Supporting Information for:

Automated Identification of Stratifying Signatures in Cellular Sub-Populations

Bruggner RV., Bodenmiller B., Dill DL., Tibshirani RJ., Nolan GP.

S1 SI Materials and Methods

S1.1 Analyzed Datasets

PBMC's measured by Mass Cytometry

Data consisted of sixteen samples of PBMC's from 8 healthy donors, 8 of which were unstimulated and 8 that were stimulated with BCR/FCR cross-linker for 30 minutes prior to measurement. Data was measured as described in Bodenmiller *et al.* [1]. BCR/FCR stimulation was performed using a mixture of anti-IgG, anti-IgM, anti-IgK and anti-IgL at 10 μg/ml each. Data was downloaded from `http://reports.cytobank.org/105/v2`. Debris were removed prior to analysis as described in [1]. In accordance Supplementary Figure S2 of [2], all analyzed markers were transformed using the arcsin-hyperbolic transformation with a cofactor of 5 before analysis.

FlowCAP-I Datasets

FlowCAP-I datasets were downloaded from `http://flowcap.flowsite.org/Availability.html`. Data were transformed and cleanup gates were applied by competition organizers prior to download.

United States Military HIV Natural History Study

There is substantial variation in time to AIDS development among HIV-infected patients. Thus, it would be useful to identify markers of high-risk individuals who would benefit from early initiation of highly active antiretroviral therapy (HAART). The United States Military HIV Natural History Study, a prospective observational cohort of HIV-infected patients, measured estimated HIV seroconversion dates and AIDS acquisition dates for enrolled subjects. A subset of 466 patients had PBMC's collected within 18 months of their estimated seroconversion. All samples were measured by florescence-based flow cytometry using markers KI67, CD3, CD28, CD45RO, CD8, CD4, CD57, VIVID / CD14, CCR5, CD19, CD27, CCR7, and CD127.

Flow cytometry samples and patient metadata was downloaded from `http://flowrepository.org/id/FR-FCM-ZZZK`. Compensation was applied and samples were singlet, viability, and CD3$^+$-gated as described in [3]. Samples having fewer than 3,000 CD3$^+$ events or a negative reported AIDS-acquisition time were discarded, leaving 416 patients for analysis. These remaining patients were partitioned into training (275 patients) and testing (141 patients) cohorts for model training and evaluation respectively. All measurements were standardized with $\mu = 0$ and $\sigma = 1$ on a per-marker basis prior to clustering.

FlowCAP-II Datasets

FlowCAP-II datasets were downloaded from `http://flowcap.flowsite.org/Availability.html`.

Challenge 2: AML

Transformed data was provided by the FlowCAP-II competition organizers. Prior to analysis, measurements were standardized with $\mu = 0$ and $\sigma = 1$ on a per-marker basis.

Challenge 2: HVTN

Transformed data was provided by the FlowCAP-II competition organizers. Prior to analysis, bimodal landmark normalization was used to normalize marker distributions between samples. Normalization was performed using the *warpSet* function of the *flowStats* package version 3.20.0 [4].

S1.2 Analytical Steps of Citrus

S1.2.1 Overview

Citrus is comprised of several steps. (i) Cells from N samples are combined and clustered in a semi-unsupervised manner to automatically identify C clusters of related cells. (ii) Descriptive statistics characterizing various properties of each cluster (cluster features) are extracted on a per-sample basis. (iii) Extracted cluster features are used in conjunction with a user-specified endpoint of interest to train a supervised model. (iv) Internal cross-validation is used to evaluate model fit and select the appropriate regularization threshold for a final model. (vi) Model features are plotted as a function of endpoint of interest and cluster phenotypes are determined by density plots of markers used for clustering.

S1.2.2 Identification of Phenotypically-Similar Cells Using Hierarchical Clustering

N training samples are collected from patients and are measured by flow cytometry. An equal number of events (or a subset of events) from all samples are randomly selected and combined (Fig. 1, i) and clustered using agglomerative hierarchical clustering (Fig. 1, ii), producing groups of phenotypically similar cells. The dissimilarity between any two cells is specified by Euclidean distance between clustering markers and Wards linkage used as the agglomeration method. Rather than cutting the dendrogram at a fixed height to identify clusters, all clusters C in the hierarchy of merged clusters larger than a user-specified size are retained for subsequent analysis.

S1.2.3 Calculation of Descriptive Cluster Statistics

After clustering the aggregated data, cells from all samples are now assigned to one or more clusters that are comparable between samples. Next, the following F cluster features are calculated for every cluster, on a per-sample basis (Fig. 1, iii):

- The percentage of a sample's cells that are assigned to that cluster.
- The median value of each functional marker for a sample's cells in that cluster.

If each sample has been measured in a basal state and under one or more perturbed conditions, the following additional metrics may also be calculated for each cluster:

- The difference in cluster abundance between the basal and perturbed states.
- The difference in cluster functional marker median values between the basal and perturbed states. unstimulated state.

This results in a $N \times (F^*C)$ matrix of cluster features with each row corresponding to a sample and each column describing a single property of a cluster in that sample (Table S1).

S1.2.4 Model Construction: Classification

When identifying cluster features that differ between sample groups, each sample is assigned by the user as belonging to one of two or more groups (Fig. 1, iv). Next, regularized classification models are constructed with calculated cluster features acting as regressors of sample group. Importantly, prior knowledge suggests that only a subset of calculated cluster features will be useful in differentiating sample groups. For this reason, classification models are constructed using the nearest shrunken centroid and lasso-regularized logistic regression methods, both of which build a series of predictive model using automatically selected informative subsets of supplied regressors. The number of features included in any give model is limited by a regularization threshold, λ. As it is unknown which subset of cluster features best stratify the user-specified sample group, a set of i models are built using a range of i regularization thresholds, $\lambda_1 ... \lambda_i$. The optimal model from this set is then selected by performing cross validation on all models and then selecting the simplest one that meets user accuracy constraints.

S1.2.5 Model Construction: Survival regression

When identifying cluster features predictive of sample survival time, the survival time and censoring status for each sample is specified by the user (Fig. 1, iv). Next, many lasso-regularized Cox proportional-hazards models are constructed with calculated cluster features acting as regressors of sample survival time. The number of features included in any give model is limited by a regularization threshold, λ. As it is unknown which subset of cluster features best predict patient risk, a set of i models are built using a range of i regularization thresholds, $\lambda_1 ... \lambda_i$ and evaluated using K-fold cross validation. The regularization threshold $\hat{\lambda}$ producing the model with the best total goodness of fit (as described in section S1.2.8) is used to constrain the final model.

S1.2.6 Analysis of a new sample

To predict the group of a new, unlabeled sample using a model constructed from training data, the same cluster statistics used to build the initial model must first be calculated for the new sample. Before this can be done, cells from the new sample must be first mapped to the clusters identified in the training data. This is done by taking a random subset of cells (preferably the same number drawn from each training sample) from the new sample and identifying the nearest neighbor of each in the training data by Euclidean distance. Then, cells from the new sample are assigned to all clusters of their nearest neighbor in the training data. After cluster assignments have been made, the previously described cluster statistics are calculated for the new sample and the existing model is used to predict its phenotypic class or relative survival risk.

S1.2.7 Selection of an optimally regularized model: Classification

To identify an optimal model regularization threshold $\hat{\lambda}$, a set of predictive models is constructed using a fixed range of regularization thresholds $\lambda_1 ... \lambda_i$, each having an differing level of complexity and accuracy. Internal K-fold cross-validation is used to estimate the model error rate at each regularization threshold (Fig. 1, v). Cross validation is performed by assigning samples randomly to K groups. Samples from all but one of the groups are used to build models at fixed range of i regularization thresholds as described in sections S1.2.2, S1.2.3, and S1.2.4. Class labels of samples in the left-out group are then predicted as described in section S1.2.6 using models at every regularization threshold. This process is repeated for all K groups, resulting in class predictions for all samples at each regularization threshold. The predicted class of each sample is compared to its true class, providing an estimated model error rate for each threshold.

S1.2.8 Selection of an optimally regularized model: Survival regression

To identify an optimal model regularization threshold $\hat{\lambda}$, a set of predictive models is constructed using a fixed range of regularization thresholds $\lambda_1 ... \lambda_i$. Samples are partition into K groups and the goodness of fit of each model was calculated as described by Simon *et al.* (Fig. 1, v)[5]. In more detail, samples are assigned randomly to K groups. Samples from all but the K'th group are used to build models at fixed range of i regularization thresholds as described in sections S1.2.2, S1.2.3, and S1.2.5. The goodness of fit for the K'th part and regularization threshold λ_i is defined as $\hat{CV}_k(\lambda_i) = \ell(\beta_{-k}(\lambda_i)) - \ell_{-k}(\beta_{-k}(\lambda_i))$ where ℓ_{-k} is the log-partial likelihood of the model excluding part K of the data and $\beta_{-k}(\lambda_i)$ is the optimal β for the non-leftout data. The total goodness of fit $\hat{CV}(\lambda_i)$ for a given λ_i is the sum of all total $\hat{CV}_{-k}(\lambda_i)$. The regularization threshold $\hat{\lambda}$ maximizing this total goodness of fit is selected to constrain the final model.

S1.2.9 Result Assessment & Selection of a Final Regularization Threshold

Model error rate plots should be used to assess the quality of results. If a constructed model has small estimated error rate, it necessarily follows that this model has identified some subset of cluster features that are robust predictors of a sample's class. These features, in turn, have a behavior that is unique for that class and are hence, stratifying subsets of interests. Conversely, if a model has a high error rate, it's likely that the features selected by the model do not consistently differ between sample classes and hence, are not useful stratifying features. Thus, users should interpret features and clusters only from models having an acceptable error rate. Examples of models having low (good) and high (bad) cross validation error rates and corresponding features from each shown in Figure S7. The same interpretation holds for the survival regression case, excepting that the model fit is evaluated by its partial likelihood deviance and features selected by the model are predictive of a patients survival risk.

When one or more models with acceptable error/likelihood deviance rates have been constructed, a user must choose a regularization threshold that will be used to constrain the final model constructed from all sample features. When seeking to identify the smallest but most informative subset of features that differ between between classes, the regularization threshold $\hat{\lambda}$ resulting in the simplest model with an acceptable error rate should be selected. When seeking to identify many or all features differentially expressed between sample classes, a regularization constraint should be selected that produces the most complex model with an acceptable estimated error and feature false discovery rate. The estimated feature false discovery rate for the nearest shrunken centroid model is calculated as follows:

For each regularization threshold λ_i used in cross validation:

1. Randomize the class labels assigned to each sample.
2. Train a new model m_i, constrained by λ_i, to predict randomized labels from step 1.
3. Count the number of non-zero features in m_i.

Steps 1-3 are repeated 1,000 times, producing a distribution of estimated of feature false discovery rates for each regularization threshold. A final model regularization threshold $\hat{\lambda}$ with an acceptable median false discovery rate is then selected and used to constrain a final model constructed from all sample features.

