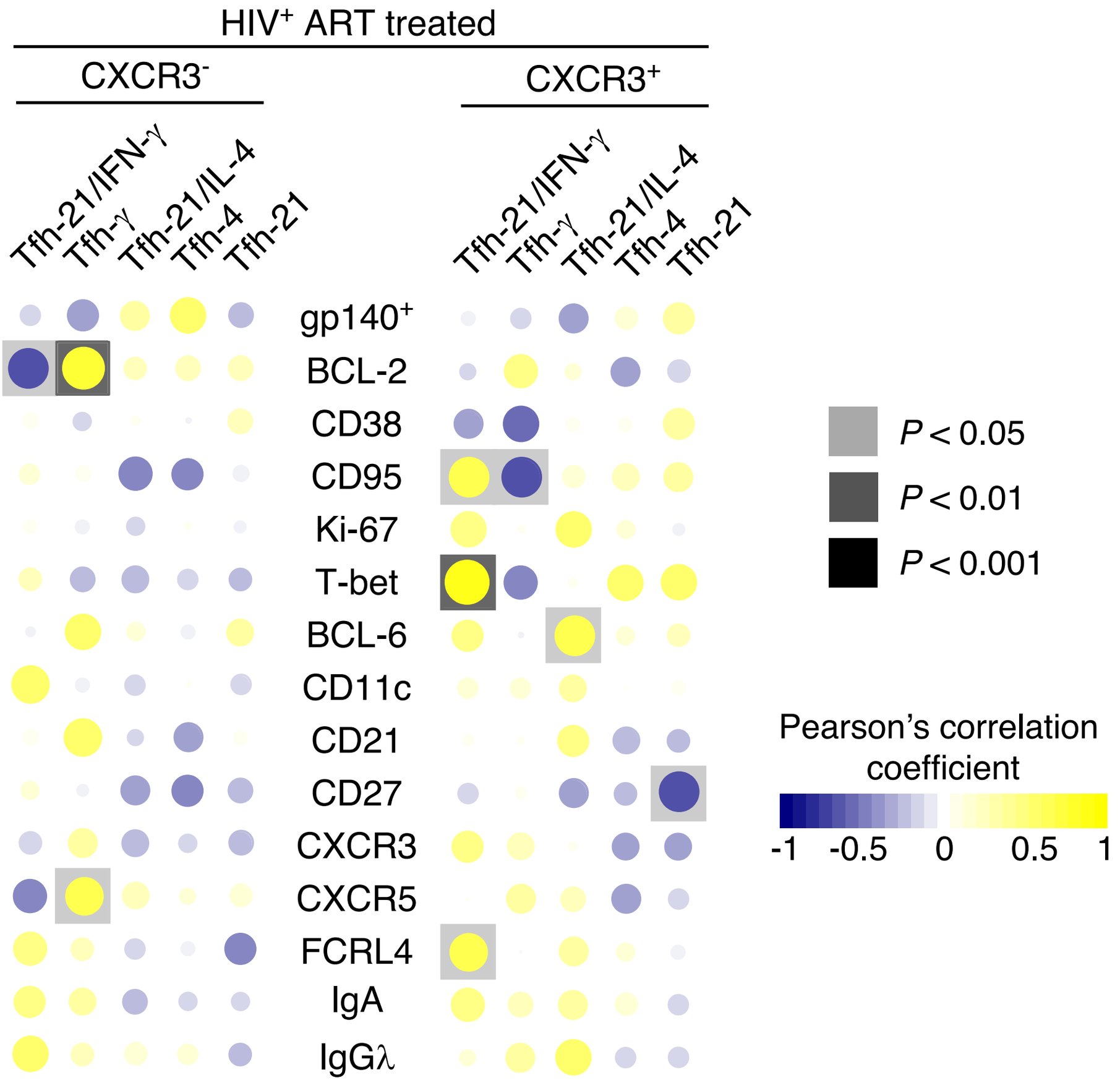



b
HIV+ ART treated
HIV+ viremics
r = 0.82
P < 0.0001
% of GC B cells
% of gp140-specific GC B cells
r = -0.79
P < 0.0001
% of switched memory B cells
% of gp140-specific GC B cells
r = 0.61
P = 0.037
Viral load (copies/ml)
% of gp140-specific GC B cells
r = -0.74
P = 0.007
Viral load (copies/ml)
% of gp140-specific switched memory B cells

