Making sense of user guided and fully automated clustering outcomes in multidimensional flow and mass cytometry data

Stephen Meehan[1,#], Gleb A. Kolyagin[2,#], David Parks[1], Justin Youngyunpipatkul[1], Leonore A. Herzenberg[1], Guenther Walther[3], Eliver E. B. Ghosn[4], and Darya Y. Orlova[1,*]

[1] Department of Genetics, Stanford University School of Medicine, Stanford, CA 94305, USA.

[2] Independent Researcher, Menlo Park, CA 94025, USA.

[3] Department of Statistics, Stanford University, Stanford, CA 94305, USA.

[4] Department of Medicine, Lowance Center for Human Immunology, Emory University School of Medicine, Atlanta, GA 30322, USA.

[#] The authors contributed equally to the work

*Corresponding author:

E-mail: orlova@stanford.edu, dyorlova@gmail.com (DYO)

Abstract

When examining datasets of any dimensionality, researchers frequently aim to identify individual subsets (clusters) of objects within the dataset. Interpreting these clustering results is a vital step in cluster analysis. The advent of multidimensional data makes the replacement of user guided clustering with fully automated clustering a compelling need. To address this need, many clustering methods have recently been implemented. However, methods that facilitate interpretation and reproducibility assessment of fully automatically generated clusters are still largely lacking.

To address these issues, we developed a pipeline of fully automated, statistically robust cluster matching and data visualization tools applicable to high-dimensional data generated with flow/mass cytometry and other technologies. We designed this pipeline to automatically (and intuitively) generate two-dimensional representations of high-dimensional datasets that are safe from the curse of dimensionality. This new approach allows more robust and reproducible data analysis facilitating development of new gold-standard practices across different laboratories and institutions.

Key terms

High-dimensional flow/mass cytometry; data clustering; fully automated clustering; cluster analysis; cluster matching; data visualization; curse of dimensionality; projection pursuit.

Introduction

The traditional approach to locating clusters (subsets) in high-dimensional (Hi-D) datasets such as those acquired by flow cytometry is to reduce the dataset dimensionality, usually by linear and/or nonlinear one/two dimensional mapping or projection strategies. This "Projection Pursuit" approach has proven to be very efficient for analyzing high dimensional data in a way that avoids a common pitfall, i.e., the curse of dimensionality ([1,2], and supplementary materials). Indeed, much of what is known about stem cells, blood cells and diseases such as leukemia and AIDS relies on flow cytometry data analyzed with these manual sequential Projection Pursuit approaches, including the widely-used methods offered by FlowJo (www.flowjo.com). Usually cell subsets identified in such user guided manner are readily biologically interpretable. However, the resolution of such subsets with manual analysis tools is by no means routine. In fact, since these manual analysis methods ultimately rely on user skills to define subset boundaries, subset identification and quantitation is still more appropriately recognized as art rather than science. Automating this data analysis process and making it more objective is clearly desirable.

Several groups have recently developed fully automated computational approaches that operate simultaneously in four or more dimensions to identify the subsets (clusters) within a given Hi-D dataset [3]. These attempts are well motivated from a functionality point of view. However, there are several issues associated with the fully automated Hi-D clustering approach. First, the reproducibility of clusters automatically generated from simultaneous analysis of multiple dimensions is proving challenging [4]; as we have shown previously, this irreproducibility is partially caused by the curse of dimensionality ([5] and supplementary materials). Second, there is no widely-accepted analytical framework to distinguish spurious

clusters from more stable entities, and presumably more biologically relevant ones [4]. Finally, there is a lack of tools to readily interpret fully automated clustering outcomes.

To facilitate statistical and biological inference from fully automated (and user guided) clustering outcomes we introduce a pipeline of multidimensional cluster matching and display methods. We based our pipeline on the quadratic form (QF) distance metric and adaptive binning [6]. We previously demonstrated [6] that a computationally efficient distance metric such as QF, which takes into account changes in both location and frequency rather than just changes in one or the other, is the most suitable and accurate method for comparing multivariate non-parametric flow/mass cytometry data distributions. In addition, by coupling QF metric with adaptive binning we avoid the curse of dimensionality in both cluster matching and data visualization.

Together with clustering algorithm, our methods provide a complete pipeline for cluster (subset) recognition, display and characterization. The analysis pipeline we describe is readily applicable to any number of dimensions and to any method that enables valid identification of cellular (or other) subsets. Here we emphasize that it is crucially important to fuse/apply cluster matching and visualization modules to valid (i.e., it avoids the curse of dimensionality) methods of subset identification. We avoid the curse in cluster identification here by coupling cluster matching and data visualization tools with fully automated Exhaustive Projection Pursuit (EPP) clustering approach available at www.cytogenie.org; http://cgworkspace.cytogenie.org/GetDown2/demo/bCellMacrophageDiscoveryDemo.pdf; (manuscript in preparation).

This clustering method relies on the same principles underlying the previous automated two-dimensional (2D) Exhaustive Projection Pursuit approaches [7]. Briefly, the Exhaustive

Projection Pursuit (EPP) works in the following stepwise approach: 1) Hi-D data is presented as a collection of 2D linear projections; 2) every 2D projection is then characterized by a numerical index that indicates the amount of structure that is present [7,8]; 3) this index is then used as the basis for a heuristic search to locate the most "useful" 2D projection; 4) once the projection with the most useful structure has been found, this structure is then segmented and each portion is recursively analyzed until there is no remaining structure detectable.

In general, Projection Pursuit methods are a big step forward towards solving the problem of Hi-D data analysis because they avoid the curse of dimensionality. However, the approaches advanced thus far have some key limitations. For example, what constitutes structures in data and how to make inferences from such identified structures is neither obvious nor trivial to specify [9]. To overcome these limitations, we are developing (manuscript is in preparation) fully automated Exhaustive Projection Pursuit (EPP) method (its implementation is available at www.cytogenie.org) that uses the smallest misclassification error across a decision boundary between identified clusters (using the DBM approach [9]) as an index to identify the most profitable 2D projection.

Here, we apply these statistically robust clustering and data visualization tools to both simulated and previously published flow/mass cytometry datasets and emphasize that they are readily applicable to similar single- or multi-dimensional data generated with other technologies.

Results

Cluster analysis and data visualization pipeline

In a simplified example shown on Figure 1 we illustrate steps of the cluster analysis and data visualization pipeline that we develop.

Here we used fully automated EPP clustering (www.cytogenie.com) to locate subsets. The basic strategy underlying the EPP methods is a search for an orthogonal 2D projection in which the data are cleanly split into subsets. Applied recursively, EPP carries out this strategy, idenfying subsets until no further splits are available (Figure 1B). Thus, for a set of measurements, EPP:

- Examines all possible 2D projections using the density based merging (DBM) [10] clustering method and assigns all unclustered data (e.g., outliers) to the nearest cluster;
- Finds all suitable candidate decision boundaries;
- Ranks them by estimated classifier error;
- Separates the data across the top ranked decision boundary to define two subsets;
- Repeats the above on each of the two subsets until no further splits are found.

To align (match) subsets identified by EPP (or other clustering algorithm) in two or more comparable samples (e.g., samples A and B on Figure 1A), we use a quadratic form (QF)-based cluster matching algorithm (QFMatch) that we described previously [6,11]. However, here we extended the previous version of QFMatch by adding an exhaustive cluster merging step. Figure 1C-F illustrates the application of QFMatch to a three-dimensional dataset. Matched subsets in samples A and B are highlighted with the same color (Figure 1G). As we show in the next section, the QFMatch can be applied to match subsets identified by different

clustering approaches (e.g., user guided versus fully automated, or to compare outcomes between different fully automated algorithms) within one sample.

To visualize clustering outcomes in a composite figure, we developed two data display alternatives that can supplement each other:

- We use a multidimensional scaling (MDS) method [12] that allows placement of each object (cluster) in two-dimensional space such that the overall between-object distances in high-dimensional space are well-preserved. To make the results more visually interpretable, we apply this MDS method to the matrix of distances between median values calculated for each of the identified clusters (Figure 1G). This reduces the effect of the "crowding problem" [13] and, importantly, allows computationally efficient application of MDS.
- We also created a tree structure data display (Figure 1H) that allows agglomerative arrangement of identified clusters based on their (dis)similarity in the space of measured parameters. This data display method builds the hierarchy from the individual clusters identified within one sample by progressively merging clusters. In order to decide which clusters should be merged a measure of dissimilarity between sets of observations is required. We used a combination of multidimensional quadratic form (QF) score described in our previous paper [6] and Euclidean distance between clusters' medians as a dissimilarity measure to combine identified clusters in a "bottom up" manner: the branching diagram starts by placing clusters with the smallest pairwise dissimilarity scores in the lowest branches of the diagram; these pairs of clusters are further progressively merged in the next branching level of the diagram and further considered as one cluster; dissimilarity scores are then

recalculated for all of the clusters on this branching level and the merging process is repeated. This process is sequentially repeated until all of the clusters identified within the sample are merged together. We named this tree- structure data display as QF-tree.

We refer to the above computational pipeline (i.e., clustering for subset identification, QFMatch for high-dimensional cluster matching, and MDS or QF-tree for data display) as the "subset identification and characterization" (SIC) pipeline. The algorithms constituting this pipeline are available as parts of the AutoGate software which is freely available for download to not-for-profit users (.edu, .org, .gov) at www.cytogenie.org .

We also provide a source code (python implementation) for the prototypes of high-dimensional cluster matching and data display algorithms (MDS and QF-tree) at https://github.com/dyorlova/QFMatch_MDS_dendrogram . The python implementation provides alternative choices for MDS data display, including the use of median values or adaptive bins that are calculated for each of the identified clusters.

SIC pipeline identifies the well-known immune cell subsets within the mouse peritoneal cavity (PerC).

To validate the SIC pipeline in an fully automated manner, we applied it to a previously published dataset [14] shown, by manual gating, to contain cells from the myeloid (small and large peritoneal macrophages, and dendritic cells), granuloid (eosinophils and neutrophils), and lymphoid (T, B, NK, and NKT cells) lineages (Figure 2A). We show that the standard cell subset measurements (i.e., median fluorescence values and cell frequencies) generated automatically by the SIC pipeline (Figure 2B and supplementary material for Figure 2B) are in strong agreement with the measurements described by the traditional manual gating method

(user guided clustering) performed by highly skilled investigator [14]. We also show that the SIC pipeline consistently detects the same immune cell subsets in the PerC of another wild-type mouse strain (BALB/c) even when a different staining panel is used (see Supplementary Figure 5).

In addition to identifying well-established immune cells subsets, the SIC pipeline was able to identify other cell subsets that were not considered within the established manual gating strategy. For example, using the same set of parameters as in the manual gating strategy, the SIC pipeline identified two subsets dendritic cells (DC) based on the expression levels of surface CD11b (Figure 2C).

Figure 2D shows agglomerative arrangement (the QF-tree) of the cell subsets (identified by user guided clustering) according to their (dis)similarity in the space of measured parameters (Forward Scatter, CD11b, CD11c, CD19, CD5, F4/80, IgD, and IgM). In other words, QF-tree organized cells in a hierarchy of related phenotypes. Although QF-tree can reliably recapitulate patterns of hematopoiesis from high-dimensional cytometry data, its utility is limited by the choice of markers that are measured in the experiment. For instance, if the tree structure is built with a marker set that is not related to cellular progression, one might not expect to recover the known lineage relationships.

To further test the SIC pipeline performance we challenged its ability to detect missing lymphocyte populations in the PerC of RAG knockout (RAG-/-) mice. Using QFMatch we aligned cell subsets identified in the wild-type mice (BALB/c) by the user guided clustering (Figure 2A) with the cell subsets identified in the knockout mice (RAG-/-) by the EPP clustering. The QFMatch algorithm readily matched the non-lymphoid cells present in both

BALB/c and RAG-/- samples and correctly detected the *lack* of T and B lymphocytes in the RAG-/- sample (Figure 2E).

SIC pipeline identifies various subsets of human peripheral B lymphocytes

Using two samples of human peripheral blood stained with the same panel of surface markers (Figure 3A), we explored the SIC pipeline ability to consistently detect the various lymphoid, myeloid, and granuloid subsets. We used manual gating strategy shown on Figure 3A to identify T cells, neutrophils, monocytes, naïve B cells, memory B cells, class-switched B cells, and transitional B cells. We further used QFMatch algorithm to align these subsets with the subsets identified by fully automated EPP clustering.