S1.2.10 Interpretation of Results

Nonzero features of the final model, as determined by the regularization parameter $\hat{\lambda}$, are the set of population features that best differentiate sample groups or predict patient survival risk. For inter-group analyses, the values of these relevant features for each sample are shown in box plots, plotted, grouped by class (Fig. 1, vi). For survival regression, stratifying features are plotted as a function sample survival time. Equally important as the identified stratifying features are the phenotypes of corresponding clusters. To determine cluster phenotype of any single cluster, densities plots or scatter plots of lineage markers in cluster cells are shown, along with plots of the same markers in all cells, permitting an investigator to see comparatively how much each marker is enriched in cluster cells (Fig. 1, vii). Related stratifying subsets that have similar behavior may be identified by highlighting relevant subsets in plots of the clustering hierarchy (Note S4.3).

S1.3 Validation of Stratifying Signals in PBMCs

In addition to measuring PBMC's from 8 different patients under 12 stimulation conditions, Bodenmiller *et al.* also measured PBMC's from a single patient under 12 different stimulation conditions in the presence of increasing concentrations of 27 different inhibitors. Cells from unstimulated and BCR-stimulated samples in the presence of no inhibitor, DMSO, and varying concentrations of Dasatinib were mapped to clusters of interest as described in Section S1.2.6. Median levels of functional markers were calculated from cells in clusters of interest on a per sample basis. Functional markers levels were plotted for each cluster and experimental condition (Fig. S15).

S1.4 Quantification of Clustering Sensitivity

S1.4.1 The F_1 measure

For a given subset of cells c_a identified a clustering algorithm and subset of cells c_m identified by manual gating, the F_1 score measures the overlap between c_a and c_m. F_1 measure values range from 0 to 1 with a 1 indicating that cells assigned to c_a by an algorithm are the same cells that were assigned to c_m by manual gating. The F_1-measure is the harmonic mean of a clustering's precision and recall and its use as a metric for evaluating clustering performance was described by *Aghaeepour et al.* [7]. The formal definition of the F_1-measure is:

$$2 \cdot \frac{P \cdot R}{P + R} \tag{1}$$

where P and R represent the precision and recall for a single cluster respectively. For clustered cells, c_a, and manually gated cells c_m, precision measures the proportion of cells in c_a that are comprised of cells from c_m. Recall measures the proportion of cells in c_m that were found in c_a. Alternatively:

$$P = \frac{|c_m \cap c_a|}{|c_a|} \tag{2}$$

and

$$R = \frac{|c_m \cap c_a|}{|c_m|} \tag{3}$$

S1.4.2 Measuring Clustering Sensitivity: Scoring Algorithm Clusters vs. Manually-Defined Populations In a Single Sample

If a single sample contains a set of n manually gated populations $P = \{p_1, p_2, ..., p_n\}$ and a clustering of that same data produces a set of m clusters $C = \{c_1, c_2, ..., c_m\}$, the sensitivity of the clustering is defined as:

$$\frac{1}{n} \sum_{p_i \in P} \max_{c_j \in C} F_1(p_i, c_j) \tag{4}$$

In words, for a manually-defined population p_i, the F_1 score is calculated for all m identified clusters and the maximum of those m measures is reported as the F_1 measure for that population. Maximum F_1 scores are calculated for all n manually gated populations and the sensitivity of a clustering is the average of those scores. Notably, this approach differs from that reported by *Aghaeepour et al.* who weighted F_1 measures by population size. Specifically, the unweighted approach employed here better reflects algorithm performance on smaller, more rare populations of cells.

S1.4.3 Measuring Clustering Sensitivity Across Many Samples

Each FlowCAP-I Dataset D consists of S_D samples with many populations of cells gated within each sample. The sensitivity measure of a clustering of dataset D is defined as:

$$\frac{1}{|S_p|} \sum_{s_k \in S_D} \sum_{p_i \in P_k} \max_{c_j \in C_k} F_1(p_i, c_j) \tag{5}$$

Here, $|S_p|$ represents the total number of manually gated populations found in all samples of dataset D. In words, for a given sample $s_k \in S_D$, the maximum F_1 score is computed for all manually gated populations P_k found in s_k. The average of all maximum F_1 scores from every manually gated population in every sample is reported as the dataset-specific clustering sensitivity score.

S1.4.4 Sensitivity Measures of Hierarchical Clustering on FlowCAP-I datasets

Hierarchical clustering using Euclidean Distance and Ward's linkage method was run on each FlowCAP-I dataset. Dimensions used for clustering are listed in (Table S2) and up to 10,000 events were selected for clustering from each sample. The minimum cluster size was set at 0.5% of the number of clustered events. Sensitivity measures were computed for each dataset as described above.

S1.4.5 Sensitivity Measures of FlowCAP-I measures on FlowCAP-I datasets

Clustering assignments from FlowCAP-I competition methods were downloaded from the FlowCAP-I website (`http://flowcap.flowsite.org/Availability.html`). Using supplied clustering assignments, clustering sensitivity measures were computed as described above.

S1.4.6 Sensitivity Measures in Rare Populations

To evaluate clustering performance on rare populations of cells, sensitivity measures were calculated only for hand-gated populations that contained fewer than 5% of a sample's total events. Results are shown in Figure S9.

S1.5 Identification of Prognostic Cell Subsets in HIV-infected Patients

Data from the United States Military HIV Natural History Study was analyzed using Citrus and *flowType*. Prior to analysis, data were partition into training and testing-set cohorts, balanced by the number of AIDS events in each cohort. Cell subsets in training data were identified using Citrus and *flowType* and used to train a model AIDS-free survival risk. AIDS-free survival risk was then estimated in testing-set patients using method models, enabling a comparison of model prognostic performance.

S1.5.1 Identification of Cell Subsets Using Citrus

Up to 3,000 cells were selected from training set samples sample and combined together for clustering. Cells were clustered based on the expression the following markers: KI67, CD3, CD28, CD45RO, CD8, CD4, CD57, VIVID / CD14, CCR5, CD19, CD27, CCR7, and CD127. Cluster abundances were calculated on a per-sample basis and clusters having an average abundance above 0.5% of events per sample were retained for further analysis. Cluster abundances from training-set samples were used to train a model of AIDS-free survival risk. To calculate testing set features, up to 3,000 cells from testing samples were mapped to training-set clusters as described in Section S1.2.6. After mapping testing data to training clusters, cluster abundances were calculated for testing patients. Cluster abundances from testing set patients were used to evaluate the performance of the survival risk model.

S1.5.2 Identification of Cell Subsets Using *flowType*

flowType was used to identify cell subsets in each sample (testing and training). Cells were partitioned using the following markers: KI67, CD28, CD45RO, CD8, CD4, CD57, CCR5, CD19, CD27, CCR7, and CD127. Cells were pre-gated on markers CD14/VIVID and CD3 prior to analysis and were not used for phenotype identification as the number of identified phenotypes would increase from 177,147 to 1,594,323 and L1-regularized Cox proportional-hazards model could not be fit on a feature set of this size using the R glmnet package. Partition boundaries were determined using the *flowMeans* method. Cell abundances from training set patients were used as train a model of AIDS-free survival risk. Cluster abundances from testing-set patients were used to evaluate the performance of the survival risk model

S1.5.3 Modeling of AIDS-Free Survival Risk

Training features were used to construct many L1-penalized Cox proportional-hazards models of AIDS-free survival risk at a range of regularization thresholds. 10-fold cross validation was used to select an optimal regularization threshold $\hat{\lambda}$. For each fold of cross validation, fold features were used to train a series of L1-penalized Cox proportional-hazards models and the partial likelihood deviance of each was calculated a fixed range of regularization thresholds. Additionally, the predicted relative risk of left-out patients was calculated for each threshold. This operation was repeated for all 10-folds. Partial likelihood deviances at each regularization threshold were averaged across 10 patients. A final regularization threshold $\hat{\lambda}$ was selected that had the minimum average partial likelihood deviance across all 10 cross-validation folds. Patient risk estimated by cross-validation models constrained by $\hat{\lambda}$ was used to assess training model performance. A final predictive model constrained by $\hat{\lambda}$ was constructed from all training patient data and used to estimate the relative risk of patients in testing-set patients.

S1.5.4 Model evaluation

Time-dependent ROC curves were used to quantify model performance on training-set and test-set predictions. For training-set data, ROC curves were constructed from estimations of relative training patient risk quantified during cross-validation. For testing-set data, ROC curves were constructed from estimations of relative testing patient risk made using the final predictive model constructed from all training patient data.

Briefly, a time-dependent ROC curve is constructed at a landmark time t and has sensitivity and specificity measures of $\Pr[M > c|T < t]$ and $\Pr[M < c|T > t]$ respectively where M is the marker of interest (predicted patient risk), T is survival time and c is the threshold of positivity [6]. Time-dependent ROC curves and estimated confidence intervals were calculated using the timeROC package for R, version 0.2 (http://cran.r-project.org/web/packages/timeROC/). Sensitivity and specificity were calculated at $t = 1025$ days, the mean event-free survival time of all patients.

Testing-set patients were assigned into high and low-risk groups if their predicted relative risk was higher than the mean relative risk for all testing patients and vice versa. Kaplan-Meier curves were constructed for high and low risk groups and the significance of differences between group curves was computed using the log-rank test. All calculations were performed using the survival package for R, version 2.37-4.

S1.5.5 Reporting of Model Features

Robust prognostic cell subsets were identified by recording the number of times each cell subset was included in the 10 $L1$-penalized Cox proportional-hazards models constructed during cross-validation. Cell subsets that were selected by models in more than two-thirds of models were reported as prognostic subsets of interest.

S1.6 Classification of Samples in FlowCAP-II Datasets

Datasets from the FlowCAP-II competition were used to evaluate the classification performance of Citrus. Each dataset consisted of labeled training data and unlabeled testing data. The objective of each challenge was to construct a classification model using training data and then predict the labels of testing data. Data are fully described by *Aghaeepour et al.* [7]. Citrus was applied to each dataset and performance was quantified using precision, recall, accuracy, and F-measures as described in [7].

S1.6.1 Challenge 2: AML

Samples were each measured using 7 different panels of markers. Data from each panel was analyzed independently. For a given panel, 2,500 cells were selected from panel training samples and combined, producing a total of 447,500 cells. Combined events were clustered using all measured markers including forward and side-scatter channels and subset abundances were calculated on a per-sample basis. Cell subsets containing at least 4,475 events (1% of the clustered dataset size) were retained for further analysis. Panel subset abundances were used to train L1-regularized logistic regression models of sample disease state (AML present or absent) at a range of regularization thresholds. Ten-fold cross validation was used to evaluate model accuracy at each regularization threshold. A final regularization threshold (λ_{1se}) was selected that had an error rate within 1 standard error of the minimum cross-validation threshold. Error rates for panel models were estimated using 10-fold cross validation (Fig. S4). The model constructed from cell subsets identified by panel 4 had the lowest estimated error among all panels. A final model constrained by λ_{1se} was constructed from all training samples measured using panel 4.

