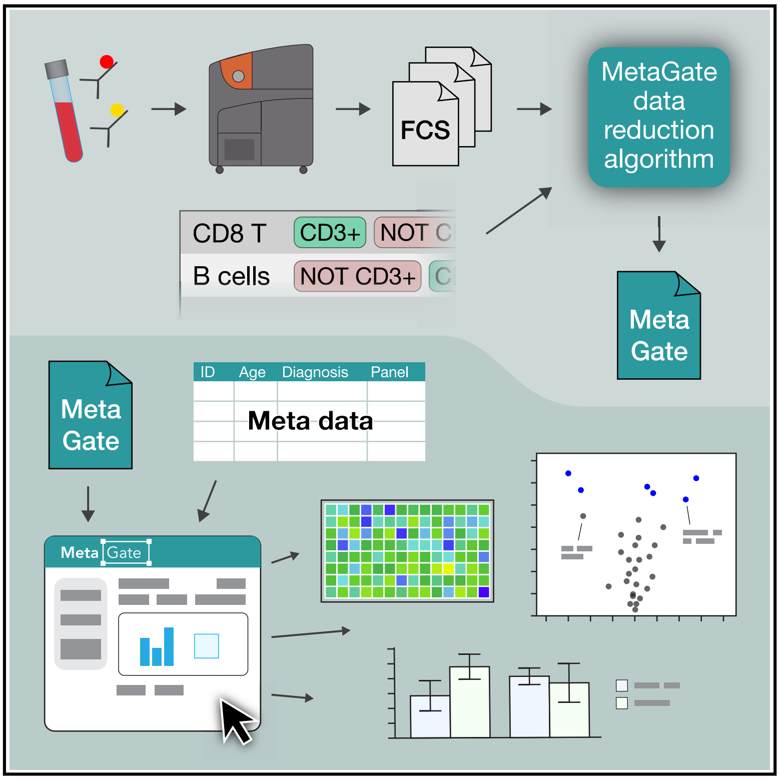
Patterns

Descriptor

MetaGate: Interactive analysis of high-dimensional cytometry data with metadata integration

Graphical abstract



Highlights

- Interactive statistical analysis and visualization of high-dimensional cytometry data
- Compact, standardized data files enhance sharing and comparability
- Integration of metadata for fast analysis through a web-based user interface

Authors

Eivind Heggernes Ask,
Astrid Tschan-Plessl, Hanna Julie Hoel,
Arne Kolstad, Harald Holte,
Karl-Johan Malmberg

Correspondence

k.j.malmberg@medisin.uio.no

In brief

MetaGate enhances the analysis of complex immune cell data generated in advanced flow cytometry platforms. Through a data reduction algorithm, this new bioinformatics tool generates a compact, standardized file for fast and interactive statistical analysis and data visualization. The customized integration of clinical metadata facilitates the identification of important immune cell population changes, contributing to our broader understanding of immune-related diseases. By offering an intuitive interface, MetaGate supports researchers in sharing and analyzing data without specialized knowledge or equipment.

Ask et al., 2024, Patterns 5, 100989
July 12, 2024 © 2024 The Authors. Published by Elsevier Inc.
https://doi.org/10.1016/j.patter.2024.100989

CellPress

Patterns

CellPress
OPEN ACCESS

Descriptor

MetaGate: Interactive analysis of high-dimensional cytometry data with metadata integration

Eivind Heggernes Ask,[1,2] Astrid Tschan-Plessl,[1,3] Hanna Julie Hoel,[1] Arne Kolstad,[4] Harald Holte,[5,6] and Karl-Johan Malmberg[1,2,7,8,*]

[1]Department of Cancer Immunology, Institute for Cancer Research, Oslo University Hospital, Oslo, Norway
[2]The Precision Immunotherapy Alliance, University of Oslo, Oslo, Norway
[3]Division of Hematology, University Hospital Basel, Basel, Switzerland
[4]Department of Oncology, Innlandet Hospital Trust Division Gjøvik, Lillehammer, Norway
[5]Department of Oncology, Oslo University Hospital, Oslo, Norway
[6]K.G. Jebsen Centre for B Cell Malignancies, Institute of Clinical Medicine, Faculty of Medicine, University of Oslo, Oslo, Norway
[7]Center for Infectious Medicine, Department of Medicine Huddinge, Karolinska Institutet, Stockholm, Sweden
[8]Lead contact
*Correspondence: k.j.malmberg@medisin.uio.no

https://doi.org/10.1016/j.patter.2024.100989

THE BIGGER PICTURE The ability to characterize, denominate, and group cells is crucial to our understanding of biology. Flow cytometry allows thousands of cells to be assessed simultaneously by detecting the presence of multiple specified molecules in each cell. This technology is in widespread use in basic research, routine diagnostics, and clinical studies. Technological advances over several decades have increased the number of molecules that can be assessed at the same time, allowing cells to be divided into smaller subclasses with highly diversified phenotypes. While greatly enhancing resolution, these advances also make bioinformatical analysis increasingly complex. This is particularly evident in the context of clinical studies, where there are often large numbers of samples and copious clinical information. Better bioinformatical analysis strategies could help the implementation of advanced cytometry technologies for biomarker discovery and treatment evaluation.

SUMMARY

Flow cytometry is a powerful technology for high-throughput protein quantification at the single-cell level. Technical advances have substantially increased data complexity, but novel bioinformatical tools often show limitations in statistical testing, data sharing, cross-experiment comparability, or clinical data integration. We developed MetaGate as a platform for interactive statistical analysis and visualization of manually gated high-dimensional cytometry data with integration of metadata. MetaGate provides a data reduction algorithm based on a combinatorial gating system that produces a small, portable, and standardized data file. This is subsequently used to produce figures and statistical analyses through a fast web-based user interface. We demonstrate the utility of MetaGate through a comprehensive mass cytometry analysis of peripheral blood immune cells from 28 patients with diffuse large B cell lymphoma along with 17 healthy controls. Through MetaGate analysis, our study identifies key immune cell population changes associated with disease progression.

INTRODUCTION

Fluorescence-based flow cytometry was invented in the late 1960s and has since gained widespread popularity in basic research, routine diagnostics, and clinical trials. Modern flow cytometers allow simultaneous quantification of more than 40 antigens with single-cell resolution, and the introduction of mass cytometry has further increased this number.[1] This has enabled detailed functional and phenotypic characterization of very complex subsets of cells within highly heterogeneous sample material, such as peripheral blood or tumor tissue.

In response to the massive advances in cytometry technology, a vast collection of clustering and dimensionality reduction algorithms has been implemented for data analysis and visualization, including t-distributed stochastic neighbor embedding (t-SNE), PhenoGraph, spanning-tree progression analysis of density-normalized events (SPADE), and FlowSOM.[2–6] Although representing major advances in our ability to explore and understand high-dimensional single-cell data, the output of these algorithms can be unpredictable due to experiment-specific marker selection, technical variation, or inherent properties of different clustering methods.[7] Therefore, cytometry data analysis is still usually carried out by manually defining biologically relevant cell populations by setting cutoff values for multiple antigen markers. This strategy, termed manual gating, allows consideration of known biology, internal controls, and experiment-specific technical issues in the data analysis. However, stratification of samples, statistical analysis, and visualization of summarized data typically involve multiple data-handling steps in different software packages, potentially reducing throughput and data traceability. To alleviate these problems, we developed the MetaGate R package. Through its graphical user interface, MetaGate provides a platform for statistical analysis and visualization of complex cytometry datasets from raw data via feature selection to publication-ready figures, based on manual gating performed in two of the most popular flow cytometry analysis software packages: FlowJo and Cytobank.

Along with genomics, proteomics, and immunological imaging techniques, cytometry remains a crucial tool for assessing the immune system in cancer, both within the tumor microenvironment and at the global level. Such understanding is important for cancer prevention, diagnostics, prognostics, and development of novel treatment strategies. To display the capabilities of MetaGate, we performed a broad mass cytometry characterization of peripheral blood from a cohort of 28 patients with diffuse large B cell lymphoma (DLBCL) alongside 17 healthy blood donors.

