CRITICAL ASSESSMENT OF AUTOMATED FLOW CYTOMETRY DATA ANALYSIS TECHNIQUES

Nima Aghaeepour1, Greg Finak2, The FlowCAP Consortium3, The DREAM Consortium3, Holger Hoos4, Tim R. Mosmann5, Raphael Gottardo2,*, Ryan Brinkman1,*, and Richard H. Scheuermann6,*

1Terry Fox Laboratory, British Columbia Cancer Agency, Vancouver, British Columbia, Canada
2Fred Hutchinson Cancer Research Center, Seattle, Washington, USA
4Department of Computer Science, University of British Columbia, Vancouver, British Columbia, Canada
5School of Medicine and Dentistry, University of Rochester, Rochester, New York, USA
6J. Craig Venter Institute, San Diego, California, USA

Abstract

Traditional methods for flow cytometry (FCM) data processing rely on subjective manual gating. Recently, several groups have developed computational methods for identifying cell populations in multidimensional FCM data. The Flow Cytometry: Critical Assessment of Population Identification Methods (FlowCAP) challenges were established to compare the performance of these methods on two tasks – mammalian cell population identification to determine if automated algorithms can reproduce expert manual gating, and sample classification to determine if analysis pipelines can identify characteristics that correlate with external variables (e.g., clinical outcome). This analysis presents the results of the first of these challenges. Several methods performed well compared to manual gating or external variables using statistical performance measures, suggesting that automated methods have reached a sufficient level of maturity and accuracy for reliable use in FCM data analysis.

Flow cytometers provide high dimensional quantitative measurement of light scatter and fluorescence emission properties of hundreds of thousands of individual cells within each analyzed sample. Flow cytometry (FCM) is used routinely both in research labs to study normal and abnormal cell structure and function, and in clinical labs to diagnose and
monitor human disease, and response to therapy and vaccination. In a typical FCM analysis, cells are stained with fluorochrome-conjugated antibodies that bind to cell surface and intracellular molecules. Within the flow cytometer, cells are passed sequentially through laser beams that excite the fluorochromes. The emitted light, which is proportional to the antigen density, is then measured. The latest flow cytometers can analyze 20 different characteristics for individual cells in complex mixtures\textsuperscript{1}, and recently developed mass spectrophotometry-based cytometers have the potential to dramatically increase this number\textsuperscript{2–4}.

A key step in the analysis of FCM data is the grouping of individual cell data records (i.e., events) into discrete populations based on similarities in light scattering and fluorescence. This analysis is usually accomplished by sequential manual partitioning (a.k.a. gating) of cell events into populations through visual inspection of plots in one or two dimensions at a time. Many problems have been noted with this approach to FCM data analysis, including its subjective and time-consuming nature, and the difficulty in effectively analyzing high dimensional data\textsuperscript{5}.

Beginning in 2007 there has been a surge in the development and application of computational methods to FCM data in an effort to overcome these serious limitations in manual gating-based analysis, with successful results reported in each case\textsuperscript{6–24}. However, it has been unclear how the results from these approaches compared with each other and with traditional manual gating results because every new algorithm was assessed using distinct datasets and evaluation methods. To address these shortcomings, members of the algorithm development, FCM user, and software and instrument vendor communities initiated the Flow Cytometry: Critical Assessment of Population Identification Methods (FlowCAP) project (http://flowcap.flowsite.org). The goals of FlowCAP are to advance the development of computational methods for the identification of cell populations of interest in FCM data by providing the means to objectively test and compare these methods, and to provide guidance to the end user about how best to use these algorithms. Here we report the results from the first two FlowCAP-sponsored competitions, which evaluated the ability of automated approaches to address two important use cases - cell population identification and sample classification.

**FlowCAP I: cell population identification challenges**

The goal of these challenges was to compare the results for assigning cell events to discrete cell populations from computational tools with manual gates produced by expert analysts. Algorithms competed in the four following challenges: Challenge 1: completely automated —comparison of completely automated gating algorithms for exploratory analysis. Software used in this challenge either did not have any tuning parameters (e.g. skewing parameters, density thresholds), or if there were tuning parameters, the values were fixed in advance and used across all datasets; Challenge 2: manually tuned—comparison of semi-automated gating algorithms with manually adjusted parameters tuned for individual datasets; Challenge 3: assignment of cells to populations with pre-defined number of populations—comparison of algorithms when the number of expected populations was known; and Challenge 4: supervised approaches trained using human-provided gates—similar to
challenge 2 with 25% of the manual gates (i.e., population membership labels) for each dataset provided to participants for training/tuning their algorithms.

Four human datasets (Graft versus Host Disease (GvHD, Diffuse Large B-cell Lymphoma (DLBCL) Symptomatic West Nile Virus (WNV); Normal Donors (ND)) and one mouse dataset (Hematopoietic Stem Cell Transplant (HSCT)) were used for these challenges. For details see the Online Methods section Cell population identification - dataset descriptions.

For these challenges the current standard practice for FCM data analysis of manual gating performed by expert analysts from the laboratory that generated datasets was used for comparison against cell population membership defined by each automated algorithm. The F-measure statistic (the harmonic mean of precision and recall; see the Online Methods section Cell population identification - clustering F-measure) was used for this comparison. An F-measure of 1.0 indicates perfect reproduction of the manual gating result with no false positive or false negative events.

Algorithm performance

Fourteen research groups submitted 36 analysis results (see a list of all participating programs in Table 1 and details of each algorithm in Supplementary Note 1). The results of the cell population identification challenges are summarized in Table 2 and Supplementary Figure 1. Not all algorithms were applied in all challenges. For example, supervised classification methods, like Radial SVM, require training data to establish classification rules, and therefore were not appropriate for Challenges 1–3. Algorithms were sorted by their rank performance score for each challenge (see Online Methods section Cell population identification challenge - rank score). Many algorithms performed well in multiple challenges on multiple datasets, with F-measures exceeding 0.85. Some algorithms were always in the top group (i.e., were not significantly different from the top algorithm) (e.g. ADICyt in Challenges 1 – 3, Sam SPECTRAL in Challenge 3), some were in the top group for some of the datasets (e.g. flowMeans, FLOCK, FLAME in Challenge 1), and some were never in the top group (e.g. flowKoh).

Allowing participants to tune algorithmic parameters did not result in much improvement as the highest overall F-measure did not increase (0.89 for both completely automated and manually tuned algorithms); only three of the six algorithms that participated in both Challenge 1 and Challenge 2 (Sam SPECTRAL, CDP, flowClust/Merge) demonstrated a modest improvement in overall F measure, and in some cases the F-measures actually decreased after human intervention (e.g., FLAME). In contrast, providing the number of cell populations sought in Challenge 3 made predictions more accurate for seven of the eight algorithms that participated in both Challenge 1 and Challenge 3, with five algorithms achieving overall F-measures greater than 0.9 (ADICyt, Sam SPECTRAL, flowMeans, TCLUST, FLOCK). In addition, providing a set of example results for algorithm training and parameter tuning in Challenge 4 improved the results of flowClust/Merge by 0.13, and allowed the Radial SVM approach to outperform the fully automated algorithms used in Challenge 1 in four of the five datasets. Taken together, these results suggest that estimating
the correct number of cell populations (as defined by manual gates) remains a challenge for most automated approaches, and providing training data improves performance.

Table 2 and Supplementary Figure 2 show the estimated runtimes of the algorithms on single core CPUs or GPUs (for CDP only). Runtimes ranged from 1 second to >4 hours per sample. ADICyt, which had the highest rank score in the first three challenges, also required the longest runtimes. flowMeans, FLOCK, FLAME, Sam SPECTRAL, and MM&PCA needed substantially shorter runtimes and still performed reasonably well in comparison with ADICyt. Note that, due to hardware and software differences, these numbers may not be precisely comparable; the information is provided to give some sense of the differences in time requirements of these specific implementations.

