GateFinder: Automatic Identification of Gating Strategies for Polychromatic and Mass Cytometry Data

April 16, 2018

Contents

1 Example One

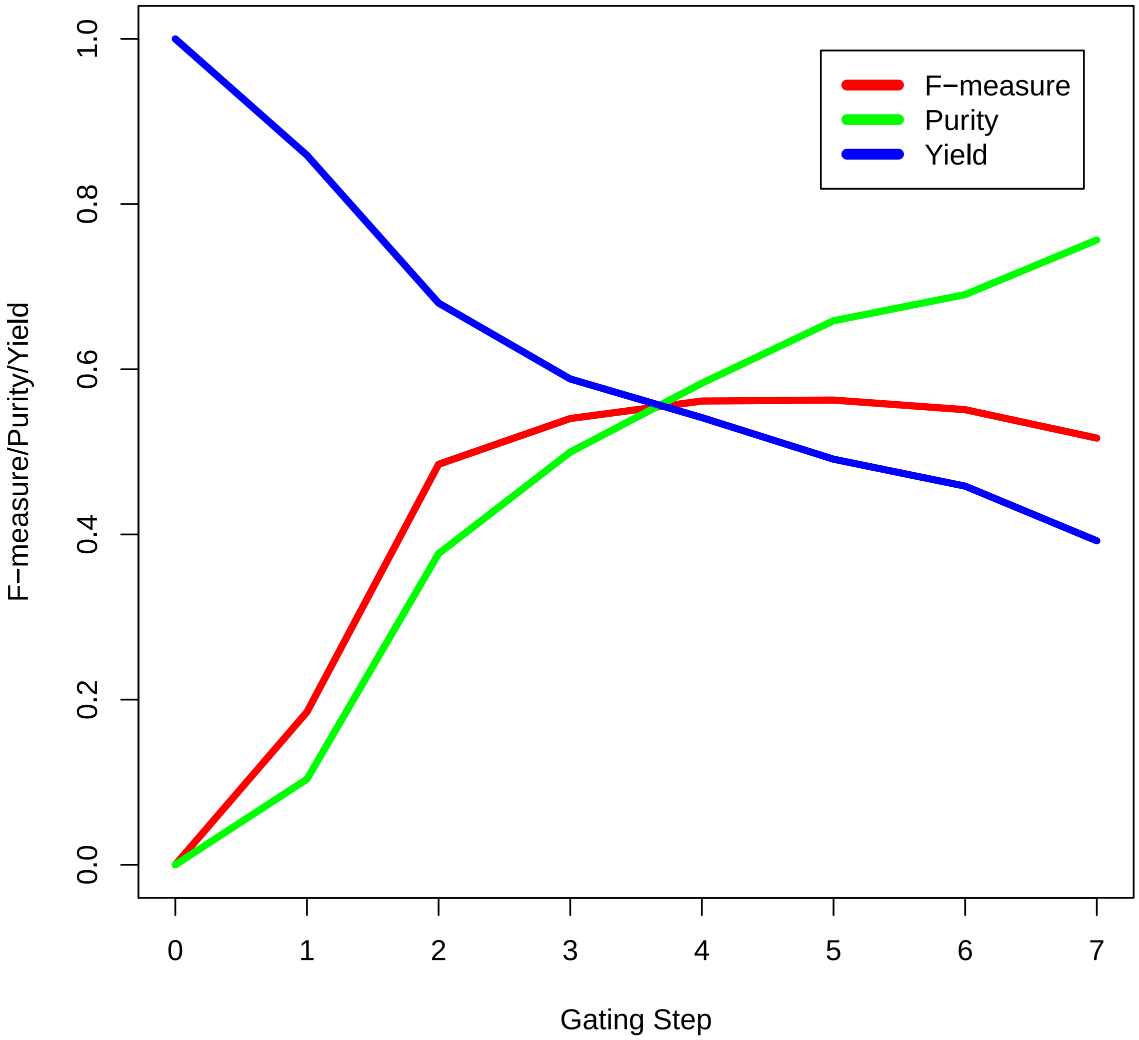


Figure S1: The F-measure, Precision, and Recall of different steps of the gating strategy in Figure 1B.

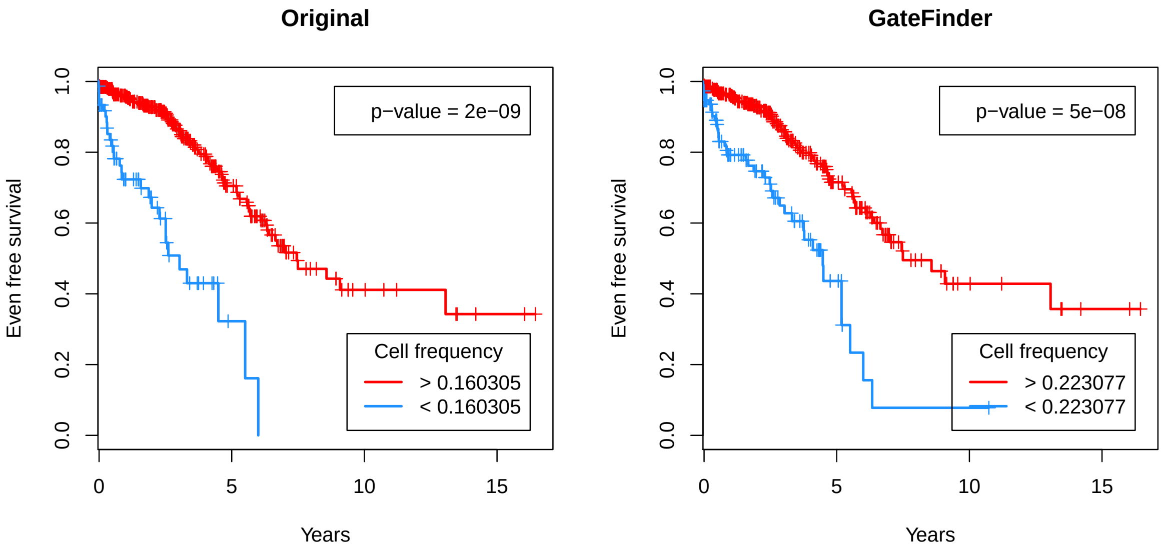


Figure S2: GateFinder (right) reproduces the Kaplan Meier curves of the original high-dimensional cell population identified by k-means clustering (left).

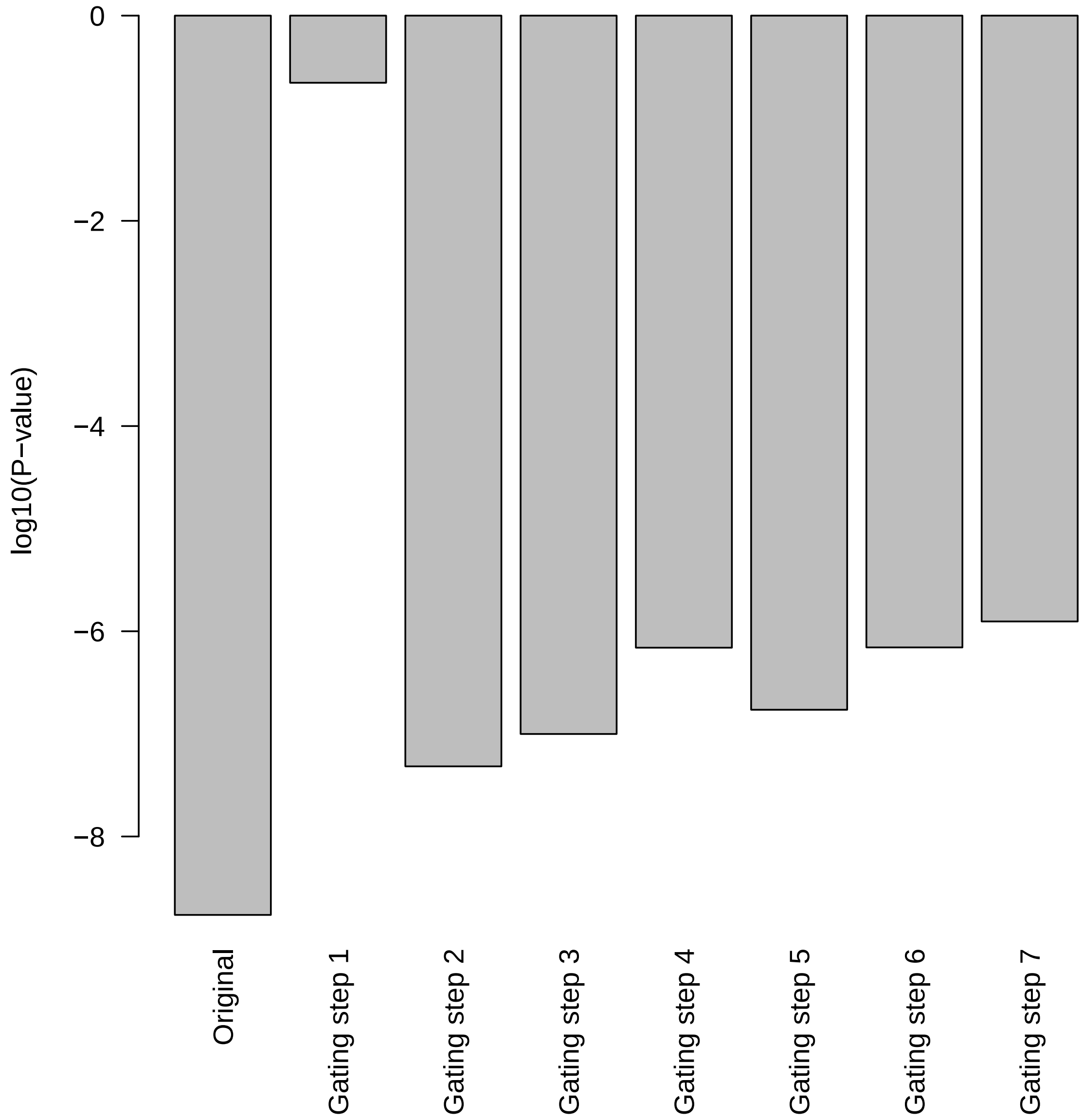


Figure S3: Correlation with clinical outcome size as a function of gating step. The Y-axis measures the correlation (Cox Proportional Hazards Ratio P-value) between the GateFinder-defined cell population and clinical outcome in the HIV dataset from Figure 1B. The P-value was recalculated for the population defined at each step of the GateFinder output. The first column represents the original Cluster 3 cell population as identified by the clustering algorithm. After only two steps most of the correlation with outcome is recovered. This is consistent with the observation in Figure S1 that the F-measure does not increase after two gating steps.