QFMatch algorithm successfully aligned the immune cell subsets that were identified in user guided manner with those that were identified by fully automated EPP. QFMatch also reported additional subsets there were not identified manually, but were readily discriminated by EPP clustering (marked as red squares on MDS display, Figure 3B). The SIC pipeline consistently detected all the cell subsets that were identified by the manual gating strategy (Figure 3B and 3C).

SIC pipeline can be applied to CyTOF datasets to identify differences in cell subset representation in clinical samples.

We tested the SIC pipeline on the publicly available CyTOF (mass cytometry) dataset collected from patients with Acute Myeloid Leukemia (AML) pathophysiology study [15] to illustrate one of the possible applications of SIC pipeline in clinical/biomedical studies- detection and quantification of a difference in subset representation between healthy controls and AML patients samples.

We randomly selected three healthy controls (H) and three AML patients (SJ) samples from the original study [15] and compared the representation of $CD11b^{hi}CD33^{low}$ and $CD11b^{hi}CD33^{hi}$ myeloid cell subsets between these samples. Figure 4A (Healthy control) shows the gating strategy used to identify the two myeloid cell subsets in sample H4 with user guided clustering. We then ran fully automated EPP clustering on all six samples (H4, H5, H6, SJ11d, SJ15d, SJ16d) individually and matched the clustering outcomes for these samples with the user guided clustering outcomes for sample H4. Figure 4B shows the difference (relative frequency) in representation of $CD11b^{hi}CD33^{low}$ and $CD11b^{hi}CD33^{hi}$ myeloid cell subsets between healthy controls and AML patients. To validate the SIC pipeline performance we further performed user guided clustering according to the gating strategy shown Figure 4A (AML patient) to verify whether some AML patients indeed completely lack $CD11b^{hi}CD33^{low}$ and $CD11b^{hi}CD33^{hi}$ myeloid cells.

Discussion

Modern multidimensional flow and mass cytometry data undoubtedly requires automation of its analysis. However, despite the decent amount of efforts that has been recently made to automate subset identification and characterization in flow/mass cytometry data, these automated methods were not widely adopted among biologists/clinicians. The majority of flow/mass cytometry users still prefer manual gating (e.g., using FlowJo) to automated clustering. One of the main reasons (apart from the vulnerability to the curse of dimensionality) for the lack of adoption of automated methods is their inability to display and align clustering outcomes in a way to allow automatic extraction of meaningful and readily applicable biologically/biomedically information. It is indeed a nontrivial challenge to present the Hi-D clustering results in a way that is easy to understand, interpret and further align these clustering outcomes between samples or between different clustering algorithms.

Several methods creating two-dimensional visualization of high (or low)-dimensional clustering outomes (e.g., viSNE/tSNE [16], SPADE [17]) have been developed to aid biologists interpreting Hi-D cytometry data. These methods can provide rich information about the high-dimensional relationship in the data. However, they have some significant drawbacks. Both viSNE/tSNE [16] and SPADE [17] are prone to suffer from the curse of dimensionality (see supplementary materials, Supplementary Figure 4) and most importantly they require user-defined input parameters that significantly affect clustering and visualization outcomes. Additionaly, as noted by Yang et al. [18] "…[viSNE and SPADE] are not intuitive to biologists who are accustomed to the two-dimensional nested gating representations."

Aiming to make clustering outcomes more intuitive to biologists Yang et al. [18] have recently developed the C2G data visualization method. This method is able to generate a

gating hierarchy that captures the target populations (identified by any clustering method) and present the hierarchy in nested two-dimensional gating sequences that resemble the conventional manual gating analysis. Presenting Hi-D data in two-dimensional nested tree structure may indeed help biologists who are accustomed to the two-dimensional nested gating representations. However, conventional (domain knowledge-driven) gating strategies not always follow the “best separation” path that underlies the gating strategy principle applied by the C2G method [18]. Thus, gating strategy that is built according to the “best separation” principle may not be readily interpretable to biologists. Moreover, it is not clear how to readily align the trees that followed different gating strategies (e.g., trees that were built from two different samples).

To facilitate statistical and biological inference from automated clustering outomes we present a pipeline of fully automated, statistically robust cluster matching and data visualization tools applicable to high (or low)-dimensional data generated with flow/mass cytometry and other technologies. The analysis pipeline that we developed here consists of three modules: (1) cluster matching with QFMatch, and (2-3) two-dimensional display of cluster identification and cluster matching results with MDS and/or QF-tree. We designed the cluster matching and data visualization algorithms in a way that automatically produce intuitive representations of Hi-D single-cell data while avoiding the curse of dimensionality. Also, these methods do not require user-defined tuning parameters and their utility is mainly limited by the choice/availability of markers unsed in a particular staining panel (e.g., QF-tree will not retrieve the known lineage relationships if the chosen/available stainset panel does not include key markers related to the cellular progression).

We successfully applied the SIC pipeline to both user guided and fully automated clustering outcomes using both flow and mass cytometry datasets from mouse and human/clinical samples. We implemented this pipeline in AutoGate (www.cytonegie.org) software package that supports graphical user interface and provided Python source code at https://github.com/dyorlova/QFMatch_MDS_dendrogram .

Materials and Methods

Experiment overview

We use DBM [10] or EPP (www.cytonegie.org) to identify cell subsets in simulated and flow/mass cytometry data, QFMatch [6] to align subsets between relevant samples (same staining panels) and MDS [12] or QF-tree to visualize user guided and fully automated clustering outcomes within the biological/biomedical datasets described below.

Flow/mass sample description

The mouse peritoneal cavity (Figure 2) and human bone marrow (Figure 4) datasets were generated in previously published studies (see [14,15] for complete materials and methods).

The human peripheral blood dataset (Figure 3) was generated using a combination of 16 monoclonal antibodies (Hi-D 18-parameter flow cytometry panel): B220-PE, CD5-PECy5, CD10-PECy5.5, CD19-BV786, CD20-BV650, CD23-APCCy7, CD27-BV421, CD38-APC, CD43-AF700, CD95-BV605, CD132-BV711, CD3/CD14/CD16 (Dump)-BV570, CD45-AF488, IgD-PECy7, IgM-PECF594, and Aqua Amine (viability). After informed consent, 10 mL of blood was drawn from adult healthy volunteers. Blood samples were collected (under IRB review) in evacuated tubes containing EDTA (K2) (Vacutainer, BD Biosciences). All samples were de-identified. Data were collected for about 0,5x10^6 cells.

Instrument details

Information about instruments used to collect human and mouse samples can be found in [14,15]. Human peripheral blood cells were analyzed on the BD LSRII instrument at Stanford Shared FACS Facility.

Data analysis details

The proposed workflow for analyzing all three datasets used in this manuscript consists of three steps:

1) Transform the compensated data (FlowJo v.10, fluorescence flow cytometry data only) with the Logicle transformation [19], and cluster the transformed data with DBM (user guided) [10] or EPP (fully automated) clustering methods. Data transformation and clustering utilities are available in AutoGate (www.cytonegie.org). Data were pregated for live singlets before EPP clustering run. See figures for gating sequences. The flow/mass cytometry data pocessing methods used here do not require user input for parameters such as number of clusters, number of grid bins, etc.

2) Use QFMatch to align cell populations between samples or between different clustering outcomes for the same sample. The QFMatch (QF-based cluster matching algorithm) is integrated into AutoGate (www.cytogenie.org).

3) Use MDS and/or QF-tree to display clustering and cluster matching outcomes. Both visualization tools are integrated into AutoGate (www.cytogenie.org).

QFMatch, MDS and QF-tree require only one user input configuration parameter, that is the set of markers (or/and light scatter signals) selected to match and display clustering outcomes. Of note, QFMatch, MDS and QF-tree work independently of how the populations

(clusters) were pre-defined. For example, the clusters could be defined by using domain knowledge-driven manual gating, a sequential automated clustering approach, or a simultaneous clustering approach.

Data Availability

The datasets generated during and/or analysed during the current study (Figures 2-4) are available in the FlowRepository and Cytobank:

https://flowrepository.org/id/RvFrin9QJBrBl7euVYafJg8MBtow5TSn0Cbf6ibJFTQbutUCP8VbTKi70DJD7TJg

https://flowrepository.org/id/RvFrmp0uY05bFrRfQW6XgcLV360pTCjz5ieEKzaHHGsTDoWEWpBspy21QVrQhFxz

https://www.cytobank.org/nolanlab/reports/Levine2015.html

Code Availability

The analysis pipeline described here is implemented in AutoGate (www.cytonegie.org) software package that supports graphical user interface; Python source code is available at https://github.com/dyorlova/QFMatch_MDS_dendrogram

Glossary

DBM – Density-Based Merging clustering algorithm allowing sequentially cluster high-dimensional flow/mass cytometry data in two-dimensions at a time in a user guided manner.

EPP – Exhaustive Projection Pursuit approach allowing sequentially cluster high-dimensional flow/mass cytometry data in two-dimensions at a time in a fully automated manner.

MDS – Multidimensional Scaling method allowing placement of each object (cluster) in two-dimensional space such that the overall between-object distances in high-dimensional space are well-preserved.

QFMatch – multidimensional cluster matching algorithm that is based on quadratic form as a measure of clusters (dis)similarity.

QF-tree – tree structure data display allowing agglomerative arrangement of identified clusters based on their (dis)similarity in the space of measured parameters. Combination of quadratic form score and Euclidean distance between clusters' medians is used as a measure of clusters (dis)similarity.

References

1. Hastie, T., Tibshirani, R., Friedman, J. The elements of statistical learning. (Springer-Verlag, 2009).

2. Altman, N., Krzywinski, M. The curse(s) of dimensionality. Nat Methods. 15(6), 399-400 (2018). doi: 10.1038/s41592-018-0019-x.

3. Saeys, Y., Gassen, S.V., Lambrecht, B.N. Computational flow cytometry: helping to make sense of high-dimensional immunology data. Nat Rev Immunol. 16(7), 449-62 (2016). doi: 10.1038/nri.2016.56.

4. Melchiotti, R., Gracio, F., Kordasti, S., Todd, A.K., de Rinaldis, E. Cluster stability in the analysis of mass cytometry data. Cytometry A. 91(1), 73-84 (2016). doi: 10.1002/cyto.a.23001.

5. Orlova, D. Y., Herzenberg, L. A., Walther, G. Science not art: statistically sound methods for identifying subsets in multi-dimensional flow and mass cytometry datasets. Nat Rev Immunol. 18, 77 (2018). doi:10.1038/nri.2017.150.

6. Orlova, D.Y., Meehan, S., Parks, D., Moore, W., Meehan, C., Zhao, Q., Ghosn, E.E., Herzenberg, L.A., Walther, G. QFMatch: multidimensional flow and mass cytometry samples alignment. Sci Rep. 8(1):3291 (2018). doi: 10.1038/s41598-018-21444-4.

7. Friedman, J.H., Tukey, J.W. A Projection Pursuit algorithm for exploratory data analysis. IEEE Transactions on Computers. C–23 (9), 881–90 (1974).

8. Herzenberg, L.A., Moore, W.A., Parks, D.R., Walther, G., Meehan, S., Orlova, D.Y., Meehan, C. Cluster Processing and Ranking Methods Including Methods Applicable to Clusters Developed Through Density Based Merging. United States Patent Application 20150293992.

9. Huber, P. Projection Pursuit. Ann. Statist. 13(2), 435 (1985).

10. Walther, G., Zimmerman, N., Moore, W., Parks, D., Meehan, S., Belitskaya, I., Pan, J., Herzenberg, L. Automatic clustering of flow cytometry data with density-based merging. Adv. Bioinformatics. 686759 (2009). doi: 10.1155/2009/686759.

11. Orlova, D.Y., Meehan, S., Moore, W.A., Walther, G., Parks, D.R., Herzenberg, L.A. Systems and methods for cluster matching across samples and guided visualization of multidimensional cytometry data. United States Patent Application. United States Patent Application 62/363,109.

12. Kruskal, J.B., Wish, M. Multidimensional Scaling. Sage University Paper Series on Quantitative Applications in the Social Sciences, No. 07-011 (1978). Sage Publications, Newbury Park. http://dx.doi.org/10.4135/9781412985130.