To estimate the disease status of testing-set samples, 2,500 cells were selected from each testing sample and mapped to the training clusters as described in section S1.2.6. Cluster abundances were calculated on a per-sample basis and the panel 4 classification model was used to predict the disease status of testing-set samples.

S1.6.2 Challenge 3: HVTN

Up to 10,0000 cells were selected from ENV and GAG-stimulated patient samples and combined, resulting in a total of 540,000 combined cells. Combined events were clustered using lineage markers CD3, CD4 and CD8 and subset abundances were calculated on a per-sample basis. Cell subsets containing at least 2,700 events (0.5% of the clustered dataset size) were retained for further analysis. Subset abundances were used to train L1-regularized logistic regression models of sample stimulation group (ENV or GAG) at a range of regularization thresholds. Ten-fold cross validation was used to estimate the model error rate at each regularization threshold. The regularization threshold λ_{1se} from the model with an error rate within 1 standard error of the minimum model was selected to constrain a final model constructed from all training samples.

To estimate the sample stimulation group of testing set samples, 10,000 cells were selected from each testing sample and mapped to the training clusters as described in section S1.2.6. Cluster abundances were calculated on a per-sample basis for testing set samples. The final classification model constructed from training set samples was used to predict the likelihood of GAG stimulation for testing patient sample pair. Of the two samples measured in a patient, the one having the highest predicted likelihood of GAG stimulation was labeled as such and the other was assigned the ENV label.

S1.7 Citrus Sensitivity Analysis

S1.7.1 Clustering Sensitivity As A Function Of Cells Selected Per Sample

Citrus selects and combines together an equal number events from each biological sample in order to ensure that each sample is equally represented in the clustered data. To evaluate the effect of number of events selected per sample on a clustering's sensitivity measure, a varying number of events was selected and combined from each sample, clustered, and the clustering's sensitivity measure was calculated for each dataset from the FlowCAP-I competition (Fig. S11). The maximum difference between any two clustering sensitivity measures run with different sample sizes was found to be 0.061. The average maximum difference across all FlowCAP-I datasets was found to be less than 0.02. This supports the conclusion that clustering performance is largely not affected by events selected per sample and recommend using 10,000 events as a default. Notably, in circumstances where the number of events to be selected from each sample was greater than the number of measured events in a sample, all cells from the sample were included but events were not included multiple times in order to reach the desired sample size. Thus, sampling larger numbers of events (i.e. 20,000 events) in datasets with a smaller number of measurements per sample will have no effect on a clustering's sensitivity measure.

S1.7.2 Stratifying Subsets Detected As A Function Of MCST

When running a Citrus analysis, investigators may specify the MCST based on a combination of prior biological knowledge and the number of events the select from each sample. Setting a smaller MCST includes smaller (but does not remove larger) clusters from an endpoint regression analysis. In other words, all features from an analysis run with a larger MCST analysis are included in an analysis having a smaller MCST. In an experiment having adequate statistical power, all stratifying features identified in an analysis run with a larger MCST would be identified in an analysis run using a smaller MCST. In practice however, the increased number of cluster features included in the regression model weakens the model's power to detect stratifying features due to corrections for multiple hypothesis testing and limited availability of samples.

The relationship between the MCST and Citrus' power to detect stratifying cell subsets was evaluated in the *Bodenmiller* PBMC dataset. Data were clustered as described in Section 2.2. Descriptive cluster properties were calculated for that clustering using MCST's of 5.0%, 2.5%, 1%, 0.75%, and 0.5%. Smaller MCST's were not included due to limited number of cells measured in some samples. The median expression of non-experimentally biased functional markers was calculated for cell subsets at each MCST. Subset descriptive properties calculated using different MCST's were used as regressors of sample stimulation group and stratifying cell subsets were identified as described in Section 2.2. The relationship between the MCST and Citrus' ability to detect stratifying cell subsets was measured by quantifying the proportion of stratifying cell subsets identified by an analysis run using a larger MCST that were reported by an analysis run using a smaller MCST (Table S4). As a general trend, analyses run using smaller MCST's identify more-rare stratifying cell subsets but lose power to detect more subtle differences between sample groups. Approaches described in SI Appendix S4.1 could help limit this sensitivity loss in future analyses.

S2 Supplemental Figures

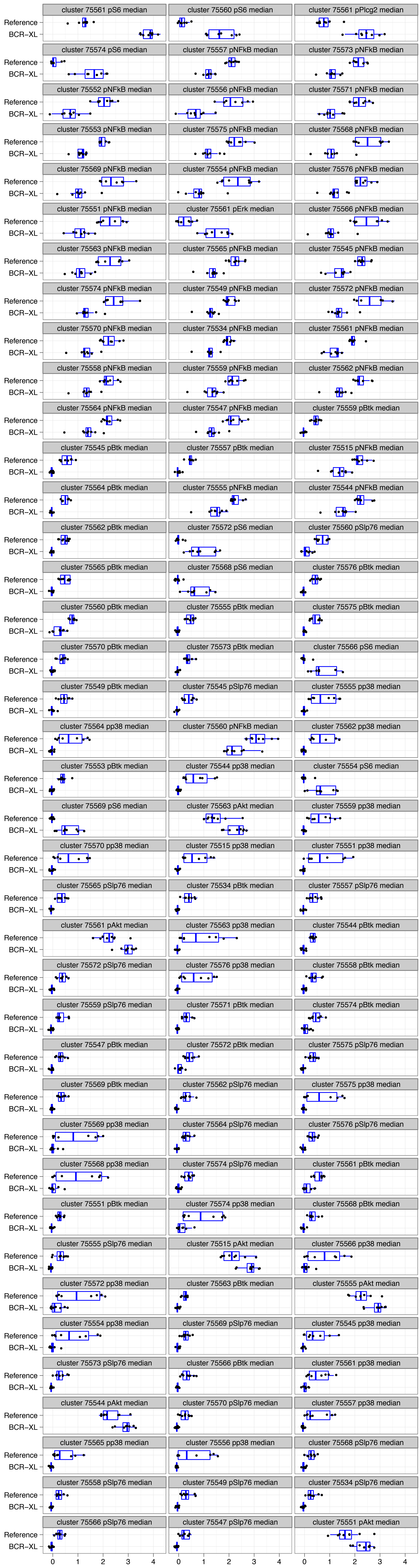

Figure S1: All features selected by the nearest shrunken centroid model as putatively differing between stimulated and unstimulated PBMCs. Shown here are feature values for each sample, grouped by stimulation group. The phenotype of corresponding clusters is determined by density plots shown in Figure S2. Features are ordered by decreasing model weight from top left to bottom right.

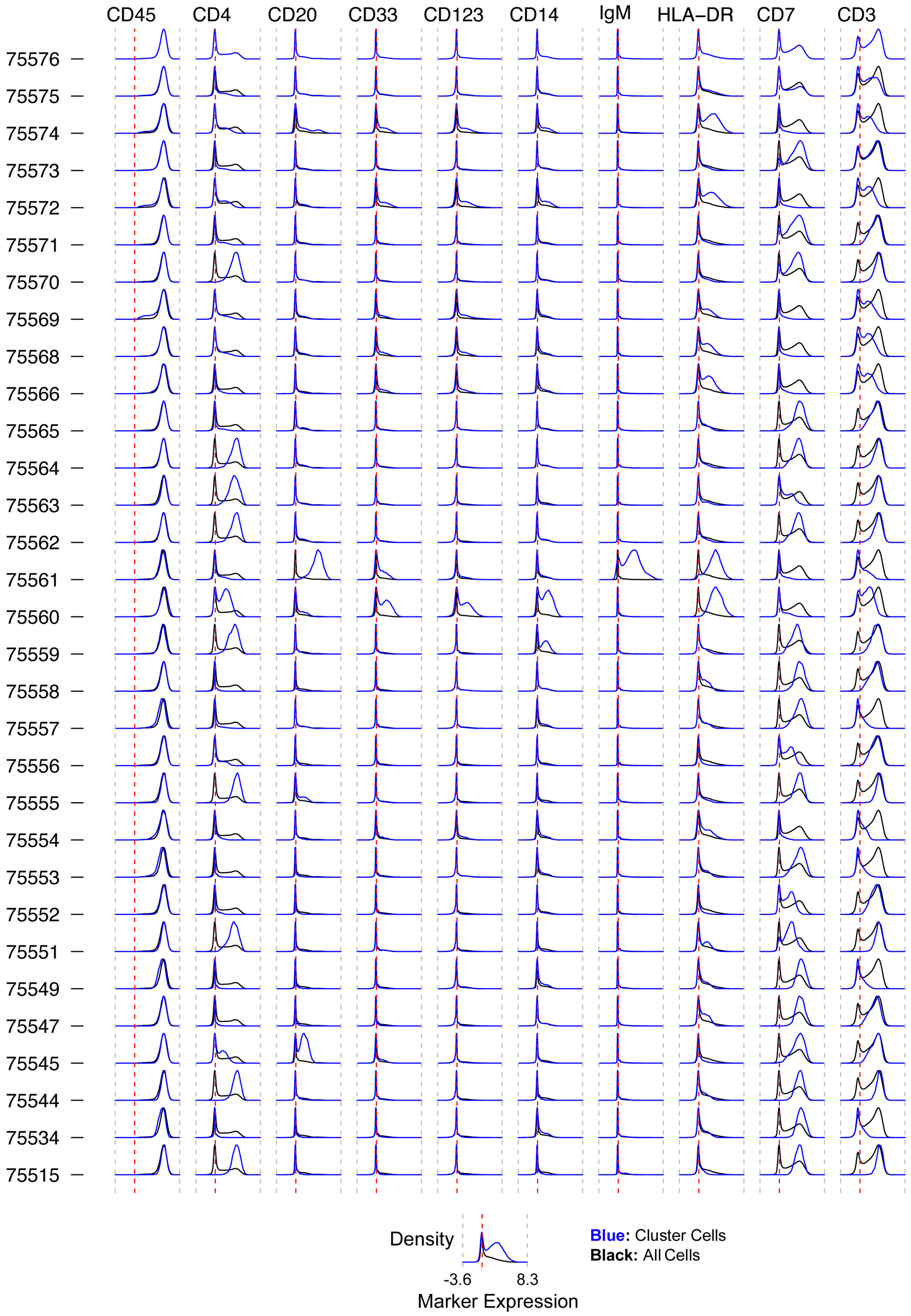

Figure S2: Corresponding phenotypes for clusters whose behavior differs between the two stimulation groups (Fig. S1). Density of lineage marker expression values in cluster cells is shown in blue. Density of lineage marker expression values for all sample cells is shown in black. Scales for all plots range from -3.6 to 8.3. The dotted red line is zero.