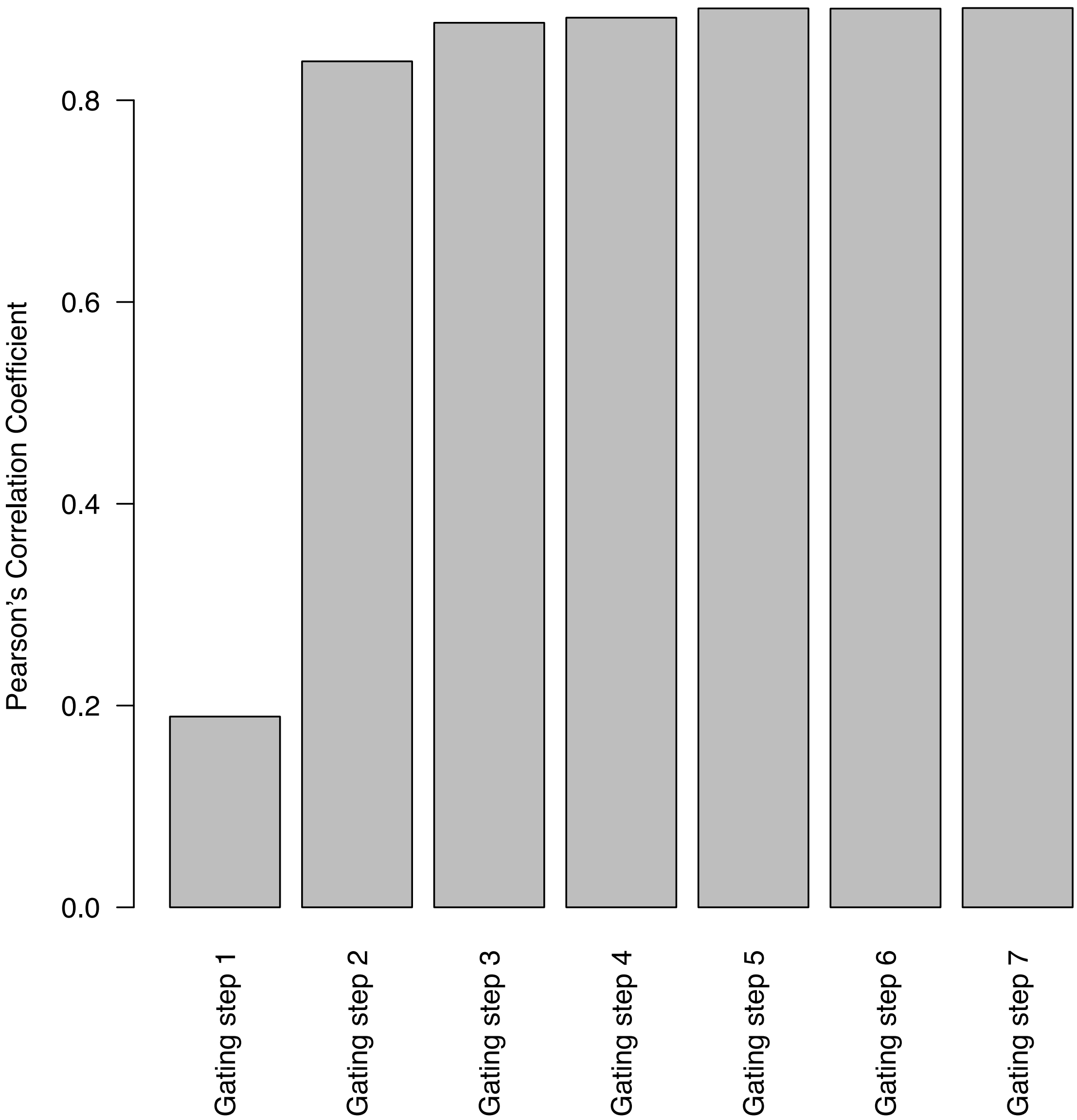


Figure S4: Correlation with target population size as a function of gating step. The Y-axis measures the correlation (Pearson's r) between the size of the GateFinder-defined cell population and the size of the original Cluster 3 cell population as identified by the clustering algorithm in the HIV dataset from Figure 1B. After only two steps, most of the correlation with the original cell population was recovered. This is consistent with the observation in Figure S1 that the F-measure does not increase after two gating steps.

2 Example Two

One application of GateFinder with great practical utility is to identify a gating strategy that uses surface markers to enrich for a cell subset that was identified based on an intracellular marker. Such a gating strategy would allow the researcher to design follow-up experiments using FACS to isolate live cells with the desired intracellular phenotype. To demonstrate this, GateFinder was used to identify the optimal combination of cell surface markers that could discriminate those cells within a human bone marrow sample in which levels of the activated transcription factor STAT5 (pY694) ("pSTAT5") increased after short-term exposure to granulocyte colony-stimulating factor (G-CSF). This cell population is of clinical interest in the context of adult acute myeloid leukemia [1]. A surface marker-based signature for the pSTAT5+ cells would allow prospective isolation of viable cells for further characterization. Density gradient-enriched bone marrow cells were treated with G-CSF or vehicle control and incubated with a panel of antibodies against 30 surface markers and pSTAT5 (see Supplemental Methods Table S1 for full list of markers). The pSTAT5+ G-CSF-responsive cells were gated manually (Figure S5, left panel) and specified as the target population in GateFinder. While the computed gating strategy included 15 gates, only two gates (four markers) were required to isolate this cell subset before the F-measure plateaued at 0.63 (Figure S6). This indicated that the first two gates had exhausted the unique information contributed by the 30 surface markers (Figures S7 and S8). These two gates (CD45+/CD36+ and CD44+/CD64+) described a monocytic phenotype of the G-CSF-responsive cells. After both gates were applied, 61.4% of the captured cells were pSTAT5+, compared with 13.5% in the ungated sample, representing a 4.5-fold enrichment (Figure S5, right panel) even though pSTAT5 itself was not used in the gating. Application of these gates to mass cytometry data from a second healthy donor required manual adjustment of the gates due to staining differences, but ultimately achieved similar enrichment of the cells (Figures S9, S10, S11, and S12), suggesting that the gating strategy is generalizable across individuals.

To test whether the computed mass cytometry gating strategy can guide prospective isolation of GCSF-responsive cells by FACS, five-color optical flow cytometry analysis of bone marrow from a third healthy donor was performed (Figure S13A). Although the population distributions in the flow cytometry data differed from the mass cytometry data in terms of signal-to-noise and spectral overlap, the gating strategy could be translated using the shape of the cell populations as a visual landmark (Figure S13B). As in the two mass cytometry datasets, the gated CD45+/CD36+/CD44+/CD64+ population achieved a 4.5-fold enrichment in frequency of G-CSF-responsive pSTAT5+ cells (Figure S13C). In this second example, GateFinder distilled a 30-parameter dataset into a concise 4-parameter phenotype for a single target population, enabling downstream experiments to further understand the function of these cells.

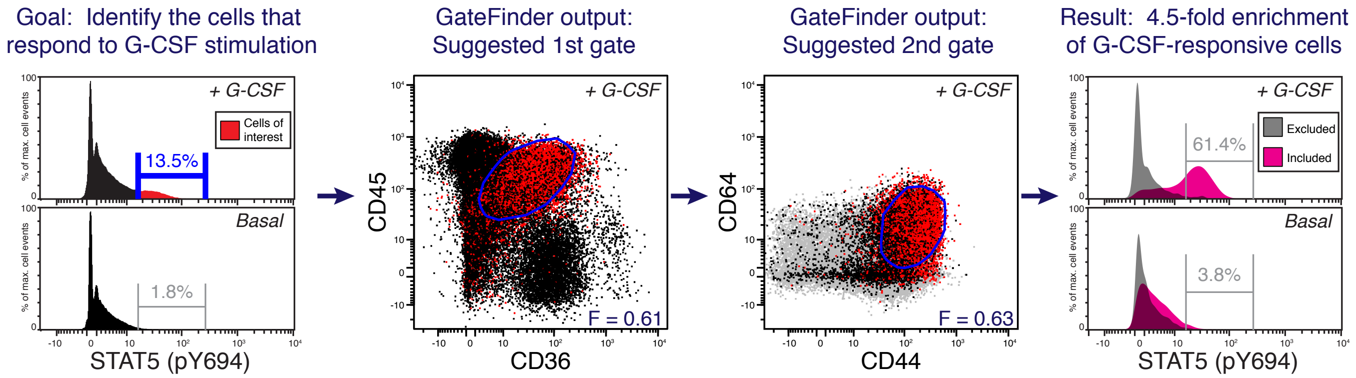


Figure S5: In Example Two, the target population was defined as the bone marrow cells that activate phosphorylated STAT5 in response to short-term stimulation with G-CSF (left panel). GateFinder searched all pairwise combinations of 30 surface parameters in this mass cytometry dataset and selected an optimal gating strategy (center panels). Applying only the first two gates, the frequency of G-CSF-responsive cells was enriched by 4.5-fold relative to the ungated dataset (compare top histograms in left and right panels).

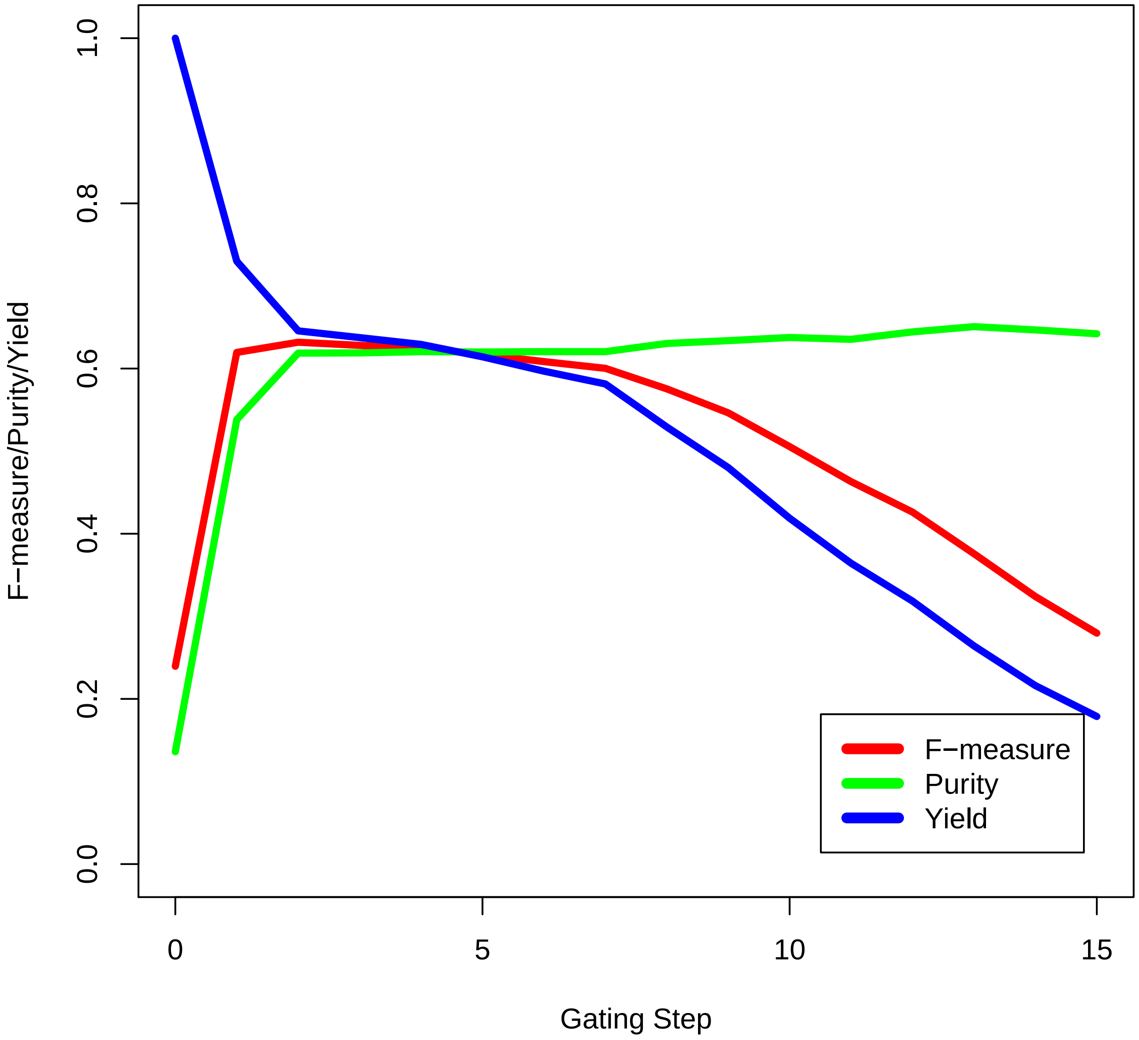


Figure S6: The F-measure, Precision, and Recall of each gating step.