DLBCL is the most common group of non-Hodgkin's lymphoma, with an incidence in the United States of around 7 cases per 100,000 persons per year.[8] First-line treatment usually includes multi-agent chemotherapy in combination with the anti-CD20 monoclonal antibody rituximab. Two main subtypes, germinal center B cell (GCB) and activated B cell (ABC), have been identified, correlating fairly well with histological features and explaining some of the outcome variation.[9] However, the highly diverse presentation and outcomes, which cannot be fully explained by existing clinical, histological, or biochemical markers, remains a major clinical challenge.[10] Therefore, to improve diagnostics, prognostics, and treatment of this disease, there is a need for a better understanding of the heterogeneity of its presentation and immunological responses.

The mass cytometry data from this study, which, in part, has been published previously,[11] are analyzed using MetaGate and describes a substantial impact on the immune system from both the disease and its treatment. All data figures and statistical analyses are generated in the MetaGate user interface. The MetaGate R package and source code are made publicly available, along with all mass cytometry data and metadata, enabling anyone to reproduce the analysis as well as further develop or use MetaGate for other datasets.

RESULTS

Generating a MetaGate dataset

MetaGate is based on manual gating, which can be performed in either the FlowJo or Cytobank software packages. Blood samples or other cell suspensions are analyzed using a mass or flow cytometer (Figure 1A), which generates Flow Cytometry Standard (FCS) files. These are imported in FlowJo or Cytobank. After quality control, exclusion of unwanted events and adjustment of compensation, biologically relevant gates are set. The gate definitions are then exported as a FlowJo Workspace file or GatingML file from FlowJo or Cytobank.

The FlowJo or GatingML file is then imported into MetaGate, which parses the file and produces a list of defined gates (Figure 1B). In the MetaGate graphical user interface, the user can then define populations by combining the gates; e.g., defining "CD8 T cells" as events inside the "CD3+" and "CD8+" gate but outside the "CD19+" gate. The MetaGate data reduction algorithm is then applied, using the definitions of gates and populations along with raw data from FCS files to calculate mean, median, and geometric intensity values and frequencies of all populations in each population. Given P populations and M markers, the algorithm will output (3 * M + P) * p values for each sample. Assuming 100,000 events, 40 markers, 100 populations, and 4 bytes per value, MetaGate will generate 86 kB of data from a 15-MB FCS file. These data are then stored as a data file that is used for all subsequent data analyses (Figure 1C).

Data analysis in MetaGate

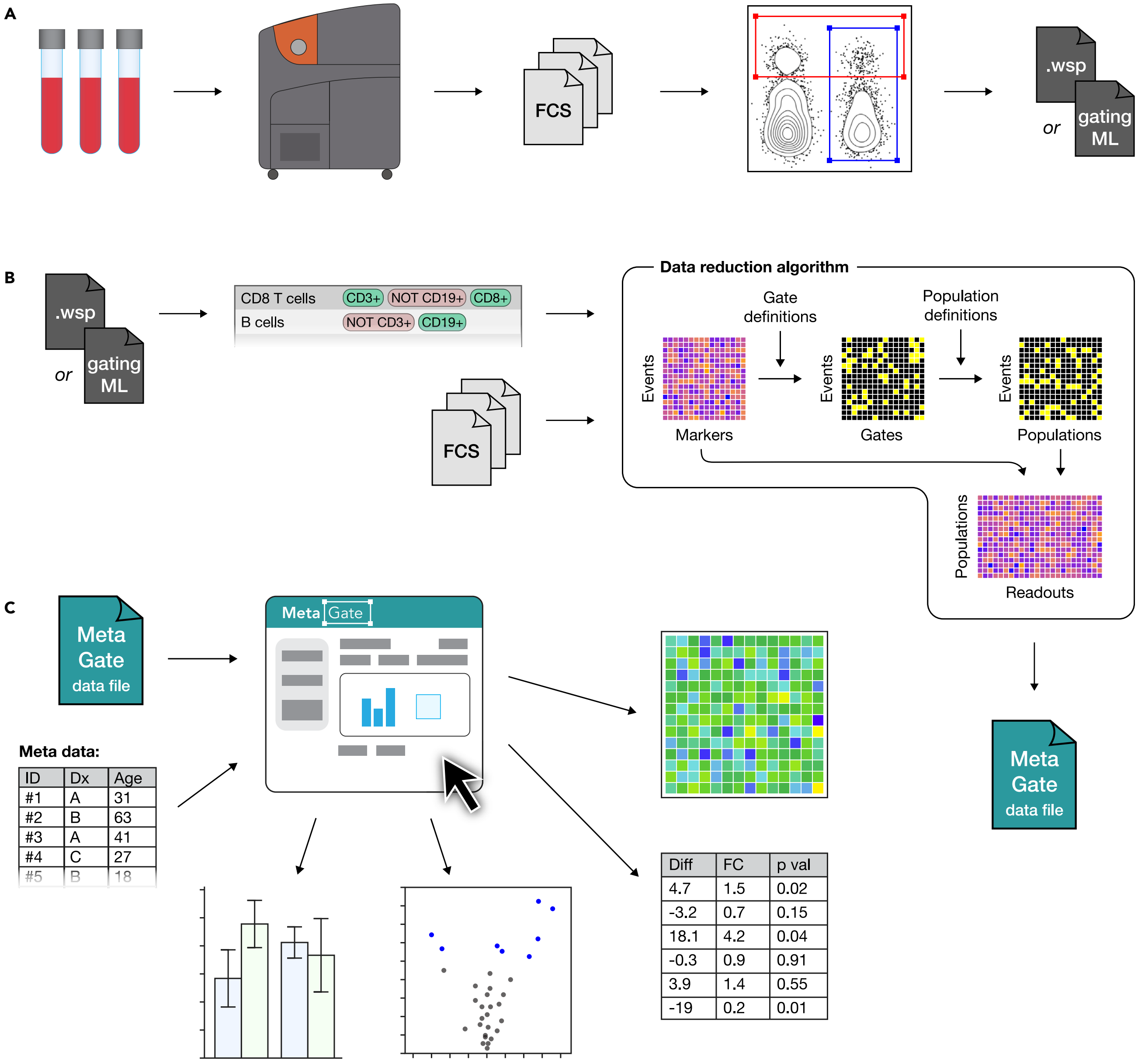
After loading the MetaGate data file in the MetaGate graphical user interface, the user can upload sample metadata, such as clinical features, experimental conditions, or sample time points (Figure 1C). Sample groups are then defined interactively by selecting features based on the metadata.

The metadata should include information about which panel is used for each sample. By setting this as a panel variable, MetaGate will automatically make sure that the same individual is not included twice in a comparison in cases where both panels would provide the same data. In projects that contain paired samples, such as multiple perturbations or time points, a variable should be included that uniquely identifies each patient or healthy donor. MetaGate will then use this variable to perform paired statistical analyses. All metadata and group definitions are stored in the MetaGate file but can be modified at any time in downstream analysis.

To demonstrate the main features of MetaGate, a previously partially reported dataset of immune cell characterization in DLBCL was analyzed. Peripheral blood mononuclear cells (PBMCs) from a total of 28 DLBCL patients and 17 age- and sex-matched healthy controls (Table 1) were investigated using two mass cytometry panels (Figure 2; Tables S1 and S2). To evaluate the effect of therapy, patients were sampled both at the time of diagnosis and after treatment with rituximab and chemotherapy. For each of the two panels separately, gating was performed in FlowJo. The two resulting MetaGate data files were then merged. All plots and statistical calculations in Figures 3, 4, and 5 and the accompanying supplementary tables were produced in MetaGate.

Large impact of DLBCL on peripheral blood immune cell phenotypes

MetaGate allows creation of three main types of heatmaps. Using the first type, which shows marker expression for multiple



ID	Dx	Age
#1	A	31
#2	B	63
#3	A	41
#4	C	27
#5	B	18

Diff	FC	p val
4.7	1.5	0.02
-3.2	0.7	0.15
18.1	4.2	0.04
-0.3	0.9	0.91
3.9	1.4	0.55
-19	0.2	0.01

Figure 1. MetaGate data analysis workflow

(A) A biological sample, such as patient blood, is analyzed using a mass or flow cytometer, which produces FCS data files. Manual gating is performed in FlowJo or Cytobank, creating a data file with specifications of each gate.