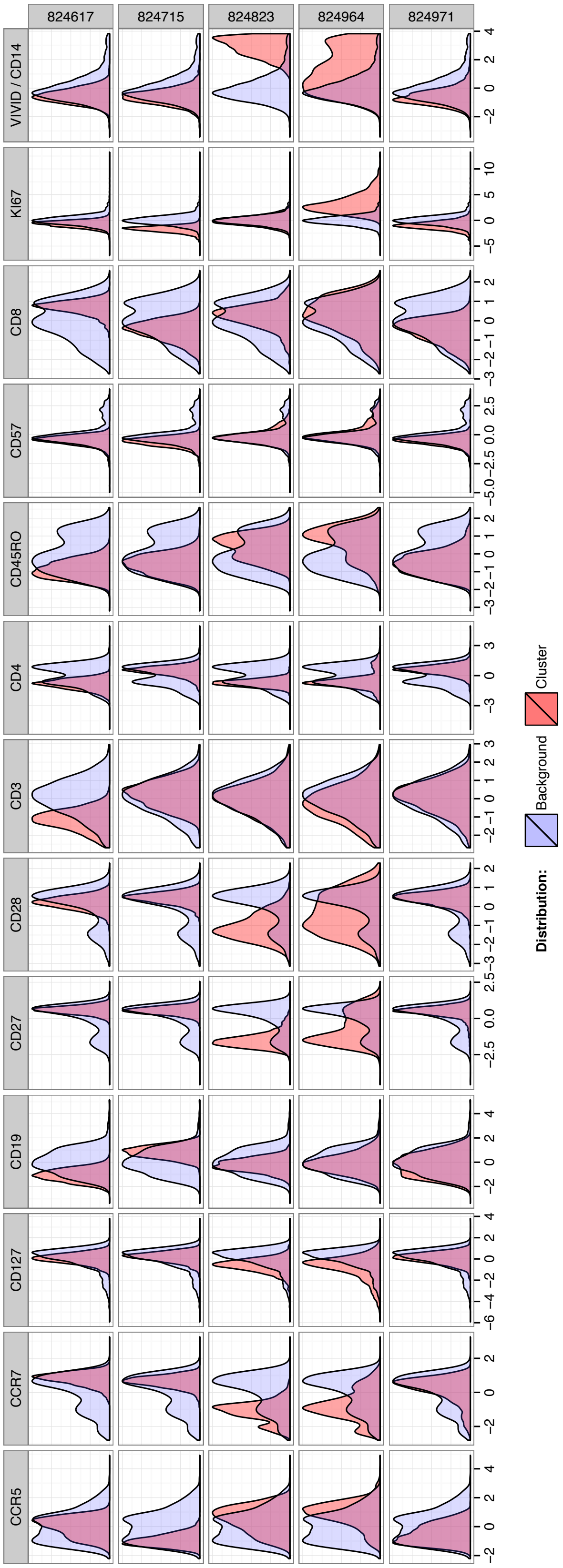

Figure S3: Clusters selected by at least two-thirds of models during cross validation. In addition to Naive $CD8^{+}$ T-cells (cluster 824617) and Ki-67^{+}, $CCR5^{+}$, $CCR7^{-}$, $CD4^{-}$,$CD45RO^{+}$ cells (824964), Citrus also identified two clusters of $CCR5^{-}$,$CCR7^{+}$, $CD27^{+}$, $CD28^{+}$, $CD4^{+}$,$CD45RO^{-}$ cells (clusters 824715 and 824971) which have a phenotype of Naive $CD4^{+}$ T-Cells. The last cluster, 824823, shares a nearly identical phenotype to cluster 824964 but does not express Ki-67.

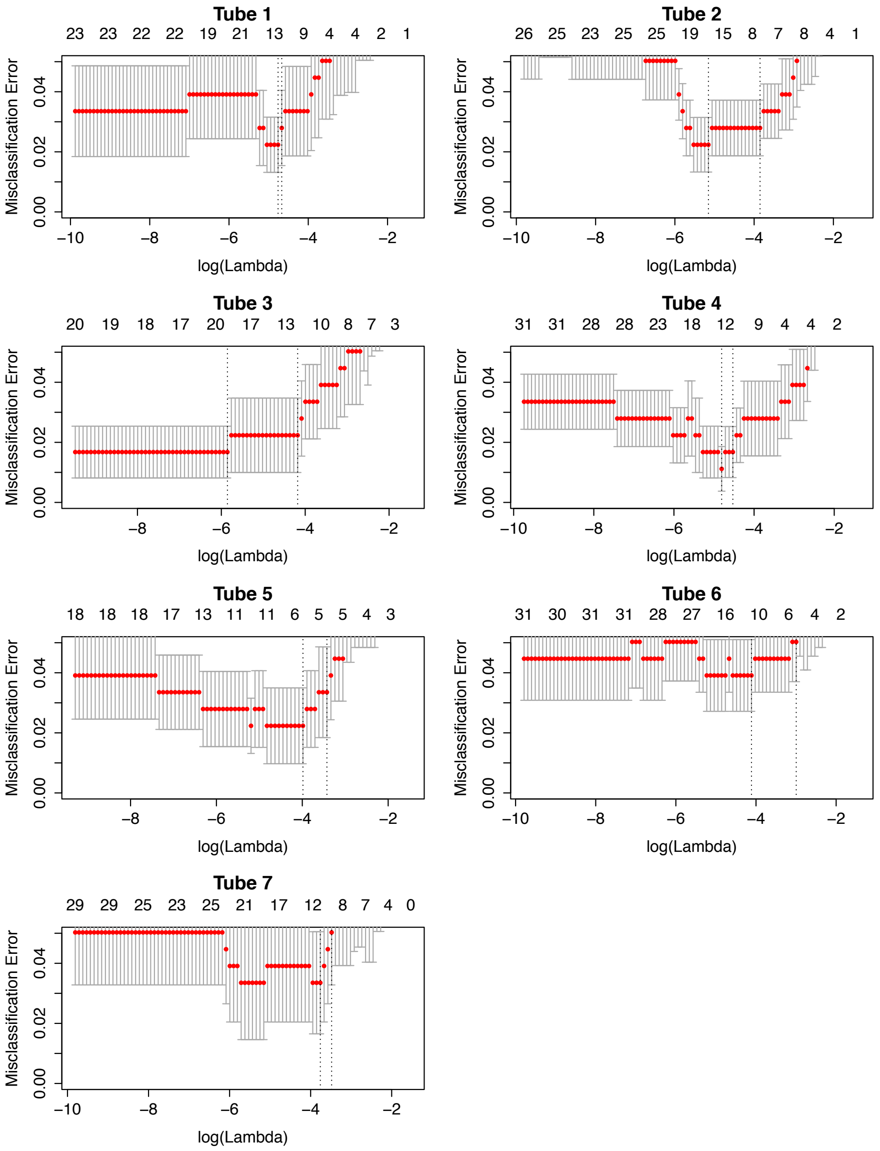

Figure S4: Cross-validation error rates as a function of regularization threshold for 7 measured panels in AML patients. For each panel, the regularization thresholds minimizing the cross-validation error rate (λ_{min}) and within 1 standard error of the minimum model (λ_{1se}) are shown by the left and right-most dotted lines respectively. The model constructed from panel 4's cell subsets had the lowest error rate among models constrained by (λ_{1se}). This model was used to predict the disease status of testing-cohort samples.

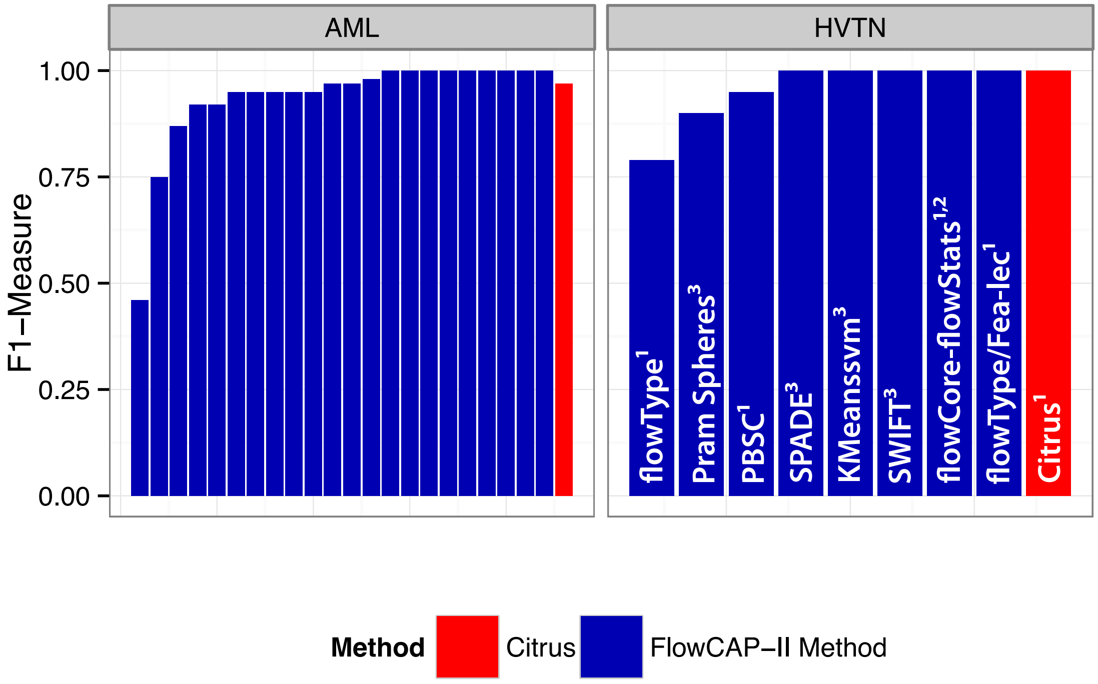

Figure S5: Classification performance of Citrus and FlowCAP-II methods in FlowCAP-II datasets. Participants analyzing the HVTN dataset were asked to submit a list of features that enabled stratification of sample classes. Methods marked with a (1) reported stratifying subsets and behaviors driving classification. Methods marked with a (2) did not perform unsupervised identification of cell subsets. Methods marked with a (3) did not report relevant stratifying subsets.

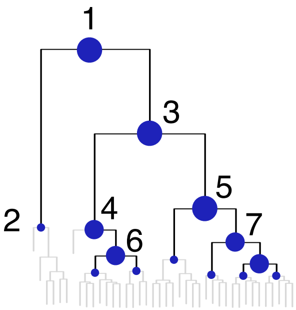

Figure S6: An example dendrogram depicting a hierarchy of identified clusters. All clusters shown as blue dots are larger than the user-specified minimum cluster size and are examined for stratifying signal. Clusters higher in the dendrogram hierarchy (i.e. cluster 3) are more likely to isolate abundant populations while those lower in the hierarchy (i.e. clusters 2,6,7) are more likely to isolate rare populations.