13. van der Maaten, L., Hinton, G. Visualizing data using t-SNE. Journal of Machine Learning Research. 9, 2579–2605 (2008).

14. Ghosn, E.E.B., Cassado, A.A., Govoni, G.R., Fukuhara, T., Yang, Y., Monack, D.M., Bortoluci, K.R., Almeida, S.R., Herzenberg, L.A., Herzenberg, L.A. Two physically, functionally, and developmentally distinct peritoneal macrophage subsets. Proc. Natl. Acad. Sci. U. S. A. 107(6), 2568–73 (2009). doi: 10.1073/pnas.0915000107.

15. Levine, J. H., et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. Cell. 162(1), 184-197 (2015). doi: 10.1016/j.cell.2015.05.047.

16. Amir, el-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. Nat Biotechnol. 31(6), 545-552 (2013). doi: 10.1038/nbt.2594.

17. Qiu, P., Simonds, E.F., Bendall, S.C., Gibbs, K.D. Jr., 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. Nat Biotechnol. 29(10), 886-891 (2011). doi: 10.1038/nbt.1991.

18. Yang, X., Qiu, P. Automatically generate two-dimensional gating hierarchy from clustered cytometry data. Cytometry A. 2018. doi: 10.1002/cyto.a.23577.

19. Moore, W.A., Parks, D.R. Update for the logicle data scale including operational code implementations. Cytometry A 81(4), 273–277 (2012). doi: 10.1002/cyto.a.22030.

Acknowledgments

We thank John Mantovani for excellent administrative help.

Author contributions

SM, GAK, DP, LAH, GW, EEBG, DYO: Conception and design; Analysis and interpretation of the data; Drafting of the article; Critical revision of the article for important intellectual content.

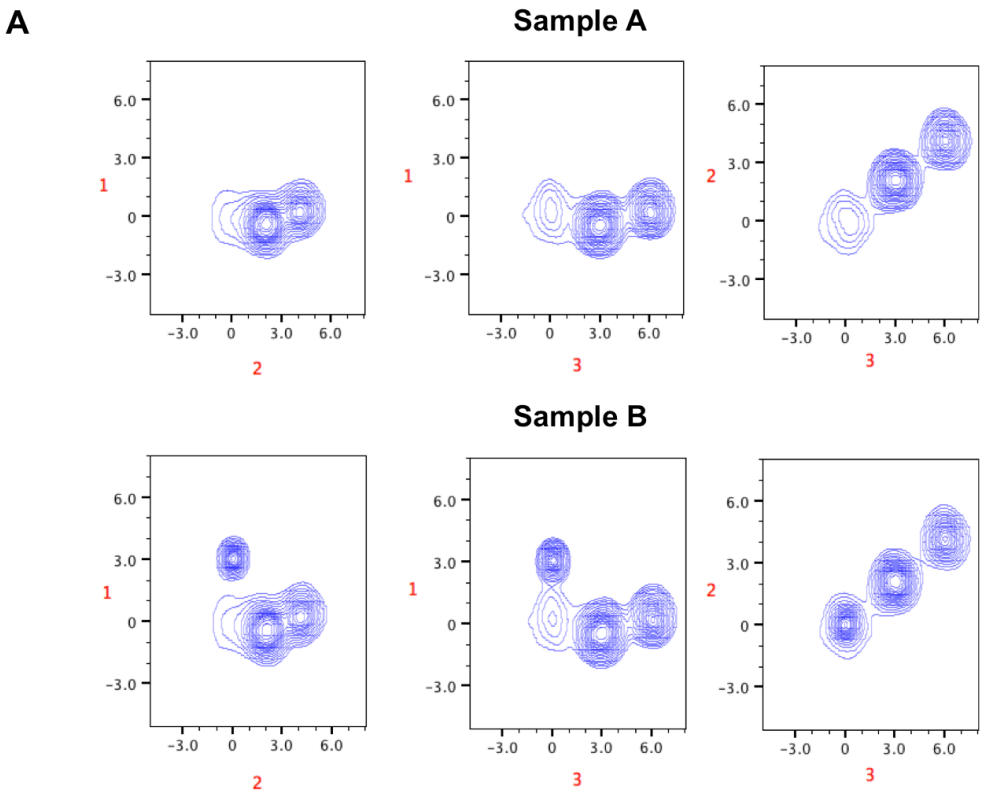
JY: Analysis and interpretation of the data; Drafting of the article; Critical revision of the article for important intellectual content.

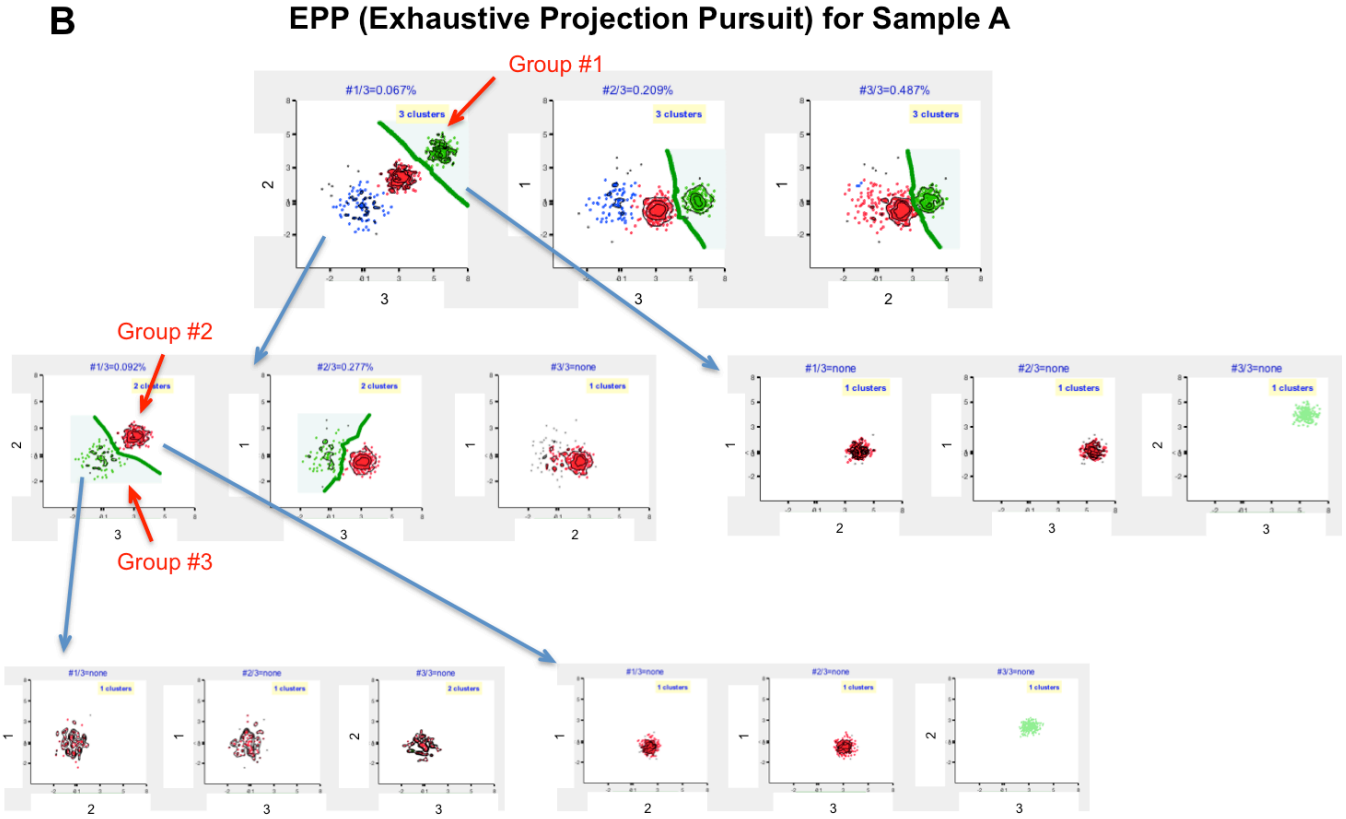
SM and LAH: AutoGate implementation of the algorithms described here.

GAK and DYO: Python implementation of the algorithms described here.