(B) Gate data and FCS files are imported into MetaGate, where a graphical user interface allows defining populations based on combinations of gates. Through a data reduction algorithm, a MetaGate data file is created, which contains marker expression and event frequencies of combinations of populations.

(C) The self-containing MetaGate data file is opened in the MetaGate graphical user interface for interactive production of statistics and plots, such as heatmaps, volcano plots, and bar plots.

populations in one group, the defining expression patterns of the key included populations can be visualized (Figure 3A).

Volcano plots are useful for quickly identifying main differences between two groups, as they provide a graphical representation of both statistical significance and magnitude of difference for multiple readouts in the same plot. In MetaGate, volcano plots can be generated based on data from multiple panels and explored interactively by holding the cursor over each dot. Using a volcano plot to compare sizes of major cell subsets between healthy donors and DLBCL patient samples before therapy reveals multiple large differences (Figure 3B; Table S5). Most significantly, HLA-DR$^-$ CD14$^+$ CD19$^-$ CD3$^-$ CD56$^-$ cells, indicative of monocytic myeloid-derived suppressor cells,[12] are greatly expanded in patients (Figure 3C). Inversely, the T cell fraction of all CD45$^+$ is lower in patients, but T cells also constitute a smaller fraction of lymphocytes (Figure 3D). As mass cytometry, in contrast to flow cytometry, does not allow distinction of lymphocytes by morphology, the

Table 1. Patients and healthy controls

	Healthy controls	Patients
Number of individuals	17	28
Female	9 (53%)	12 (43%)
Median age	67	65
Subtype		
GCB DLBCL	–	13 (46.4%)
Non-GCB DLBCL	–	11 (39.3%)
Other	–	4 (14.3%)
Stage		
Stage I	–	3 (10.7%)
Stage II	–	7 (25%)
Stage III	–	3 (10.7%)
Stage IV	–	15 (53.6%)

lymphocyte population is here defined as the sum of T, B, and natural killer (NK) cells. In patients, the CD56^{bright} cells constitute a smaller part of the NK cell compartment relative to the more mature CD56^{dim} cells (Figure 3E).

The second main type of heatmaps that MetaGate can produce enables two-group comparisons of multiple markers in multiple populations (Figure 3F). Markers can represent both marker intensities and percentages of positive cells, and data from multiple panels can be displayed in the same plot. Using colors for displaying the *p* values from multiple non-parametric tests and the direction of change, these plots give a fast overview of potentially significant findings. MetaGate furthermore produces a complete table of all statistics and allows this to be exported as a Microsoft Excel file. Most strikingly, T cells of DLBCL patients display higher levels of CD38, Ki-67, PD-1, and TIM-3 (Figure 3G).

Immune cell subset dynamics through the course of treatment

In addition to slightly varying chemotherapy regimens, the anti-CD20 antibody rituximab was given to all patients. As expected, peripheral blood B cells were virtually non-detectable in post-treatment samples, while B cell numbers before treatment did not differ significantly from those of healthy controls (Figure 4A). As illustrated here, MetaGate automatically selects appropriate statistical tests based on the number of groups compared.

The observed B cell depletion highlights the importance of assessing absolute cell counts, in contrast to the relative subset sizes usually provided by cytometry assays. If absolute counts of a population are available, then MetaGate automatically calculates absolute counts of all subpopulations. By linking clinical lymphocyte counts to the lymphocyte population in MetaGate, absolute counts of key T, B, and NK cell subsets could be assessed. Most significantly, patients displayed larger numbers of the CD56^{bright} NK cells after therapy, while several subsets of the more mature CD56^{dim} NK cells decreased in size

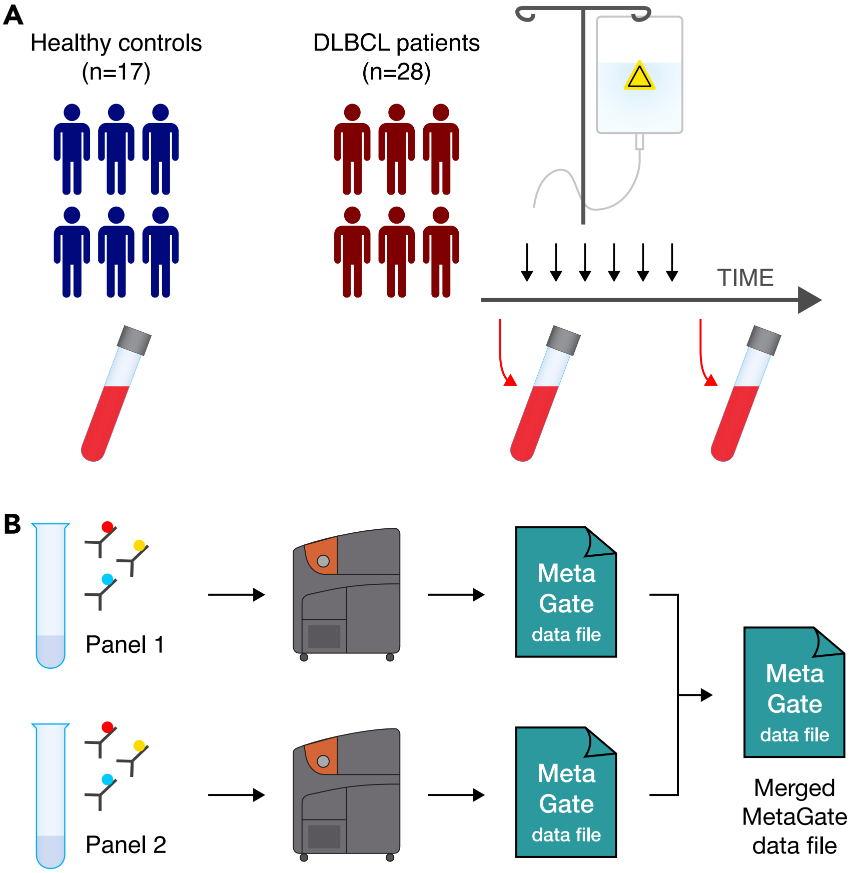


Figure 2. DLBCL immune characterization workflow

(A) Peripheral blood was collected from healthy blood donors (*n* = 17) and from patients diagnosed with DLBCL (*n* = 28) before and after chemotherapy.
(B) Blood samples were split and analyzed using two mass cytometry panels. Data from each panel were imported separately in MetaGate and later merged.