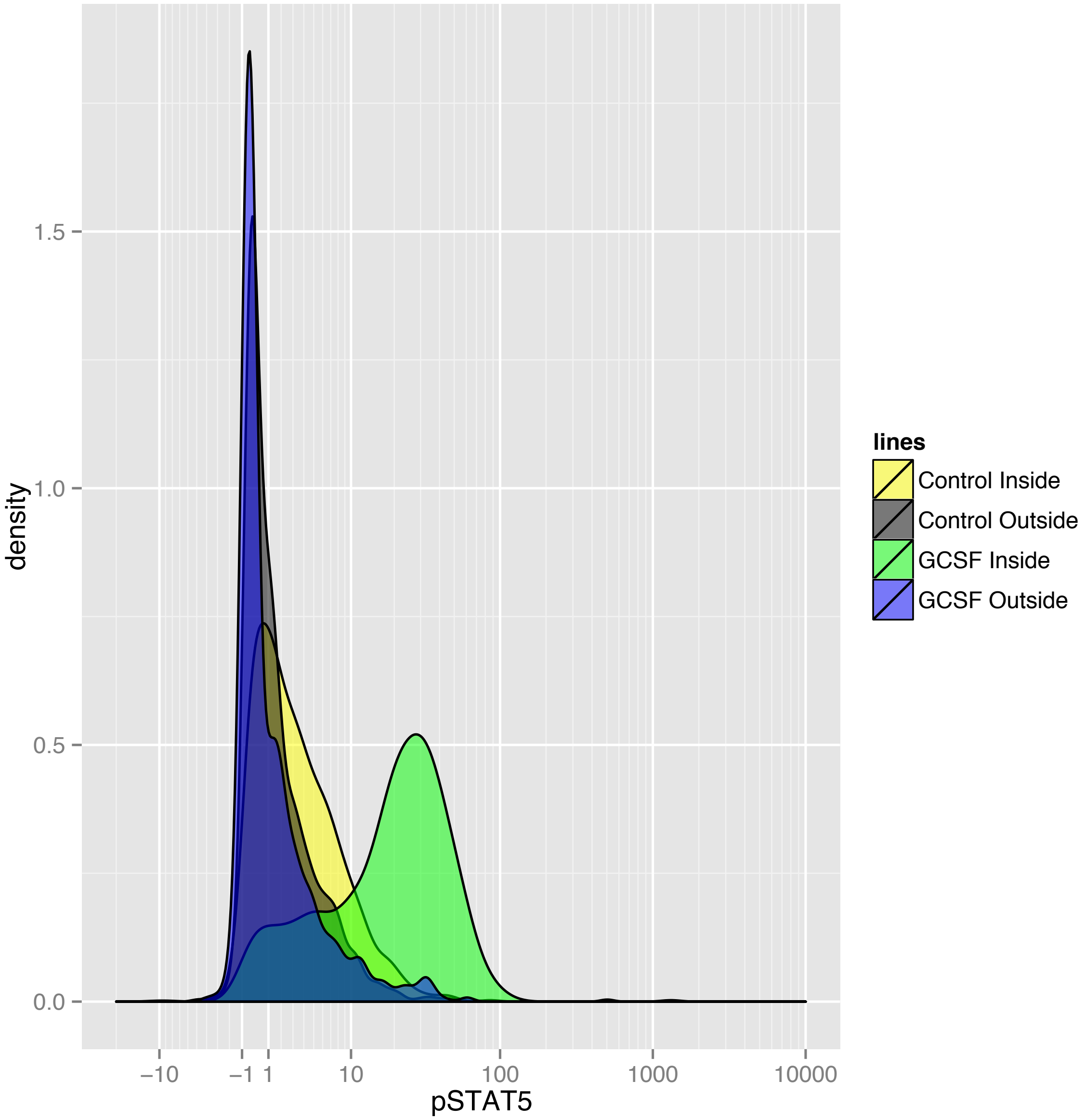


Figure S7: The density of pSTAT5 for four different cell populations after two gating steps. The cells selected using the GateFinder polygons have a significantly higher median expression than control cells as well as GCSF stimulated cells outside the polygons. The gating strategy was limited to two steps. A similar analysis using all gates produced similar results (Figure S8).

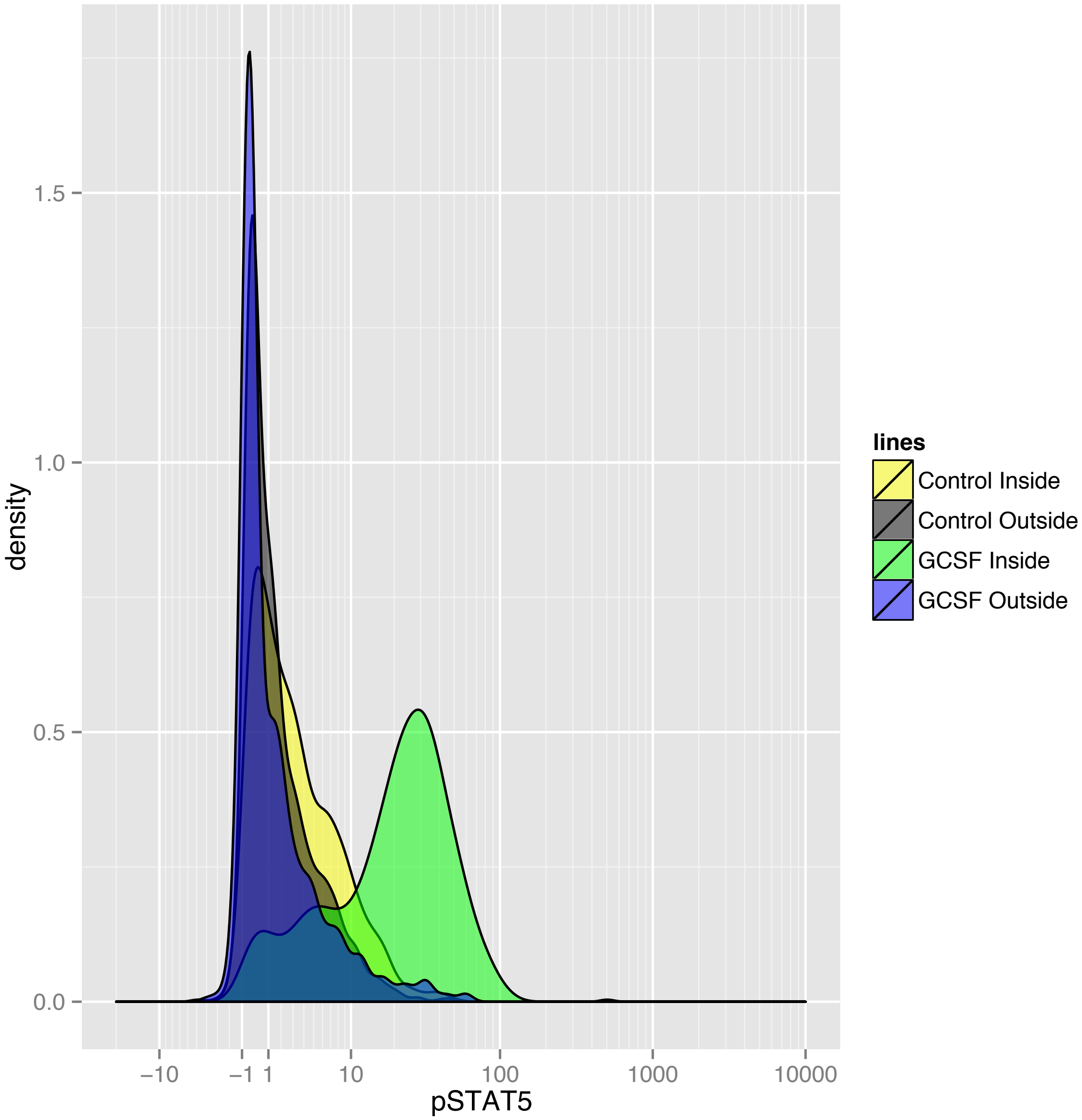


Figure S8: The analysis in Figure S7 reproduced using all gating steps.

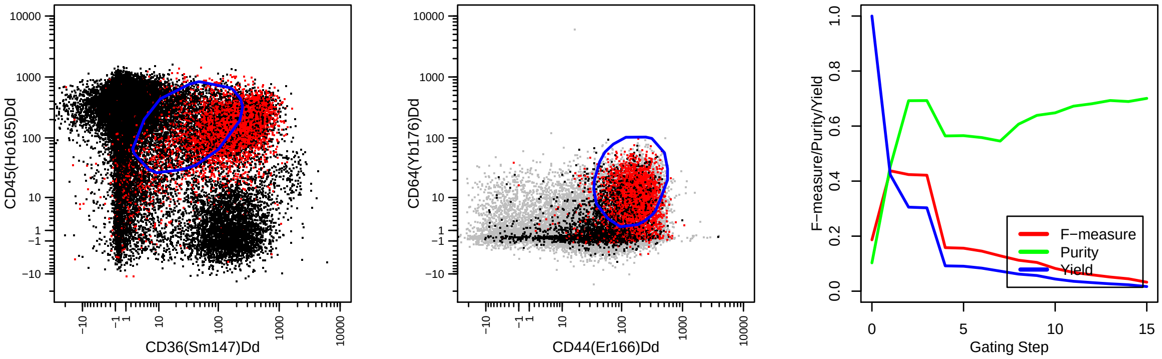


Figure S9: When the CD45+/CD36+ and CD44+/CD64+ gates were applied to a bone marrow sample from a second healthy donor, the initial alignment of the first gate was poor. The poor fit is due to variable staining between experiments. This common occurrence can be corrected either manually or semi-automatically. Manual adjustment of the gates to match the staining patterns in the second donor, achieved a similar enrichment of the target cells without using the information from pSTAT5 (Figures S10 and S11). Alternatively, semi-automatic adjustment of the gates can be performed by specifying the pSTAT5+ target population in the second donor and configuring GateFinder to mimic the first gating strategy (Figure S12). In both cases, the identified cell population validated the original findings with a 4.5-fold enrichment for the target cell population.

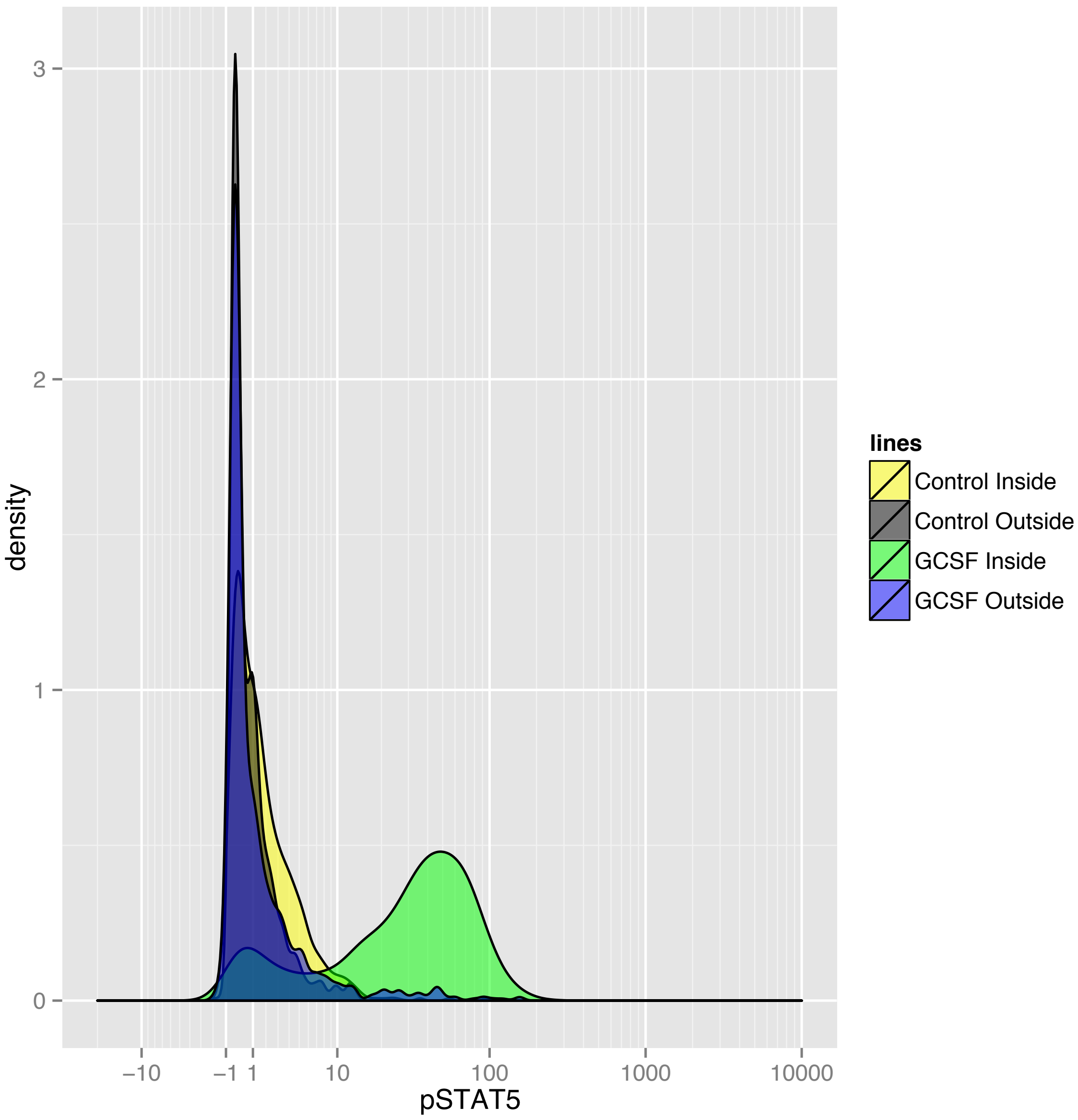


Figure S10: The density of pSTAT5 for the cell population identified using the manual gates in Figure S11 to confirm the signaling patterns of the identified population in a new donor.