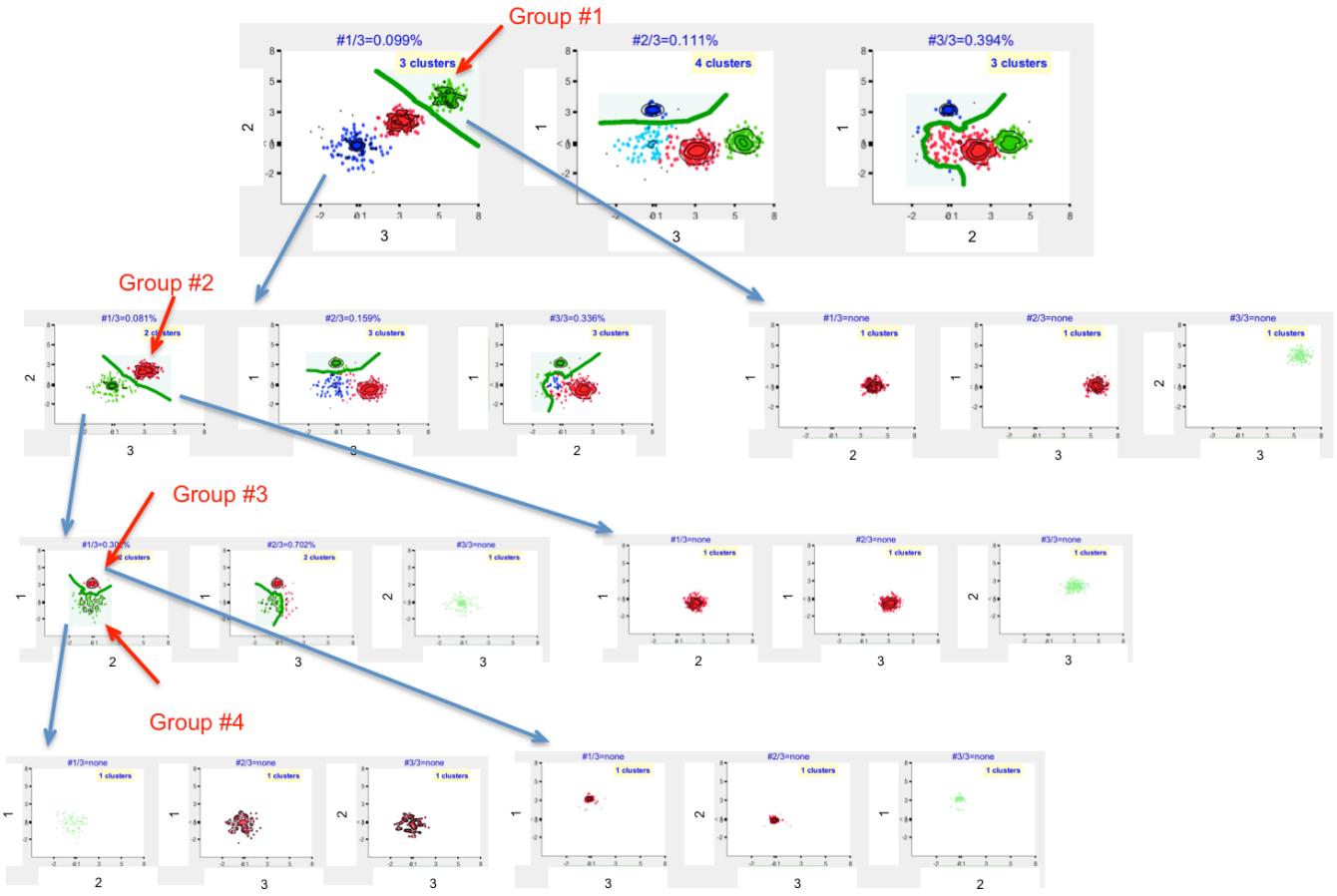
Conflict of Interest

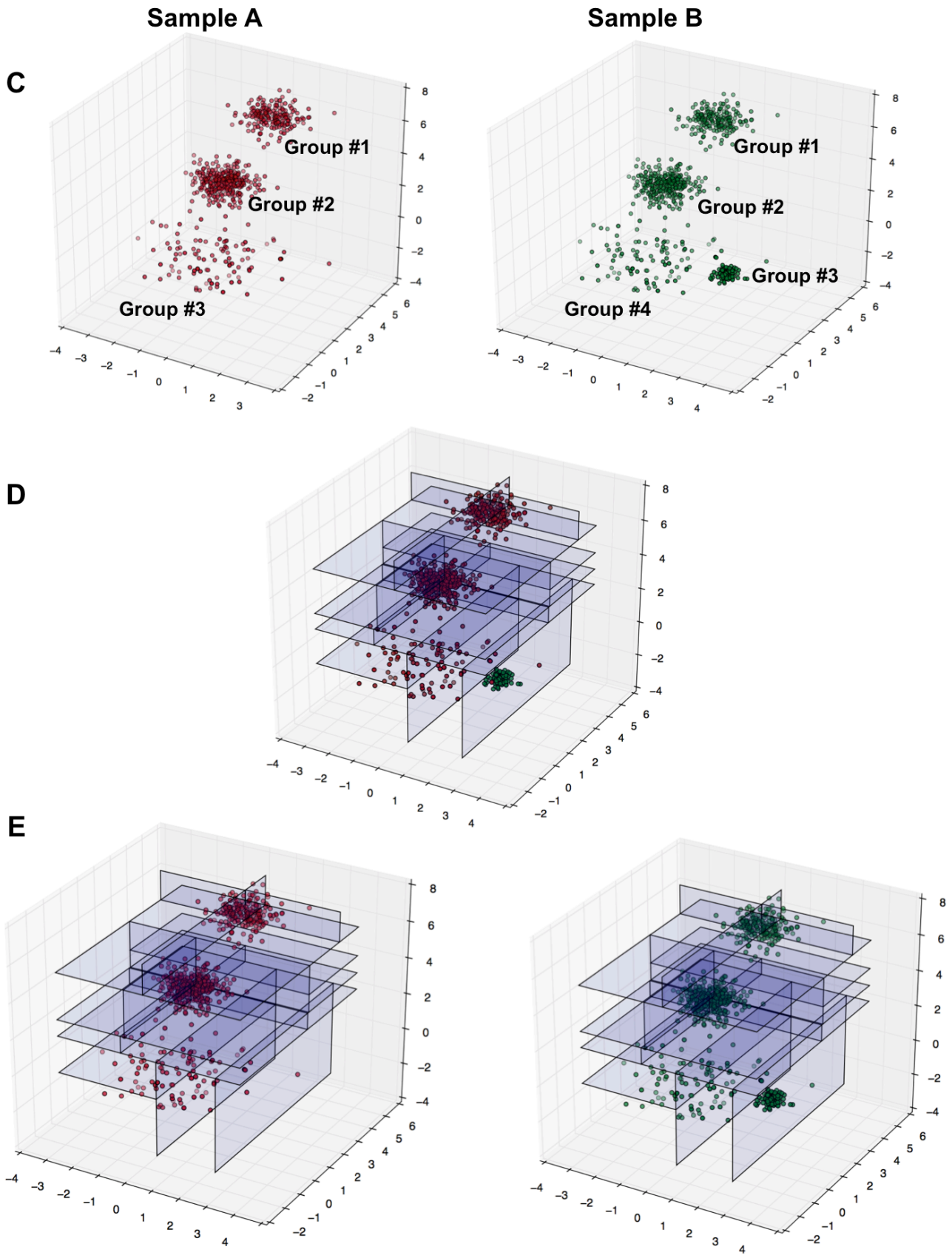
The authors declare no competing financial interests.


A
Sample A
Sample B
6.0
3.0
0
-3.0
1
2
3


B
EPP (Exhaustive Projection Pursuit) for Sample A
Group #1
Group #2
Group #3
#1/3=0.067%
#2/3=0.209%
#3/3=0.487%
#1/3=0.092%
#2/3=0.277%
#1/3=none
#2/3=none
#3/3=none
3 clusters
2 clusters
1 clusters

EPP for Sample B




Sample A
Sample B
C
Group #1
Group #2
Group #3
Group #1
Group #2
Group #3
Group #4
D
E

F

Sample B

Sample A

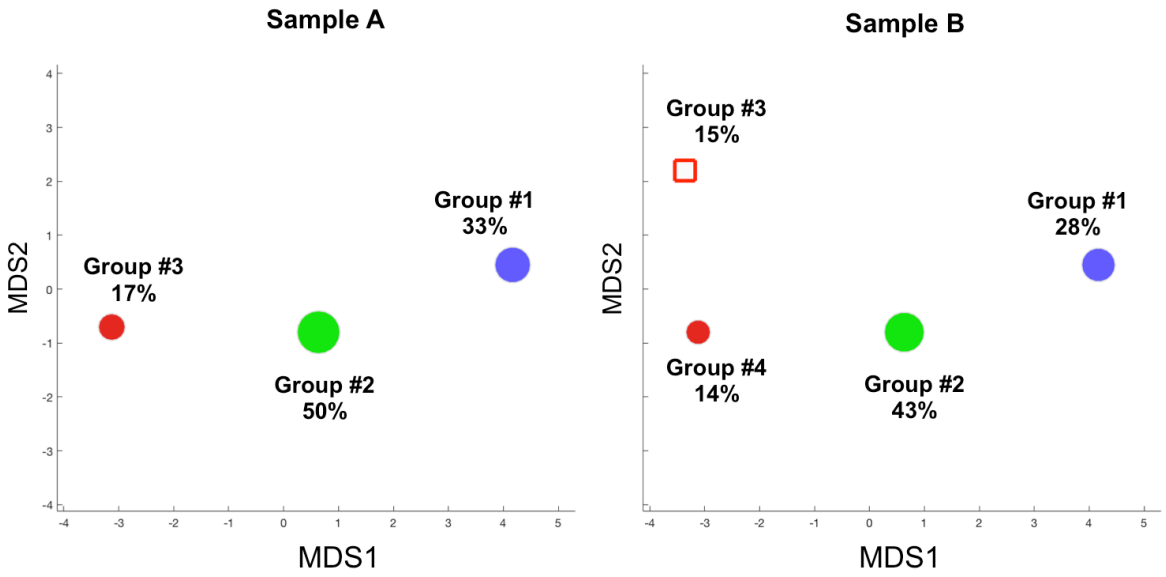
Group ID	1	2	3	4
1	0.0003			
2		0.00004		
3		0.86	0.73	0.00016

Sample B

Sample A

Group ID	1	2	3+4
1			
2			
3			0.17

G



Matrix of distances between groups' medians in Sample B

Group ID	1	2	3	4
1	-	3.75	7.77	7.36
2		-	5.01	3.74
3			-	2.90
4				-

H

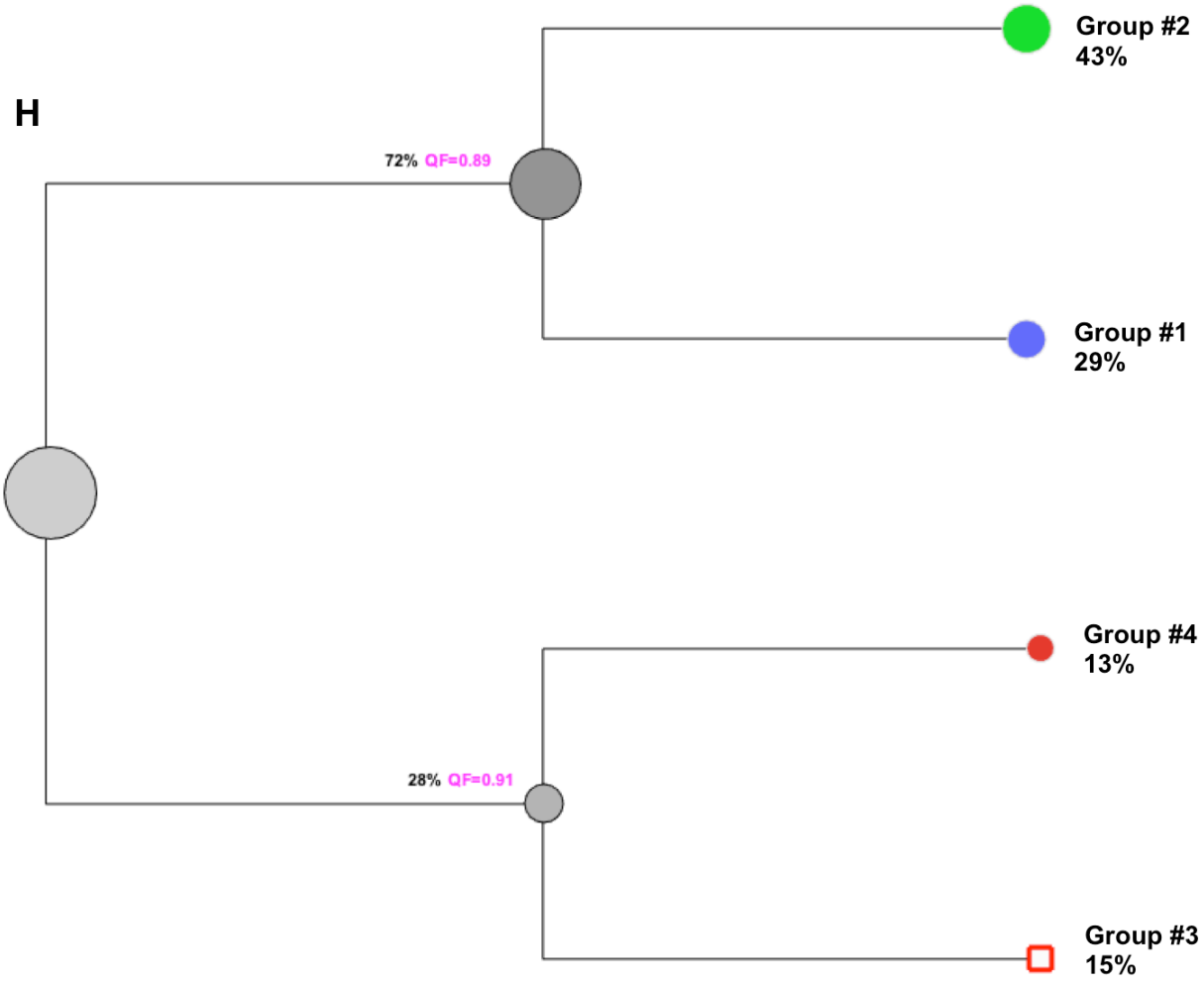
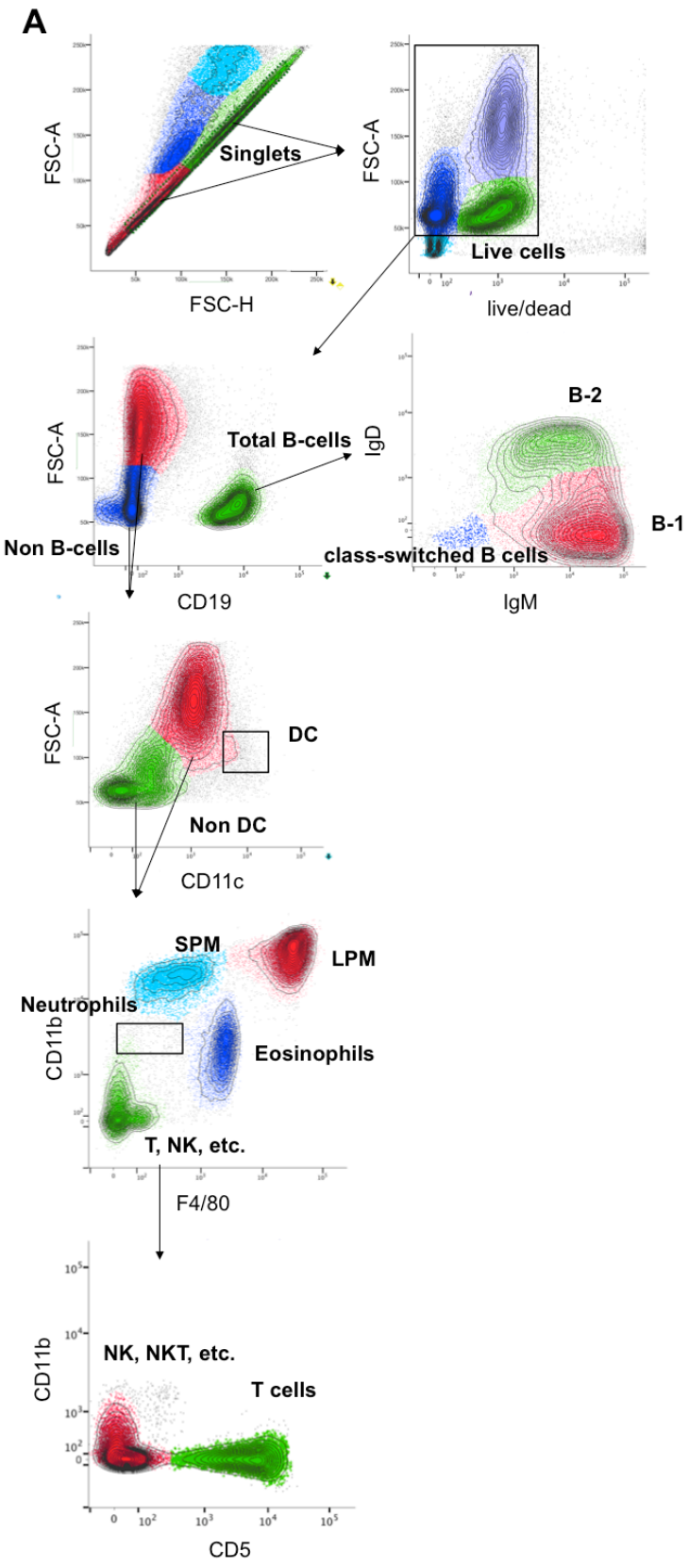
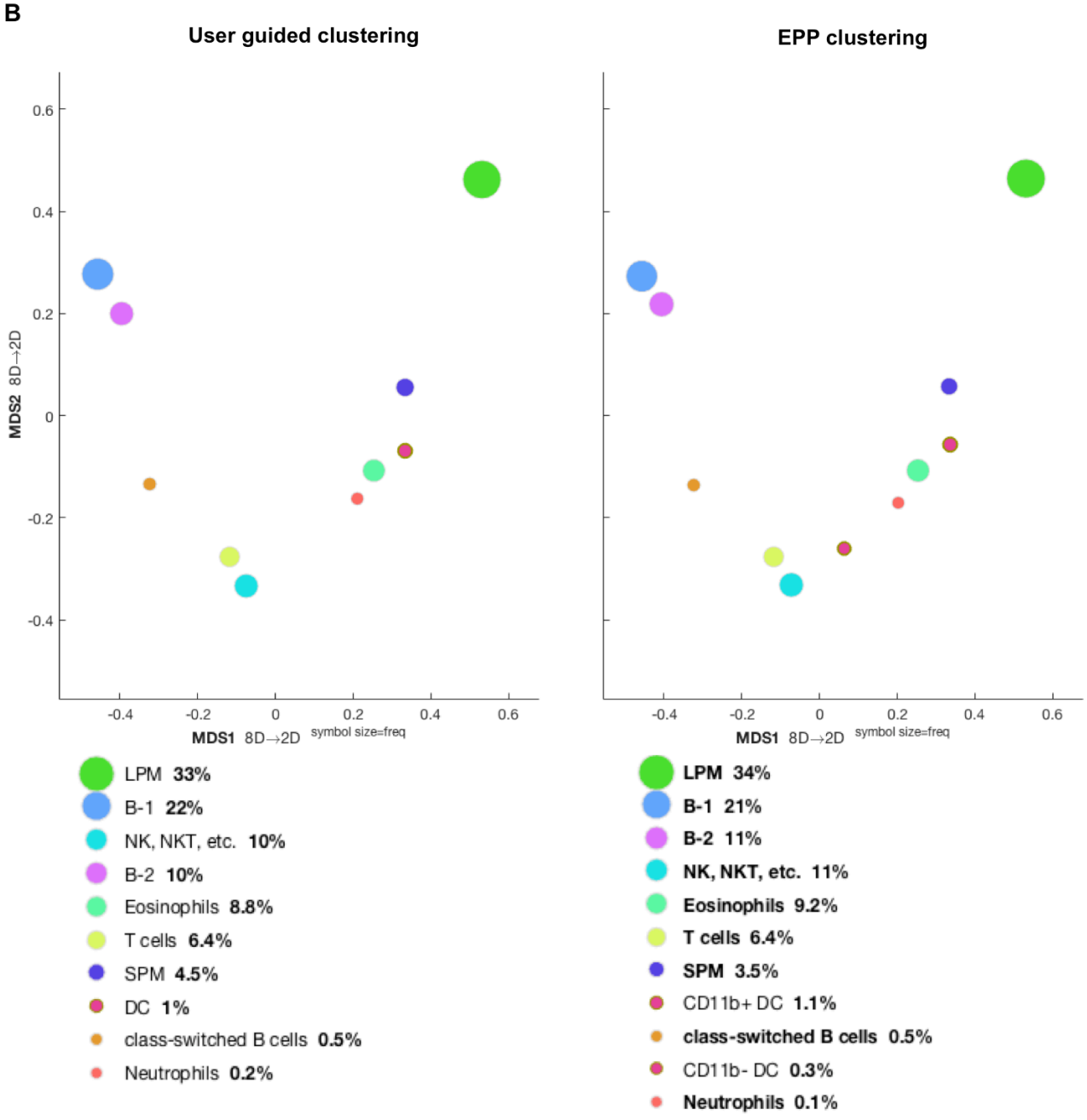


