Cytosplore: Interactive Visual Single-Cell Profiling of the Immune System

T. Höllt[1,2], N. Pezzotti[2], V. van Unen[3,4], Na Li[3], F. Koning[3], E. Eisemann[2], B.P.F. Lelieveldt[5], and A. Vilanova[2]

[1]Leiden Computational Biology Center, Leiden University Medical Center, Leiden, The Netherlands
[2]Computer Graphics and Visualization Department, TU Delft, Delft, The Netherlands
[3]Leiden University Medical Center, Department of Immunohematology, Leiden, The Netherlands
[4]Institute for Immunity, Transplantation, and Infection, Stanford University School of Medicine, Stanford, CA, USA
[5]Leiden University Medical Center, Department of Radiology, Leiden, The Netherlands

Abstract

Recent advances in single-cell acquisition technology have led to a shift towards single-cell analysis in many fields of biology. In immunology, detailed knowledge of the cellular composition is of interest, as it can be the cause of deregulated immune responses, which cause diseases. Similarly, vaccination is based on triggering proper immune responses; however, many vaccines are ineffective or only work properly in a subset of those who are vaccinated. Identifying differences in the cellular composition of the immune system in such cases can lead to more precise treatment. Cytosplore is an integrated, interactive visual analysis framework for the exploration of large single-cell datasets. We have developed Cytosplore in close collaboration with immunology researchers and several partners use the software in their daily workflow. Cytosplore enables efficient data analysis and has led to several discoveries alongside high-impact publications.

CCS Concepts

• ***Human-centered computing*** → ***Information visualization; Visualization theory, concepts and paradigms;***

1. Introduction

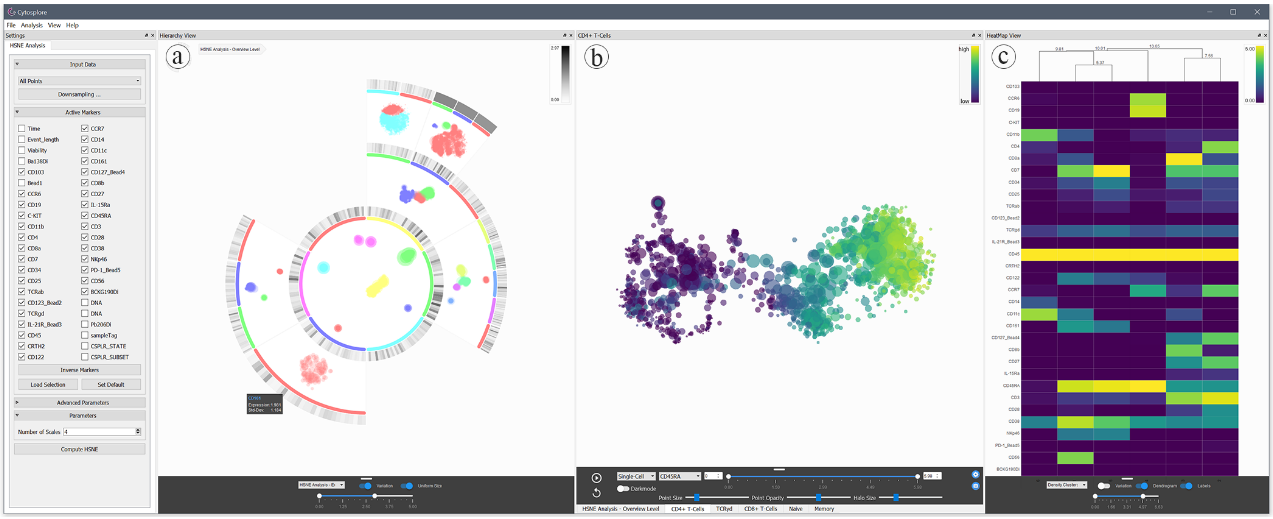
The rapid development of high-throughput single-cell acquisition techniques based on transcriptional and proteomic profiling enable the comprehensive classification of cell types. The power of these techniques led to broad efforts to explore and understand the cellular composition of the human body [RTL*17]. However, the complex and large data pose considerable challenges for analysis.