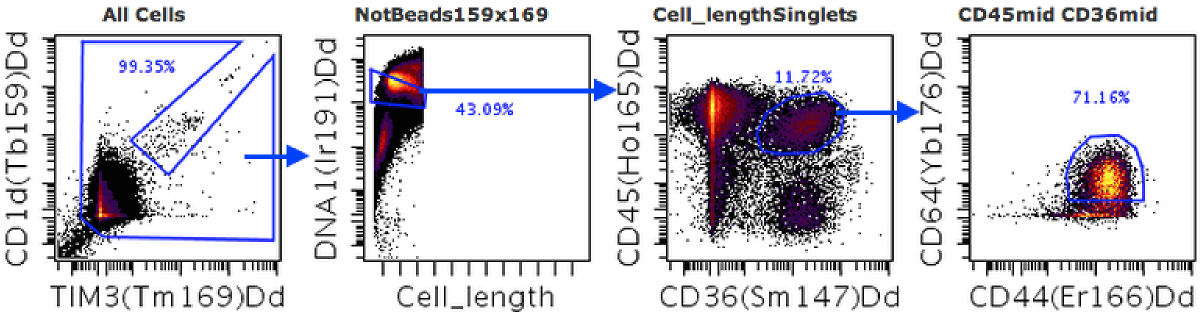


Figure S11: The cell population from Figure 1B identified in a new donor using manual gates guided by Figure S9 to account for cross-sample variations.

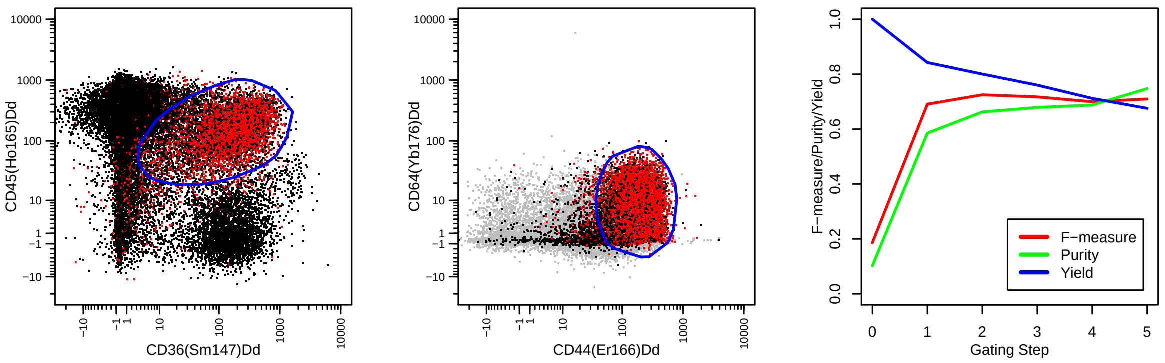


Figure S12: The cell population from Figure 1B identified in a new donor by providing the order of the scatter plots in the original samples to GateFinder and recalculating the gates to automatically adjust the gates for technical variations.

A

Antibody	Vendor	Catalog No.	Clone	Staining performed
V450 anti-human CD44	BD Biosciences	561292	G44-26	After fix
FITC anti-human CD36	BD Biosciences	555454	CB38	After fix
PE anti-human STAT5 (pY694)	BD Biosciences	612567	47	After fix + perm
Alexa647 anti-human CD64	Biolegend	305012	10.1	After fix
APC-Cy7 anti-human CD45	Biolegend	304014	HI30	After fix + perm

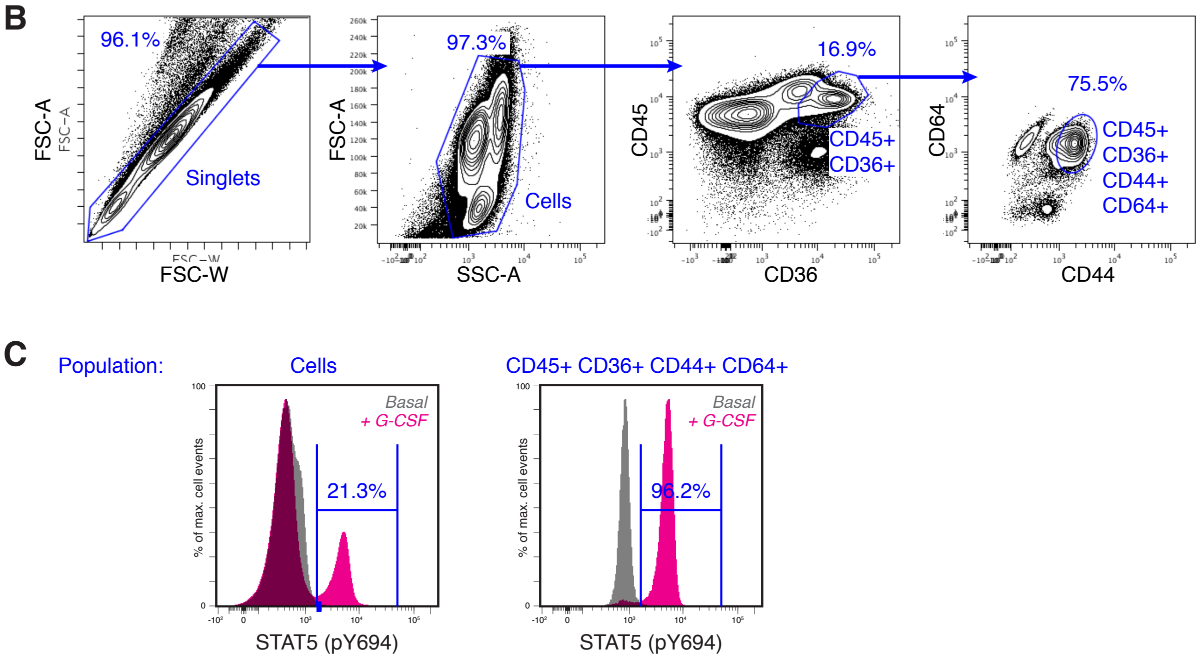


Figure S13: (A) Five-color fluorescence staining panel for validation of GateFinder gating strategy suggested in Figure 1B. (B) Fluorescence cytometry analysis of independent BMMC sample. Data was collected on an LSR-II and gates were drawn based on the GateFinder suggested gates from Figure 1B in the main text. Data from Basal sample is shown. Identical gates were used for the G-CSF stimulated sample. (C) Enrichment of G-CSF-responsive cells in fluorescence cytometry data. Populations were defined by serial gating strategy shown in (B). Frequencies of pSTAT5+ cells in the G-CSF-stimulated sample are shown.

3 Example Three

In more complex experiments, the researcher may seek a concise set of markers that can capture multiple target populations. Examples may include designing a FACS experiment to sort multiple populations from a single sample, or selecting the minimal set of markers to use in a clinical diagnostic or drug companion biomarker assay. To illustrate the utility of GateFinder in such scenarios, we used GateFinder to identify a gating strategy that captures 5 well-defined immunological cell types in healthy human peripheral blood. The 5 target populations were first selected from a 31-marker mass cytometry dataset using Spanning Tree Progression Analysis of Density Normalized Events (SPADE)[2], a high-dimensional clustering and visualization tool for cytometry (Figure S14, left panel).

To select an optimal subspace of markers for distinguishing these 5 cell types, we set the GateFinder algorithm to undertake a supervised random forest-based feature selection step (Figures S15, S16, and S17). With GateFinder configured to select the best 4 markers (of the available 31) to identify these 5 cell populations, CD8, CD16, CD33, and HLA-DR were selected by the algorithm. Restricted to this subspace of 4 markers, the gating strategy suggested by GateFinder captured all 5 target populations at a purity of >90% purity (Figure S18). A purity matrix defining the contribution of each SPADE-gated population to each gated region can be represented as a heatmap (Figure S14, right panel - marker distributions are available in Figure S19). Including more surface markers further improved the separation between these populations (Figure S18). We note here that to compute a minimalist gating strategy for five target populations using only four markers, the algorithm has excluded canonical markers of the target populations to replace them with markers that can contribute to identification of all target cell types. For example, rather than selecting CD19 or CD20 for identification of B cells, the algorithm selected a combination of two non-canonical markers (HLA-DR and CD33) because these markers were also useful for identifying monocytes. Designing concise staining panels and gating strategies for multiple cell populations is a challenging task, even for well-defined systems such as peripheral blood. In this third example, GateFinder automated the process of finding an efficient and practical gating strategy.

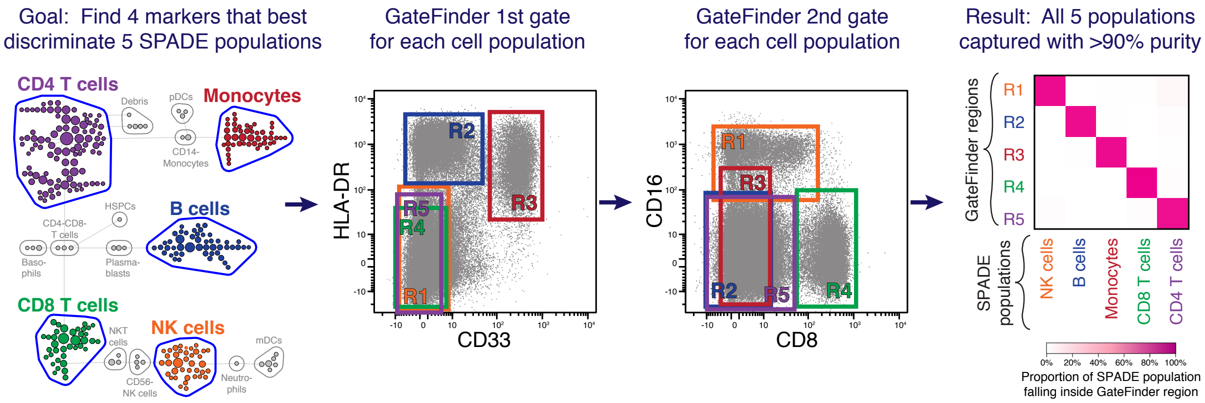


Figure S14: In Example Three, the SPADE clustering algorithm was used to select 5 mutually exclusive target cell populations in a mass cytometry dataset of human peripheral blood (left panel). GateFinder identified a gating strategy for each cell population, using a consistent set of only 4 markers determined using a machine-learning approach (center panels). Rectangular gates are shown here for clarity (see Figure S18 for actual gates). Using only two gates and a restricted set of markers, each population was discriminated with >90% purity. A heatmap summarizes the degree of overlap between each gold-standard SPADE population and each GateFinder-defined population (right panel).