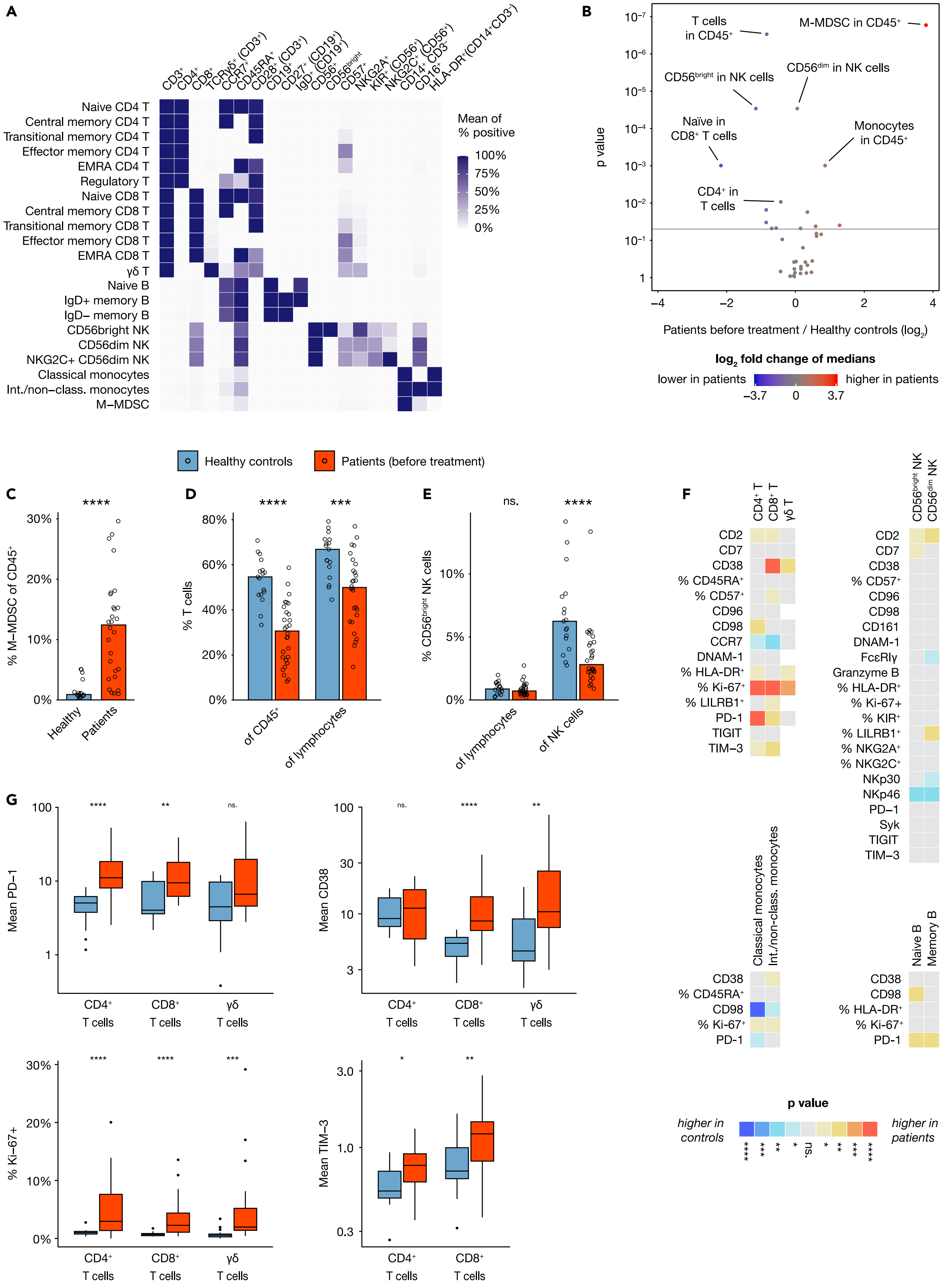


Figure 3. Peripheral blood immune cell composition in DLBCL

(A) Heatmap showing expression of key markers in subsets of analyzed cell types, visualizing how subsets were defined for downstream analysis.

(B) Volcano plot showing size differences of 36 key immune cell types between healthy donors and all patients before chemotherapy.

(legend continued on next page)

(Figures 4B–4D). The NK cell subset dynamics can be further investigated by utilizing the third type of heatmap available in MetaGate, which allows visualization of multiple readouts across more than two groups (Figure 4E). In addition to the expansion of the CD56^{bright} NK cells, the CD56^{dim} compartment displays a shift toward less mature cells with more NKG2A-expressing and fewer CD57-expressing cells. Looking at changes in marker expression after therapy, this is corroborated by the observed increase in NKp30 and NKp46 expression (Figure 4F). Furthermore, a clear increase in CD38 expression is observed in NK cells, consistent across all major subsets (Figure 4G).

Prediction of disease outcome

Using provided metadata, MetaGate allows simple and dynamic creation of sample groups for visualization and statistical testing. Looking at absolute cell counts of key lymphocyte populations in patient samples taken at the time of diagnosis, no clear differences were seen based on major age and subtype groups (Figures 5A and 5B). However, advanced disease (Ann Arbor stage III or IV) was somewhat associated with lower numbers of CD4^{+} T cells and CD56^{bright} NK cells (Figures 5C–5E). Only five patients experienced disease progression during the follow-up time. Still, this group showed an association with lower absolute counts of CD56^{dim} NK cells and higher numbers of immunoglobulin D (IgD)$^{-}$ memory B cells (Figures 5F–5H).

DISCUSSION

The continuously increasing complexity of cytometry data warrants new strategies for data analysis. We developed MetaGate, allowing interactive and fast statistical analysis and visualization of complex cytometry datasets. In this paper, we visualize the novel features of MetaGate through the analysis of a previously partly published broad multi-panel mass cytometry characterization of peripheral blood immune cells in a cohort of 28 DLBCL patients.

All plots and statistical analyses throughout this paper were generated in MetaGate, illustrating many of the most important features of the software package. Modern cytometry datasets often contain large numbers of readouts for comparison, and assessing all of them manually can be very laborious, especially when there is a need to stratify the data on multiple clinical variables. Volcano plots, which are routinely used in genomics and proteomics, allow both statistical significance and the magnitude of change to be displayed in one graphical representation, which in MetaGate can be explored interactively. Conversely, heatmaps allow more than two groups to be compared or multiple readouts to be assessed in multiple populations. Importantly, when comparing two groups, MetaGate heatmaps can also display statistical significance and direction of change, which can be particularly useful when assessing marker expression across multiple cell subsets. Such large-scale statistical testing introduces a considerable risk of type I errors. While MetaGate offers several p value correction techniques that can partly alleviate this problem, the use of p values in heatmaps and volcano plots in MetaGate should primarily be considered as a data exploration method, useful for highlighting potential findings of interest. Such findings can then be further explored using bar plots, which also allow multi-group comparisons and visualization of other metadata. In all plots, MetaGate automatically selects appropriate non-parametric statistical tests.

In cytometry experiments with clear groups of samples (for example, perturbation and controls), cytometry data can be managed relatively easily manually for statistics and visualization. However, studies involving clinical data often include multiple variables of metadata, such as age, sex, diagnosis, sampling time point, and treatment response. In this case, appropriate sample groups and comparisons may be numerous and not necessarily obvious early in the data analysis workflow. This can make manual data handling laborious and prone to errors. MetaGate seeks to alleviate this by mapping metadata from separate data files to samples and allowing groups to be created through a point-and-click query system in which the user selects features from the imported metadata. As both metadata and group definitions can be modified at any time, data exploration becomes simple and efficient. All data analyses in MetaGate are based on manual gating of the data, meaning that cell types are defined by manually setting presumed biologically relevant cutoffs for marker expression in several one- or two-dimensional data plots. Manual gating allows knowledge about biology and technical aspects of specific experiments, sample batches, or individual batches to be considered in the bioinformatical analysis. Furthermore, gating strategies can easily be standardized across experiments or studies. However, this strategy also has multiple drawbacks. The reliance on visual inspection of data by a trained professional introduces potential operator bias and confirmation bias.[13,14] Furthermore, with the increasing complexity of cytometry data, manual gating represents a laborious analysis strategy.[1,15,16] The majority of novel analysis algorithms created to handle this complexity is based on unsupervised clustering.[2] This includes popular tools like t-SNE, SPADE, Phenograph, and FlowSOM.[3–6] These prove particularly useful for exploring novel or complex cell subsets but may not produce results that are easily compared between different studies or experimental batches.[7] DeepCyTOF and flowLearn are examples of algorithms that address these obstacles by automating the manual gating procedure through machine learning, thereby attempting to reduce the laboriousness of manual gating while preserving most of its benefits.[17,18] While MetaGate relies on gating of cells, there is no intrinsic requirement for these gates to be created manually by humans. Therefore, MetaGate can be further developed to allow

(C–E) Bar plots showing median percentages of (C) M-MDSCs (defined as HLA-DR^{-} CD14^{+} CD19^{-} CD3^{-} CD56^{-} cells), (D) T cells, and (E) CD56^{bright} NK cells within various parent populations in healthy controls (n = 17) and all patients before therapy (n = 28).

(F) Heatmap showing differences in marker expression between healthy controls (n = 17) and patients before therapy (n = 21–28) within multiple immune cell subsets, with colors indicating direction of difference and statistical significance from nonparametric tests without p value adjustment. Values are mean intensity values unless otherwise indicated.

(G) Boxplots showing selected readouts from (F).

All p values are calculated using the Mann–Whitney U test.