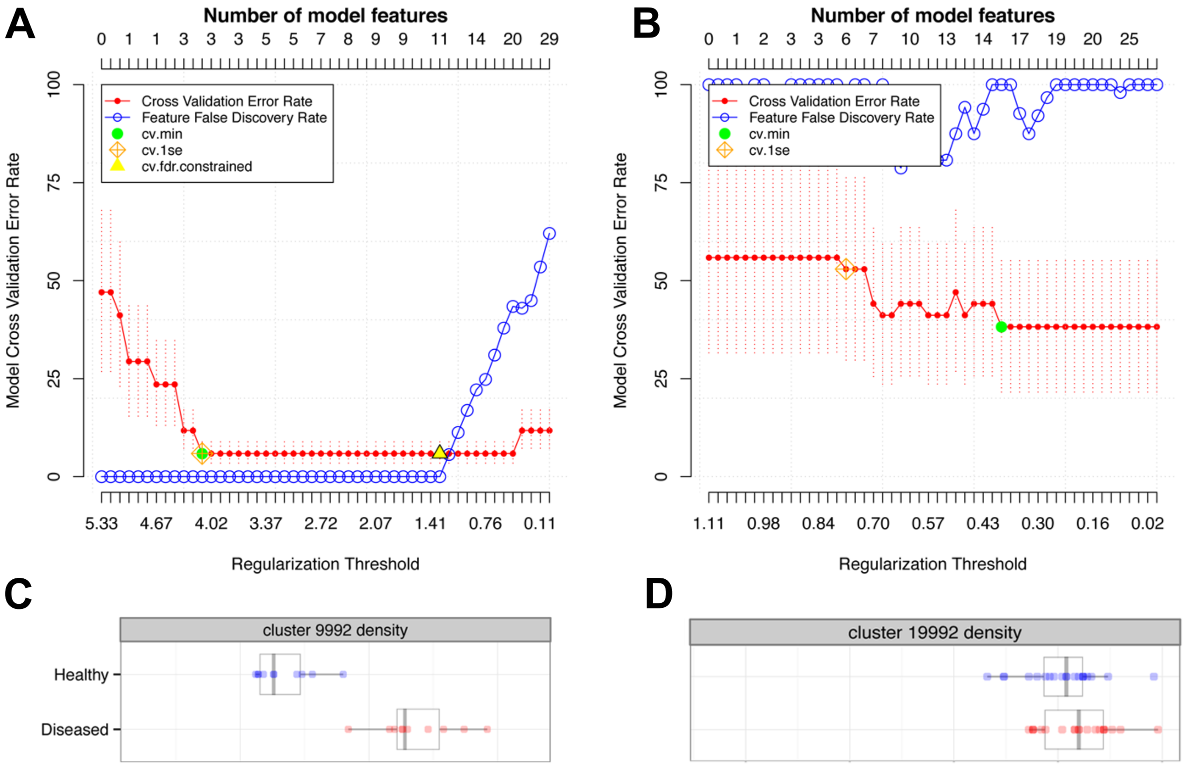

Figure S7: Example cross validation error plots from two simulated data sets. (A) An example analysis producing models with low (good) cross validation error rates. (B) An example analysis producing models with high (bad) cross validation error rates. (C) An example of a good stratifying feature identified by analysis (A). Such features are likely robust stratifiers of samples. (D) An example of a bad stratifying feature identified by analysis (B). Such features are likely spurious and thus a poor predictor of sample class.

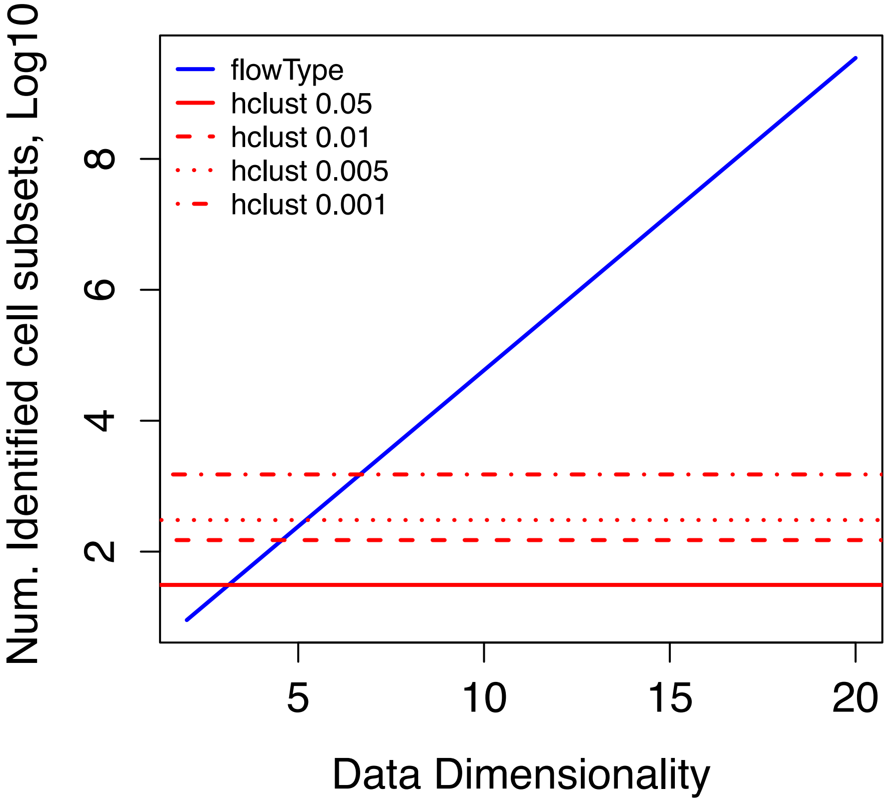

Figure S8: Theoretical number of clusters identified by *flowType* and hierarchical clustering at several MCST's as a function of data dimensionality. The number of cell subsets identified by hierarchical clustering is a function of the MCST and thus scales well to higher dimensions. *flowType* identifies fewer cell subsets than hierarchical clustering on lower dimensional data.

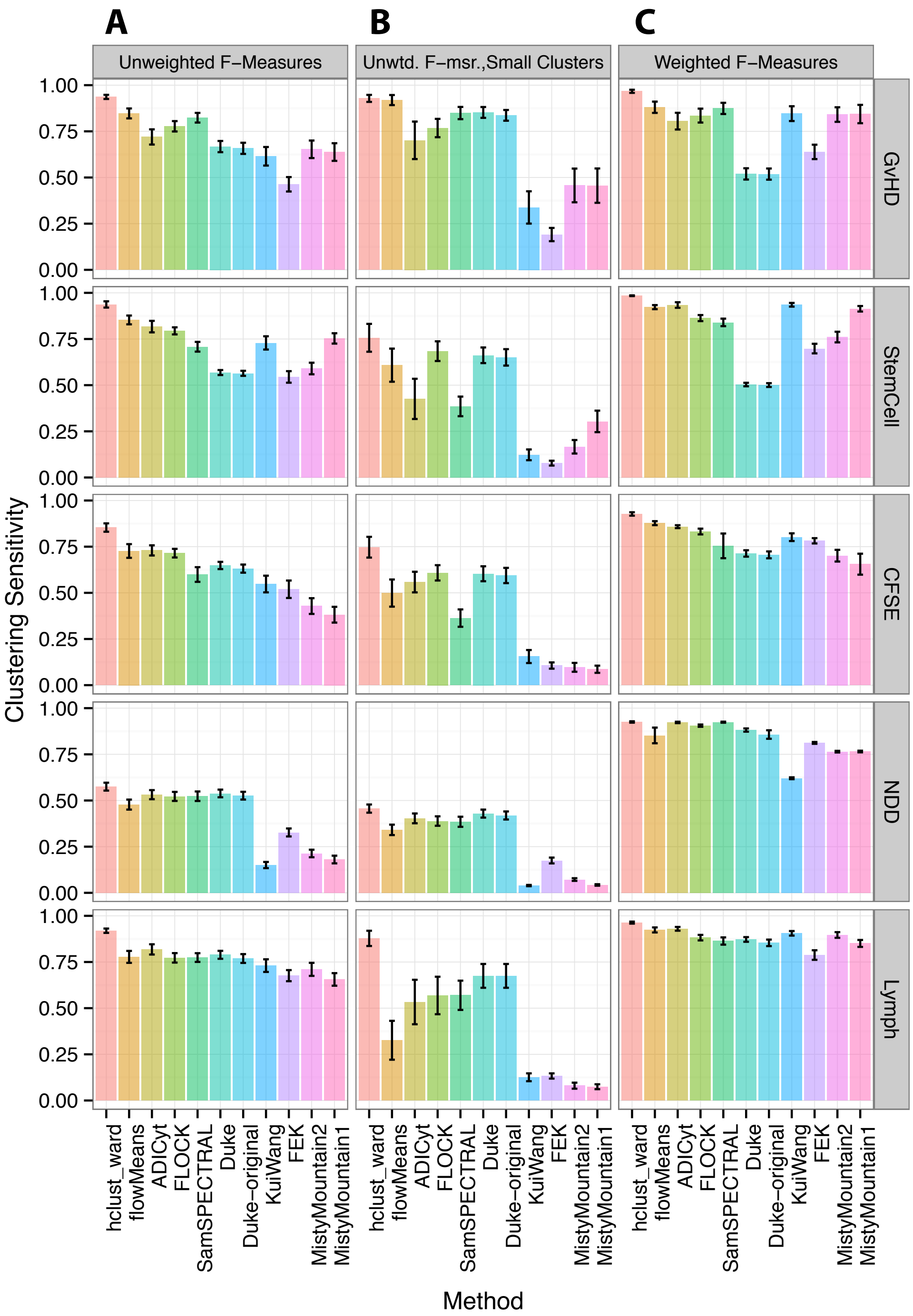

Figure S9: Clustering sensitivity measures for hierarchical clustering. (A) Clustering sensitivity measures computed from all hand-gated populations. (B) Clustering sensitivity measures computed against manually gated populations that contained fewer than 5% of sample events. (C) Sensitivity measures computed with F-measures weighted by population size.