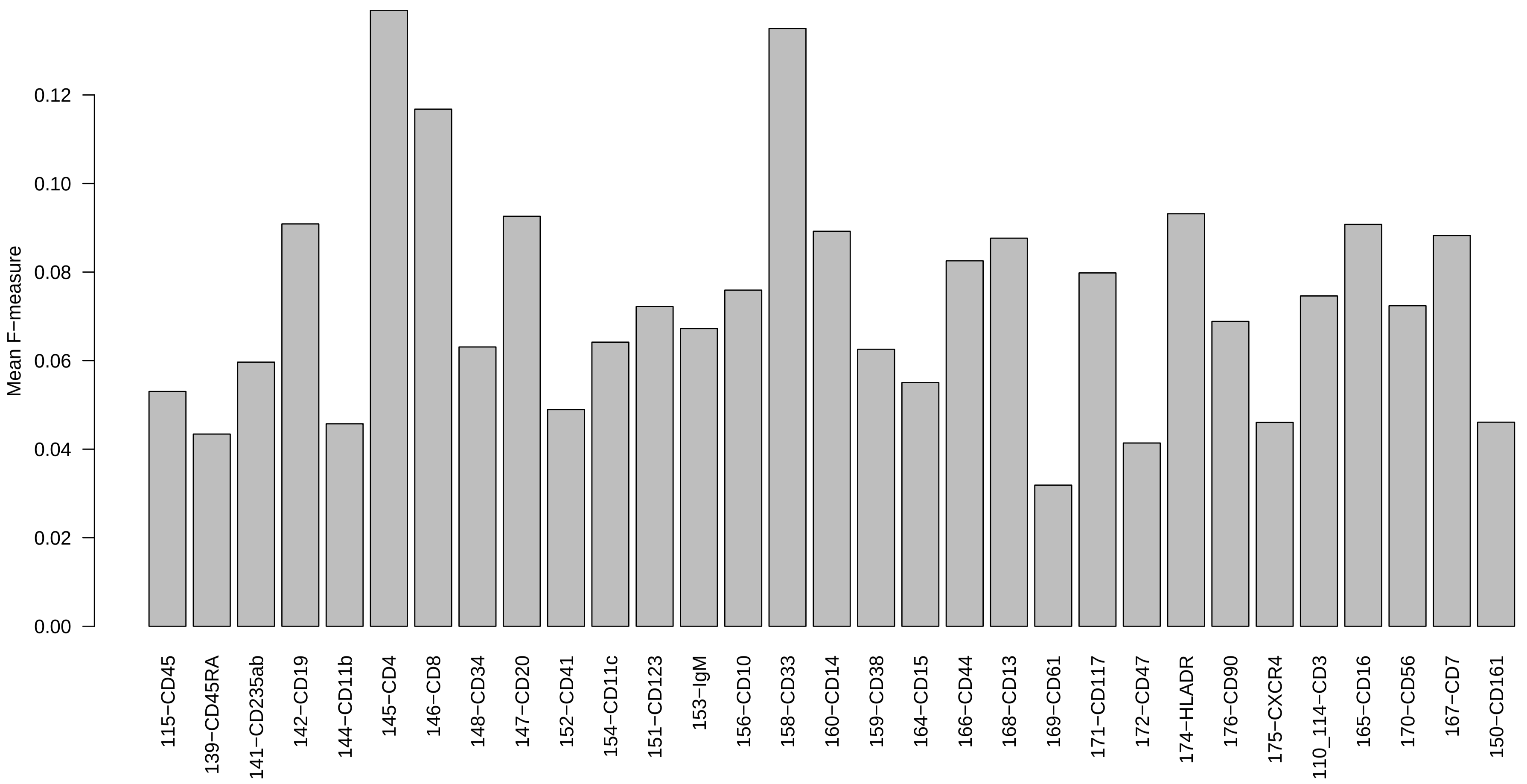


Figure S15: The importance of each marker to a supervised random forest model for separating the target population. However, there could be redundancies between the information provided by combinations of these markers. A backwards-elimination process was used to address this issue. For this example, a gating strategy using four markers was desired. Eight top marker from this plot were selected for further analysis using the backwards-elimination algorithm to eliminate redundant markers (see Figure S16).

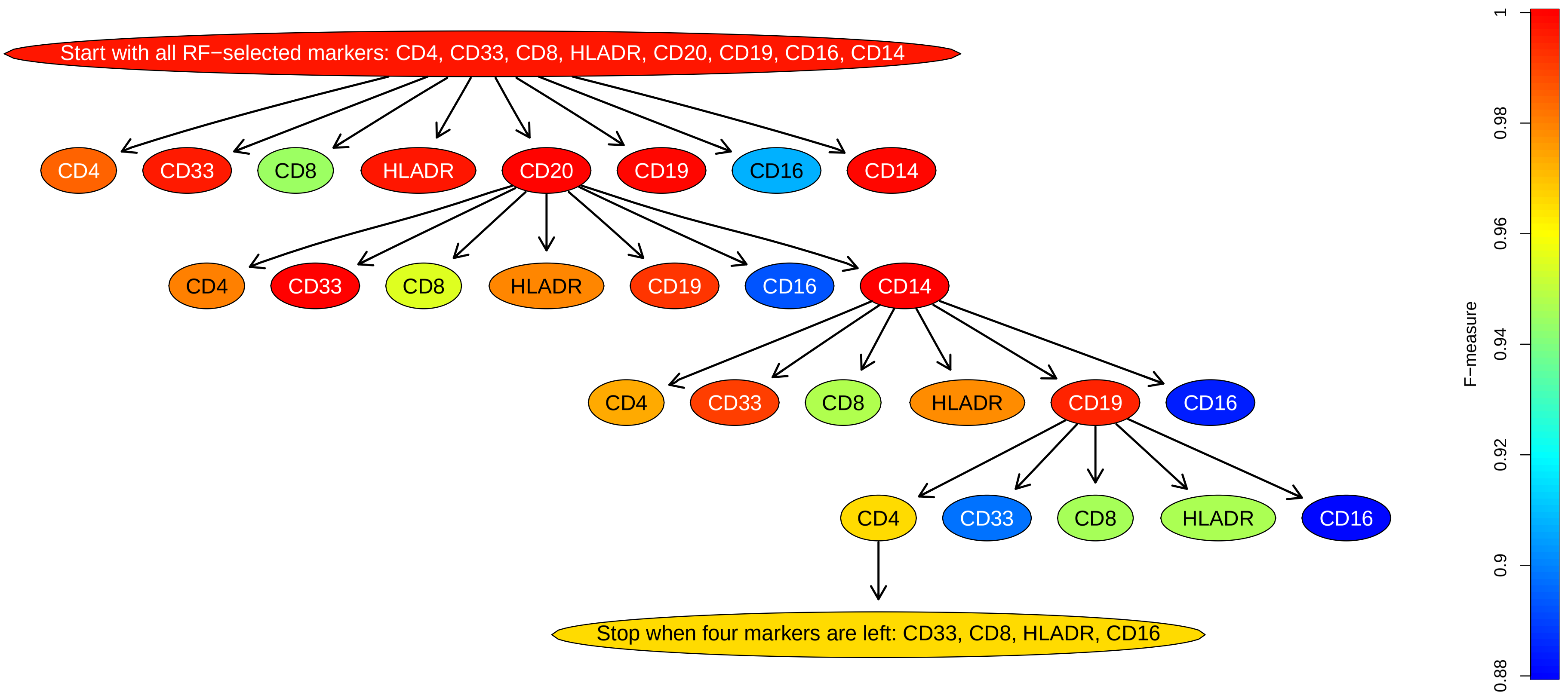

Start with all RF−selected markers: CD4, CD33, CD8, HLADR, CD20, CD19, CD16, CD14
CD4
CD33
CD8
HLADR
CD20
CD19
CD16
CD14
CD4
CD33
CD8
HLADR
CD19
CD16
CD14
CD4
CD33
CD8
HLADR
CD19
CD16
CD4
CD33
CD8
HLADR
CD16
Stop when four markers are left: CD33, CD8, HLADR, CD16
F−measure
1
0.98
0.96
0.94
0.92
0.9
0.88

Figure S16: Backwards elimination of the redundant markers selected by the feature selection algorithm. The analysis starts will all selected markers included. Each intermediate node corresponds to an iteration of the analysis using all remaining markers except for one (the node's label). The color of each node represents the average F-measure of the gated target populations using the remaining markers. The process continues until a pre-specified number of markers (in this case, 4) are left.

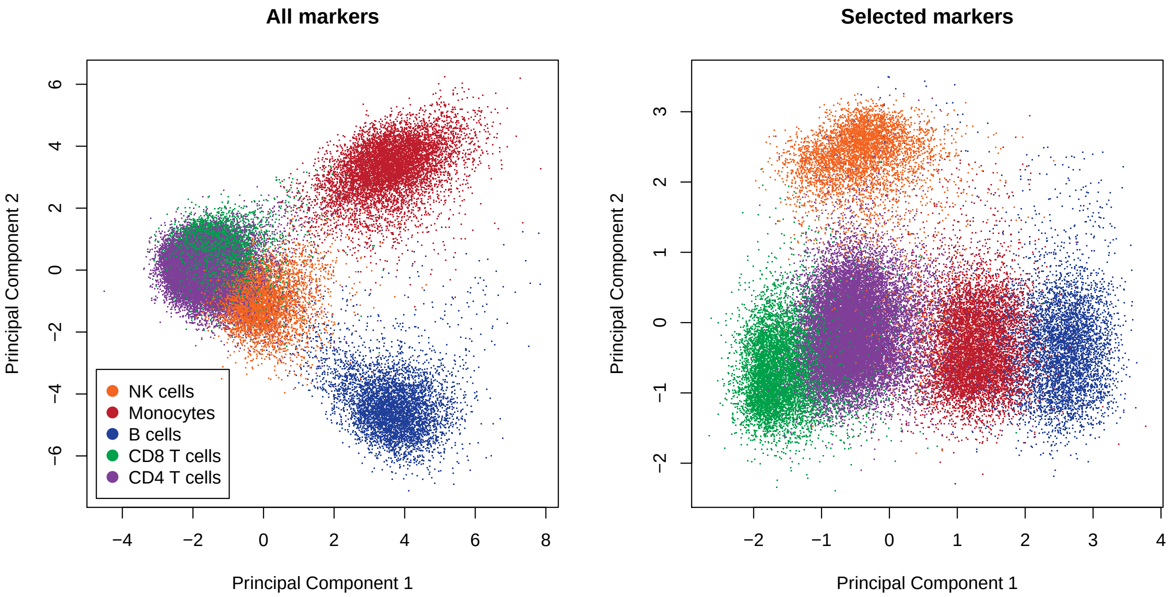


Figure S17: PCA visualization of all available surface markers (left) and versus the four markers which best separate the target populations (right). This demonstrates that the sub-space selected by RF and backwards elimination better separates the target populations than even all available markers combined.

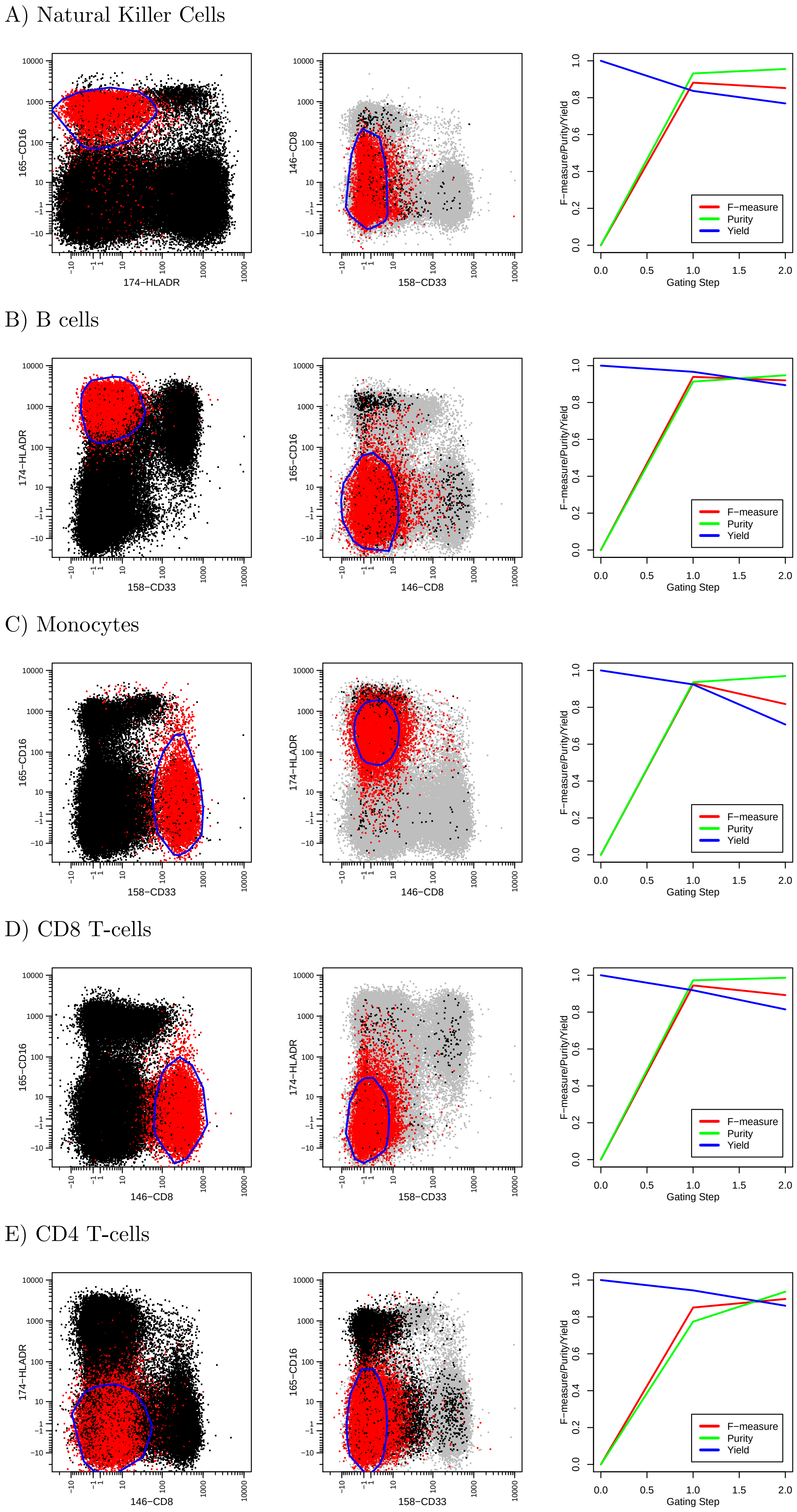


Figure S18: The identified gating strategies for the five populations in Figure S5.