Figure 1. Subset identification and characterization pipeline. Panel A. We simulated two three-dimensional sets of data (Sample A and Sample B). **Panel B.** To identify clusters in these datasets we applied Exhaustive Projection Pursuit approach that recursively 1) projects the data in a collection of 2D linear projections, 2) characterizes every given 2D projection by a numerical index that indicates the smallest misclassification error across a decision boundary (green line) between identified clusters, 3) separates the data across the top ranked decision boundary to produce two subsets, 4) repeats on both subsets until no splits are found. Subsets that have no further splits (groups#1-4) are final clusters identified by the EPP approach. **Panels C-F.** The steps of the QFMatch algorithm as applied in aligning clusters identified by the EPP approach. Merge the beforehand clustered samples (panel C, samples were clustered as described on panel B) and perform adaptive binning (panel D) insuring 2lnN events per bin, where N is the number of events in the smallest sample. Split samples back preserving the binning pattern (panel E). Calculate QF dissimilarity [6] between each possible combination of cluster pairs (Group #1-#4 from Sample A and Sample B) which medians are located no more than four standard deviations apart in every dimension (panel F). Pairs with the smallest dissimilarity scores are marked in green. The merging candidate is marked in blue. If there is more than one merging candidate then all possible permutations of merging candidates are considered. If as a result of the merging process the initial dissimilarity score decreases then the presence of a cluster split is indicated, if not then the unmatched cluster is considered as missing. **Panel G.** Display of the EPP clustering outcomes using MDS method. Each circle represents one subset identified by the EPP approach. The size of the circle directly correlates with the relative frequency of the subset in the sample. Subsets that match (identified using QFMatch) between Sample A and Sample B are highlighted with the same color. X and Y axes

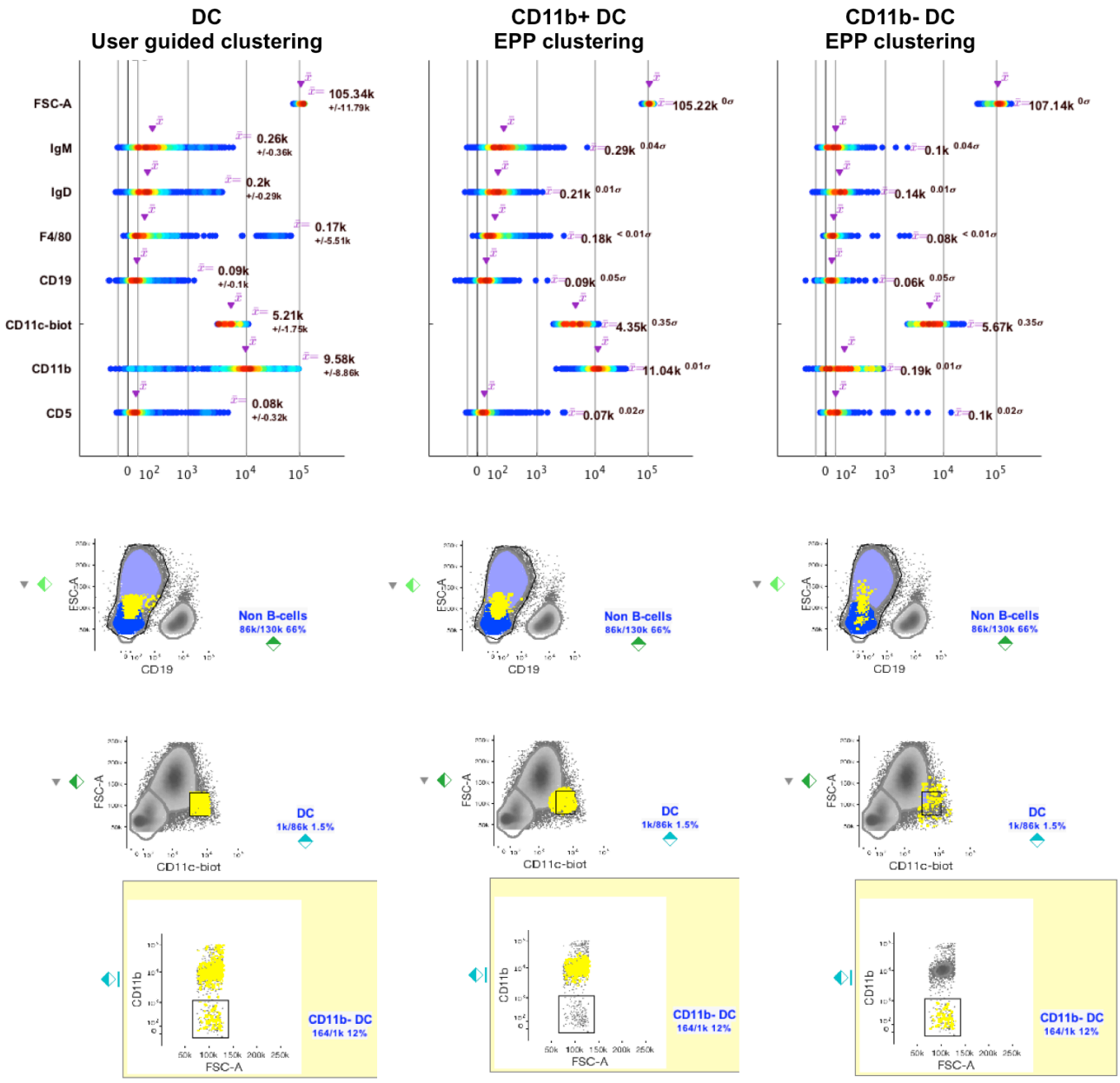
are MDS coordinates. We ran MDS on a mixture of Sample A and Sample B to display them in the same X/Y scale. Relative location of identified subsets in MDS space corresponds well with the Euclidean distances between subsets' (groups') medians presented in the table. **Panel H.** QF (quadratic form)-tree built for Sample B. To build this hierarchical tree from individual clusters we used the following modification of a multidimensional QF score [6] as a measure of dissimilarity to progressively merge clusters: QF+*c**DM, where DM is the Euclidean distance between clusters' medians and *c* is a scaling factor ensuring that the smallest QF and the biggest DM are numbers of the same order of magnitued. This branching diagram starts by placing clusters with the smallest pairwise dissimilarity scores in the lowest branches of diagram; these pairs of clusters are further progressively merged in the next branching level of the QF tree and further considered as one cluster; dissimilarity scores are then recalculated for all of the clusters on this branching level and the merging process is repeated. This process is sequentially repeated until all of the clusters identified within the sample are merged together. We named this tree structure data display as QF tree.