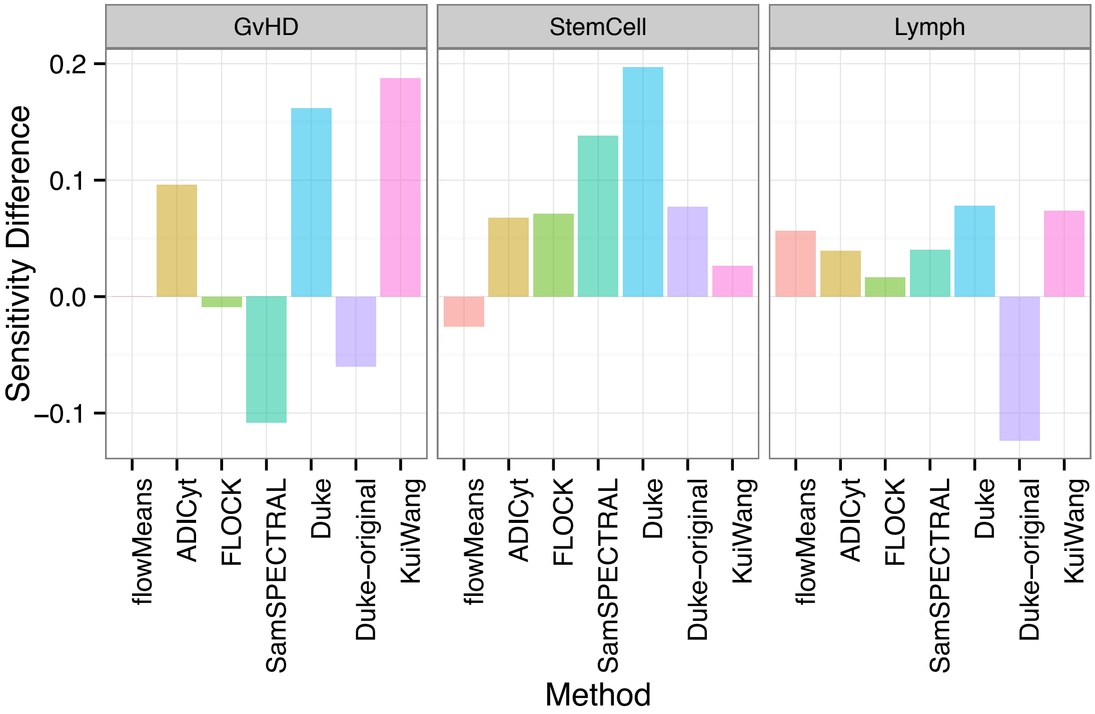

Figure S10: Gains in clustering sensitivity of FlowCAP-I algorithms when supplied with the number of manually-gated populations in a sample. Performance gains are computed as the difference in clustering sensitivity between challenge 1 results in which algorithms estimated the number of clusters in a sample and challenge 3 results in which algorithms were supplied with the the number of manually gated populations in a sample. The sensitivity of most clustering algorithms improves when provided the number of clusters in a dataset indicating that estimation of this parameter remains a challenge for clustering algorithms that produce a fixed partition of the data.

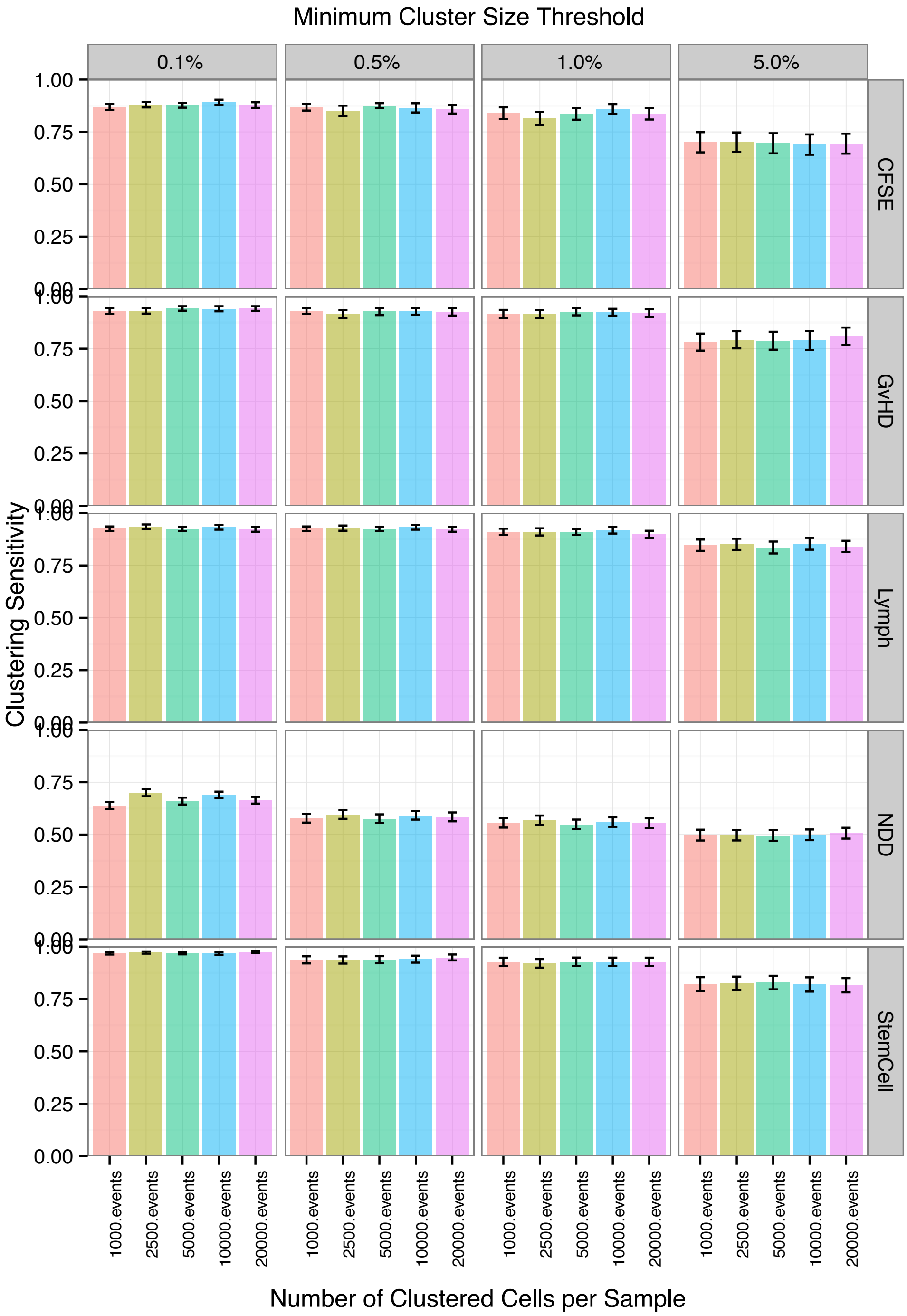

Figure S11: Clustering sensitivity as a function of number of events sampled per file in FlowCAP-I datasets. The clustering sensitivity of hierarchical clustering does not appear to be greatly affected by the number of events sampled per file.

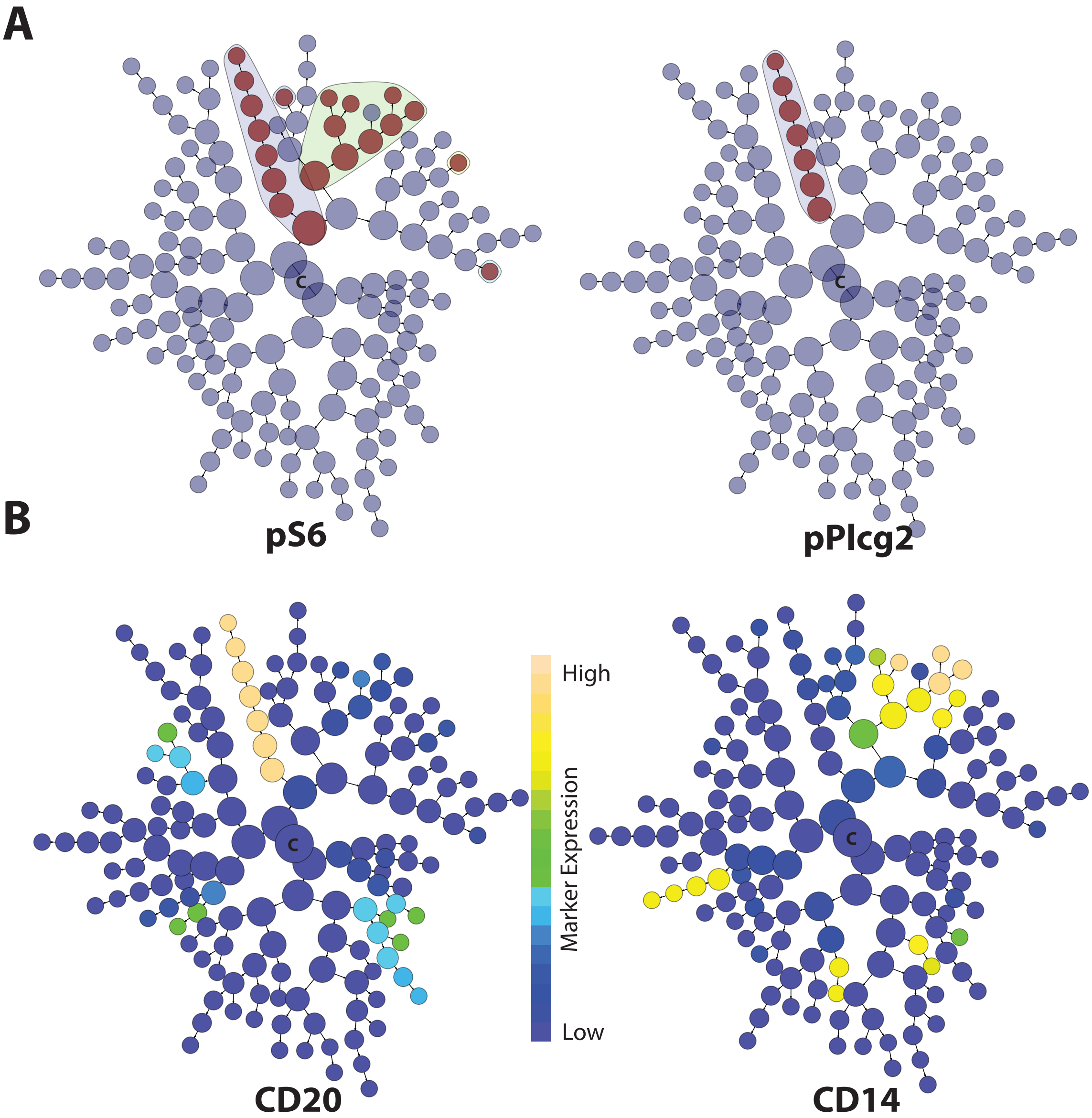

Figure S12: Hierarchy plots of stratifying features and markers from a Citrus analysis of the *Bodenmiller* PBMC dataset. All plots show the same clustering hierarchy. The cluster in the center of the graph marked with a 'C' is the root cluster that contains all cells. Each cluster is divided into two smaller clusters that are further towards the periphery of the graph. Clusters that are below the MCST of 1% are not displayed. (A) Cell subsets that display differential expression levels of phosphorylated S6 and PLCγ2 are shown in red. Contiguous branches of the clustering hierarchy having similar behavior (i.e. all branch subsets show differential expression of phosphorylated S6) are encircled. (B) Hierarchy plots colored by the median level of clustering markers CD20 and CD14. Brighter colors indicate higher levels of marker expression. Phosphorylated S6 levels differ between BCR-stimulated and unstimulated patients in cell subsets expressing high levels of CD20 and high levels of CD14. Figures S13 and S14 show hierarchy plots for all detected stratifying biological features and expression of all clustering markers respectively.