In immunology research, recently-introduced single-cell mass cytometry [OKB*08] has gained considerable traction as the preferred data acquisition tool. The functionality of immune cells mostly relates to a set of proteins expressed on the cells' surface and, at the moment, mass cytometry allows to measure the expression of approximately 50 different proteins per cell simultaneously. This is a significant increase over the clinical standard, flow cytometry, which typically allows for simultaneously measuring up to 15 proteins However, this number is still orders of magnitude smaller than the estimated 10,000 immune-system-wide available proteins. Consequently, researchers are required to select an often unique subset of proteins for their studies. While this enables the discovery of previously unknown cell types, it requires the identification and classification of different cell types to be carried out in a data-driven fashion by studying data heterogeneity rather than applying prior knowledge. The classified cells can then be used to facilitate further analysis at different levels, ranging from unravelling developmental pathways at the cellular level to comparison of the cellular composition of the immune system at a patient level.

Traditionally, flow cytometry data is being analyzed manually through a set of two-dimensional scatterplot visualizations showing two user-selected proteins at a time. The analyst plots the data according to two proteins of interest as the axes of a scatterplot, then selects a subset of interest of the data, typically by dividing the two axes into low and high expression regions and subsequently visualizes this subset according to two different proteins in a new plot. This process is repeated until all proteins of interest have been inspected. This strategy, known as hierarchical gating, has several problems. Most importantly, covering the complete combinatorial space, even for only 15 proteins is infeasible; therefore, analysts need to focus on specific combinations of interest, which requires prior knowledge, thus implictly biasing the results and limiting the potential for new discoveries.

Here, we present Cytosplore [CSP17], an interactive visual analysis framework for the exploration of mass cytometry data. Cytosplore is designed around dimensionality reduction to facilitate unbiased exploration without prior knowledge. Cytosplore is being developed in close collaboration with immunology researchers at Leiden University Medical Center (LUMC). Together, we designed and implemented interactive, data-driven workflows for different analysis tasks, alongside improved and new progressive analytics methods to facilitate these workflows. Cytosplore is used throughout LUMC in a wide array of clinical research projects and the public version has been downloaded over 1,500 times since its release in November 2017.

DOI: 10.2312/egm.20191032



Settings

HSNE Analysis

Input Data

All Points

Downsampling ...

Active Markers

□ Time	☑ CCR7
□ Event_length	☑ CD14
□ Viability	☑ CD11c
□ Ba138Di	☑ CD161
☑ CD103	☑ CD127_Bead4
□ Bead1	☑ CD8b
☑ CCR6	☑ CD27
☑ CD19	☑ IL-15Ra
☑ C-KIT	☑ CD45RA
☑ CD11b	☑ CD3
☑ CD4	☑ CD28
☑ CD8a	☑ CD38
☑ CD7	☑ NKp46
☑ CD34	☑ PD-1_Bead5
☑ CD25	☑ CD56
☑ TCRab	☑ BCKG190Di
☑ CD123_Bead2	□ DNA
☑ TCRgd	□ DNA
☑ IL-21R_Bead3	□ Pb206Di
☑ CD45	□ sampleTag
☑ CRTH2	□ CSPLR_STATE
☑ CD122	□ CSPLR_SUBSET

Inverse Markers

Load Selection | Set Default

Advanced Parameters

Parameters

Number of Scales 4

Compute HSNE

***Figure 1: Cytosplore Screenshot** with multiple views open; a) CyteGuide, b) HSNE plot, and c) cluster heatmap.*

2. Mass Cytometry

Mass cytometry is a relatively new, *mass* spectrometry-based technique for characterizing protein expression on cells (*cytometry*) at single-cell resolution. In short, antibodies, selected to bind to specific proteins of interest are conjugated with heavy-metal reporters which can be measured in a time-of-flight mass spectrometer to quantify the selected proteins on a per-cell basis. Mass cytometry currently enables the simultaneous analysis of approximately 50 proteins. The resulting data is a table where each row represents a cell and each column one of the measured proteins. The continuous values correspond to the expression of the protein for the given cell. The proteins, i.e., columns are typically interpreted as the axes of a high-dimensional space and accordingly each cell can be interpreted as a high-dimensional data point in the constructed space encoding the different proteins. The cellular composition of the immune system can be described by a hierarchy consisting of a few main compartments, or cell lineages, which divide into more fine-grained subsets. The increased number of proteins measured in mass cytometry, compared to flow cytometry, allows investigating several cell lineages simultaneously, resulting in a broad coverage of the immune system. Analysis techniques can potentially exploit the hierarchical structure of the resulting data.