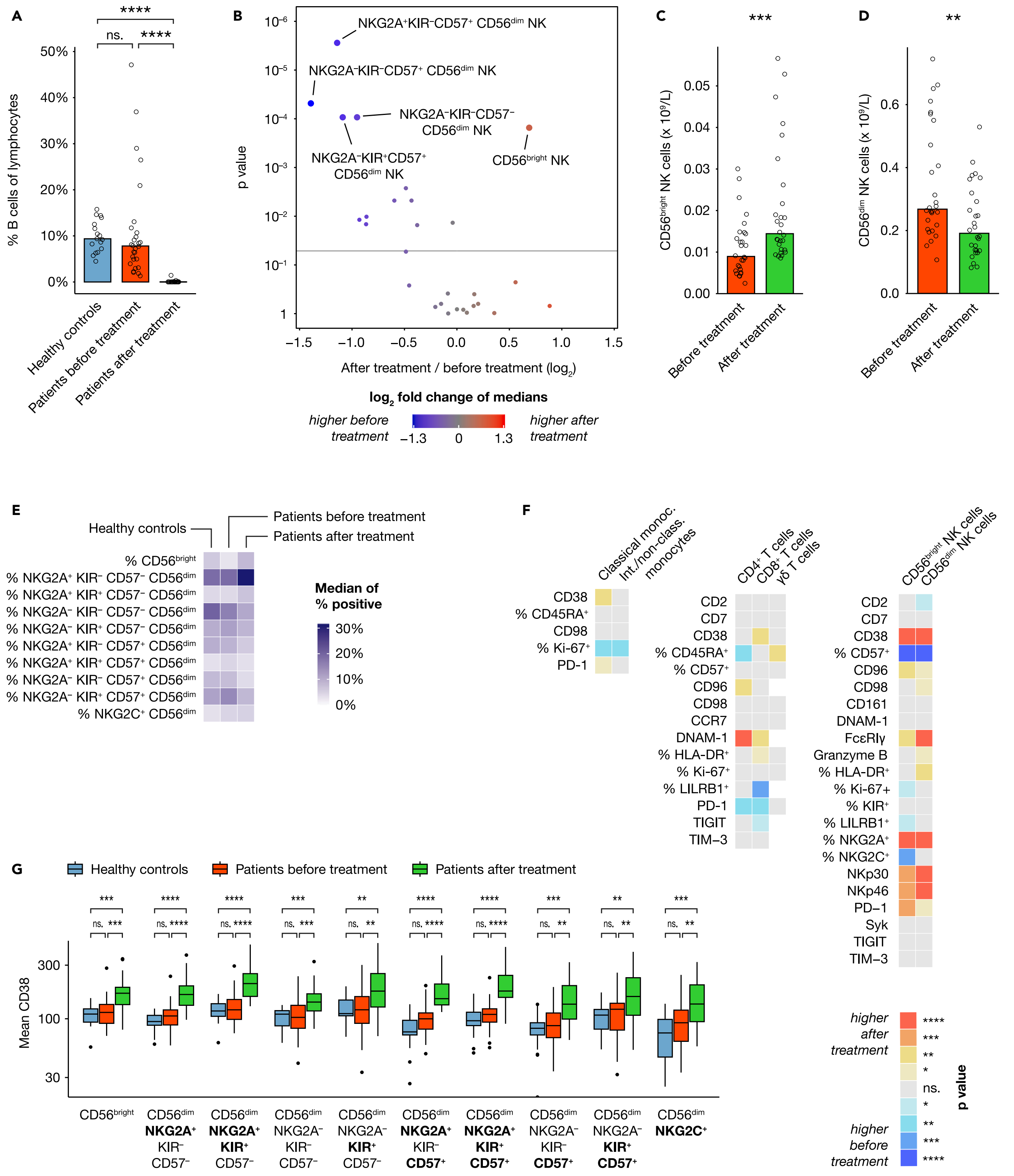


Figure 4. Effects of treatment on immune cell phenotypes

(A) B cell frequencies as percentage of all CD45$^+$ in healthy controls (n = 17) and all patients (n = 28) before and after treatment. Bar height represents median.

(B) Volcano plot showing differences in absolute counts of 28 cell subsets before and after treatment (n = 28).

(C and D) Selected comparisons from (B). Bar height represents median.

(legend continued on next page)

(semi-)automatic gating by any of these algorithms upstream of the interactive statistical analysis in MetaGate.

The MetaGate data reduction algorithm works by calculating mean intensity values and sizes of all defined populations for each sample, producing a very condensed dataset that can be used for downstream analysis without access to the raw data. Consequently, MetaGate can only generate plots and statistics based on predefined populations, limiting its usefulness for exploration of novel cell subsets. However, there are multiple benefits to this strategy. Because cytometry data consist of single-cell measurements of multiple parameters, datasets are typically large. A theoretical set of 100 files with one million events and 40 parameters in each would create around 15 gigabytes of data, which exceeds the available memory of most common workstations. Furthermore, the computational expensiveness of gating is increasing with the number of events and parameters. By performing all of the memory- and processor-consuming tasks in the MetaGate data import procedure, the downstream analysis in MetaGate becomes comparably very fast. Fixing gates, population definitions, and sample selections at one point and making these visible to the user also enhances the traceability of the analysis. The intuitive data processing steps and the small size of the data file simplifies data sharing, making data analysis possible without in-depth experimental knowledge, powerful computers, or access to other specialized software.

MetaGate is fully written in the R programming language, utilizing the shiny[19] package to provide a web browser-based user interface. This strategy allows MetaGate to take advantage of the large selection of available R packages, including the OpenCyto framework,[20] which provides a wide range of functionalities for cytometry data analysis. As a shiny-based application, MetaGate can either run locally on the user's computer or be run on a remote server and accessed through the internet. As internet connection is not required, and all source code is open and without need of compilation, MetaGate can also be used in secure data environments where custom software installation is prohibited, as long as R is available.

BCyto, CYANUS, and CytoPipelineGUI are examples of shiny-based R packages that focus on various aspects of cytometry data analysis.[21–23] MetaGate is mainly distinguished from these by its data reduction algorithm and tight integration with metadata. However, these features also give rise to the main limitations of MetaGate. All gates and populations must be defined during data import, and observations made during statistical analysis and visualization cannot easily be verified in the raw data. Furthermore, the easy integration of metadata allows high-throughput statistical testing, which can give rise to type I statistical errors. In conclusion, the complexity of cytometry data may, in many cases, demand data analysis using multiple tools with different features and limitations.

While demonstrating some of the most important features of MetaGate, the mass cytometry analysis of 28 DLBCL patients and matched controls reveals marked effects on the peripheral blood immune system of DLBCL patients. Although current therapy induces remission in a large majority of DLBCL patients, incomplete remission or relapses are seen in around one-third of the patients, and a better understanding of the immune responses could potentially lead to improved prognostics and treatment customization.[10] Monocytic myeloid-derived suppressor cells (M-MDSCs) are pathologically activated monocytes that have been associated with immunosuppression and poor outcome in multiple cancer settings.[24] Our data show high numbers of M-MDSCs among DLBCL patients, which has been reported previously and linked to immunosuppression,[25,26] potentially explaining why monocytosis was identified as a negative prognostic marker in DLBCL.[27] Furthermore, the increased expression of Ki-67, CD38, PD-1, and TIM-3 on T cells represents a phenotype consistent with exhaustion and potential dysfunctional activation.[28,29]

Apart from the expected near-total depletion of B cells, the most marked effect of chemotherapy on peripheral blood immune cell phenotypes was seen for NK cells. After chemotherapy, NK cells displayed lower expression of the maturation marker CD57, while higher expression was seen for the inhibitory receptor NKG2A and activating receptors NKp30 and NKp46, which is in line with observations of reconstitution of NK cell subsets after hematological stem cell transplantation.[30] The broad upregulation of CD38 expression across all NK cell subsets suggests a systemic immune activation following chemo-immunotherapy, possibly reflecting homeostatic recovery. Corroborating previous DLBCL studies, our data showed a positive correlation between NK cell counts before initiation of therapy and beneficial outcome.[31,32]

In conclusion, we present a new bioinformatical tool for high-throughput statistical analysis and visualization of cytometry data. The features of this software are displayed through the analysis of a mass cytometry characterization of peripheral blood from 28 DLBCL patients and matched controls, highlighting large immunophenotypic effects of both the disease and chemoimmunotherapy treatment, corroborating previously published reports. The initial manual gating of data, data reduction algorithm, and dynamic integration with metadata, simplify feature selection, data sharing, and generation of publication-ready statistics and plots. Published as an open-source R package, MetaGate can be improved, customized, and integrated in existing workflows, potentially allowing researchers to more easily tackle the continuously increasing complexity of cytometry data.