**Improving algorithmic performance by combining predictions**

Similar to other data analysis settings (see Yang *et al.* for a review), combining results from different cell population identification methods provides improved accuracy over any individual method. For all four cell population identification challenges, Ensemble Clustering (EC), which combines the results of all the submitted algorithms (see Online Methods section *Cell population identification challenge - ensemble clustering*), resulted in a higher overall F-measure and rank score than any individual algorithm (Table 2, Supplementary Figure 3 and Supplementary Figure 4). In addition, EC gave a higher F-measure for each of the individual datasets in each challenge, with only four exceptions in Challenge 4.

In addition to identifying cell populations more accurately, ensemble clustering can provide an alternative approach for evaluating algorithms by measuring their contribution to the combined predictions by using ablation analysis. For example in Challenge 3, when only 4 algorithms were included in the ensemble (i.e., TCLUST, ADICyt, FLAME, and SWIFT), the F-measure was still close to 0.95 (Supplementary Figure 5). Adding two more algorithms to the set resulted in only a minor improvement. Similar patterns were observed in the other challenges. Although the absolute order differed in the ablation analysis, algorithms with higher F-measures tended to be removed later (i.e., they had larger contribution to the ensemble). We also performed the ablation analysis in the reversed order (i.e., the algorithm with maximum contribution was removed first). As expected, the algorithms with a higher F-measure tend to be excluded earlier (Supplementary Figure 6).

**Algorithm performance with refined manual gates**

In the population identification challenges, pre-defined populations identified by human experts corresponded to a single set of manual gates prepared by the original data providers for comparison. However, manual gating is known to be subjective and potentially error-prone even in the hands of domain experts. Without detailed guidance on the goals of FlowCAP, the data providers tended to focus gating only on cells considered relevant to the goals of their study and therefore provided incomplete population delineation in some cases. In addition, by relying on a single set of gates, inconsistencies in manual gating between different analysts were not taken into account (see Supplementary Note 2). To address these deficiencies, eight individuals from five different institutions were instructed to identify all
cell populations (i.e., exhaustive gating) discernible within the HSCT and GvHD datasets (see Supplementary Note 2 for manual analysis instructions). These datasets were selected since they had the highest and lowest overall F-measures representing the best and worst cases for the automated methods, respectively.

A consensus of the eight manual gates was first constructed as a reference (see Online Methods section Cell population identification – consensus of manual gates). Algorithm comparison against this reference started with cell populations in the entire dataset that demonstrated the best match across all eight manual gates and then gradually proceeded to include more cell populations with weaker matches between the human analysts (Fig. 1). Including cell populations with less agreement between the human experts resulted in a gradual reduction in F-measures for both individual manual gates and algorithms, suggesting that certain populations were more difficult to resolve by both manual and automated analysis, especially for the GvHD dataset. However, the overall relative performance of algorithms for both datasets using these multiple sets of exhaustive gates were generally consistent with the initial results. For example, the top four algorithms for the HSCT dataset were FLAME, ADICyt, flowMeans, and MM&PCA for both the initial and the consensus manual gates (Supplementary Table 1). In addition, ensemble clustering performs well within the range of manual results especially for the most consistent populations.

As an alternative to the overall F-measures, consensus manual clusters were used as a reference in a per-population analysis (see Online Methods section Cell population identification – per population analysis) to determine if certain cell populations were responsible for high or low algorithm performance by determining F-measures for each cell population separately (Fig. 2, Supplementary Figure 7, and Supplementary Figure 8). For most populations in both samples, the high F-measure values highlight the close agreement between manual and automated results. For example, Cell Population #3 in the HSCT dataset demonstrates high pairwise F-measures between all of the algorithms and manual gates, indicating that this cell population was easily identified manually and algorithmically. In contrast, Cell Population #5 was only effectively identified by the manual gates and a few of the algorithms – SWIFT, ADICyt, CDP and FLOCK. Similar conclusions were reached for the GvHD dataset (Supplementary Figure 9 and Supplementary Figure 10).

**Practical considerations**

The F-measure analysis provides a rigorous quantitative measure of algorithm performance for population identification. Based on this analysis, while several algorithms performed well on individual datasets, combining the results of a subset of the algorithms produced better results than individual algorithms in almost every case. The per-population analysis showed that the best-matching algorithms were not always the same for each population suggesting that different algorithms may have different abilities to resolve populations depending on the exact structure of the data. This was not surprising given the wide range of strategies utilized by the different algorithms and motivates the recommendation for using an ensemble approach over any single algorithm for optimal performance.
Further demonstration of the practical utility of using ensemble clustering of automated algorithm results is provided through a visual example using the HSCT dataset (Fig. 3). Cell population classification by ensemble clustering was compared against consensus manual gating in two- and three-dimensional dot plots. Two samples were selected as examples of both strong and weak agreement between the computational and manual results. For both samples shown, cell events determined to be members of the same cell population by ensemble clustering were nearly always located within a single polygon from manual gating. CD45.1 and CD45.2 are allotype markers of murine hematopoietic cells that are frequently use to distinguish between donor and recipient cells after transplantation, with CD45.1 marking recipient cells and CD45.2 marking donor cells in this case. In one sample (Fig. 3 a,b), ensemble clustering identified some CD45.2 positive cells that were Ly65/Mac1 positive (granulocyte/monocyte; in green) and others that were Ly65/Mac1 negative (lymphocytes; in red), indicating repopulation of both major hematopoietic lineages and successful hematopoietic stem cells engraftment. In contrast, while the other sample (Fig. 3 c,d) was found to contain CD45.2 positive Ly65/mac1 negative lymphocytes, no CD45.2 positive Ly65/mac1 positive monocytes/granulocytes were observed, indicating unsuccessful stem cell engraftment. Thus, ensemble clustering was found to be an excellent method for automated assessment of hematopoietic stem cell engraftment using CD45 allotype markers in mouse models.

FlowCAPII: sample classification challenges

Another important use case for FCM analysis is the use of biomarker patterns in FCM data for the purposes of sample classification. We assembled a benchmark of three datasets in which the subjects/samples were associated with an external variable that could be used as an independent measure of truth for sample classification. The benchmark consisted of three datasets for: (1) studying the effect of HIV exposure on African infants that were HIV-exposed in utero, but uninfected (HEU) vs. unexposed (UE); (2) diagnosis of acute myeloid leukemia (AML) using AML and non-AML samples from a reference diagnostic laboratory; (3) discriminating between two antigen stimulation groups of post-HIV vaccination T-cells (Gag- vs. Env-stimulated) from the HIV Vaccine Trials Network (HVTN) - detailed descriptions can be found in the Online Methods section Sample classification – challenge descriptions. For each dataset, half of the correct sample classifications were provided to participants for training purposes; the other half was used for independent testing/validation. For the AML challenge, additional results where submitted through the DREAM (Dialogue for Reverse Engineering Analysis and Methods) initiative.

Algorithm Performance

We received a total of 43 submissions (algorithm descriptions are provided in Table 1 and Supplementary Note 1), including 14 through the DREAM project (see Supplementary Note 3). The results of this challenge are summarized in Table 3, Supplementary Figure 11 and Supplementary Table 2. The precision, recall, accuracy, and F-measure values on the test set show that for two of the datasets (AML and HVTN) many algorithms were able to perfectly predict the external variables. For example, flowCore-flowStats, flowType-FeaLect, Kmeanssym, PRAMS, SPADE and SWIFT all gave perfect classification accuracy (i.e. F-
measure = 1.0) on the HVTN dataset. For the third dataset (HEUvsUE), despite mostly accurate predictions on the training data, none of the algorithms performed well on the test data. The lack of good performance of any algorithm on this dataset combined with a theoretical consideration of the underlying biology (non-productive HIV exposure several months before sampling may not lead to long term changes in peripheral blood cell populations) suggests that these samples may be unclassifiable based on the FCM markers used.