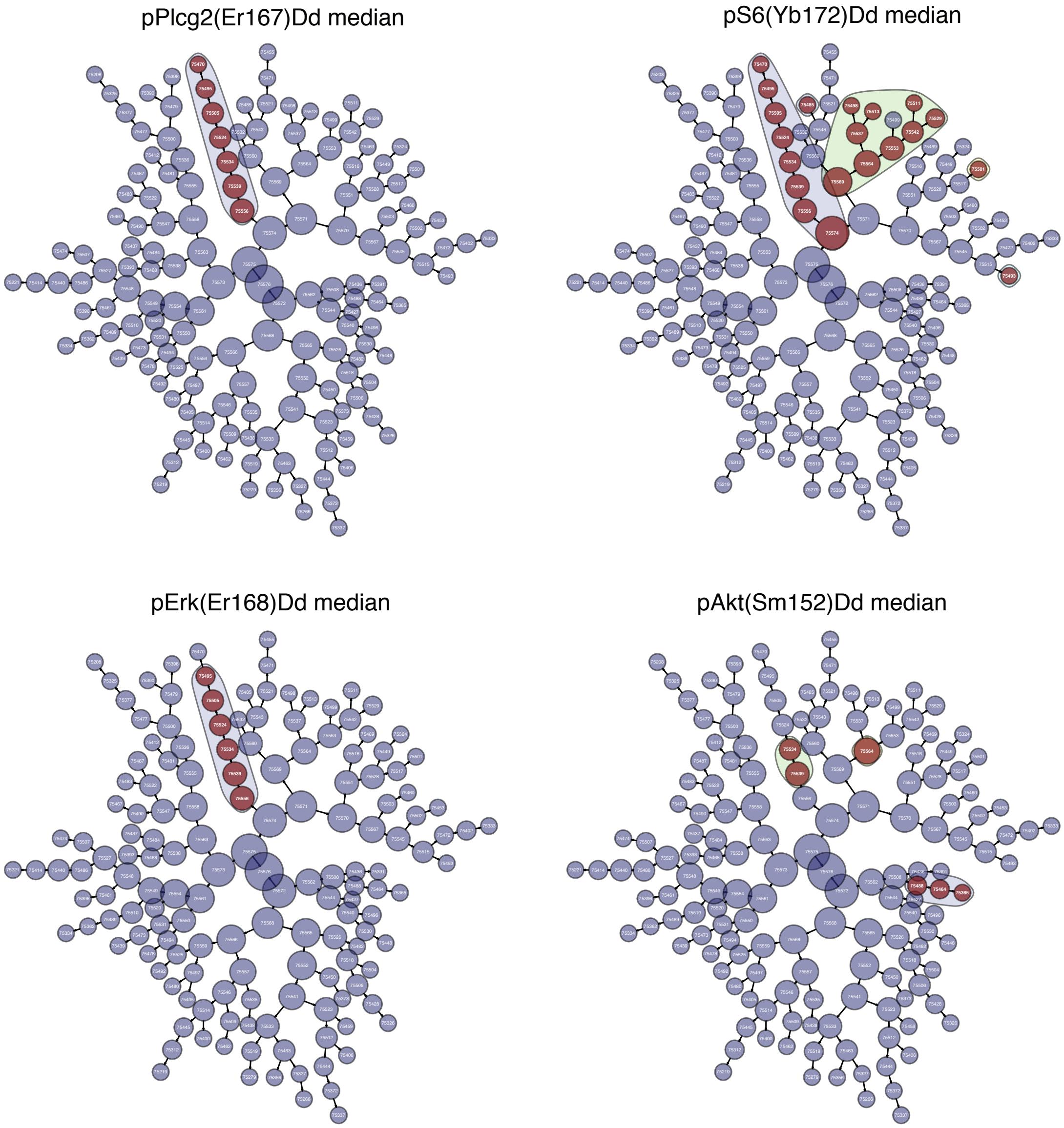

Figure S13: PBMC subsets that respond to BCR/FCR cross-linking, shown in the context of the clustering hierarchy. Phenotype plots for these clusters are shown in Figure S14.

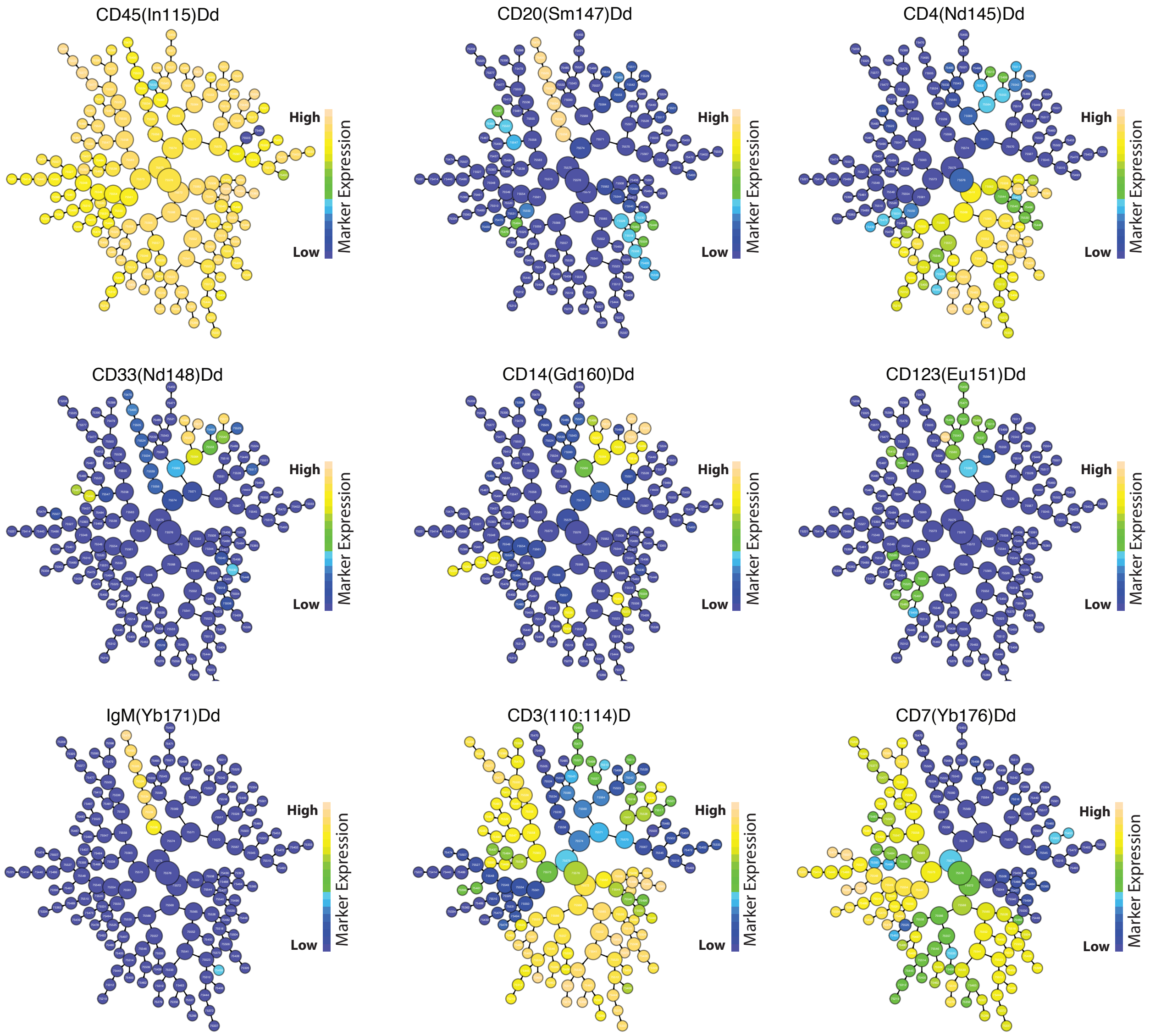

Figure S14: Clustering hierarchy of clustered PBMC data with clusters colored by the median value of each clustering marker in cluster cells. Functional responses to BCR/FCR cross-linking in these cell subsets are highlighted in Figure S13.

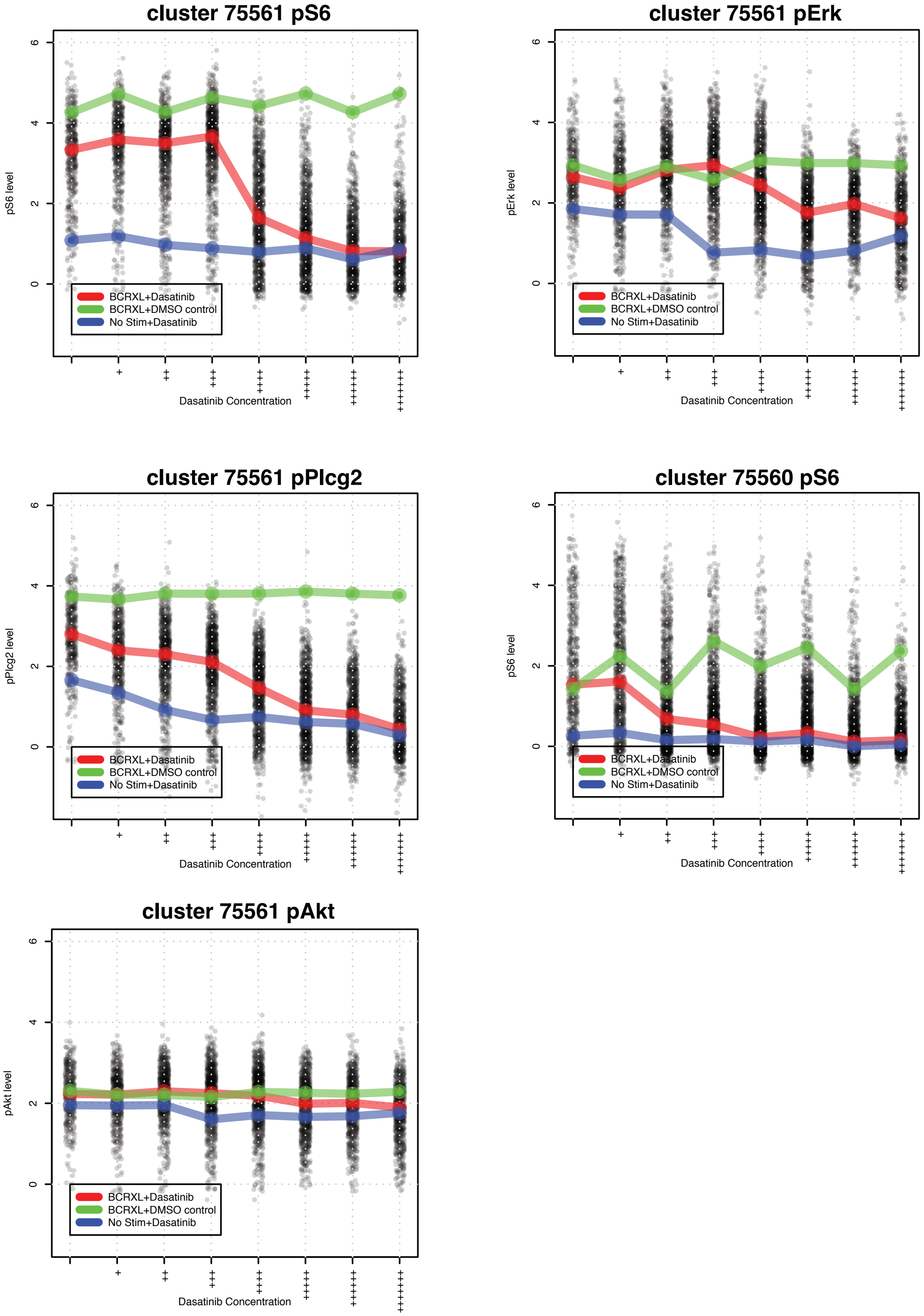

Figure S15: Effects of the inhibitor Dasatinib on functional responses identified by Citrus in B-cells (Cluster 75561) and monocytes (Cluster 75560).