Typical data sizes greatly vary, depending on the type and goal of the related study. In our collaborations the smallest mass cytometry dataset in terms of the number of cells after preprocessing consists of 220,000 cells, distributed over 7 tissue samples [LvUH*18], while the largest contains 33 million cells over 75 blood samples taken from 25 donors at 3 time points [dJ17, Chapter 7].

The basis for any analysis and main goal for Cytosplore is a reliable cell phenotype identification and classification. In this context, neighborhood-preserving dimensionality-reduction techniques, such as t-SNE [vdMH08] are commonly used. By preserving the high-dimensional neighborhoods, such techniques group cells with similar protein expression (similar types) in a two-dimensional visualization, while maintaining access to single cells. This enables the easy identification and classification of known cell types as well as the discovery of new phenotypes.

3. Cytosplore

Cytosplore (Figure 1) aims at providing interactive exploration for cell phenotype identification and discovery through a combination of different clustering and dimensionality-reduction techniques in combination with a set of linked visualizations for interactive detail inspection. However, many of these techniques are computationally intensive and the large data sizes described in Section 2 pose significant challenges. In fact, most existing tools are not feasible for interactive exploration and often resort to downsampling in some part of the analysis pipeline, posing the risk of information loss.

Through a flexible design, Cytosplore allows the analyst to combine the implemented computational tools as desired and set up a workflow specific to the needs of the goal and size of the given study. For cell phenotype identification such a workflow usually consists of grouping similar cells into clusters followed by inspecting, refining, and labeling those clusters. In Section 3.1 we highlight some examples of different workflows.

3.1. Interactive Cell Phenotype Identification

Cytosplore has been made possible through a number of technical and workflow contributions, described in the following.

A-tSNE. At the center of any of the implemented workflows is a neighborhood embedding. The goal of the embedding is to provide a visualization of the data that preserves local structure (i.e., clusters) of the high-dimensional space. Such clusters in the high-dimensional space can be assumed to represent cells with a similar expression over all proteins, that is, cells of the same type. While t-SNE [vdMH08] is a widely used neighborhood-preserving dimensionality reduction technique, it is computationally extremely demanding and would severely limit interactivity. Therefore, we developed A-tSNE [PLvdM*17] a variant of t-SNE following the progressive visual analytics paradigm [TPB*18]. A-tSNE approximates the neighborhoods in the high-dimensional space, reducing preprocessing time by up to two orders of magnitude. The embedding is then iteratively optimized allowing us to visualize the process and interact during the optimization. We visualize the embedding in a two-dimensional scatterplot, where the points that are close together indicate phenotypically similar cells.

Embedding Clustering. While the resulting embedding allows the analyst to visually identify groups of similar cells, a subsequent step is necessary to define and label these groups. Early works proposed manual selection [ADT*13], similar to the traditional gating, we propose to use automated clustering. Therefore, we implemented a GPU-based version of the Gaussian mean-shift algorithm [HPvU*16]. Mean-shift clustering effectively maps visual clusters to logical clusters, including non-linear separations between clusters. With our GPU-based implementation it can be applied in real-time, even for hundreds of thousands of data points. The resulting clusters can then be inspected, adjusted, and labeled in a linked heatmap visualization (Figure 1c).

Hierarchical Exploration. Even though A-tSNE is much more scalable than the original t-SNE, it is still not feasible to embed millions of cells in one go. For such cases we have two options, both exploiting the hierarchical nature of the data. In early work [vULM*16, HPvU*16], we have used SPADE clustering [QSB*11] to partition the data into the main lineages. SPADE is a relatively fast but imprecise clustering technique making it suitable for roughly partitioning the data. Then, for each of the lineages, the analyst can create an A-tSNE embedding and define the cell types within the lineage as described above. While this combination provides some relief in terms of computational complexity it still provides only limited scalability. In fact, for the original study [vULM*16] it was necessary to downsample the main lineages and were only able to classify approximately one million of the original 5.2 million cells. Furthermore, researchers need to learn and understand multiple tools and algorithms, their parameters, and interpretation of the resulting visualizations.