Supplementary Figure 5



Supplementary Figure 6



Supplementary Figure 7

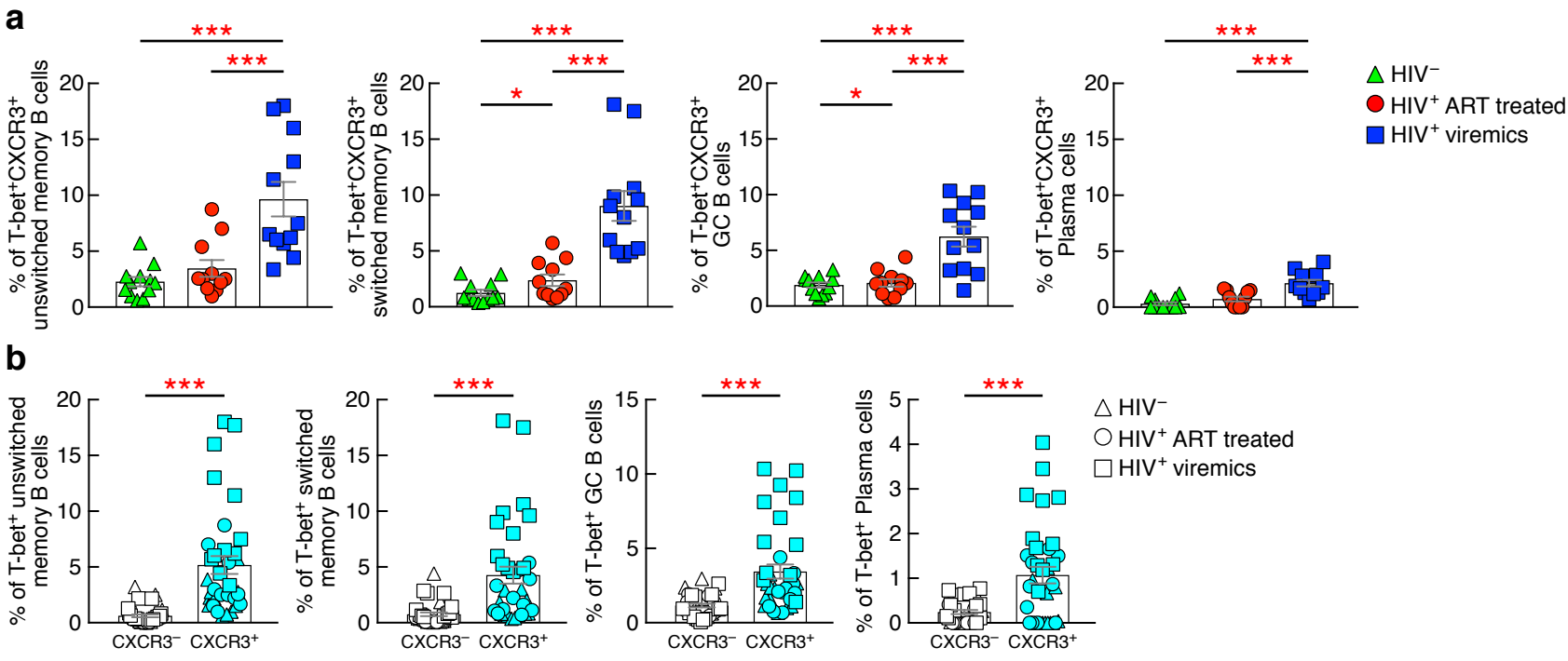


Table S1

a Mass cytometry T cell panel.