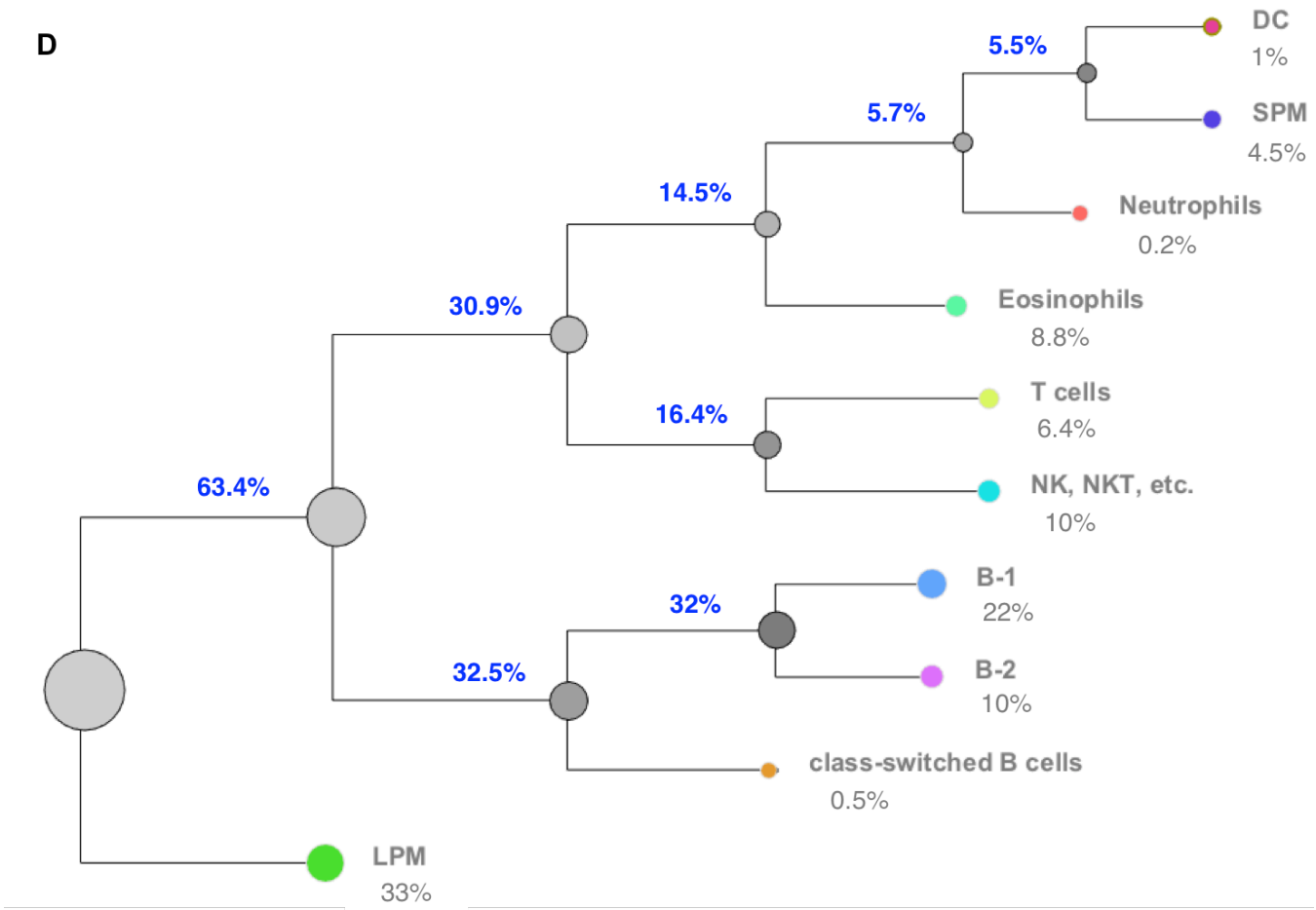

A
FSC-A
FSC-H
Singlets
FSC-A
live/dead
Live cells
FSC-A
CD19
Total B-cells
Non B-cells
IgD
IgM
B-2
B-1
class-switched B cells
FSC-A
CD11c
DC
Non DC
CD11b
F4/80
SPM
LPM
Neutrophils
Eosinophils
T, NK, etc.
CD11b
CD5
NK, NKT, etc.
T cells

B


User guided clustering
EPP clustering
MDS2 8D→2D
MDS1 8D→2D
symbol size=freq
LPM 33%
B-1 22%
NK, NKT, etc. 10%
B-2 10%
Eosinophils 8.8%
T cells 6.4%
SPM 4.5%
DC 1%
class-switched B cells 0.5%
Neutrophils 0.2%
LPM 34%
B-1 21%
B-2 11%
NK, NKT, etc. 11%
Eosinophils 9.2%
T cells 6.4%
SPM 3.5%
CD11b+ DC 1.1%
class-switched B cells 0.5%
CD11b- DC 0.3%
Neutrophils 0.1%

C


DC
User guided clustering
CD11b+ DC
EPP clustering
CD11b- DC
EPP clustering
FSC-A
IgM
IgD
F4/80
CD19
CD11c-biot
CD11b
CD5
105.34k
+/-11.79k
0.26k
+/-0.36k
0.2k
+/-0.29k
0.17k
+/-5.51k
0.09k
+/-0.1k
5.21k
+/-1.75k
9.58k
+/-8.86k
0.08k
+/-0.32k
105.22k
0.29k
0.21k
0.18k
0.09k
4.35k
11.04k
0.07k
107.14k
0.1k
0.14k
0.08k
0.06k
5.67k
0.19k
0.1k
Non B-cells
86k/130k 66%
DC
1k/86k 1.5%
CD11b- DC
164/1k 12%


D
DC
1%
5.5%
SPM
4.5%
5.7%
Neutrophils
0.2%
14.5%
Eosinophils
8.8%
30.9%
T cells
6.4%
16.4%
NK, NKT, etc.
10%
63.4%
B-1
22%
32%
B-2
10%
32.5%
class-switched B cells
0.5%
LPM
33%

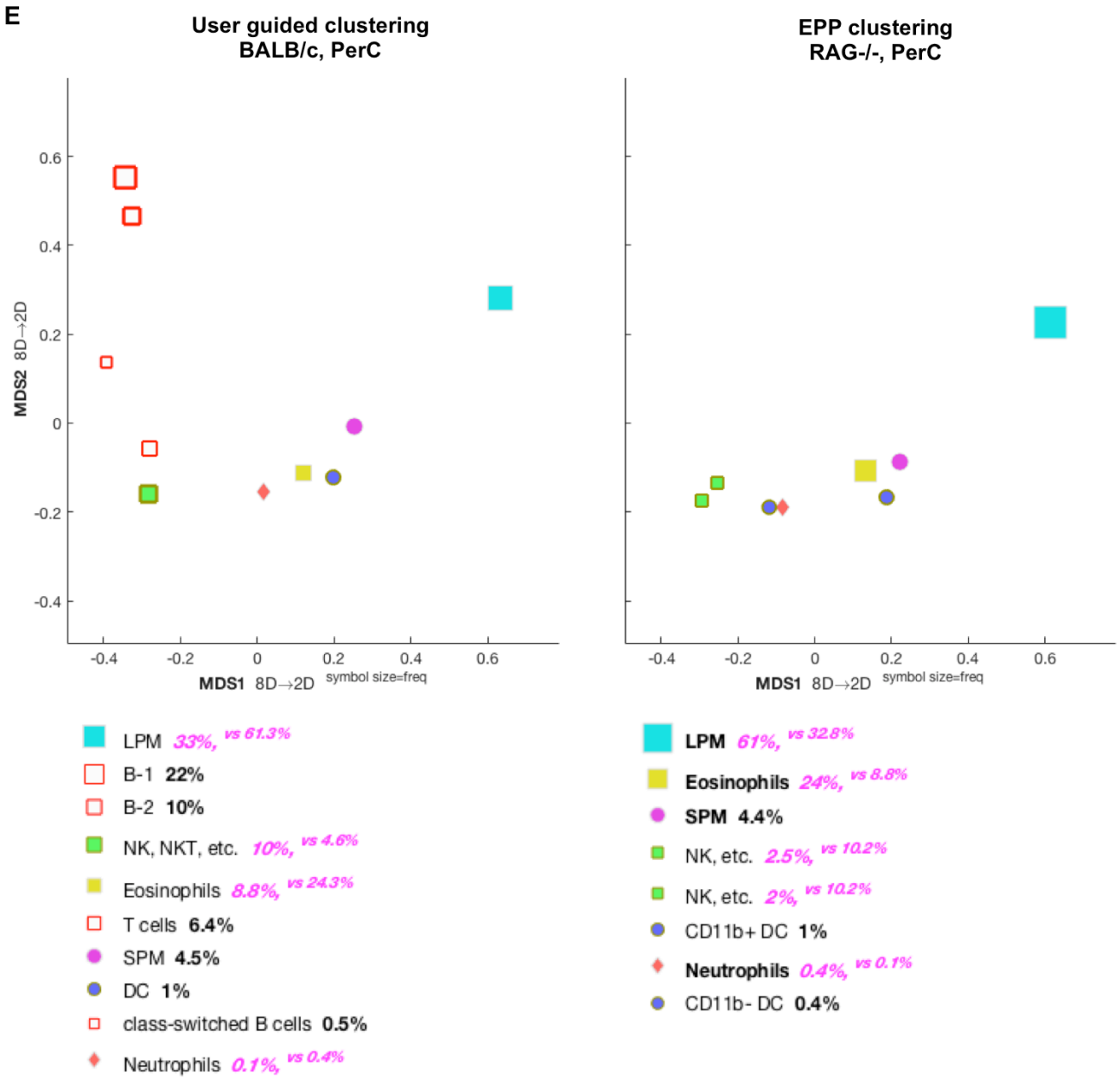
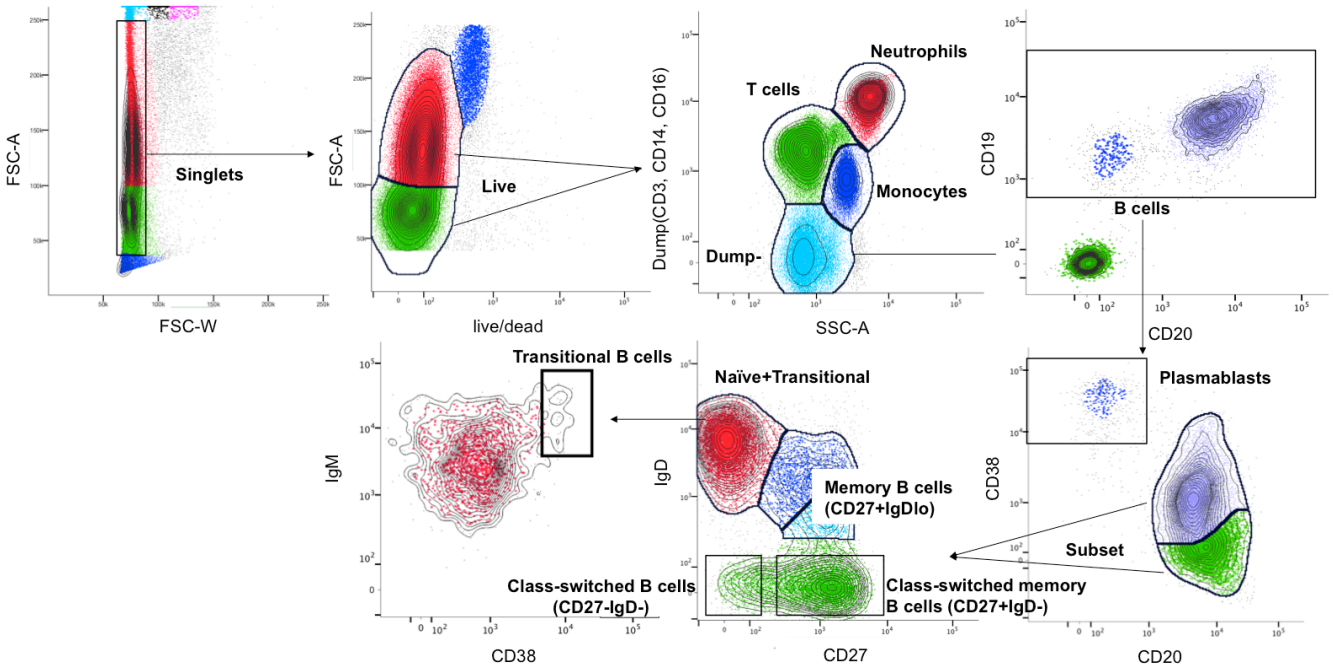