Outlier Analysis

In all datasets, the misclassifications were uniformly distributed across the test sets (Fig. 4a, Supplementary Figure 12, and Supplementary Figure 13), with only a single exception (sample #340 of the AML dataset), suggesting that no systematic problems were causing misclassifications. Visualization of FCM data from the sample #340 outlier in comparison with typical AML and non-AML subjects suggested that the outlier, like typical AML cases, had a sizable CD34+ population, however, the forward scatter values overlap with those of normal lymphocytes (Fig. 4 b-g). Obtaining additional information on this patient was not possible. However, independent evaluation of the FCM results by a hematopathologist suggested alternative explanations for why this sample was an outlier: The forward scatter (roughly proportional to the diameter of the cell) of the blasts was lower than that found in other AML patients. Leukemic blast size shows wide variations from patient to patient, and even within a given patient, being medium to large in size in most\(^{31}\), and very small (“microblastic”) in rare patients (e.g.,\(^{32,33}\)). The other possibility is that given the lower blasts frequency (16.7%), this patient may have been diagnosed with high grade myelodysplasia (blasts 10–19%), a preleukemic condition, rather than AML, which requires a blast count of >20% for diagnosis. Alternately, the patient may have AML by morphological blast count, but FCM may be underestimating the blast frequency because of hemodilution of the bone marrow specimen or presence of cell debris or unlysed red blood cells\(^{34}\).

Predictive Cell Populations Identified

Previous manual gating-based analysis of the HVTN data identified the CD4+/IL2+ T–cell subpopulation as discriminative between Env- and Gag-stimulated samples, with the proportion of CD4+/IL2+ cells in the Env-stimulated samples systematically higher than in the Gag-stimulated samples (data not shown). This effect was not observed in manually gated placebo data, indicating that it is vaccine specific, and consistent with the gp120 Env protein boost given to study participants. Interestingly, examination of the features selected by automated methods for classification between Env- and Gag-stimulated samples revealed that, of the eight methods that directly identified predictive features, four selected features containing the CD4+/IL2+ phenotype. The sample classifications using the CD4+/IL2+ population gated manually were slightly less accurate than the automatic results obtained from the same population. Post-hoc examination of the data revealed that several of the control and stimulated samples in the data set were matched from different experimental runs, suggesting a possible run–specific effect. When these samples were filtered out of the analysis, manual gating was able to perform as accurately as the algorithms, suggesting that...
the algorithmic approaches were actually more robust to the technical variation than the manual analysis. For more details see Supplementary Note 4.

**Practical considerations**

Of the three datasets assembled to test algorithms in the sample classification challenge, the AML dataset represents an important real world patient classification use case. FCM is the laboratory method of choice for the diagnosis of acute leukemia since it not only allows for the identification of abnormal cell populations in comparison with normal blood or bone marrow but also allows for the classification of the disease into different subtypes with different prognoses and treatment options. Of the 25 algorithms that participated in the AML sample classification challenge, 12 provided perfect classification of all 359 patient samples (F-measure = 1.00) into the AML versus non-AML categories using data from 2872 separate FCM staining samples. An additional 8 algorithms were only discrepant on Sample #340 classification, which although labeled as a non-AML sample appears to be a borderline case. This impressive result, in which 80% of the automated methods performed near perfectly in the classification of acute leukemia indicates that these methods can now be incorporated into diagnostics pathology laboratory workflows for the diagnosis of AML, and possibly other neoplastic diseases, thereby eliminating the labor-intensive, subjective and error-prone features of manual analysis.

The HVTN challenge represented a relatively difficult problem of distinguishing between T cell responses to two viral antigens present in the same HIV vaccine. Based on the modest results of previous manual analysis (data not shown), we were surprised by the high performance of classification algorithms in the HVTN challenge. This was an important conclusion of this part of FlowCAP - that several sample classification algorithms performed much better than expected. Importantly, two of the four algorithms that provided results for both of the datasets (flowType-FeaLect and SPADE) gave perfect classifications for both, suggesting that automated methods perform very well in sample classification, even for datasets that were challenging for manual analysis.

**Discussion**

The FlowCAP project represents a community effort to develop and implement evaluation strategies to judge the performance of computational methods developed for FCM data analysis. Two sets of benchmark FCM data were assembled to evaluate automated gating methods based on their ability to either reproduce cell populations defined through expert manual gating, or their ability to classify samples based on external variables. Seventy-seven different computational pipeline/challenge combinations were evaluated through these efforts. Every approach to automated FCM analysis published in the last five years, as well as several unpublished methods, participated in at least one of the challenges. Participation by the flow informatics community was not only widespread, it was also collaborative, including the sharing of ideas and the distribution of work to avoid duplication of efforts. The recent establishment of the flow informatics discipline has also coincided with the growth of the open source software philosophy, which has been widely adopted by the flow informatics community. This open access philosophy has most certainly contributed to the
rapid maturation of these novel methods. One of the sample classification challenges was organized in collaboration with the DREAM (Dialogue for Reverse Engineering Analysis and Methods) initiative, which aims at nucleating the systems biology community around important computational biology problems. Given the growing use of FCM data in systems biology research, the collaboration between DREAM and FlowCAP was natural and fruitful.

One of the major goals of the FlowCAP project was to determine if automated algorithms had reached a level of maturity that they could be considered practically useful for routine FCM data analysis. While none of the individual methods provided perfect results for all use cases and sample sets, the results clearly show that automated methods are now practical for many FCM use cases. From the Cell Population Identification challenges it is now clear that many of the individual algorithmic techniques provide excellent delineation of many different cell populations in diverse datasets. Since users are often focused on the analysis of well-defined subsets of cell populations in a given experiment, many high-ranking techniques (especially those that can learn from manual gating examples) appear to be well suited for this purpose.

In addition, ensemble clustering provides further improvement by combining the best results from multiple methods, giving excellent performance across all of the cell population identification datasets. The mean F-measure values and rank scores showed that the combined predictions obtained by ensemble clustering were more accurate than the results from individual algorithms and individual manual gates. This is important because in practice it may not be feasible to solicit multiple experts for manual gating, however it is realistic to run multiple automated methods at minimal cost. The ablation analysis (presented in Supplementary Note 3) confirmed that increasing the number of algorithms in the ensemble resulted in improved predictions up to a certain point. In cases where algorithms with high scores were more frequent, the ensemble clustering performed better and was less sensitive to the exclusion of several of the algorithms (Challenges 1 and 3). This suggests that having a number of good algorithms is necessary to obtain good ensemble results, but there might be a point after which adding more algorithms does not significantly improve the results. Particularly, when a large number of algorithms with high F-measures were available (the entire HSCT dataset and the top 50 most consistently identified populations in the GvHD dataset), the ensemble clustering out-performed the individual algorithms. When the individual algorithms were performing poorly (the remaining cell populations in the GvHD dataset), the ensemble clustering’s performance decreased as well. However, it remains to be determined if this reflects a poor performance of the automated methods or poor performance of manual gating.

In the sample classification challenges, many individual methods provided perfect sample classification accuracy for two different representative datasets, with the leukemia classification use case being an important practical example. The excellent performance of automated methods, even with the relatively challenging HVTN dataset, was somewhat surprising but indicates that automated methods can perform well on sample classification use cases, detecting useful biomarkers in FCM data. While this result is promising, it will be important to obtain additional sample classification datasets for future FlowCAP challenges.
in order to determine if they have reached a level of maturity for broad routine use, especially for clinical diagnosis applications. The third dataset (HEUvsUE), in which none of the algorithms performed well, revealed an additional interesting outcome from the sample classification challenges – situations in which algorithms consistently perform well on training data but poorly on test data may indicate sample sets that are not classifiable given the data provided.