EXPERIMENTAL PROCEDURES

Resource availability

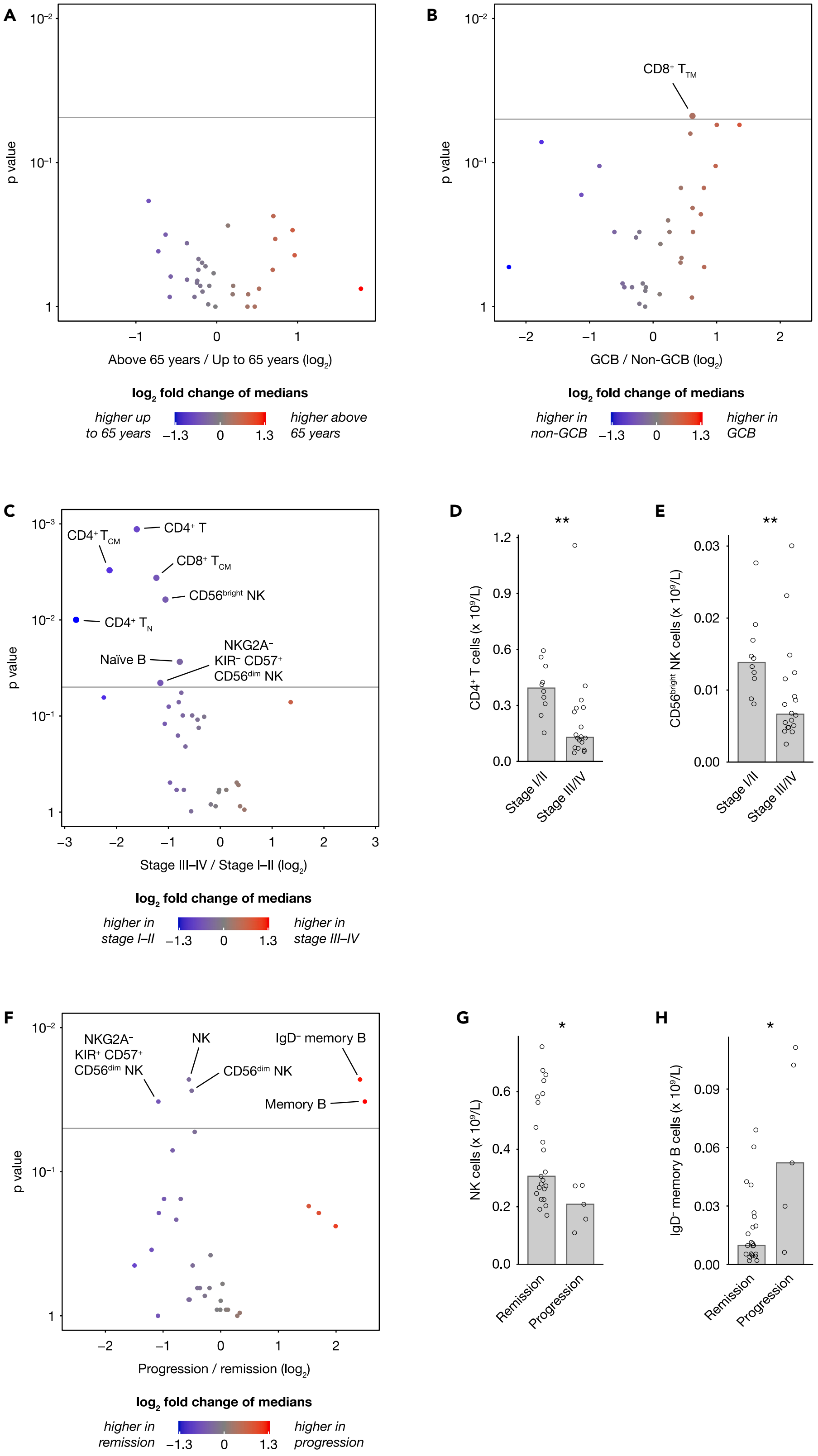
Lead contact

Requests for further information and resources should be directed to and will be fulfilled by the lead contact, Karl-Johan Malmberg (k.j.malmberg@medisin.uio.no).

(E) Heatmap showing median frequencies of key NK cell subsets as percentage of bulk NK cells in healthy controls (*n* = 17) and patients (*n* = 28) before and after therapy.

(F) Heatmap showing differences in marker expression within multiple immune cell subsets between patients before and after treatment (*n* = 20–28), with colors indicating direction of difference and statistical significance from paired nonparametric tests without *p* value adjustment.

(G) Boxplots showing mean CD38 expression in multiple NK cell subsets of healthy controls (*n* = 15–17) and patients (*n* = 25–28) before and after treatment. *p* values are calculated using the Dunn test (A and G) or Wilcoxon signed-rank test (B–D and F).



(legend on next page)

Materials availability

There are restrictions to the availability of the patient material described in this manuscript due to ethics and patient consent. Requests directed to the lead contact will be considered on an individual basis.

Data and code availability

The full MetaGate source code is published and available under an open-source GNU GPLv3 license at https://github.com/malmberglab/metagate and Zenodo (https://doi.org/10.5281/zenodo.10871021).[33] Documentation and installation instructions are available at https://metagate.malmberglab.com. Raw data for the included dataset are available at Figshare (https://doi.org/10.6084/m9.figshare.24542173.v1).[34] The MetaGate file used to generate all statistics and figures can be downloaded from https://metagate.malmberglab.com.

Methods

Development of MetaGate

MetaGate is developed as an R[35] package with a web browser-based graphical user interface implemented using the shiny package.[19] Interaction with FlowJo workspaces, GatingML files, and FCS files is implemented with the use of the flowWorkspace, CytoML, flowCore, and flowUtils packages.[36–39] Plots are generated using the ggplot2 package.[40]

Patient samples and clinical data

The use of patient and healthy donor blood samples and clinical data was approved by the regional ethics board in Norway (refs. 2012/1143, 2015/2142, 2018/2482, and 2018/2485). Patients were selected from a lymphoma patient biobank established in January 2015 at Oslo University Hospital. Fully informed written consent was obtained from all healthy donors and patients. The study included 17 healthy donors and 28 patients. Median age was 65 for healthy donors and 67 for patients, while the percentages of female subjects were 53% and 43%, respectively. PBMCs were collected from patients directly before initiation and after completion of first-line chemotherapy, while healthy donor samples were collected at one time point. Inclusion diagnoses were DLBCL, high-grade B cell lymphoma (HGBCL) with MYC and BCL2 and/or BCL6 rearrangements (or based on the 2008 World Health Organization [WHO] classification of lymphoid neoplasms, "B-cell lymphoma, unclassifiable, with features intermediate between diffuse large B-cell lymphoma and Burkitt lymphoma"), and T cell/histiocyte-rich large B cell lymphoma (THRLBCL). All patients were treated with a combination of rituximab and chemotherapy regimens containing cyclophosphamide, doxorubicin, vincristine, etoposide, and prednisolone (CHOP/EPOCH/CHOEP). The Hans algorithm was used for subtype classification of GCB and non-GCB DLBCL. For patients, absolute numbers of lymphocytes were retrieved from diagnostic white blood cell differential counts, while such data were not available for healthy donors.

Mass cytometry

PBMCs from patients and healthy blood donors were isolated by density gradient centrifugation using Lymphoprep (Axis-Shield, Oslo, Norway). Cells were subsequently aliquoted and cryopreserved in 10% DMSO, 70% fetal calf serum (Sigma-Aldrich, St. Louis, MO, USA) and 20% RPMI 1640 (Thermo Fisher Scientific, Waltham, MA, USA). Upon experiments, PBMCs were thawed and rested overnight in RPMI 1640 with 10% fetal calf serum.