S3 Supplemental Tables

Sample ID	Cluster 1 Abundance	...	Cluster 1 Feature F	...	Cluster C Abundance	...	Cluster C Feature F
Sample 1	1.34	...	85.3	...	0.22	...	4.31
Sample 2	1.52	...	24.4	...	0.18	...	6.81
...	...	...	...	...	...	...	...
Sample N	1.10	...	59.1	...	0.73	...	7.42

Table S1: An example data matrix with example values from F cluster features in each of the C clusters in N patients.

Dataset	# Samples	# Pops.	Dim.	Clustering Dims.	Min Cluster Size	Max Cluster Size
CFSE	13	51	8	PE-A, PE-Cy5-A, PE-Cy7-A, APC-A, Alexa Fluor 700-A, CFSE-A	0.91	67.58
GvHD	12	46	6	FL1.H, FL2.H, FL3.H, FL4.H	0.53	87.54
Lymph	30	88	5	FL1.LOG,FL2.LOG,FL4.LOG	2.08	71.20
NDD	30	210	12	FITC-A, PerCP-Cy5-5-A, Pacific Blue-A, Pacifc Orange-A, QDot 605-A, APC-A, Alexa 700-A, PE-A, PE-Cy5-A, PE-Cy7-A	0.06	63.70
StemCell	30	99	6	FL1-H, FL2-H, FL3-H, FL4-H	1.53	98.19

Table S2: Summary of FlowCAP-I Datasets. Minimum and maximum cluster size are reported as a proportion of a sample's events with a maximum value of 100. Column 3 (# Pops.) reports the total number of manually gated populations in all dataset samples.

Dataset	Recall	Precision	Accuracy	F-Measure
AML	0.95	1.0	0.99	0.97
HVTN	1.0	1.0	1.0	1.0

Table S3: Summary of Citrus classification performance on FlowCAP-II datasets

MCST Baseline	MCST Comparison	Relative Number of Features	Retained Significance Percentage
5.0%	2.5%	1.91	89%
2.5%	1.0%	2.57	67%
1.0%	0.75%	1.36	86%
0.75%	0.5%	1.53	85%

Table S4: Relative sensitivity of PMBC analyses run using different MCST's. Each row compares results between two Citrus analyses of the *Bodenmiller* PBMC dataset that were run with different MCST's. The *Retained Significance Percentage* measure summarizes the percent of stratifying features identified in a baseline Citrus analysis that were re-identified in an analysis run using a smaller MCST. Row 1 summarizes a comparison between a Citrus analysis run with an MCST of 5% and a Citrus analysis run with an MCST of 2.5%. The latter analysis included nearly twice as many features and detected 89% of significant features reported by the former.

S4 Supplementary Notes

S4.1 Note: Preparing Data and Setting Analysis Parameters

Data Preprocessing

Prior to analysis with Citrus, data should be cleaned and transformed in a manner that is consistent with standard manual analyses. For example, doublets and debris should be removed using scatter channels, dead cells removed using a viability marker, and data transformed using the logicle or other appropriate transformation. Additionally, if measured markers have different dynamic ranges, measurements may be standardized on a per-marker basis to ensure that each marker has an equal influence on clustering.

If analysis is to be restricted to a particular lineage of cells based on existing biological knowledge, cells from other lineages should be removed. If looking for responses in subsets of T-cells for instance, non T-cells should be gated and removed prior to analysis as this will reduce clustering time and increase statistical power to detect relevant populations within the T-cell compartment. See further comments in the section on setting the minimum cluster-size threshold.

Number of Cells Selected Per Sample

Citrus selects and combines an equal number of cells from each sample for analysis. Notably, the clustering sensitivity does not appear to be greatly affected by the number of events selected per sample (Fig. S11). However, descriptive statistics that are derived from clusters may have low precision if the number of cells in a cluster is small. For instance, the precision of a median phosphoprotein measure for a cell subset is likely poor if the subset contains only 5 cells. Thus, selecting more events per sample is likely to lead to more stable estimates of clusters features. As a default, Citrus selects 5,000 events per sample. However, one should adjust this number based on their minimum cluster size of interest. As a general rule, one can select a number of events per sample such that the minimum cluster of interest will have on average, 50 events. Thus, if considering a minimum cluster-size threshold of 1%, one would select 5,000 events per sample. Selection of more events per sample results in longer analysis runtimes as clustering runtime is a function of the number of events that must be clustered. More information on the runtime of clustering may be found on the Citrus GitHub page.

Choosing Clustering Markers

For common usage, one may cluster cells by the same markers that would be used to gate the data by hand. In many scenarios, this will simply be cell surface markers. However, if there are additional functional markers that also distinguish subpopulations of cells (i.e. activated B-Cells), one may cluster on those markers as well. In this scenario, abundance features would be used to measure the presence or absence of cells defined by such functional markers.

Selecting a minimum cluster-size threshold

The minimum cluster-size threshold (MCST) parameter controls the number of cell subsets included in the endpoint regression analysis. This parameter is expressed as a percentage of the number of total cells that have been clustered. An MCST of 1% specifies that a cluster must contain at least 1% of the total clustered events in order to be included in the regression analysis. Citrus specifies a very conservative default value of 5% although this value should be adjusted based on the number of cells selected per sample for clustering and existing biological knowledge.

Setting a large MCST will result in fewer cell subsets being included in the regression analysis, but will exclude more rare cell subsets. Setting a small MCST will include more rare subsets in the regression analysis but decrease statistical power. To optimize statistical power when searching for signal in many and/or cell subsets, users may either include more samples in the analysis or limit the number of features that are included in the regression analysis. There are several strategies for the latter approach that make use of existing biological knowledge. First, if the investigator has prior knowledge that suggests that the informative cell subset is rare, Citrus may be instructed to ignore more abundant subsets (i.e. cell subsets that contain more than 5% of cells). Additionally, if prior knowledge suggests that informative signal lies within particular lineages of cells (i.e. T-Cells), all non T-Cells may be removed from the dataset prior to analysis and a larger MCST may be used.

Selecting a classification model

Users may choose to build regression models using either or both of the nearest shrunken centroid and lasso-regularized regression methods. Importantly, the former employs univariate approach for classification while the latter uses a multivariate regression model. In other words, the nearest shrunken centroid approach evaluates the prognostic utility of each feature independently while the regularized regression approach builds a model based on a combination of signals found in different cell subsets. Accordingly, the nearest shrunken centroid method should be used when seeking to identify all clusters whose behavior differs between samples. Conversely, the L_1-regularized regression model should be used when identifying clusters that are combinatorially informative. The L_1-regularized regression model may also be used to identify cell subsets that are prognostic of continuous or time-valued clinical endpoints.

Evaluation of Results

Citrus estimates model accuracy using cross validation. Users may use these plots to assess the quality of results reported by Citrus. If a model has low cross-validation error, the user may have confidence that the cell subsets identified by Citrus have a unique behavior within each sample group. Examples of good and poor cross-validation results along with accompanying features are shown in Figure S7.

The investigator must determine what an acceptable error rate is on a per-experiment basis. For instance, a cell subset that uniquely differentiates patients 70% of the time may be acceptable for some experimental situations while other may require accuracy above 90%.

S4.2 Note: Differences in detected responses between Bodenmiller *et al.* and Citrus

Bodenmiller used 27 repeated measures (one from each inhibitor plate) from a single patient to determine phosphoprotein responses induced by BCR cross-linking. This approach provided a high-confidence assessment of responses in a single patient but was not necessarily reflective of responses found in all measured samples. Citrus did not detect pERK and pPLCγ2 responses in Dendritic cells reported by Bodenmiller *et al.*. To assess whether these responses were seen across all 8 measured patients, unpaired T-tests were used to compare the median value of both phosphoproteins from manually gated dendritic cells in stimulated and unstimulated patients. Differences in levels of ERK and PLCγ2 were not found to be statistically significant (P-values of 0.1453 and 0.5966 respectively) suggesting that the effect reported by Bodenmiller *et al.* was specific to a single patient. Alternatively, responses may be statistically insignificant due to inter-patient variability in which case measurement of additional samples could produce a significant result.

S4.3 Note: Visualization of Subsets of cells that exhibit similar behavior

Sets of correlated descriptive cluster properties may be derived from related cell subsets that have similar behavior. To enable the investigator to quickly identify which cell subsets display similar behavior, Citrus plots the clustering hierarchy and highlights stratifying cell subsets with similar responses. Fig. S12A shows an example of such a plot from a Citrus analysis of the *Bodenmiller* PBMC dataset with an MCST of 1.0%. Additionally, hierarchy clusters may be colored by the median value of a lineage marker in that cluster, providing investigators with an another means for determining the phenotype of stratifying clusters (Fig. S12B). Figures S13 and S14 show hierarchy plots for all detected stratifying biological features and expression of all clustering markers respectively.

Supplemental References

1. Bodenmiller B, et al. (2012) Multiplexed mass cytometry profiling of cellular states perturbed by small-molecule regulators. *Nat Biotechnol* 30(9):858–867.

2. Bendall SC, et al. (2011) Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. *Science* 332(6030):687–696.

3. Weintrob AC, et al. (2008) Increasing age at HIV seroconversion from 18 to 40 years is associated with favorable virologic and immunologic responses to HAART. *J Acquired Immune Deficiency Syndromes* 49(1):40–47.

4. Hahne F, Gopalakrishnan N, Khodabakhshi AH, Wong CJ, Lee K flowStats: Statistical methods for the analysis of flow cytometry data.

5. Simon N, Friedman JH, Hastie T, Tibshirani R (2011) Regularization paths for Cox's proportional hazards model via coordinate descent. *J Stat Softw* 39(5)

6. Heagerty PJ, Lumley T, Pepe MS (2000) Time-dependent ROC curves for censored survival data and a diagnostic marker. *Biometrics* 56(2):337–344.

7. Aghaeepour N, et al.; FlowCAP Consortium; DREAM Consortium (2013) Critical assessment of automated flow cytometry data analysis techniques. *NatMethods* 10(3):228–238.