In conclusion, the FlowCAP project has provided a valuable venue for comparison of computational methods for FCM data analysis. While there is still much to be done to make these methods optimally useful and broadly adopted (see Supplementary Note 5 for future FlowCAP challenges), the results presented here are promising and suggest that automated methods will soon supplement manual FCM data analysis methods. The ability to rapidly, objective and collaboratively compare these methods through FlowCAP should catalyze rapid progress in the flow informatics field.

**Online Methods**

**Availability**

To promote reproducible research, the detailed methodologies for all approaches participating in FlowCAP are included by reference to free, open source software packages, algorithms, or through detailed descriptions (as pseudocode) as described in the Supplementary Note 1. The display items presented in this manuscript can be fully reproduced using the scripts provided on the FlowCAP website (http://flowcap.flowsite.org/codeanddata). Annotated raw data using MIFlowCyt descriptions is available through FlowRepository.org using the following experiment IDs: FR-FCM-ZZY2 (GvHD), FR-FCM-ZZYY (DLBCL), FR-FCM-ZZY3 (WNV), FR-FCM-ZZY6 (HSCT), FR-FCM-ZZZY (ND), FR-FCM-ZZZU (HEUvsUE), FR-FCM-ZZYA (AML), and FR-FCM-ZZZV (HVTN).

**Cell population identification - dataset descriptions**

The following datasets were used in the Cell Population Identification challenges:

**Diffuse Large B-cell Lymphoma (DLBCL)**—The DLBCL dataset consists of data from 30 randomly selected lymph node biopsies from patients treated at the British Columbia Cancer Agency between 2003 and 2008. Cell suspensions were produced from freshly disaggregated lymph node biopsies. Patients were histologically confirmed to have diffuse large B-cell lymphoma (DLBCL). This dataset was provided by Andrew Weng at the BCCRC.

**Symptomatic West Nile Virus (WNV)**—Samples are human peripheral blood mononuclear cells (PBMC) from patients with symptomatic West Nile virus infection stimulated in vitro with peptide pools representing different regions of the WNV polyprotein. This dataset was provided by Jonathan Bramson at McMaster University.

**Normal Donors (ND)**—For this dataset, the investigators examined differences in the response of a variety of cell types to various stimuli for a set of healthy donors. For the
samples used here, the time periods were relatively short, such that the surface markers would not be expected to change. The staining panel contains antibodies to surface markers and intracellular proteins. Note that these experiment were done with phosflow-fixed cells, and thus some of the populations are not as distinct or clean as would be seen with other processing methods. This dataset was provided by Hugh Rand at Amgen, Inc.

**Hematopoietic Stem Cell Transplant (HSCT)**—This dataset contains data from 30 randomly selected samples derived from hematopoietic stem cell transplant experiments done in the Terry Fox Laboratory. Suspensions were produced from bone marrow cells. The suspensions were depleted of erythroid precursors by immunomagnetic removal of biotin-conjugated anti-Ter119-labeled cells using EasySep reagents (STEMCELL Technologies, Vancouver, BC, Canada). This dataset was provided by the Connie Eaves at the BCCRC.

**Graft versus Host Disease (GvHD)**—Twelve FCM samples for finding cellular signatures to predict or correlate with early detection of GvHD. PBMC were collected from patients pre and post allogeneic blood and marrow transplantation. Cells were isolated using Ficoll-Hypaque and then were cryopreserved for subsequent batch analysis. The dataset was publicly available as part of previous research with additional analysis provided by Jill Schoenfeld at Treestar, Inc.

The protein markers evaluated are listed in Supplementary Table 4.

**Cell population identification - data preprocessing**

The following pre-processing steps were applied to these datasets before providing them to the participants: (1) compensation (to account for the overlap of emission spectra from fluorochrome labels); (2) transformation to linear space (to scale data appropriately for visualization); (3) pre-gating for removal of irrelevant cells (e.g., dead cells, as routinely performed by human analysts).

**Cell population identification - clustering F-measure**

F-measure is the harmonic mean of the precision and recall according to the equation. 

\[
F = \frac{2 \cdot Pr \cdot Re}{Pr + Re}
\]

Precision (Pr) and recall (Re) can be described in terms of a 2 × 2 contingency table comparing results for a test method, in this case the results of a cell population identification algorithm, with some reference method, in this case the results of manual gating by the subject matter expert as the current standard practice, with true positive (TP) defined as the situation in which the positive assignment of the prediction algorithm matches a positive assignment of manual gating, false positive (FP) when the positive assignment of the prediction algorithm matches a negative assignment of manual gating, and false negative (FN) when the negative assignment of the prediction algorithm matches a positive assignment of manual gating. Recall is calculated as TP/(TP + FN); precision is calculated as TP/(TP + FP). F-measure values are always in the interval [0,1], with 1 indicating a perfect prediction.

In this analysis Pr corresponds to the number of cells correctly assigned to a cluster divided by the total cells assigned to that cluster, and Re corresponds to the number of cells correctly
assigned to a cluster divided by all the cells that should have been assigned to that cluster. Given a correct set of reference clusters \( C = \{c_1, c_2, \ldots, c_n\} \), and a clustering result \( K = \{k_1, k_2, \ldots, k_m\} \), the number of matches between combinations of \( C \) and \( K \) is a matrix, \( M = [a_{ij}] \), where \( i \in [1,n] \) and \( j \in [1,m] \). Then \( Pr(c_i, k_j) = a_{ij}/|k_j| \) and \( Re(c_i, k_j) = a_{ij}/|c_i| \), where \( |c_i| \) denotes the number of elements in \( c_i \). The F-measure to compare one cluster to another is then \( F(c_i, k_j) = (2 \cdot Pr(c_i, k_j) \cdot Re(c_i, k_j)) / (Pr(c_i, k_j) + Re(c_i, k_j)) \). To calculate the F-measure of an entire clustering result, for each cluster \( c_j \) in the reference, a set of F-measures against every predicted cluster \( k_j \) is calculated, and the largest F-measure (best match), normalized by the size of \( k_j \) is reported. The sum of these scores produced a total F measure, defined as

\[
F(C, K) = \sum_{c_i \in C} \frac{c_i}{N} \cdot \max_{k_j \in K} \{F(c_i, k_j)\}.
\]

To show the relationship between F-measure and recall/precision, recall, precision, and F-measure values were plotted for flowMeans when the number of clusters was iterated from 2 to 10 (Supplementary Figure 14), using the same HSCT sample plotted in the main manuscript. For this sample, 4 populations were identified by manual gating, whereas ensemble clustering suggested that there are 5 populations. This figure provides some intuition about F-measure behavior. For example, missing one cluster (total of 3 clusters) results in a drop of less than 0.05 in F-measure, but missing two clusters (total of 2 clusters) results in a drop of 0.3. However, identifying an additional cluster (remember that the ensemble clustering suggested that there are actually 5 real populations) doesn’t decrease the F-measure. The figure also shows the trade-off between recall and precision. From 2 to 5 populations, recall and F-measure increase and precision decrease slightly. After that, precision decreases quickly while recall remains constant, resulting in a decrease in F-measure. F measure is relatively low when either recall or precision is low.

See Aghaeepour et al.,\(^38\) for a comparison of F-measure versus other metrics in the evaluation of clustering algorithms.