Target	Metal	Company	Clone
CD4	115In	Biolegend	RPA-T4
CCR6	141Pr	Fluidigm/DVS	G034E3
CD19	142Nd	Fluidigm/DVS	HIB19
ICOS	143Nd	Biolegend	C398.4A
CD8	145Nd	Biolegend	RPA-T8
IgD	146Nd	Fluidigm/DVS	IA6-2
CD7	147Sm	Fluidigm/DVS	CD7-6B7
CD57	148Nd	BD	G10F5
CCR4	149Sm	Fluidigm/DVS	205410
CXCR3	153Eu	Fluidigm/DVS	RF8B2
CD21	152Sm	Fluidigm/DVS	BL13
CXCR3	154Sm	BIolegend	G025H7
CD27	155Gd	Fluidigm/DVS	L128
CD11c	156Gd	Biolegend	3.9
CCR7	159Tb	Fluidigm/DVS	G043H7
CD25	158Gd	Biolegend	M-A251
CD14	160Gd	Fluidigm/DVS	M5E2
CD1C	161Dy	Biolegend	L161
CD32-APC	162Dy	Fluidigm/DVS	FUN2
CD20	166Er	Biolegend	2H7
CD38	167Er	Fluidigm/DVS	HIT2
CD45RA	169Tm	Fluidigm/DVS	HI100
CD40L	168Er	Fluidigm/DVS	CD40L
CD3	170Er	Fluidigm/DVS	UCHT1
CCR5	171Yb	Fluidigm/DVS	NP-6G4
HLA-DR	173Yb	Fluidigm/DVS	L243
PD-1	174Yb	Fluidigm/DVS	EH12.2H7
CXCR4	175Lu	Fluidigm/DVS	12G5
CD127	176Yb	Fluidigm/DVS	A019D5
CD16	209BI	Fluidigm/DVS	3G8
Live/Dead	195Pt	Fluidigm/DVS	Cell-ID

b Mass cytometry B cell panel.

Target	Metal	Company	Clone
IgG*	113-In	BD	G18-145
Blimp-1*	115-In	BIO-TECHNE AG	646702
PARP*	141-Pr	BD	F21-852
CD19	142-Nd	DVS	HIB19
HLADR*	143-Nd	DVS	L243
CD38*	144-Nd	DVS	HIT2
CD8	145-Nd	Biolegend	RPA-T8
IgD*	146-Nd	DVS	IA6-2
CD20*	147-Sm	DVS	2H7
IgA*	148-Sm	DVS	Polyclonal
CD79A*	149-Sm	Biolegend	gα (alpha)
CD138*	150-Nd	DVS	DL-101
IgG lambda*	151-Eu	DVS	MHL-38
CD21*	152-Sm	DVS	BL13
CXCR5*	153-Eu	DVS	RF8B2
CXCR3*	154-Sm	Biolegend	G025H7
CD27*	155-Gd	DVS	L128
gp140-PE	156-Gd	DVS	PE001
ICOS-L*	158-Gd	Biolegend	2D3
CD11C*	159-Tb	DVS	Bu15
IgG kappa*	160-Gd	DVS	MHK-49
T-bet*	161-Dy	DVS	4B10
H1N1-APC	162-Dy	DVS	APC003
BCL6*	163-Dy	DVS	K112-91
CD95*	164-Dy	DVS	FAS
CD40*	165-Ho	DVS	5C3
CD24*	166-Er	DVS	ML5
FCRL4*	167-Er	Biolegend	413D12
Ki-67*	168-Er	DVS	Ki-67
CD45RA*	169-Tm	DVS	HI100
CD3	170-Er	DVS	UCHT1
CD83*	171-Yb	Biolegend	HB15e
IgM*	172-Yb	DVS	MHM-88
BCL-2*	173-Yb	Biolegend	100
PD-1*	174-Yb	DVS	EH12.2H7
CXCR4*	175-Lu	DVS	12G5
CD4	176-Yb	DVS	RPA-T4

*used for MDS analysis