Cells were stained with Cell-ID Intercalator-Rh (Fluidigm, San Francisco, CA, USA) and GLUT1.RBD.GFP (Metafora Biosystems, Evry Cedex, France) according to the manufacturer's instructions to allow for viability testing and GLUT-1 detection, respectively. Samples were then incubated with an Fc receptor binding inhibitor polyclonal antibody (Thermo Fisher Scientific) before staining with a surface antibody cocktail (Table S1 and S2). Antibodies were either obtained pre-labeled from Fluidigm or in house conjugated using Maxpar X8 antibody labeling kits (Fluidigm). After staining, the cells were fixed using 2% paraformaldehyde in PBS without Ca and Mg and then permeabilized and barcoded using the Cell-ID 20-Plex Barcoding Kit (Fluidigm) according to the manufacturer's instructions. Samples were then pooled, resuspended in pure methanol, and stored at −20°C. On the day of mass cytometry acquisition, samples were thawed, stained with an intracellular antibody cocktail, and labeled with Cell-ID Intercalator-Ir (Fluidigm) according to the manufacturer's instructions. Immediately before acquisition, samples were supplemented with EQ Four Element Calibration Beads (Fluidigm) and acquired on a CyTOF 2 (Fluidigm) equipped with a SuperSampler (Victorian Airship, Alamo, CA, USA). The event rate was kept below 400 events per second. Samples were analyzed in 8 batches, with healthy donors and patients distributed evenly across batches and patient samples from different timepoints always included in the same batch. Due to lack of sufficient cell numbers, PBMCs from 3 of the healthy donors were not analyzed using mass cytometry panel 2.

Data preparation

FCS files were normalized using the Fluidigm Helios software and debarcoded either by manual gating or using the Helios software. The files were then imported in Cytobank (Cytobank, Santa Clara, CA, USA), where debris, doublets, and dead cells were excluded. Data were then gated on CD45$^+$ events and exported as FCS files. Files from the two panels were imported into separate FlowJo workspaces and gated according to Figures S1 and S2. In each FlowJo workspace, all samples shared identical gating hierarchies, but gates were adjusted manually for each sample. Each FlowJo workspace was then imported in MetaGate. In MetaGate, populations were defined according to Tables S3 and S4. Channels that were empty or represented intercalators or non-relevant markers were excluded (Tables S1 and S2). Furthermore, the markers GLUT-1, CD71, CD137, and NKG2D were removed due to problematic performance or batch effects. The event limit was kept at 50, meaning that populations with fewer than 50 events were excluded from calculation of marker intensities or child population sizes. No data transformation was applied in MetaGate. Gating strategy plots were generated using the CytoML and ggcyto R packages.

Statistical analysis

All statistical plots and statistical analyses were generated in MetaGate v.1.0 on macOS 13.1 running R v.4.2.2. Minor typographical changes and insertion of p value annotation were subsequently performed in Adobe Illustrator v.27.2. The Mann-Whitney U test was used for unpaired comparison of two groups (Figures 3B–3G and 5A–5F). Paired two-group comparisons were tested using the Wilcoxon signed-rank test (Figures 4B–4D and 4F). Comparison of multiple groups was done using the Kruskal-Wallis H test and, in the case of p values $\leq$ 0.05, subsequent pairwise group comparisons using the Dunn test (Figures 4A and 4G). Adjustment of p values was not performed.

p values above 0.05 were defined as not significant (ns.), while *, **, *** and **** were used to indicate p values below or equal to 0.05, 0.01, 0.001, and 0.0001, respectively. Bar plot height represents the median. In boxplots, hinges correspond to the 25th and 75th percentiles, while whiskers range to the most extreme values but no longer than 1.5 times the interquartile range, and data points outside that range were plotted individually.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2024.100989.

ACKNOWLEDGMENTS

The project was supported by The Research Council of Norway (projects 275469 and 237579), Center of Excellence: Precision Immunotherapy Alliance 332727, the Norwegian Cancer Society (projects 190386 and 223310), the South-Eastern Norway Regional Health Authority (2021-073), the EU H2020-

Figure 5. Immune cell repertoires stratified on patient characteristics

(A–C and F) Volcano plots showing differences in 33 absolute cell counts in peripheral blood of patients before therapy, stratified on (A) age above 65 (n = 15) or below or equal to 65 (n = 13), (B) GCB (n = 13) or non-GCB (n = 11) subtype, (C) stage I/II (n = 10) or III/IV (n = 18), and (F) disease progression (n = 5) or remission (n = 23) within the follow-up time.

(D, E, G, and H) Selected readouts from (C) and (F). Bar height represents median.

All p values are calculated using the Mann–Whitney U test.

MSCA Research and Innovation Program (project 801133), the Knut and Alice Wallenberg Foundation, the Swedish Foundation for Strategic Research, and US National Cancer Institute (P01 CA111412, subaward P009500901). This work was further supported by grants from the Swedish Research Council (223310), the Swedish Children's Cancer Society (PR2020-1059), the Swedish Cancer Society (21-1793Pj), Sweden's Innovation Agency, the Karolinska Institutet, and the Swiss Cancer League (to A.T.-P., grant BIL KFS-3745-08-2015). We are grateful for the support from the Flow Cytometry Core Facility at Oslo University Hospital, Radiumhospitalet.

AUTHOR CONTRIBUTIONS

A.T.-P., H.J.H., and E.H.A. conducted experiments. E.H.A. conceptualized software with input from A.T.-P. and K.-J.M. E.H.A. wrote the code. A.K. and H.H. provided clinical samples and clinical data. E.H.A. wrote the manuscript. All authors edited the manuscript.

DECLARATION OF INTERESTS

K-J.M. is a consultant at Fate Therapeutics and Vycellix and has research support from Oncopeptides for studies unrelated to this work.

Received: October 30, 2023
Revised: March 12, 2024
Accepted: April 15, 2024
Published: May 13, 2024