HSNE. To further improve scalability and usability, we developed and integrated Hierarchical Stochastic Neighbor Embedding (HSNE) [PHL*16]. As the name suggests, HSNE is a hierarchical variation of t-SNE. First, HSNE builds a hierarchy on the data. Therefore, a neighborhood graph is constructed on the data which is evaluated to find representative data points. These so-called *landmarks* are pushed to the next level of the hierarchy, where a new neighborhood graph is constructed. To maintain the non-linear structure of the data even on the coarser hierarchy levels, the similarity in the new neighborhood graph is based on the neighborhood graph of the previous level. The hierarchy is then explored top-down through a set of similarity embeddings (Figure 1b). In Cytosplore the analyst can cluster these embeddings using mean-shift clustering as described above and then select one or multiple clusters to request more detail in a new embedding using the data from the next, more detailed hierarchy level. We have shown [vUHP*17] that we were not only able to reproduce a previous study, using the SPADE/A-tSNE workflow in a fraction of the time but were able to identify several cell types, some of them previously unreported, that were missed in that original study due to downsampling. In fact, in later work, our collaborators were able to confirm some of these cell types through developmental pathway analysis in Cytosplore, followed by confirmation in the wet lab [LvUH*18].

CyteGuide. The exploration of HSNE hierarchies with multiple levels often requires tens of plots. Keeping track of origins and connections between, and navigating those plots imposes a high cognitive load on the analyst. Therefore, we designed and integrated CyteGuide [HPvU*18], to guide the analyst through the exploration, and provide a complete overview of the current state of the analysis. CyteGuide is a meta-visualization collecting and arranging all plots according to their origin in the hierarchy (Figure 1a), as they are created during the exploration. Furthermore, it augments the embeddings with information such as protein expression variation per cluster to guide the exploration.

3.2. Biomedical Discoveries Facilitated by Cytosplore

Here, we summarize results, including several significant discoveries and accompanying publications in high-impact domain journals that have been made possible by Cytosplore.

We started Cytosplore development in collaboration with a small group of domain experts from LUMC (co-authors of this paper) in 2015 and gradually extended collaborations throughout LUMC. We designed and implemented the hierarchical workflow using SPADE and t-SNE for a broad study on the immune system in relation to different gastrointestinal diseases [vULM*16] which we later reanalyzed with an HSNE-based workflow [vUHP*17]. The original study revealed unprecedented heterogeneity in the mucosal immune system and led to the identification of disease-specific immune subsets. Furthermore, using HSNE we were able to identify several subsets that were previously missed, due to the necessary downsampling. Finally, the scalability of HSNE made a large-scale clinical study on immune cell infiltrates in inflammatory bowel disease possible [vU18, Chapter 5]. In two studies on the human fetal intestine, Li, van Unen et al. [LvUH*18] first identified novel innate lymphoid cell types and their differentiation pathways, through a pathway analysis enabled by the progressive nature of our A-tSNE algorithm, allowing the visualization of intermediate results. In another study they show that memory CD4+ T cells are generated in the human fetal intestine [LvUA*19], suggesting that the immune system before birth is far more mature than previously thought. Redeker et al. [RRvdG*18] used an A-tSNE-based analysis in Cytosplore to elucidate the effect of a chronic viral infection over time, revealing the importance of the infectious dose to the extent of immune senescence. Santegoets et al. [SvHE*19] followed a similar approach to associate the Cytosplore-identified immune compositions with the anatomical location of tumors to show their impact on survival rates of cancer patients. Laban, Suwandi et al. [LSvU*18] used Cytosplore and HSNE to show the heterogeneity of circulating antigen-specific CD8 T cells in relation to type 1 diabetes in a study which would not have been possible without a hierarchical approach. Finally, de Jong et al. [dJ17, Chapter 7] used Cytosplore to investigate natural immunity to malaria. The corresponding dataset consists of 33 million cells.

In collaboration with the Allen Institute for Brain Science we developed Cytosplore Viewer, including CyteGuide, which they have used in a comprehensive study of human and mouse cortex cells acquired through single-nuclei RNA sequencing [HBM*18].

In the studies presented above, Cytosplore was used to explore and annotate the acquired data with the corresponding cell types. However, this is often only the first step in a comprehensive analysis. By following domain-specific standards, we made sure that Cytosplore integrates well into larger workflows. Beyrend et al. [BSH*18] exploited this to implement a workflow in R taking the output of Cytosplore to rapidly prepare plots on the sample composition that can be used in publications.