While mean F-measures can be used to assess the performance of each of the algorithms on each dataset, the significance of the difference in the F-measure values must be accounted for in order to truly rank the algorithms. Therefore, to measure how significant these differences were (i.e., how sensitive they are to this specific set of samples), bootstrapping was used to compute 95% confidence intervals (CIs). Bootstrapping is a non-parametric, resampling based method for measuring the accuracy of a sample estimate\(^39\). For a vector \( F \) of F-measure values produced by a given algorithm on a given dataset we produced the 95% bootstrap percentile CI for the mean as follows: (1) Repeat 10,000 times: sample from \( F \) with replacement (sample size = size of \( F \)), and calculate the mean F-measure of the sample; (2) Report the 2.5th and 97.5th percentiles of the average F-measures as the CI; (3) End. The results are presented in Supplementary Figure 1. Algorithms with overlapping CIs were subsequently considered tied (bolded in Table 2).

**Cell population identification - rank score**

To derive an overall ranking of the algorithms, we used their rank score calculated as the sum of fractional rankings of each algorithm across different datasets. Fractional ranking is based on the Borda count strategy\(^40\) - for \( N \) algorithms, the top algorithm scored \( N \) points,
the second one \( N-1 \) points, and so on. The last algorithm scored 1 point. The average number of points was used in case of ties (i.e., overlapping CIs). For \( D \) datasets, rank score values are in the \([D, N \times D]\) interval; an algorithm that scored first in every dataset would have a rank equal to \( N \times D \).

**Cell population identification - ensemble clustering**

To evaluate the hypothesis that a consensus of all methods would provide a result better than any individual method, populations that were identified by all methods were combined using ensemble clustering. The consensus clustering problem is defined as follows: given a set of partitions (the ensemble), find a new partition \( P \), that minimizes the dissimilarity between \( P \) and the partitions in the ensemble. A partition \( M \) is defined as a binary matrix with each column corresponding to a class label. The dissimilarity between a partition \( P \) and a partition element of the ensemble \( Q \) is defined as,

\[
d(P, Q) = \min_{\Pi} \| P - Q \cdot \Pi \|_p
\]

where \( \| \cdot \|_p \) is the entry-wise p-norm. The permutation matrix provides a mapping between corresponding classes. For example given three observations \( x, y, z \), one partition may label the observations as \( x \in A \), \( y \in B \), \( z \in C \) and another may label the observations (with independent labels) as \( y \in \alpha, x \in \gamma, c \in \gamma \). The partitions in fact are the same if we consider the classes as \( A = \gamma \), \( B = \alpha \), \( C = \gamma \). The permutation matrix \( \Pi \) determines how the classes in \( P \) correspond with the classes in \( Q \). When \( p = 1 \), the measure is known as the Manhattan distance. This distance can be calculated efficiently using linear programming methods.

Once a dissimilarity measure is defined, in our case the Manhattan distance with \( p = 1 \), we must solve the harder problem of finding the partition \( p^* \) that minimizes the distance for all of the partitions \( Q \) in the ensemble \( E \).

\[
P^* = \arg\min_P \sum_{Q \in E} \min_{\Pi} \| P - Q \cdot \Pi \|_1
\]

This is an NP-hard problem (multi-dimensional assignment) so we used a heuristic method\(^{41}\) that provides approximate solutions for the consensus partition problem, as implemented in the CLUE package\(^{42}\).

Ablation analysis was performed as follows. For a set of \( N \) algorithms \( A = \{a_1, a_2, \ldots, a_N\} \), and an ensemble clustering result \( EC \), the following steps were performed to measure the contribution of each individual algorithm to the \( EC \): (1) Find the algorithm \( a_i \) that results in the smallest reduction in F-measure when excluded from the \( EC \); (2) Remove \( a_i \) from \( EC \); (3) Record the F-measure of \( EC \); (4) If \( A \) is not empty, go to (1); (5) End.

**Cell population identification – consensus of manual gates**

As discussed in the main text, consensus clustering of manual gates was used to rank the algorithms in the refined manual gate analysis. For each population in the consensus clusters, the mean F-measure to the matching population in all other manual gates was
calculated. A comparison of the relationship between the score assigned to each cell population in the consensus was compared with the absolute or relative cell frequency in linear or log space (Supplementary Figure 15 – S17). This showed that there usually is considerable agreement between human experts and their consensus for large cell populations. However, for small populations there was often (although not always) considerable disagreement across the experts. For this reason, we focused our ranking on cell populations with an F-measure of higher than 0.8. For evaluation of the algorithms, we started by limiting the comparison to only those cell populations that matched strongly across all manual gates (F-measure cutoff=1) and relaxed this condition gradually (Figure 1).

After completing the comparison between these independent manual gates and the automated results it became apparent that one and perhaps two sets of manual gates were somewhat different from the others. We considered whether it might be appropriate to remove these from the ensemble of manual gates that was used in the F-measure comparison since they might be statistical outliers. However, the differences between the individual gates represent an expert’s valid interpretation of the data rather than statistical noise or outliers, a conclusion supported by the observation that the outlier effect is only observable in a subset of the cell populations. That two of the gating results diverge from the others is not a sufficient justification for calling them outliers or discarding them. Removing these two sets of manual gates would, in fact, bias the results of our study since the decision would have been made after observing the results. For this reason, we would argue that removal of an outlier set of manual gates from this analysis is not scientifically or statistically justified. Indeed, this wide variation in manual gating analysis reflects the current state of flow cytometry analysis, and provides additional support for the importance of adopting objective automated approaches.

**Cell population identification – per population analysis**

Human consensus clustering results were matched across samples to the sample with maximum number of populations. Then, the human consensus for each sample was used as a reference for matching of the automated results of that sample. Pairwise F-measures between all algorithms and manual gates for the HSCT and GvHD datasets are shown in Fig. 2 and Supplementary Figure 9, respectively. The dendrograms were calculated using the complete-linkage hierarchical clustering and Euclidean distance between the F-measures as the metric.

These results can be used to identify cell populations that are responsible for high (or low) F-measures for further visual investigation. For example, Cell Population #3 in the HSCT dataset demonstrates a high overall pairwise F-measure between all of the algorithms and manual gates (Fig. 2), suggesting that this cell population has been relatively easy to identify. This was visually confirmed in Supplementary Figure 7 and S8. In contrast, Cell Population #2 in the GvHD dataset represents a cell population that was only identified by manual gating (Supplementary Figure 9). Further evaluation shows that this population (colored in red) is generally identical to the cyan population in every channel but has a lower FSC (Supplementary Figure 10). This emphasizes the importance of designing
methodologies that can use background biological knowledge in the clustering process. In this case, the humans used their knowledge about the scatter channels to partition these cells into two different populations based on cell size despite their similarity in every other channel (see Supplementary Figure 18 for a density plot of the sample).

Sample classification – challenge descriptions

FlowCAP-II included three datasets for sample classification (markers are listed in Supplementary Table 4).

Challenge 1: HIV-Exposed-Uninfected versus Un-exposed (HEUvsUE)—The goal of this challenge was to find cell populations that can be used to discriminate between HEU (n = 20) and UE (n = 24) infants. Blood samples were taken at 6 months after birth and were left unstimulated (for control) or stimulated with 6 Toll-like receptor ligands. In addition to raw FCS files, half of the subject labels were provided for training purposes. Algorithms were to use this data to label the rest of the samples. These labels were used to evaluate algorithm performance.

Challenge 2: Acute Myeloid Leukemia (AML)—The goal of this challenge was to find cell populations that can discriminate between AML positive (n = 43) and healthy donor (n = 316) patients. Peripheral blood or bone marrow aspirate samples were collected over a 1-year period using 8 tubes (tube #1 is an isotype control and #8 is unstained) with different marker combinations. In addition to raw FCS files, half of the subject labels were provided for training purposes. Algorithms were to use this data to label the rest of the samples. These labels were be used to evaluate algorithm performance.