REFERENCES

1. Bendall, S.C., Nolan, G.P., Roederer, M., and Chattopadhyay, P.K. (2012). A deep profiler's guide to cytometry. Trends Immunol. *33*, 323–332.
2. Cheung, M., Campbell, J.J., Whitby, L., Thomas, R.J., Braybrook, J., and Petzing, J. (2021). Current trends in flow cytometry automated data analysis software. Cytometry A. *99*, 1007–1021.
3. Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. Journal of machine learning research *9*.
4. Levine, J.H., Simonds, E.F., Bendall, S.C., Davis, K.L., Amir, E.a.D., Tadmor, M.D., Litvin, O., Fienberg, H.G., Jager, A., Zunder, E.R., et al. (2015). Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell *162*, 184–197.
5. Qiu, P., Simonds, E.F., Bendall, S.C., Gibbs, K.D., Jr., Bruggner, R.V., Linderman, M.D., Sachs, K., Nolan, G.P., and Plevritis, S.K. (2011). Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nat. Biotechnol. *29*, 886–891.
6. Van Gassen, S., Callebaut, B., Van Helden, M.J., Lambrecht, B.N., Demeester, P., Dhaene, T., and Saeys, Y. (2015). FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. Cytometry A. *87*, 636–645.
7. Newell, E.W., and Cheng, Y. (2016). Mass cytometry: blessed with the curse of dimensionality. Nat. Immunol. *17*, 890–895.
8. Morton, L.M., Wang, S.S., Devesa, S.S., Hartge, P., Weisenburger, D.D., and Linet, M.S. (2006). Lymphoma incidence patterns by WHO subtype in the United States, 1992-2001. Blood *107*, 265–276.
9. Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., et al. (2000). Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature *403*, 503–511.
10. Sehn, L.H., and Salles, G. (2021). Diffuse Large B-Cell Lymphoma. N. Engl. J. Med. *384*, 842–858.
11. Ask, E.H., Tschan-Plessl, A., Gjerdingen, T.J., Sætersmoen, M.L., Hoel, H.J., Wiiger, M.T., Olweus, J., Wahlin, B.E., Lingjærde, O.C., Horowitz, A., et al. (2021). A Systemic Protein Deviation Score Linked to PD-1(+) CD8(+) T Cell Expansion That Predicts Overall Survival in Diffuse Large B Cell Lymphoma. Med (N Y) *2*, 180–195.e5.
12. Khalifa, K.A., Badawy, H.M., Radwan, W.M., Shehata, M.A., and Bassuoni, M.A. (2014). CD14(+) HLA-DR low/(-) monocytes as indicator of disease aggressiveness in B-cell non-Hodgkin lymphoma. Int. J. Lab. Hematol. *36*, 650–655.
13. Grant, R., Coopman, K., Medcalf, N., Silva-Gomes, S., Campbell, J.J., Kara, B., Braybrook, J., and Petzing, J. (2021). Quantifying Operator Subjectivity within Flow Cytometry Data Analysis as a Source of Measurement Uncertainty and the Impact of Experience on Results. PDA J. Pharm. Sci. Technol. *75*, 33–47.
14. Landay, A.L., Brambilla, D., Pitt, J., Hillyer, G., Golenbock, D., Moye, J., Landesman, S., and Kagan, J. (1995). Interlaboratory variability of CD8 subset measurements by flow cytometry and its applications to multicenter clinical trials. NAID/NICHD Women and Infants Transmission Study Group. Clin. Diagn. Lab. Immunol. *2*, 462–468.
15. Lugli, E., Roederer, M., and Cossarizza, A. (2010). Data analysis in flow cytometry: the future just started. Cytometry A. *77*, 705–713.
16. Mair, F., Hartmann, F.J., Mrdjen, D., Tosevski, V., Krieg, C., and Becher, B. (2016). The end of gating? An introduction to automated analysis of high dimensional cytometry data. Eur. J. Immunol. *46*, 34–43.
17. Li, H., Shaham, U., Stanton, K.P., Yao, Y., Montgomery, R.R., and Kluger, Y. (2017). Gating mass cytometry data by deep learning. Bioinformatics *33*, 3423–3430.
18. Lux, M., Brinkman, R.R., Chauve, C., Laing, A., Lorenc, A., Abeler-Dörner, L., and Hammer, B. (2018). flowLearn: fast and precise identification and quality checking of cell populations in flow cytometry. Bioinformatics *34*, 2245–2253.
19. Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., and Borges, B. (2021). shiny: Web Application Framework for R. R package version 1.8.1.1. https://github.com/rstudio/shiny.
20. Finak, G., Frelinger, J., Jiang, W., Newell, E.W., Ramey, J., Davis, M.M., Kalams, S.A., De Rosa, S.C., and Gottardo, R. (2014). OpenCyto: an open source infrastructure for scalable, robust, reproducible, and automated, end-to-end flow cytometry data analysis. PLoS Comput. Biol. *10*, e1003806.
21. Bonilha, C.S. (2022). BCyto: A shiny app for flow cytometry data analysis. Mol. Cell. Probes *65*, 101848.
22. Arend, L., Bernett, J., Manz, Q., Klug, M., Lazareva, O., Baumbach, J., Bongiovanni, D., and List, M. (2022). A systematic comparison of novel and existing differential analysis methods for CyTOF data. Brief. Bioinform. *23*, bbab471.
23. Hauchamps, P., Bayat, B., Delandre, S., Hamrouni, M., Toussaint, M., Temmerman, S., Lin, D., and Gatto, L. (2024). CytoPipeline and CytoPipelineGUI: a Bioconductor R package suite for building and visualizing automated pre-processing pipelines for flow cytometry data. BMC Bioinf. *25*, 80.
24. Veglia, F., Sanseviero, E., and Gabrilovich, D.I. (2021). Myeloid-derived suppressor cells in the era of increasing myeloid cell diversity. Nat. Rev. Immunol. *21*, 485–498.
25. Lin, Y., Gustafson, M.P., Bulur, P.A., Gastineau, D.A., Witzig, T.E., and Dietz, A.B. (2011). Immunosuppressive CD14+HLA-DR(low)/- monocytes in B-cell non-Hodgkin lymphoma. Blood *117*, 872–881.
26. Azzaoui, I., Uhel, F., Rossille, D., Pangault, C., Dulong, J., Le Priol, J., Lamy, T., Houot, R., Le Gouill, S., Cartron, G., et al. (2016). T-cell defect in diffuse large B-cell lymphomas involves expansion of myeloid-derived suppressor cells. Blood *128*, 1081–1092.
27. Tadmor, T., Fell, R., Polliack, A., and Attias, D. (2013). Absolute monocytosis at diagnosis correlates with survival in diffuse large B-cell lymphoma-possible link with monocytic myeloid-derived suppressor cells. Hematol. Oncol. *31*, 65–71.
28. Chow, A., Perica, K., Klebanoff, C.A., and Wolchok, J.D. (2022). Clinical implications of T cell exhaustion for cancer immunotherapy. Nat. Rev. Clin. Oncol. *19*, 775–790.
29. Verma, V., Shrimali, R.K., Ahmad, S., Dai, W., Wang, H., Lu, S., Nandre, R., Gaur, P., Lopez, J., Sade-Feldman, M., et al. (2019). PD-1 blockade in

subprimed CD8 cells induces dysfunctional PD-1(+)CD38(hi) cells and anti-PD-1 resistance. Nat. Immunol. *20*, 1231–1243.

30. Björkström, N.K., Riese, P., Heuts, F., Andersson, S., Fauriat, C., Ivarsson, M.A., Björklund, A.T., Flodström-Tullberg, M., Michaëlsson, J., Rottenberg, M.E., et al. (2010). Expression patterns of NKG2A, KIR, and CD57 define a process of CD56dim NK-cell differentiation uncoupled from NK-cell education. Blood *116*, 3853–3864.
31. Plonquet, A., Haioun, C., Jais, J.P., Debard, A.L., Salles, G., Bene, M.C., Feugier, P., Rabian, C., Casasnovas, O., Labalette, M., et al. (2007). Peripheral blood natural killer cell count is associated with clinical outcome in patients with aaIPI 2-3 diffuse large B-cell lymphoma. Ann. Oncol. *18*, 1209–1215.
32. Klanova, M., Oestergaard, M.Z., Trněný, M., Hiddemann, W., Marcus, R., Sehn, L.H., Vitolo, U., Bazeos, A., Goede, V., Zeuner, H., et al. (2019). Prognostic Impact of Natural Killer Cell Count in Follicular Lymphoma and Diffuse Large B-cell Lymphoma Patients Treated with Immunochemotherapy. Clin. Cancer Res. *25*, 4634–4643.
33. Ask, E.H. (2024). Malmberglab/Metagate. v.1.1.0 (Zenodo). https://doi.org/10.5281/zenodo.10871021.
34. Ask, E.H. (2023). Mass cytometry immune phenotyping in DLBCL (MetaGate sample data). figshare. https://doi.org/10.6084/m9.figshare.24542173.v1.
35. R Core Team (2020). R: A Language and Environment for Statistical Computing (Vienna, Austria: R Foundation for Statistical Computing).
36. Finak, G., Jiang, W., Pardo, J., Asare, A., and Gottardo, R. (2012). QUAliFiER: an automated pipeline for quality assessment of gated flow cytometry data. BMC Bioinf. *13*, 252.
37. Finak, G., Jiang, W., and Gottardo, R. (2018). CytoML for cross-platform cytometry data sharing. Cytometry A. *93*, 1189–1196.
38. Hahne, F., LeMeur, N., Brinkman, R.R., Ellis, B., Haaland, P., Sarkar, D., Spidlen, J., Strain, E., and Gentleman, R. (2009). flowCore: a Bioconductor package for high throughput flow cytometry. BMC Bioinf. *10*, 106.
39. Spidlen, J., Gopalakrishnan, N., Hahne, F., Ellis, B., Gentleman, R., Dalphin, M., Le Meur, N., Purcell, B., and Jiang, W. (2020). flowUtils: Utilities for flow cytometry. R package version 1.10.0. https://github.com/jspidlen/flowUtils.
40. Wickham, H., and Sievert, C. (2016). ggplot2 : elegant graphics for data analysis. In Data Analysis, second edition (Springer International Publishing), pp. 189–201.