4. Conclusion

We presented Cytosplore, an interactive visual analysis software for single-cell mass cytometry data. Thanks to several technical contributions such as A-tSNE and HSNE, Cytosplore scales to large datasets and enables progressive analysis. By supporting standard file formats for import and export, Cytosplore can be embedded in larger workflows of the application domain.

Throughout our collaborations with LUMC researchers, Cytosplore has enabled a significant number of biological discoveries, resulting in high-impact publications. Cytosplore has been downloaded more than $1,500$ times by researchers from all over the world, actively using the software. First publications [SHF*18] using the software outside of our direct collaborations start to appear.

By facilitating in-depth profiling of the immune system, we expect that Cytosplore can aid in the development of improved diagnostics and personalized therapeutics in the future.

References

[ADT*13] Amir E.-a. D., Davis K. L., Tadmor M. D., Simonds E. F., Levine J. H., Bendall S. C., Shenfeld D. K., Krishnaswamy S., Nolan G. P., Pe'er D.: viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nature Biotechnology 31* (2013), 545–552. doi:10.1038/nbt.2594. 3

[BSH*18] Beyrend G., Stam K., Höllt T., Ossendorp F., Arens R.: Cytofast: A workflow for visual and quantitative analysis of flow and mass cytometry data to discover immune signatures and correlations. *Computational and Structural Biotechnology Journal 16* (2018), 435 – 442. doi:10.1016/j.csbj.2018.10.004. 3

[CSP17] Cytosplore. https://www.cytosplore.org, 2017. Accessed: January 25th, 2019. 1

[dJ17] de Jong S.: *Immunological differences between urban and rural populations*. PhD thesis, 2017. 2, 3

[HBM*18] Hodge R. D., Bakken T. E., Miller J. A., Smith K. A., Barkan E. R., Graybuck L. T., Close J. L., Long B., Penn O., Yao Z., et al.: Conserved cell types with divergent features between human and mouse cortex. *bioRxiv* (2018). doi:10.1101/384826. 3

[HPvU*16] Höllt T., Pezzotti N., van Unen V., Koning F., Eisemann E., Lelieveldt B. P. F., Vilanova A.: Cytosplore: Interactive immune cell phenotyping for large single-cell datasets. *Computer Graphics Forum (Proceedings of EuroVis) 35*, 3 (2016), 171–180. doi:10.1111/cgf.12893. 3

[HPvU*18] Höllt T., Pezzotti N., van Unen V., Koning F., Lelieveldt B., Vilanova A.: Cyteguide: Visual guidance for hierarchical single-cell analysis. *IEEE Transactions on Visualization and Computer Graphics 24*, 1 (2018), 739 – 748. doi:10.1109/TVCG.2017.2744318. 3

[LSvU*18] Laban S., Suwandi J. S., van Unen V., Pool J., Wesselius J., Höllt T., Pezzotti N., Vilanova A., Lelieveldt B., Roep B. O.: Heterogeneity of circulating cd8 t-cells specific to islet, neo-antigen and virus in patients with type 1 diabetes mellitus. *PLOS one 13*, 8 (2018), e0200818. doi:10.1371/journal.pone.0200818. 3

[LvUA*19] Li N., van Unen V., Abdelaal T., Guo N., Kasatskaya S., Ladell K., McLaren J., Egorov J., Izraelson M., de Sousa Lopes S. C., Höllt T., Britanova O., Eggermont J., de Miranda N., Chudakov D., Price D., Lelieveldt B., Koning F.: Memory CD4+ T cells are generated in the human fetal intestine. *Nature Immunology* (2019). doi:10.1038/s41590-018-0294-9. 3

[LvUH*18] Li N., van Unen V., Höllt T., Thompson A., van Bergen J., Pezzotti N., Eisemann E., Vilanova A., Chuva de Sousa Lopes S. M., Lelieveldt B. P., Koning F.: Mass cytometry reveals innate lymphoid cell differentiation pathways in the human fetal intestine. *Journal of Experimental Medicine 215*, 5 (2018), 1383–1396. doi:10.1084/jem.20171934. 2, 3

[OKB*08] Ornatsky O. I., Kinach R., Bandura D. R., Lou X., Tanner S. D., Baranov V. I., Nitz M., Winnik M. A.: Development of analytical methods for multiplex bio-assay with inductively coupled plasma mass spectrometry. *Journal of Analytical Atomic Spectrometry 23* (2008), 463–469. doi:10.1039/B710510J. 1