Challenge 3: Identification of Antigen Stimulation Group of Intracellular Cytokine Staining of Post-HIV Vaccine Antigen Stimulated T-cells (HVTN)—The goal of this challenge was to correctly label the antigen stimulation group of post-HIV vaccine T-cells. The data set contains samples from 48 individuals from the HIV Vaccine Trials Network (HVTN). Each individual received an experimental HIV vaccine. Samples were collected approximately 10 months later and T-cells challenged with two antigens ENV-1-PTEG and GAG-1-PTEG. The response of CD4+ and CD8+ T-cells was measured by flow cytometry for each group. Cells were found to respond differently to the two antigen stimulations. This is essentially a classification challenge (see Supplementary Figure 19 for an example). For training purposes we provided data from 24 individuals within each group. The antigen stimulation label was provided. Participants were to correctly identify the antigen stimulation group of the test data (n = 24). The complete data set consisted of 240 FCS files. The data was compensated, transformed and partially gated (gated for singlets, live cells and lymphocytes).

Sample classification - classification F-measure

F-measure for classification is defined as the harmonic mean of precision and recall (the additional “matching” step for clustering F-measure is not required). Precision is defined as \( TP/(TP+FP) \) and recall is defined as \( TP/(TP+FN) \), where \( TP \), \( TN \), \( FP \), and \( FN \) are true
positives (e.g., and AML predicted as AML), true negatives, false positives, and false negatives, respectively.

Participants in the DREAM6/FlowCAP-II challenge were required to submit a list of subjects ordered according to the confidence assigned to the subject being affected with AML. That allowed us to compute more metrics than the ones used in the other FlowCAP challenges (see Supplementary Note 4).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The FlowCAP summits - held on the NIH campus, Bethesda, MD, United States, 2010 and 2011 - were generously sponsored by NIH/NIAID. For more information please visit the FlowCAP website (flowcap.flowsite.org). This work was partially supported by the following grants: NIH/R01EB008400, NIH/N01AI40076, NIH/R01NS067305, NIH/RC2-GM093080, CCS #700374, NSERC (HH) and by the Terry Fox Foundation and Terry Fox Research Institute (RRB). This work was partially supported by the following scholarships: CIHR/MSFHR (NA), UBC’s 4YF (NA), ISAC Scholar (NA), MSFHR Scholar (RRB), HWN Scholar (PLD), Rachford and Carlota A. Harris Professorship (GPN).

Appendix

Authors Contributions

Nima Aghaeepour1, Greg Finak2, Holger Hoos4, Tim R. Mosmann5, Raphael Gottardo2, Ryan Brinkman1, and Richard H. Scheuermann6 were responsible for the formation of the FlowCAP consortium, the development of all FlowCAP challenges, results evaluation and manuscript preparation. In addition, members of the FlowCAP consortium contributed as follows:

Population identification challenge data analysis team:

David Dougall7, Alireza Hadj Khodabakhshi8, Phillip Mah4, Gerlinde Obermose9, Josef Spidlen1, Ian Taylor10, Sherry A Wuensch5.

Population identification challenge data providers:

Jonathan Bramson11, Connie Eaves12, Andrew P. Weng12, Edgardo S. Fortuno III13, Kevin Ho13, Tobias Kollmann13, Wade Rogers14, Stephen De Rosa15.

Clinical and hematopathological consultation:

Bakul Dalal16.

Population identification challenge algorithm developers and challenge participants:

Ariful Azad17, Alex Pothen17, Aaron Brandes18, Hannes Bretschneider19, Robert Bruggner20, Rachel Finck20, Robin Jia20, Noah Zimmerman20, Michael Linderman20, David Dill20, Gary Nolan20, Cliburn Chan21, Faysal El Khettabi1, Kieran O’Neill1, Maria
Members of the DREAM consortium contributed as follows:

**Sample classification challenge designers:**

Pablo Meyer$^{34}$, Gustavo Stolovitzky$^{34}$, Julio Saez-Rodriguez$^{35}$.

**Sample classification challenge data analysis team:**

Raquel Norel$^{34}$.

**Sample classification challenge algorithm developers and challenge participants:**

Madhuchhanda Bhattacharjee$^{36}$, Michael Biehl$^{37}$, Philipp Bucher$^{38}$, Kerstin Bunte$^{39}$, Barbara Di Camillo$^{40}$, Francesco Sambo$^{40}$, Tiziana Sanavia$^{40}$, Emanuele Trifoglio$^{40}$, Gianna Toffolo$^{40}$, Slavica Dimitrieva$^{38}$, Rene Dreos$^{38}$, Giovanna Ambrosini$^{38}$, Jan Grau$^{41}$, Ivo Grosse$^{41}$, Stefan Posch$^{41}$, Nicolas Guex$^{32}$, Jens Keilwagen$^{24}$, Miron Kursa$^{13}$, Witold Rudnicki$^{43}$, Bo Liu$^{14}$, Mark Maireschein-Cline$^{45}$, Tapio Manninen$^{27}$, Heikki Huttunen$^{27}$, Pekka Ruusuvuori$^{27}$, Matti Nykter$^{27}$, Petra Schneider$^{46}$, Michael Seifert$^{24}$, Marc Strickert$^{47}$, Jose M. G. Vilar$^{48}$.

---

1Terry Fox Laboratory, British Columbia Cancer Agency, Vancouver, British Columbia, Canada. 2Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. 4Department of Computer Science, University of British Columbia, Vancouver, British Columbia, Canada. 5School of Medicine and Dentistry, University of Rochester, Rochester, New York, USA. 6J. Craig Venter Institute, San Diego, California, USA. 7Department of Pathology and Division of Biomedical Informatics, U.T. Southwestern Medical Center, Dallas, Texas, USA. 8Canada’s Michael Smith Genome Sciences Centre, Vancouver, British Columbia, Canada. 9Baylor Research Institute, Dallas, Texas, USA. 10Tree Star Inc., Ashland, Oregon, USA. 11Cancer Division, McMaster University, Ontario, Canada. 12BC Cancer Agency, Vancouver, British Columbia, Canada. 13Child & Family Research Institute, Vancouver, British Columbia, Canada. 14University of Pennsylvania, Philadelphia, Pennsylvania, USA. 15Laboratory Medicine, University of Washington, Seattle, Washington, USA. 16Vancouver General Hospital, Vancouver, British Columbia, Canada. 17Computer Science, Purdue University, West Lafayette, Indiana, USA. 18Program in Medical & Population Genetics, Broad Institute of Harvard University and MIT, Cambridge, Massachusetts, USA. 19Department of Computer Science, University of Toronto, Ontario, Canada. 20Stanford University, Stanford, California, USA. 21Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, North Carolina, USA. 22Department of Neurology, Mount Sinai School of Medicine, New York, USA. 23Program in Translational NeuroPsychiatric Genomics, Institute for the Neurosciences,
Departments of Neurology and Psychiatry, Brigham & Women’s Hospital and Harvard Medical School, Boston, Massachusetts, USA, and, Program in Medical & Population Genetics, Broad Institute of Harvard University and MIT, Cambridge, Massachusetts, USA. 24Leibniz Institute of Plant Genetics and Crop Plant Research, Gatersleben, Germany. 25Georgetown University Medical Center, Washington, District of Columbia, USA. 26ADINIS s.r.o., Bratislava, Slovakia. 27Tampere University of Technology, Tampere, Finland. 28Department of Mathematics and Institute for Molecular Bioscience, University of Queensland, Brisbane, Australia. 29Department of Electrical and Computer Engineering, University of Rochester, Rochester, New York, USA. 30British Columbia Institute of Technology, Burnaby, British Columbia, Canada. 31CR Rao Advanced Institute of Mathematics, Statistics and Computer Science, Hyderabad 500046, India and Broad Institute of MIT & Harvard University, Cambridge, Massachusetts, USA. 32University of Texas M.D. Anderson Cancer Center, Houston, Texas, USA. 33Simon Fraser University, Burnaby, British Columbia, Canada. 34IBM Computational Biology Center, IBM Research, USA. 35European Bioinformatics Institute, Hinxton, UK. 36University of Pune, Pune, India, 411007 and University of Hyderabad, Hyderabad, India, 50046. 37Johann Bernoulli Institute for Mathematics and Computer Science, University of Groningen, The Netherlands. 38EPFL, Lausanne, Switzerland. 39CITEC Center of Excellence Cognitive Interaction Technology, Bielefeld University, Germany. 40University of Padova, Padua, Italy. 41Martin Luther University Halle-Wittenberg, Von-Seckendorff-Platz 1, 06120 Halle/Saale, Germany. 42Swiss Institute of Bioinformatics, Geneva, Switzerland. 43University of Warsaw, Warsaw, Poland. 44University of Maryland-College Park, USA. 45University of Chicago, Chicago, Illinois, USA. 46Centre for Endocrinology, Diabetes and Metabolism School of Clinical and Experimental Meine University of Birmingham, United Kingdom. 47SYNMIKRO Center for Synthetic Microbiology, Philips University of Marburg, Germany. 48Ikerbasque, Basque Foundation for Science and University of the Basque Country, Bilbao, Spain.