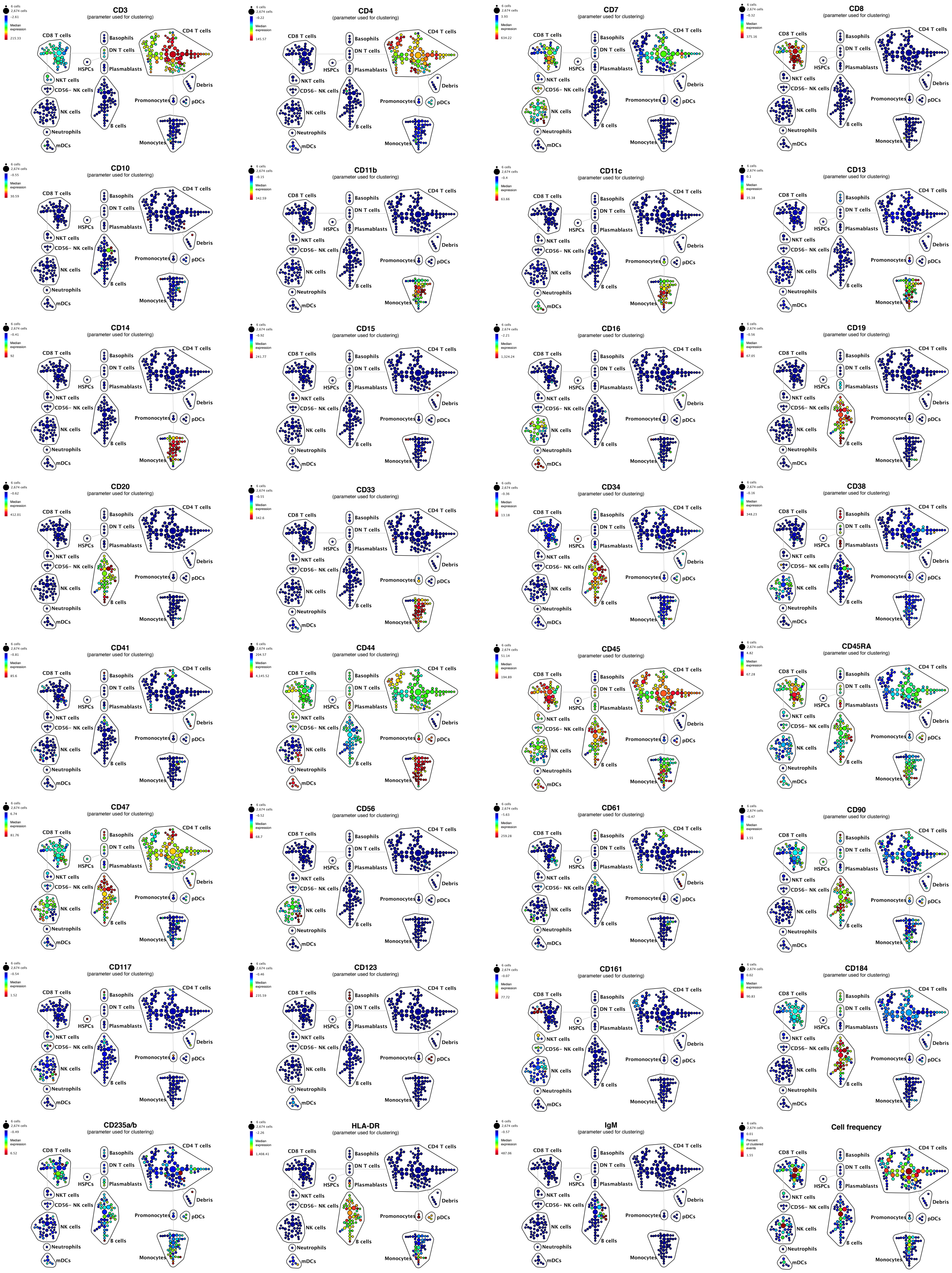
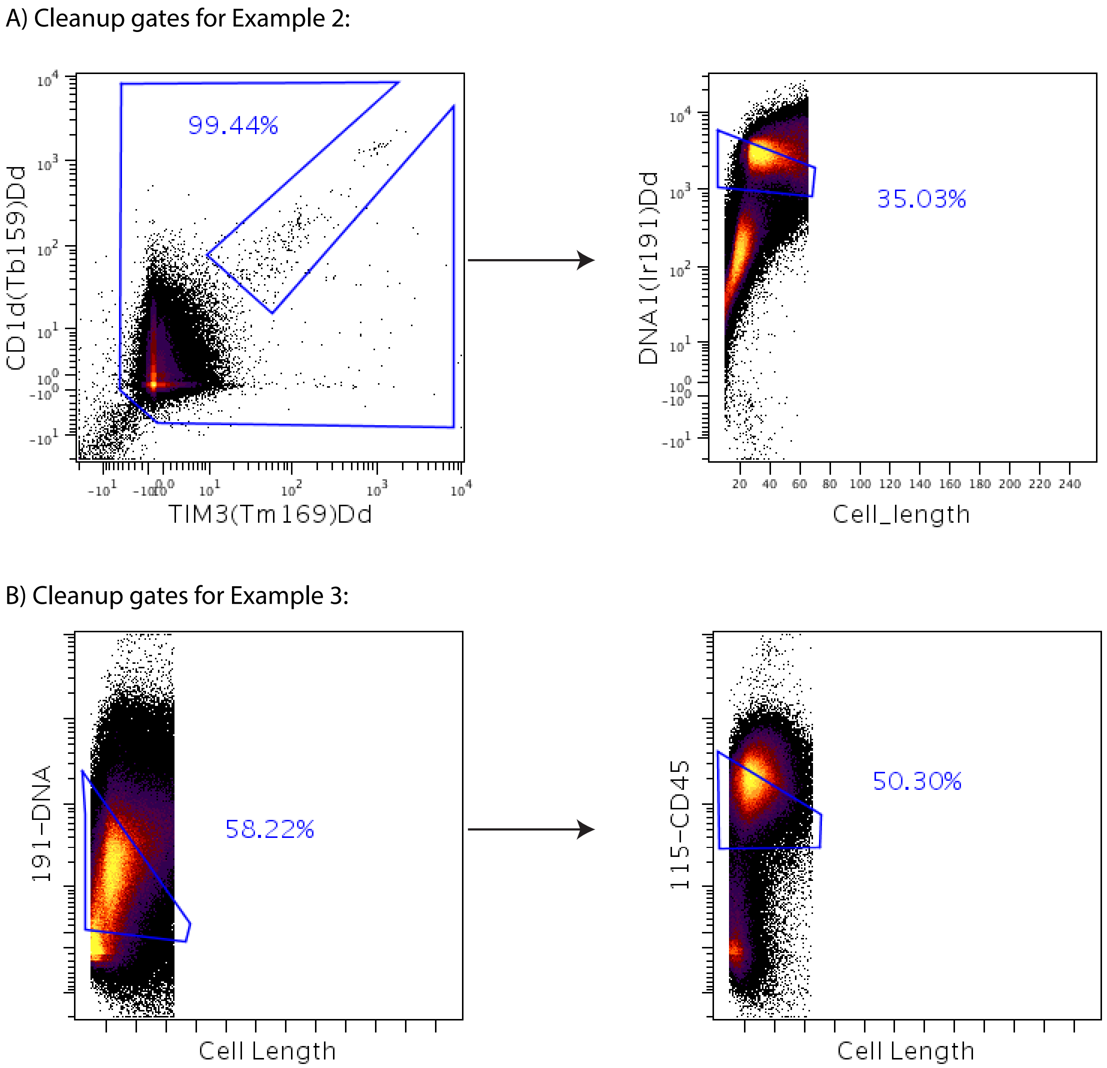


Figure S19: The SPADE plots in Figure S5, colored by expression of different markers.

4 Online Methods Figures



A) Cleanup gates for Example 2:

B) Cleanup gates for Example 3:

Figure S20: Manual gates used for pre-selecting singlet cells. A) Gates for excluding beads and then selecting the singlet events; B) Two gates for selecting singlet events.

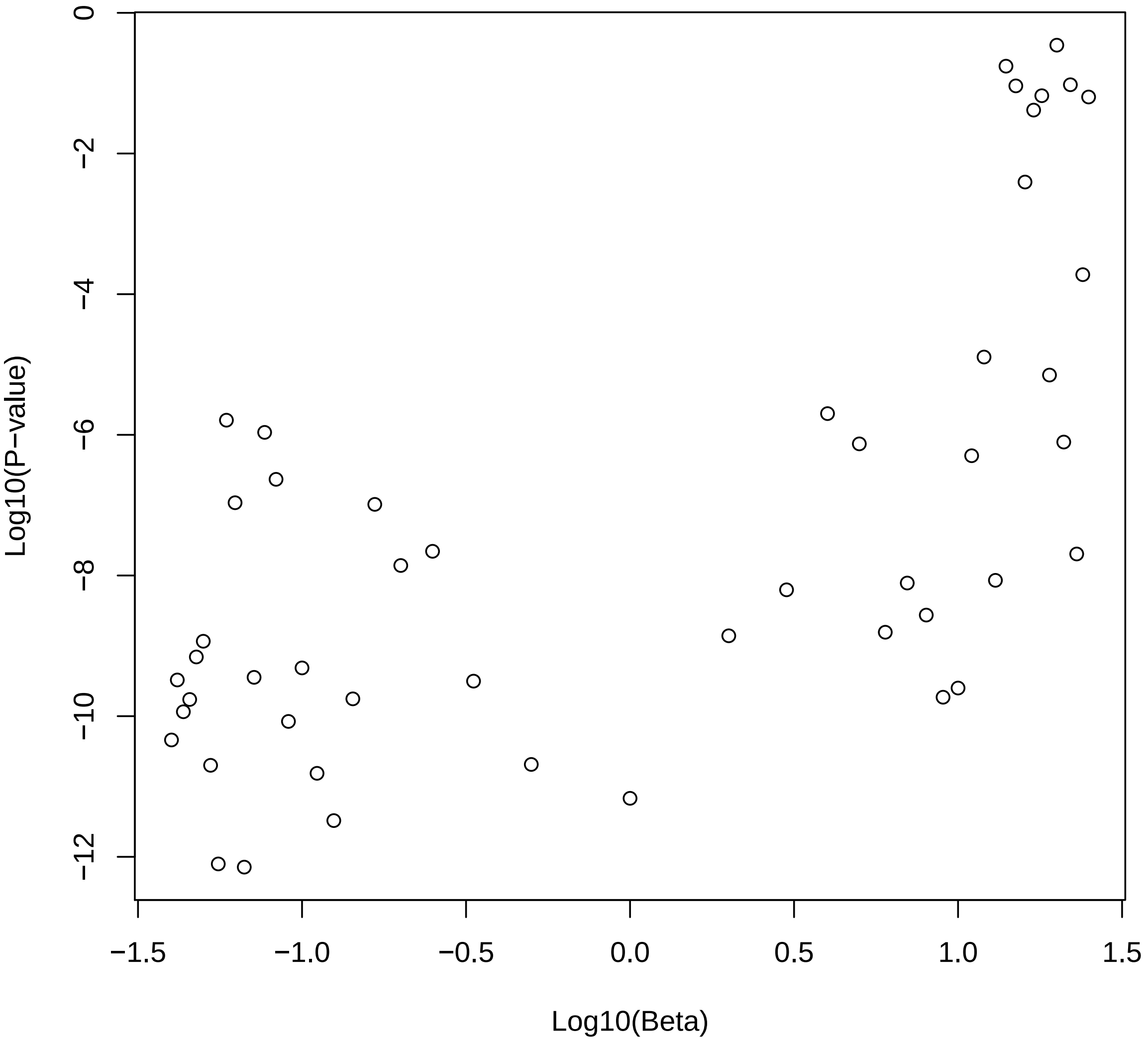


Figure S21: Correlation with clinical outcome as a function of different Beta values. The Y-axis measures the correlation (Cox Proportional Hazards Ratio P-value) between the GateFinder-defined cell population and clinical outcome in the HIV dataset from Figure 1B. Beta of 1 or smaller (log10(Beta)=0) where purity is as important or more important than specificity results in a stronger correlation with the clinical outcome.