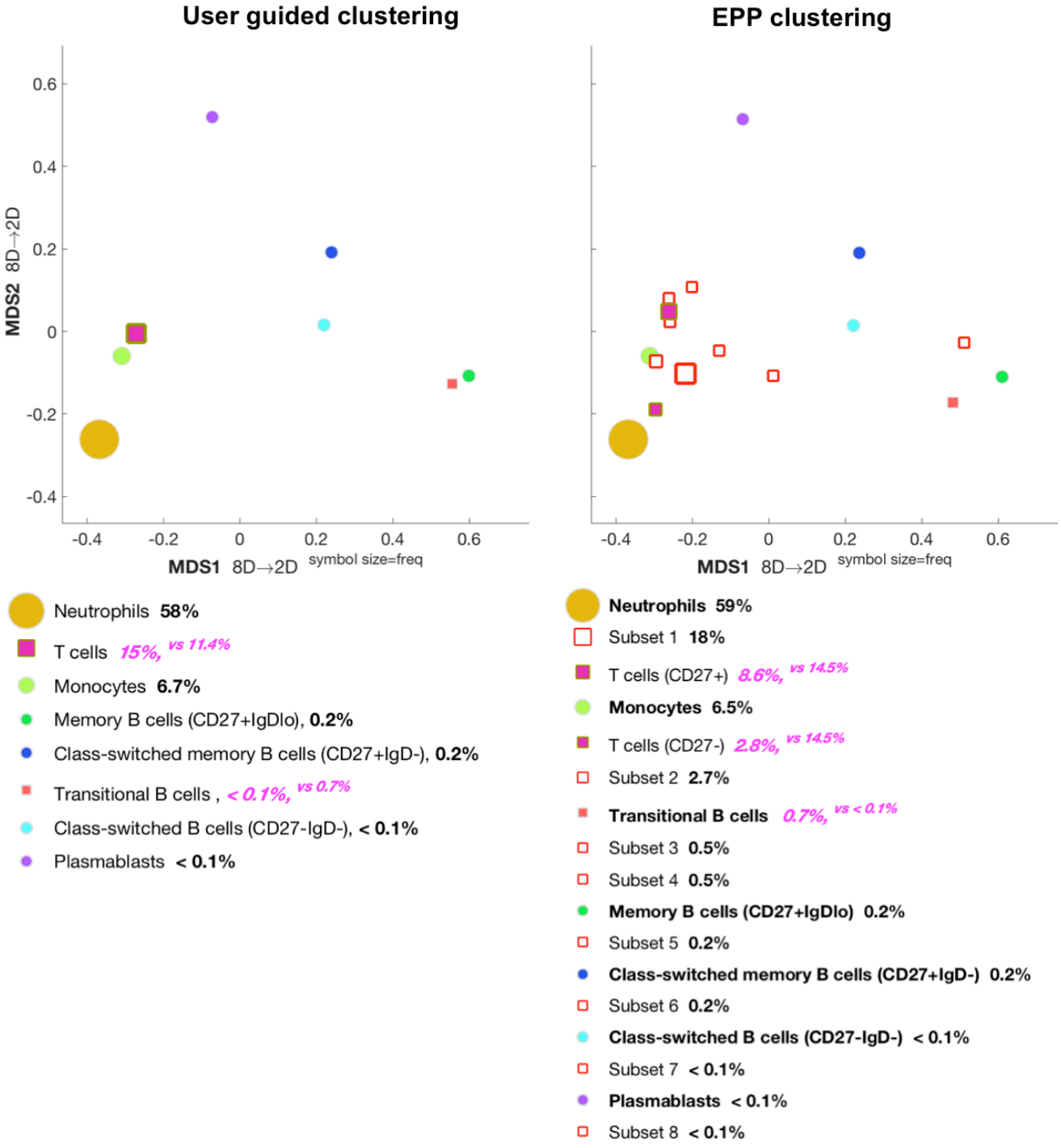
Figure 2. SIC pipeline applied to mouse peritoneal cavity flow cytometry data. Panel A. Conventional gating strategy used to identify well-known subsets of mouse peritoneal cavity (PerC) cells with user guided clustering [10]. **Panel B.** Comparison of conventional (user guided) and EPP (fully automated) clustering outcomes for wild-type (BALB/c) mouse PerC sample as displayed with MDS method (for median fluorescence values comparisons see

supplementary material for Figure 2B). EPP clusters were annotated according to multidimensional QFMatch cluster alignment between user guided and fully automated outcomes in the space of measured parameters (Forward Scatter, CD11b, CD11c, CD19, CD5, F4/80, IgD, IgM). Each circle corresponds to one identified cell subset and the size of the circle represent relative cell frequiency. Matched cell subsets are labeled with the same color. **Panel C.** Using the same set of parameters as in the manual gating strategy, SIC pipeline identified split in DC subset. “Pathfinder” tool provided by AutoGate (http://cytogenie.org/path-finder) was used to show the staining/scatter signal on measured parameters for selected cells. Pathfinder depicts each parameter with a horizontal bar that uses pseudocolor convention to show where the staining/scatter signal is most intense. Each horizontal bar is accompanied by the median value for this bar. AutoGate also supports a tool allowing to highlight (here is shown in yellow) a population of interest's location in the gating tree. **Panel D.** QF tree reflects phenotypic similarity (within the set of measured parameters) between user guided clustering results for BALB/c mouse PerC sample. The length of edges corresponds to QF score value. **Panel E.** SIC pipeline detected lack of lymphocyte compartment in $RAG^{-/-}$ mouse. Unmatched subsets (T cells, B-1, B-2 and class-switched B cells) presented as red squares. Cell subsets which medians located more than two standard deviations at least in one dimension are presented as diamond shape. Filled square shape highlights matched cell subsets with more than three percent difference in relative cell frequiency between them. Each sample (BALB/c PerC and $RAG^{-/-}$ PerC) contains 200,000 cells.

A


FSC-A
FSC-W
Singlets
FSC-A
live/dead
Live
Dump(CD3, CD14, CD16)
SSC-A
Neutrophils
T cells
Monocytes
Dump-
CD19
B cells
CD20
Plasmablasts
CD38
Subset
CD20
IgD
CD27
Naïve+Transitional
Memory B cells
(CD27+IgDlo)
Class-switched memory
B cells (CD27+IgD-)
IgM
CD38
Transitional B cells
Class-switched B cells
(CD27-IgD-)

B


User guided clustering
EPP clustering
MDS2 8D→2D
MDS1 8D→2D
symbol size=freq
0.6
0.4
0.2
0
-0.2
-0.4
-0.4
-0.2
0
0.2
0.4
0.6
Neutrophils 58%
T cells 15%, vs 11.4%
Monocytes 6.7%
Memory B cells (CD27+IgDlo), 0.2%
Class-switched memory B cells (CD27+IgD-), 0.2%
Transitional B cells , < 0.1%, vs 0.7%
Class-switched B cells (CD27-IgD-), < 0.1%
Plasmablasts < 0.1%
Neutrophils 59%
Subset 1 18%
T cells (CD27+) 8.6%, vs 14.5%
Monocytes 6.5%
T cells (CD27-) 2.8%, vs 14.5%
Subset 2 2.7%
Transitional B cells 0.7%, vs < 0.1%
Subset 3 0.5%
Subset 4 0.5%
Memory B cells (CD27+IgDlo) 0.2%
Subset 5 0.2%
Class-switched memory B cells (CD27+IgD-) 0.2%
Subset 6 0.2%
Class-switched B cells (CD27-IgD-) < 0.1%
Subset 7 < 0.1%
Plasmablasts < 0.1%
Subset 8 < 0.1%