[PHL*16] Pezzotti N., Höllt T., Lelieveldt B. P. F., , Eisemann E., Vilanova A.: Hierarchical stochastic neighbor embedding. *Computer Graphics Forum (Proceedings of EuroVis) 35*, 3 (2016), 21–30. doi:10.1111/cgf.12878. 3

[PLvdM*17] Pezzotti N., Lelieveldt B. P. F., van der Maaten L., Höllt T., Eisemann E., Vilanova A.: Approximated and user steerable tSNE for progressive visual analytics. *IEEE Transactions on Visualization and Computer Graphics 23*, 7 (2017), 1739–1752. doi:10.1109/TVCG.2016.2570755. 2

[QSB*11] Qiu P., Simonds E. F., Bendall S. C., Gibbs Jr K. D., Bruggner R. V., Linderman M. D., Sachs K., Nolan G. P., Plevritis S. K.: Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. *Nature Biotechnology 29* (2011), 886–891. doi:10.1038/nbt.1991. 3

[RRvdG*18] Redeker A., Remmerswaal E., van der Gracht E., Welten S., Höllt T., Koning F., Cicin-Sain L., Nikolich-Zugich J., ten Berge I., van Lier R., van Unen V., Arens R.: The contribution of cytomegalovirus infection to immune senescence is set by the infectious dose. *Frontiers in Immunology 8*, 1953 (2018). doi:10.3389/fimmu.2017.01953. 3

[RTL*17] Regev A., Teichmann S. A., Lander E. S., Amit I., Benoist C., Birney E., Bodenmiller B., Campbell P., Carninci P., Clatworthy M., et al.: Science forum: The human cell atlas. *eLife 6* (2017), e27041. doi:10.7554/eLife.27041. 1

[SHF*18] Smolders J., Heutinck K. M., Fransen N. L., Remmerswaal E. B. M., Hombrink P., ten Berge I. J. M., van Lier R. A. W., Huitinga I., Hamann J.: Tissue-resident memory t cells populate the human brain. *Nature Communications 9*, 4593 (2018). doi:10.1038/s41467-018-07053-9. 4

[SvHE*19] Santegoets S. J., van Ham V. J., Ehsan I., Charoentong P., Duurland C. L., van Unen V., Höllt T., van der Velden L.-A., van Egmond S. I., Kortekaas K., de Vos van Steenwijk P. J., van Poelgeest M. I., Welters M. J. P., van der Burg S. H.: The anatomical location shapes the immune infiltrate in tumors of same etiology and impacts survival. *Clinical Cancer Research 25*, 1 (2019), 240 – 252. doi:10.1158/1078-0432.CCR-18-1749. 3

[TPB*18] Turkay C., Pezzotti N., Binnig C., Strobelt H., Hammer B., Keim D. A., Fekete J.-D., Palpanas T., Wang Y., Rusu F.: Progressive data science: Potential and challenges. *arXiv* (2018). arXiv:arXiv:1812.08032v1. 2

[vdMH08] van der Maaten L., Hinton G.: Visualizing data using t-SNE. *Journal of Machine Learning Research 9* (2008), 2579–2605. 2

[vU18] van Unen V.: *Mucosal immunology revisited through mass cytometry:From biology to bioinformatics and back*. PhD thesis, 2018. 3

[vUHP*17] van Unen V., Höllt T., Pezzotti N., Li N., Reinders M. J. T., Eisemann E., Koning F., Vilanova A., Lelieveldt B. P. F.: Visual analysis of mass cytometry data by hierarchical stochastic neighbor embedding reveals rare cell types. *Nature Communications 8*, 1740 (2017). doi:10.1038/s41467-017-01689-9. 3

[vULM*16] van Unen V., Li N., Molendijk I., Temurhan M., Höllt T., van der Meulen-de Jong A. E., Verspaget H. W., Mearin M. L., Mulder C. J., van Bergen J., Lelieveldt B. P. F., Koning F.: Mass cytometry of the human mucosal immune system identifies tissue- and disease-associated immune subsets. *Immunity 44*, 5 (2016), 1227–1239. doi:10.1016/j.immuni.2016.04.014. 3