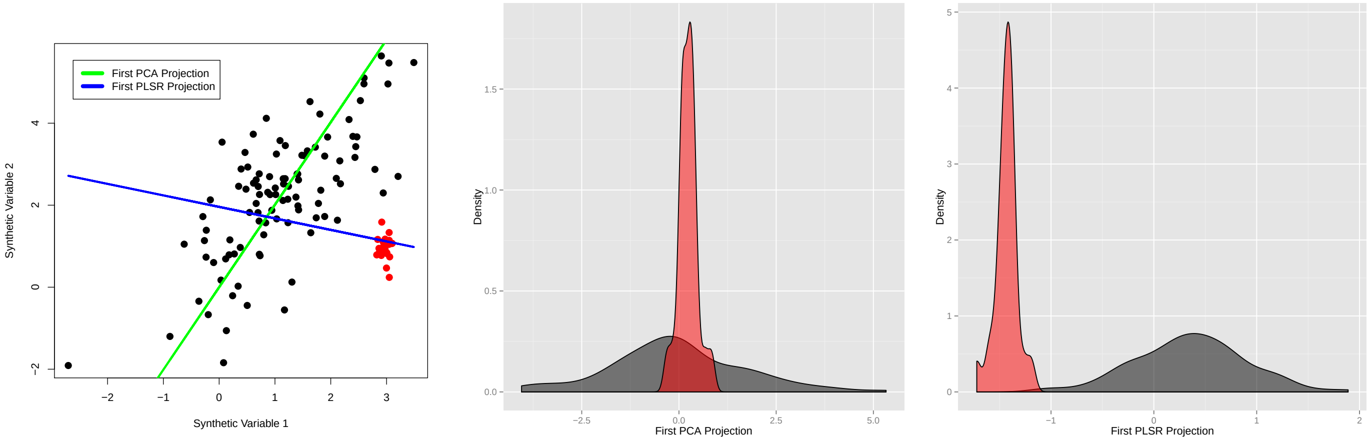


Figure S22: Comparison of unsupervised and supervised analysis for dimension reduction. A two-dimensional synthetic dataset was created by sampling 100 random points from a bi-variate normal distribution ($\mu = (1, 2)$ and $Sigma = (1, 1, 1, 2)$) and 20 points from a secondary bivariate normal distribution ($\mu = (3, 1)$ and $Sigma = (0.01, 0.01, 0.01, 1)$), demonstrated using the black and red dots, respectively. The green line demonstrates the first principal component of the data (the projection with maximum variance). This is an unsupervised analysis and therefore is blind to the colors of the dots. The blue line is a projection calculated using a Partial Least Squares Regression (PLSR) to maximize the separation between the two classes. The center and right panels demonstrate that the supervised projection separates the data points from each of the two normal distributions.

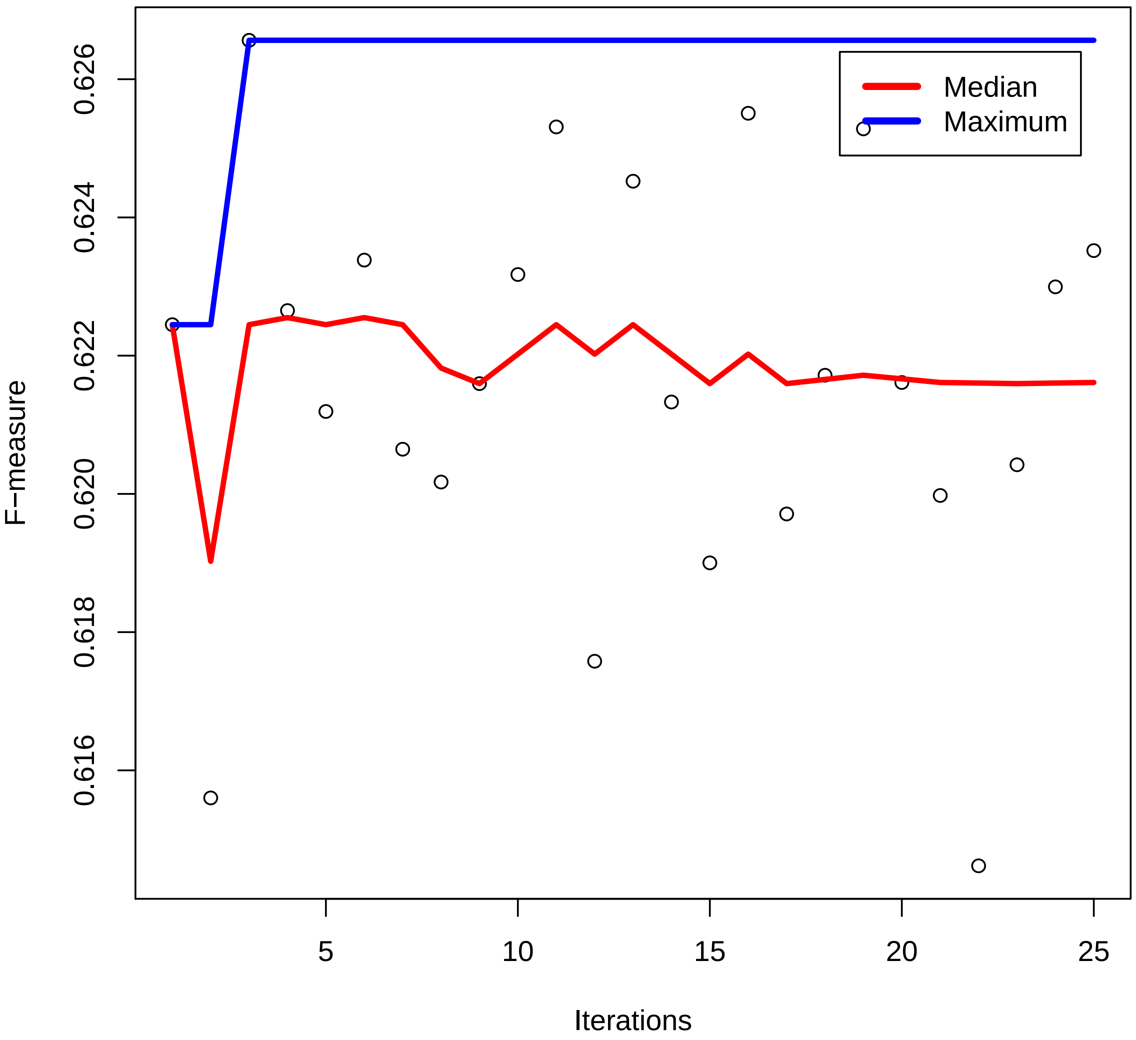


Figure S23: Stability analysis for the gating strategies. 25 independent gating strategies were identified using randomly and uniformly selected subsets of cells from the GCSF/pSTAT5 example. The F-measure of the gating strategies with both median and maximum F-measures stabilize after only three iterations.

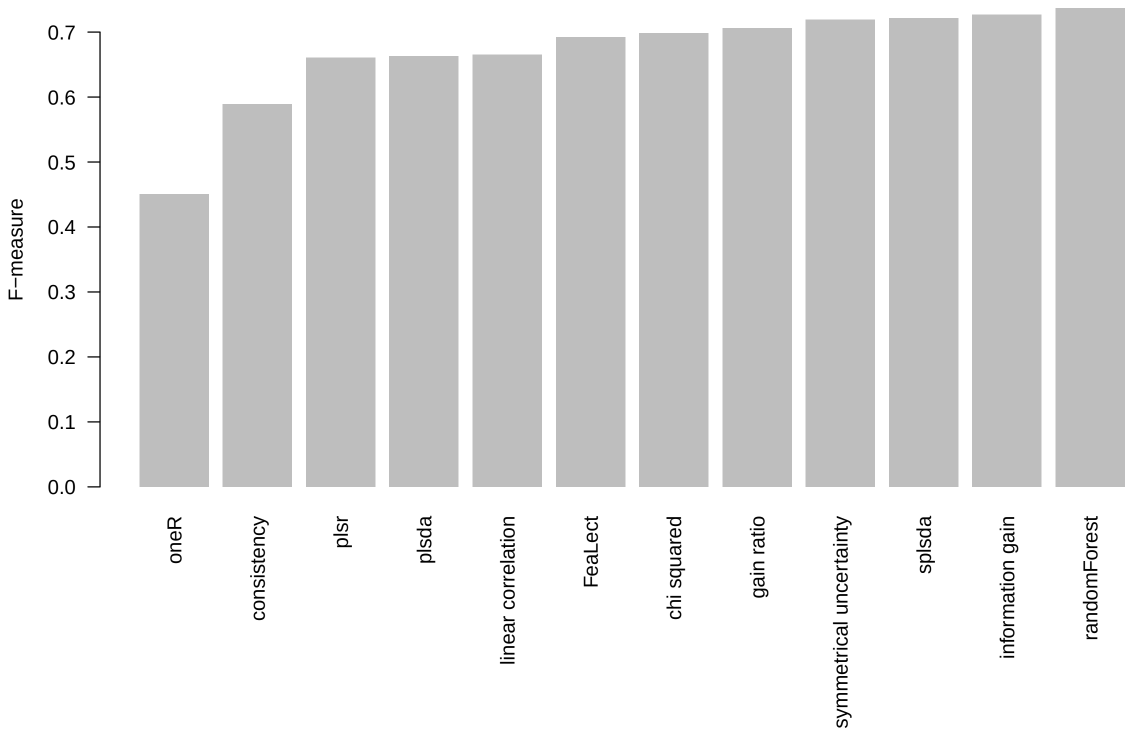


Figure S24: F-measures of gating strategies for simultaneous gating of 10 Kmeans cell populations in the PBMC dataset in subspaces selected by different feature selection algorithms.