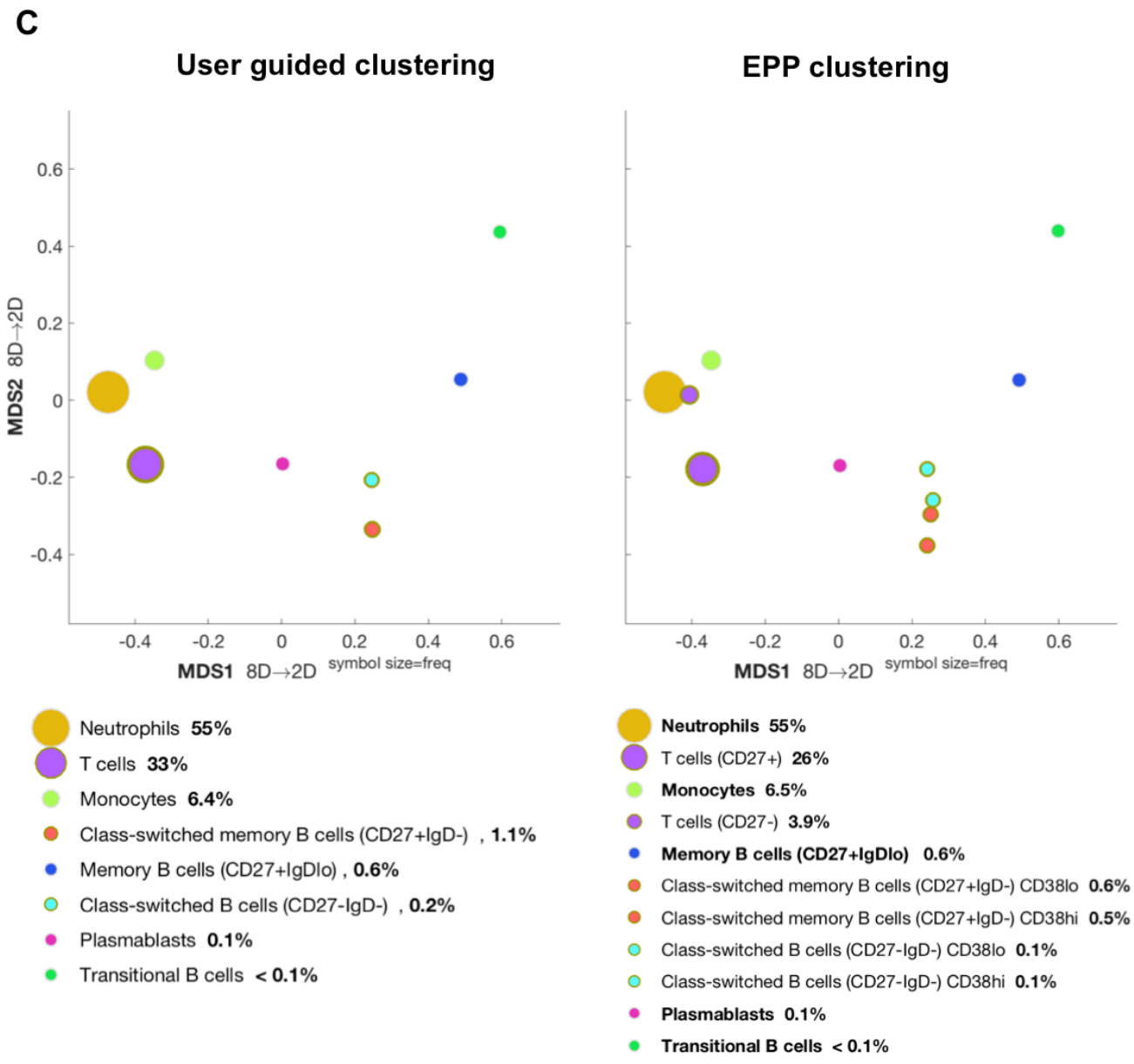
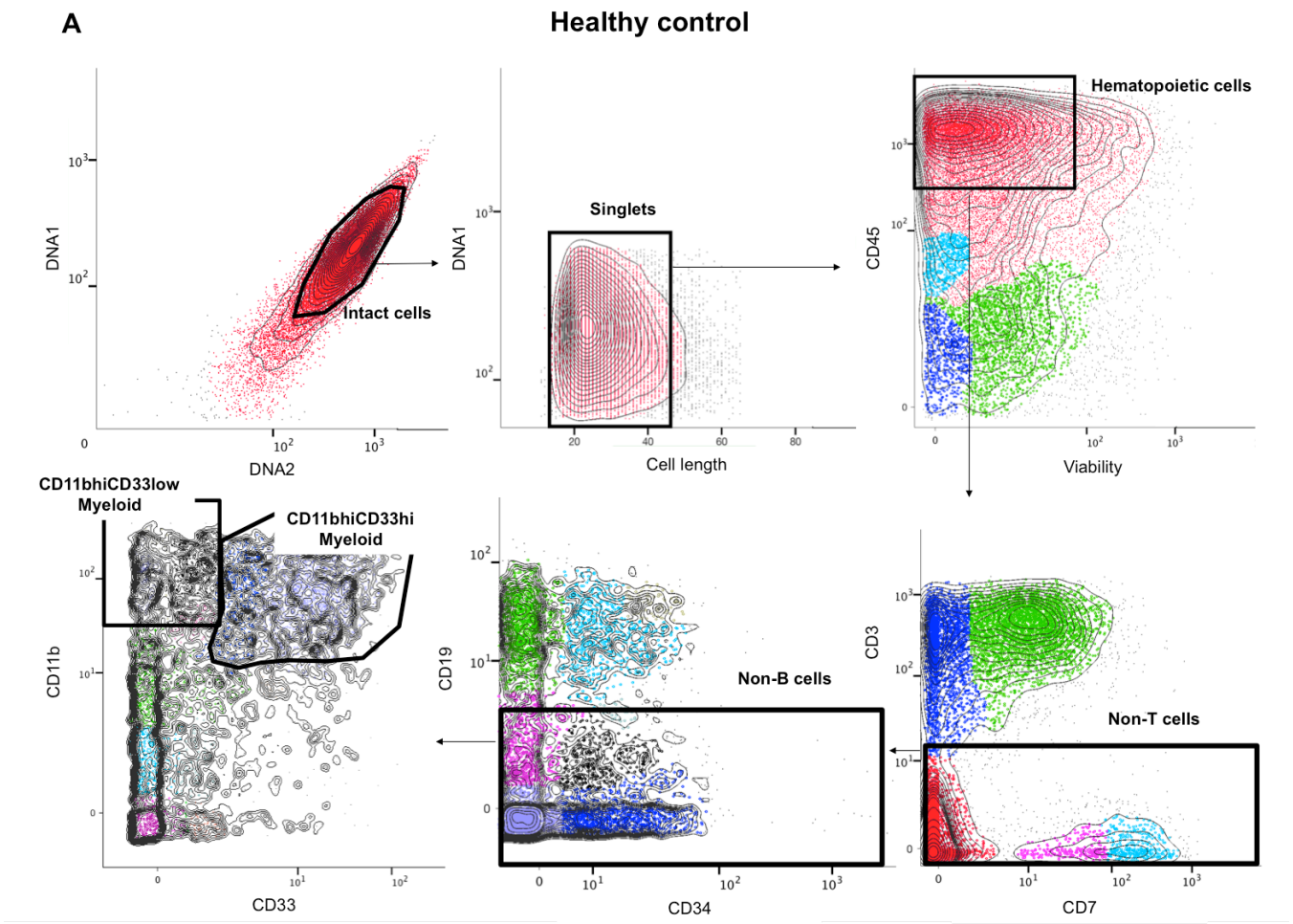
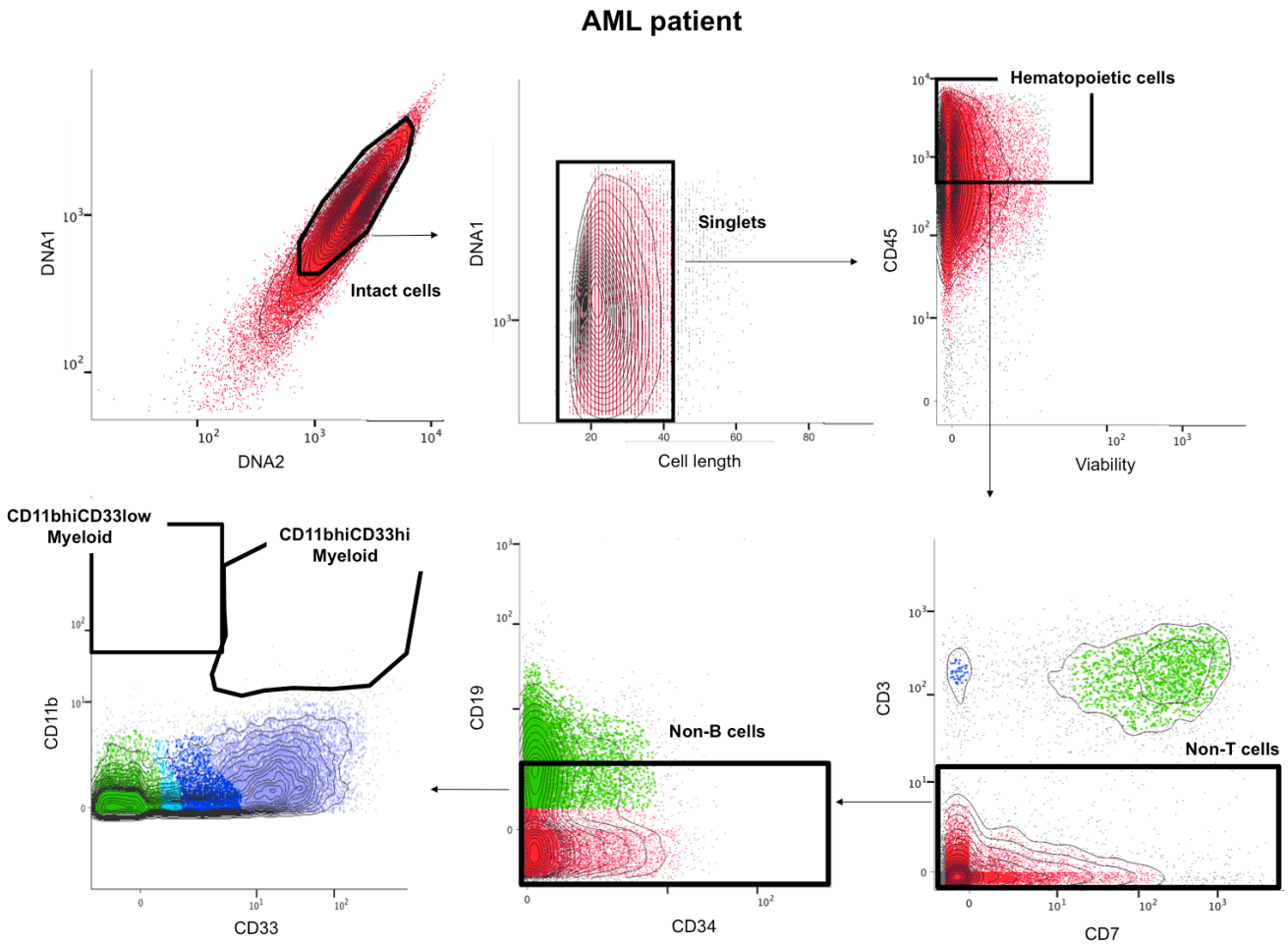


Figure 3. SIC pipeline applied to human peripheral blood flow cytometry data. Panel A. Conventional (user guided) gating strategy that we used to identify well-known human B cell subsets (naive, memory, plasmablasts, etc.) with user guided clustering. **Panel B.** Results of multidimensional QFMatch alignment between user guided and fully automated clustering outcomes for one of the samples (~200k live singlets). The following set of measured

parameters was used for user guided clustering, EPP clustering, cluster matching and data visualization of pregated live singlets: Side Scatter, Dump (CD3, CD14, CD16), CD19, CD20, CD38, CD27, IgM, IgD. Unmatched cell subsets (subsets that were not identified manually, but were identified by EPP clustering) are indicated as red squares. **Panel C.** SIC pipeline consistently identifies the main myeloid and lymphoid cell subsets in the human peripheral blood stained with the same panel of surface markers. Results are shown for a different human peripheral blood sample (~440k live singlets), and only the matched subsets are displayed.


A
Healthy control
DNA1
DNA2
Intact cells
Singlets
Cell length
CD45
Viability
Hematopoietic cells
CD11b
CD33
CD11bhiCD33low
Myeloid
CD11bhiCD33hi
Myeloid
CD19
CD34
Non-B cells
CD3
CD7
Non-T cells


AML patient
DNA1
DNA2
Intact cells
DNA1
Cell length
Singlets
CD45
Viability
Hematopoietic cells
CD11b
CD33
CD11bhiCD33low
Myeloid
CD11bhiCD33hi
Myeloid
CD19
CD34
Non-B cells
CD3
CD7
Non-T cells

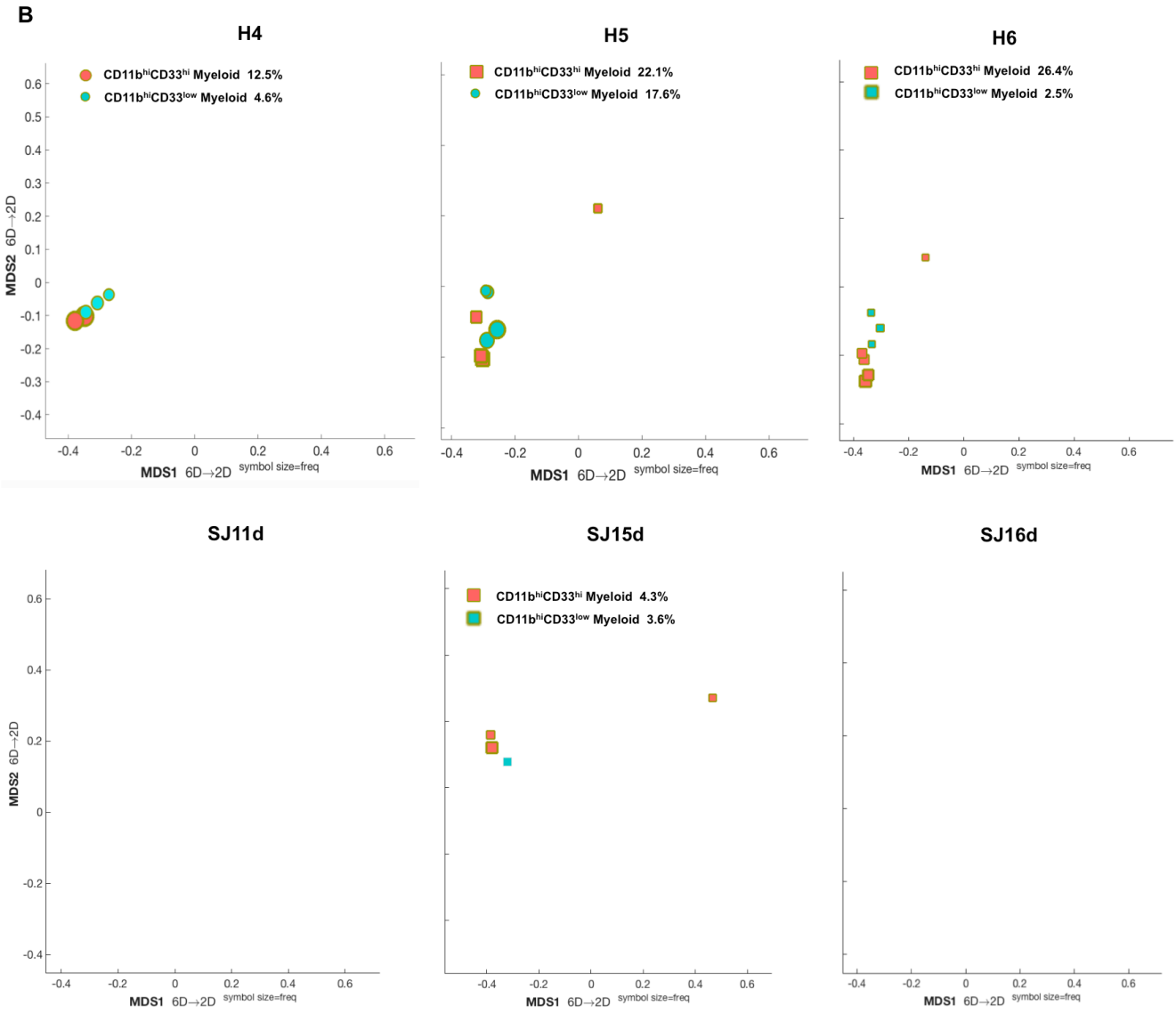


Figure 4. SIC pipeline applied to human bone marrow mass cytometry data. Panel A. Gating strategy that we used to identify CD11b^{hi}CD33^{low} and CD11b^{hi}CD33^{hi} myeloid cell subsets in healthy controls and AML patients with the user guided clustering. **Panel B.** SIC pipeline reveals the difference in CD11b^{hi}CD33^{low} and CD11b^{hi}CD33^{hi} myeloid cell subsets representation between healthy controls (H4, H5, H6) and AML patients (SJ11d, SJ15d, SJ16d).