References

1. Baumgarth N, Roederer M. A practical approach to multicolor flow cytometry for immunophenotyping. J Immunol Methods. 2000; 243:77–97. [PubMed: 10986408]
2. Tanner SD, et al. Flow cytometer with mass spectrometer detection for massively multiplexed single-cell biomarker assay. Pure Appl. Chem. 2008; 80:2627–2641.
3. Bendall SC, Nolan GP, Roederer M, Chattopadhyay PK. A deep profiler’s guide to cytometry. Trends Immunol. 2012; 33:323–332. [PubMed: 22476049]
4. Newell EW, Sigal N, Bendall SC, Nolan GP, Davis MM. Cytometry by time-of-flight shows combinatorial cytokine expression and virus-specific cell niches within a continuum of CD8+ T cell phenotypes. Immunity. 2012
5. Lugli E, Roederer M, Cossarizza A. Data analysis in flow cytometry: The future just started. Cytometry A. 2010; 77:705–713. [PubMed: 20583274]
6. Quinn J, et al. A statistical pattern recognition approach for determining cellular viability and lineage phenotype in cultured cells and murine bone marrow. Cytometry A. 2007; 71:612–624. [PubMed: 17542025]
7. Lo K, Brinkman RR, Gottardo R. Automated gating of flow cytometry data via robust model-based clustering. Cytometry A. 2008; 73:321–332. [PubMed: 18307272]
8. Finak G, Bashashati A, Brinkman R, Gottardo R. Merging mixture components for cell population identification in flow cytometry. Adv Bioinformatics. 2009; 2009:247646. [PubMed: 20049161]
9. Pyne S, et al. Automated high-dimensional flow cytometric data analysis. Proc Natl Acad Sci U S A. 2009; 106:8519–8524. [PubMed: 19443687]

10. Naumann U, Luta G, Wand MP. The curvHDR method for gating flow cytometry samples. Bmc Bioinformatics. 2010; 11:44. [PubMed: 20096119]

11. Suchard MA, et al. Understanding GPU programming for statistical computation: studies in massively parallel architectures. J Comput Graph Stat. 2010; 19:419–438. [PubMed: 20877443]

12. Zare H, Shooshtari P, Gupta A, Brinkman RR. Data reduction for spectral clustering to analyze high throughput flow cytometry data. Bmc Bioinformatics. 2010; 11:403. [PubMed: 20667133]

13. Qian Y, et al. Elucidation of seventeen human peripheral blood B-cell subsets and quantification of the tetanus response using a density-based method for the automated identification of cell populations in multidimensional flow cytometry data. Cytometry Part B-Clinical Cytometry. 2010; 78B:S69–S82.

14. Sugar IP, Sealfon SC. Misty Mountain clustering: application to fast unsupervised flow cytometry gating. Bmc Bioinformatics. 2010; 11:502. [PubMed: 20932336]

15. Aghaeepour N, Nikolic R, Hoos HH, Brinkman RR. Rapid cell population identification in flow cytometry data. Cytometry A. 2011; 79:6–13. [PubMed: 21182178]

16. Ge Y, Sealfon SC. flowPeaks: a fast unsupervised clustering for flow cytometry data via K-means and density peak finding. Bioinformatics. 2012; 28:2052–2058. [PubMed: 22595209]

17. Aghaeepour N, et al. Early immunologic correlates of HIV protection can be identified from computational analysis of complex multivariate T-cell flow cytometry assays. Bioinformatics. 2012; 28:1009–1016. [PubMed: 22383736]

18. Zare H, et al. Automated analysis of multidimensional flow cytometry data improves diagnostic accuracy between mantle cell lymphoma and small lymphocytic lymphoma. American Journal of Clinical Pathology. 2012; 137:75–85. [PubMed: 22180480]

19. Costa ES, et al. Automated pattern-guided principal component analysis vs expert-based immunophenotypic classification of B-cell chronic lymphoproliferative disorders: a step forward in the standardization of clinical immunophenotyping. Leukemia : official journal of the Leukemia Society of America, Leukemia Research Fund, U.K. 2010; 24:1927–1933.

20. Roederer M, Nozzi JL, Nason MC. SPICE: exploration and analysis of post-cytometric complex multivariate datasets. Cytometry A. 2011; 79A:167–174. [PubMed: 21265010]

21. Azad A, Pyne S, Pothen A. Matching phosphorylation response patterns of antigen-receptor-stimulated T cells via flow cytometry. Bmc Bioinformatics. 2012; 13:S10. [PubMed: 22536861]

22. Bendall SC, et al. Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum. Science. 2011; 332:687–696. [PubMed: 21551058]

23. Qiu P, et al. Extracting a cellular hierarchy from high-dimensional cytometry data with SPADE. Nat Biotechnol. 2011; 29:886–891. [PubMed: 21964415]

24. Aghaeepour N, et al. RchyOptimyx: Cellular hierarchy optimization for flow cytometry. Cytometry A. 2012; 81:1022–1030. [PubMed: 23044634]

25. Yang P, Yang H, Zhou B, Zomaya A. A review of ensemble methods in bioinformatics. Current Bioinformatics. 2010; 5:296–308.

26. Macek JF, et al. Standardization of cytokine flow cytometry assays. BMC Immunol. 2005; 6:13. [PubMed: 15978127]

27. Prill RJ, et al. Towards a rigorous assessment of systems biology models: the DREAM3 challenges. PloS ONE. 2010; 5:e9202. [PubMed: 20186320]

28. Stolovitzky G, Prill RJ, Califano A. Lessons from the DREAM2 Challenges. Annals of the New York Academy of Sciences. 2009; 1158:159–195. [PubMed: 19348640]

29. Meyer P, et al. Verification of systems biology research in the age of collaborative competition. Nat Biotechnol. 2011; 29:811–815. [PubMed: 21904331]

30. Califano A, Kellis M, Stolovitzky G. Preface: RECOMB Systems Biology, Regulatory Genomics, and DREAM 2011 special issue. J Comput Biol. 2012; 19:101. [PubMed: 22300312]

31. Bain BJ. Blood cells: A practical guide. 2006; 13:420.
32. Maddox AM, et al. Philadelphia chromosome-positive adult acute leukemia with monosomy of chromosome number seven: a subgroup with poor response to therapy. Leuk Res. 1983; 7:509–522. [PubMed: 6578392]

33. Tecimer C, Loy BA, Martin AW. Acute myeloblastic leukemia (M0) with an unusual chromosomal abnormality: translocation (1;14)(p13;q32). Cancer Genetics and Cytogenetics. 1999; 111:175–177. [PubMed: 10347560]

34. Peters JM, Ansari MQ. Multiparameter flow cytometry in the diagnosis and management of acute leukemia. Arch Pathol Lab Med. 2011; 135:44–54. [PubMed: 21204710]

35. Gentleman R, Temple Lang D. Statistical analyses and reproducible research. Journal of Computational and Graphical Statistics. 2007; 16(1):1–23.