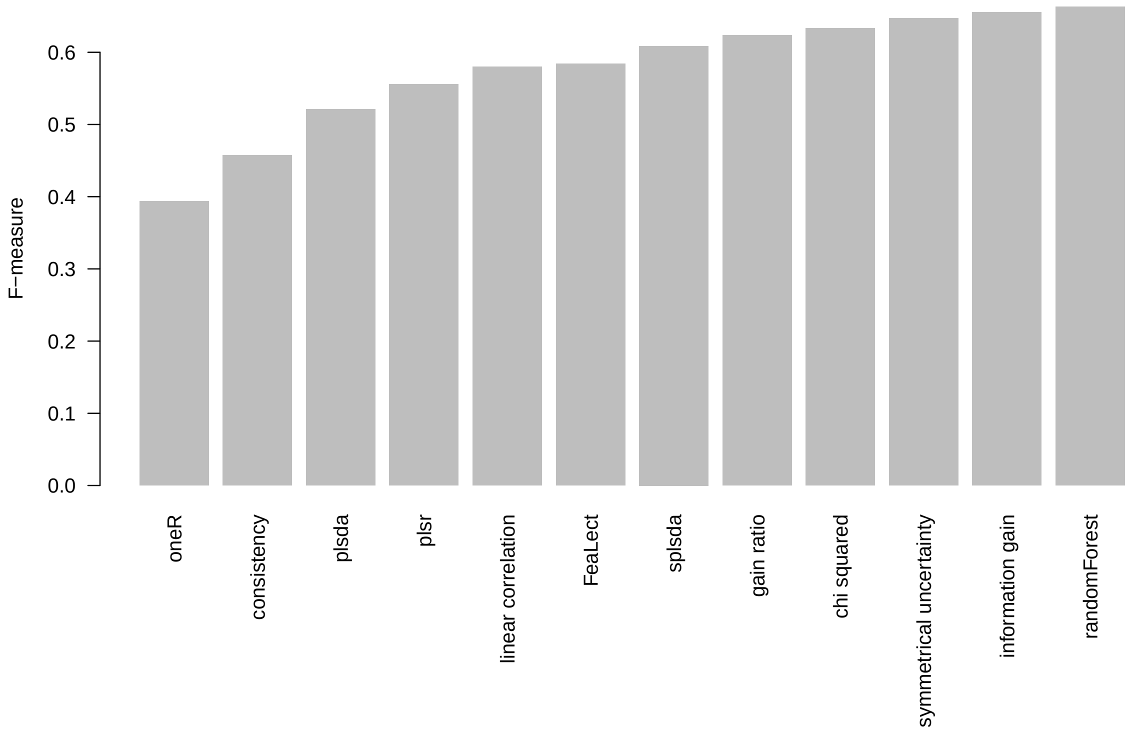
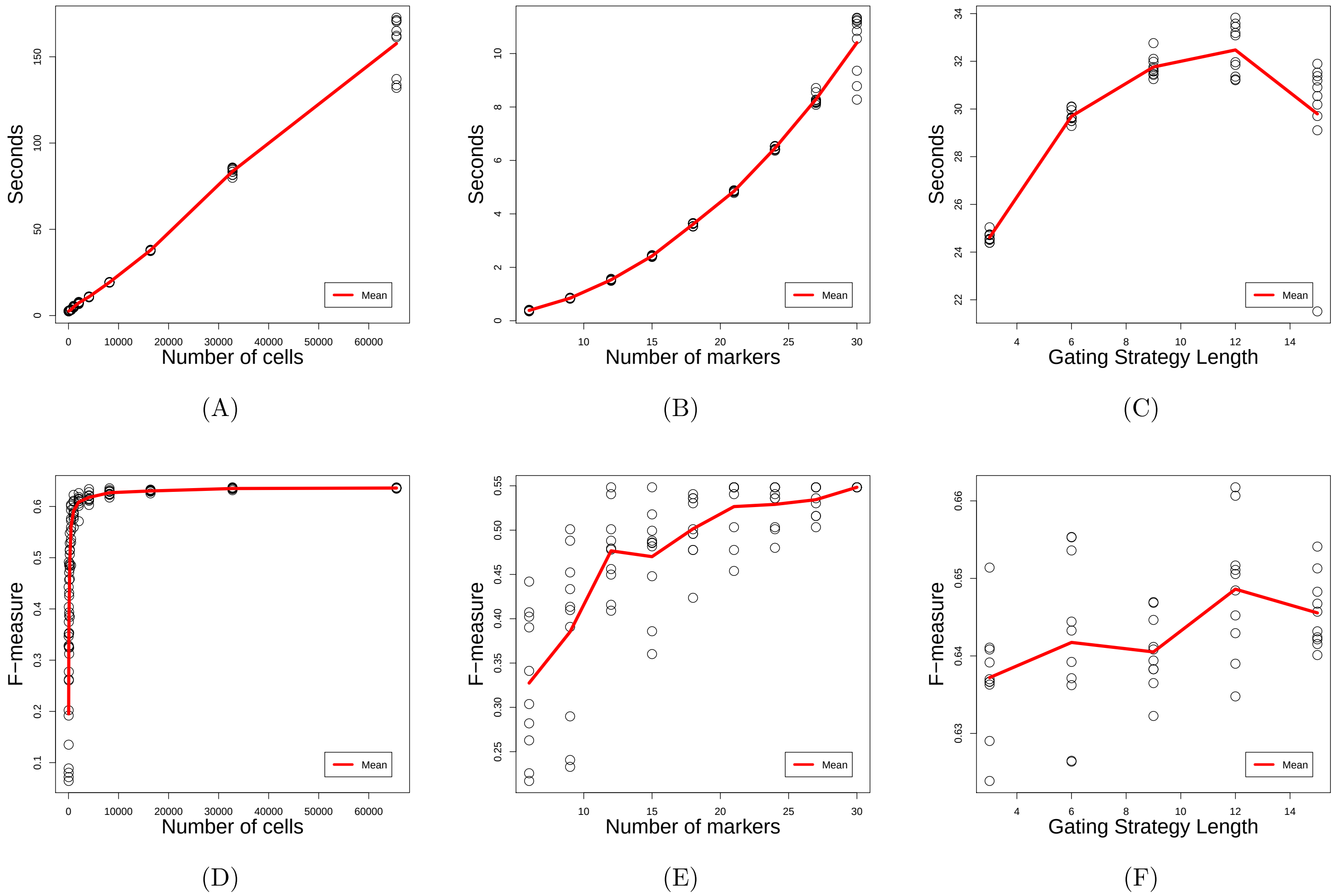


Figure S25: F-measures of gating strategies for simultaneous gating of 10 Kmeans cell populations in the bone marrow dataset in subspaces selected by different feature selection algorithms.



(A)

(B)

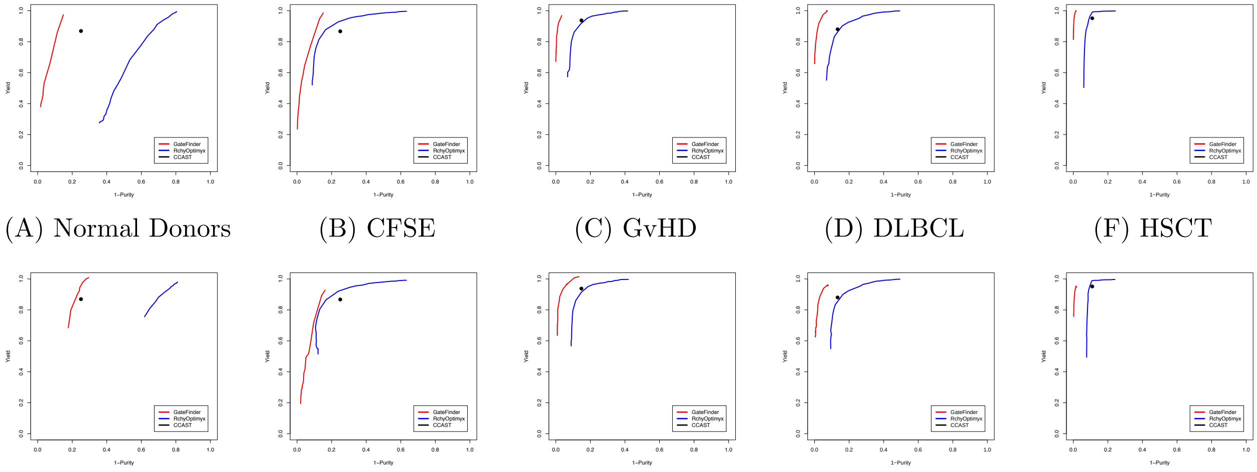
(C)

(D)

(E)

(F)

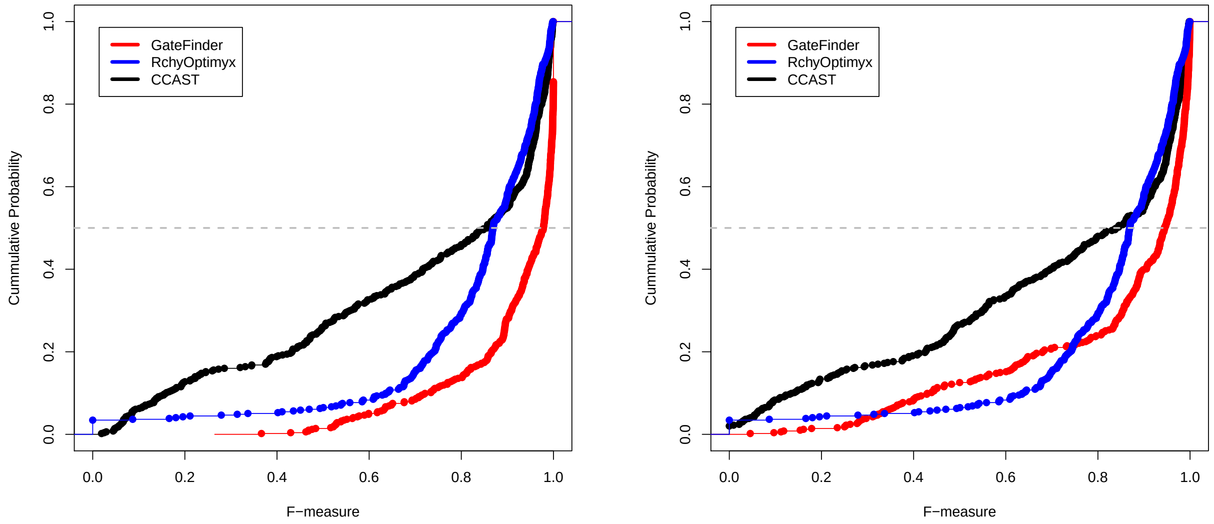
Figure S26: The impact of the number of cells and markers analyzed and the maximum length of the requested gating strategy on runtime and F-measure values. See Supplemental Methods for details.



(A) Normal Donors (B) CFSE (C) GvHD (D) DLBCL (F) HSCT

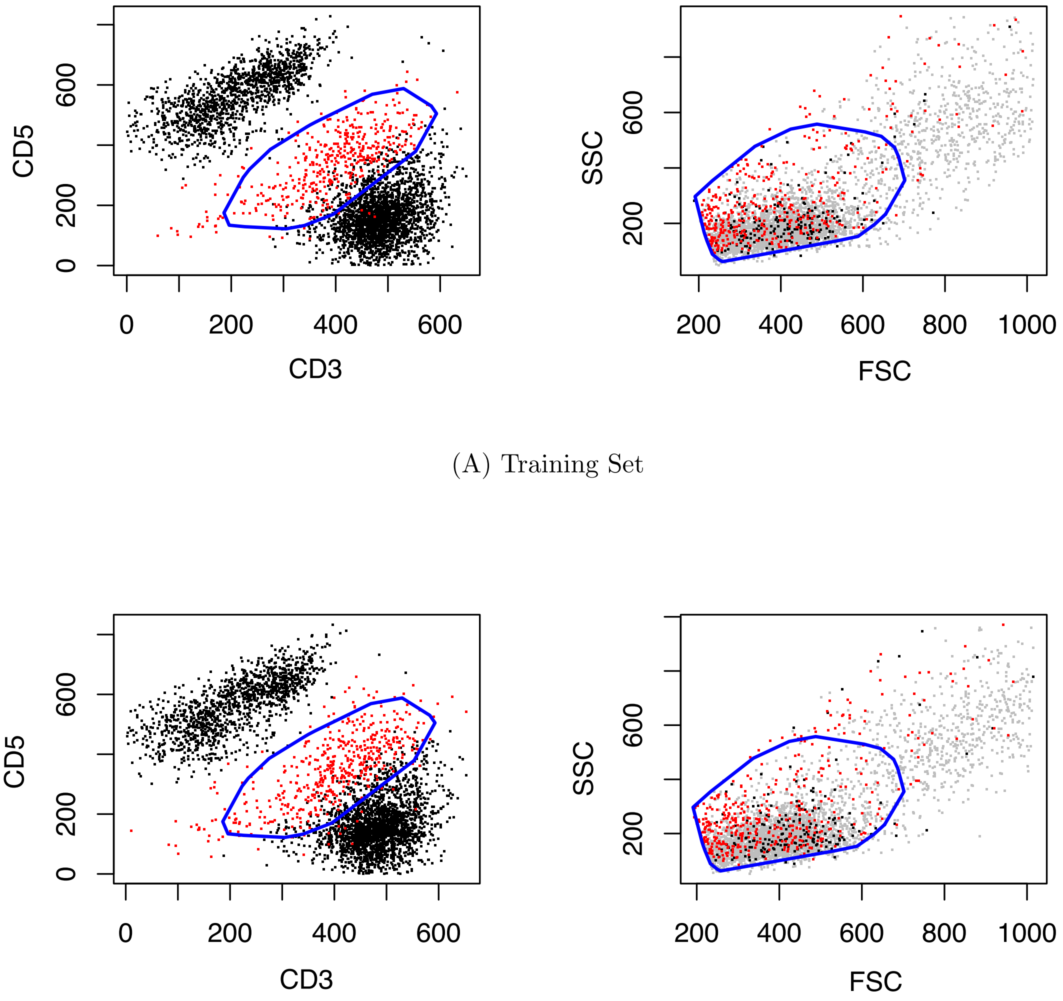
(G) Normal Donors (H) CFSE (I) GvHD (J) DLBCL (K) HSCT

Figure S27: The trade-off between purity and yield for 494 cell populations identified by expert analysts from five datasets. Each dataset was divided into two equal sets. The first set (top row) was used for training the algorithms. The second set (bottom row) was used for independently testing the identified signatures.



(A) Training Set (B) Test Set

Figure S28: Cumulative distribution functions of the three algorithms across all five datasets from Figure S27.



(A) Training Set

(B) Test Set

Figure S29: An example from the FlowCAP dataset in which GateFinder has significantly outperformed other algorithms.

References

[1] J. M. Irish, R. Hovland, P. O. Krutzik, O. D. Perez, Ø. Bruserud, B. T. Gjertsen, and G. P. Nolan, “Single cell profiling of potentiated phospho-protein networks in cancer cells,” *Cell*, vol. 118, no. 2, pp. 217–228, 2004.

[2] P. Qiu, E. F. Simonds, S. C. Bendall, K. D. Gibbs Jr, R. V. Bruggner, M. D. Linderman, K. Sachs, G. P. Nolan, and S. K. Plevritis, “Extracting a cellular hierarchy from high-dimensional cytometry data with spade,” *Nature biotechnology*, vol. 29, no. 10, pp. 886–891, 2011.