36. Lee JA, et al. MIFlowCyt: the minimum information about a Flow Cytometry Experiment. Cytometry A. 2008; 73:926–930. [PubMed: 18752282]

37. Brinkman RR, et al. High-content flow cytometry and temporal data analysis for defining a cellular signature of graft-versus-host disease. Biol Blood Marrow Transplant. 2007; 6:691–700. [PubMed: 17531779]

38. Aghaeepour N, Khodabakhshi AH, Brinkman RR. In: Ben-David, S., et al., editors. Clustering Theory Workshop, Neural Information Processing Systems (NIPS); Whistler; British Columbia, Canada. 2009.

39. Hesterberg T, Moore DS, Monaghan S, Clipson A, Epstein R. Bootstrap methods and permutation tests. Introduction to the Practice of Statistics. 2005; 47(4):1–70.

40. Dym CL, Wood WH, Scott MJ. Rank ordering engineering designs: pairwise comparison charts and Borda counts. Research in Engineering Design. 2002; 13:236–242.

41. Hornik K, Bohm W. Hard and soft Euclidean consensus partitions. Data Analysis, Machine Learning and Applications. 2008:147–154.

42. Hornik K. A clue for cluster ensembles. Journal of Statistical Software. 2005; 14:1–25.

43. Maecker HT, Rinfret A, D’Souza P, Darden J, Roig E, Landry C, Hayes P, Birungi J, Anzala O, Garcia M, et al. Standardization of cytokine flow cytometry assays. BMC Immunology. 2005; 6(1):13. [PubMed: 15978127]

44. Maecker HT, McCoy JP, Nussenblatt R. Standardizing immunophenotyping for the human immunology project. Nature Reviews Immunology. 2012
Figure 1. F-measure results of cell population identification challenges

Average manual and algorithm F-measures are represented against the manual consensus cluster as a function of the number of populations included, ranked from most consistent to least consistent. For a given population, consistency was defined as the agreement among manual gates, calculated as the average manual F-measures against the manual consensus cluster for that population. All populations across all samples were included in this calculation, and as such, the numbers on the x-axis should be multiplied by 12 and 30 (for GvHD and HSCT, respectively) to reflect the total number of populations in all samples in
the reference. Individual manual gating results are plotted as gray lines. (a) Graft versus host disease (GvHD) dataset. (b) Hematopietic stem cell transplant (HSCT) dataset.
Figure 2. Per population pair-wise comparisons of the cell population identification challenges
Average F-measures of all pairs of results for the five cell populations across all samples in the HSCT dataset represented as heat maps. The heat map color in individual squares reflects the pairwise agreement between each method for each cell population independently, and the position in the matrix reflects the pattern of agreement across all methods based on hierarchical clustering. The manual gate consensus cluster for each sample was used as a reference for matching of the automated results of that sample. Pairwise F-measures between all algorithms and manual gates for the HSCT dataset are
shown. The dendrogram groups the algorithms/manual gates based on the similarities between their pairwise F-measures.
Figure 3. Comparison of manual gate consensus and ensemble clustering results
Dots are color-coded based on population membership determined by ensemble clustering. Colored polygons enclose regions corresponding to the consensus clustering of manual gates. (a) and (b) show an example of a sample for which all of the cell populations have been accurately identified. (c) and (d) show an example of a sample in which the tail of the blue population has been misclassified as orange by the algorithms, resulting in a lower F-measure for the blue population. The red, blue, green, purple, and orange cell populations match to Cell Population #1, 2, 3, 4, and 5 of Figure 2, respectively.
Figure 4. AML subject detected as outlier by the algorithms

(A) Total number of misclassifications for each sample in the test-set (Sample #180 - #359) of the AML dataset. FSC/SSC (b-d) and FSC/CD34 (e-g) scatter plots of representative normal (b & e) and AML (c & f) samples and the outlier sample #340 (d & g), with the CD34+ cells highlighted in red (b to g). Cell proportions of the CD34+ population are reported as blast frequency percentages.
## Participating algorithms

A list of algorithms that were applied in at least one challenge

| Algorithm Name | Availability | Brief Description | SN/Ref  |
|----------------|--------------|------------------|---------|
| **Cell Population Identification** | | | |
| ADICyt | Commercially Available | Hierarchical clustering and entropy—based merging | 1.1.1/- |
| CDP | Python source–code | Bayesian non-parametric mixture models, calculated using massively parallel computing on GPUs | 1.1.2/ |
| FLAME | R package | Multivariate finite mixtures of skew and heavy-tailed distributions | 1.1.3/[9] |
| FLOCK | C source–code | Grid-based partitioning and merging | 1.1.4/[13] |
| flowClust/Merge | Two R/BioC packages | t-mixture modeling and entropy-based merging | 1.1.5/[7,8] |
| flowKoh | R source–code | Self-organizing maps | 1.1.6/- |
| flowMeans | R/BioC package | k-means clustering and merging using the Mahalanobis distance | 1.1.7/[15] |
| FlowVB | Python source–code | t-mixture models using variational Bayes inference | 1.1.8/- |
| L2kmeans | JAVA source–code | Discrepancy learning | 1.1.9/ |
| MM, MM&PCA | Windows and Linux executable | Density-based misty mountain clustering | 1.1.10/[14] |
| NMF-curmHDR | R source–code | Density-based clustering and non-negative matrix factorization | 1.1.11/[10] |
| Radial SVM | MATLAB source–code | Supervised training of radial SVMs using example manual gates | 1.1.12/[6] |
| SamSPECTRAL | R/BioC package | Efficient spectral clustering using density-based down-sampling | 1.1.13/[12] |
| SWIFT | MATLAB source–code | Weighted iterative sampling and mixture modeling | 1.1.14/- |
| **Sample Classification** | | | |
| 2DhistSVM | Pseudocode | 2D histograms of all pairs of dimensions and support vector machines | 1.2.1/- |
| admire-lvq | MATLAB source–code | 1D features and learning vector quantization | 1.2.2/- |
| biolobe | Pseudocode | k-means and correlation matrix mapping | 1.2.3/- |
| daltons | MATLAB source–code | Linear discriminant analysis and logistic regression | 1.2.4/- |
| EMMIXCYTOM, eqs | R source–code | Skew-t-mixture model and Kullback-Leibler divergence | 1.2.5/- |
| DREAM-A | Pseudocode | 2D Gaussian mixtures and support vector machines | 1.2.6/- |
| DREAM-B | Pseudocode | 2D Gaussian mixtures and support vector machines | 1.2.7/- |
| DREAM-C | Pseudocode | 1D gating and several different classifiers | 1.2.8/- |
| DREAM-D | Pseudocode | 4D clustering and bootstrapped t-tests | 1.2.9/- |
| FiveByFive | Pseudocode | 1D histograms and support vector machines | 1.2.10/- |
| flowBin | R package | High-dimensional cluster mapping across multiple tubes and support vector machines | 1.2.11/- |
| flowCoreflowStats | R source–code | Sequential gating and normalization and a Beta-Binomial model | 1.2.12/- |
| Algorithm Name   | Availability   | Brief Description                                                      | SN/Ref |
|-----------------|----------------|-----------------------------------------------------------------------|--------|
| flowPeaks/svm   | R package      | Kmeans and density-based clustering and support vector machines        | 1.2.13/16 |
| Kmeans/svm      |                |                                                                       |        |
| flowType        | Two R/BioC packages | 1D gates extrapolated to multiple dimensions and bootstrapped      | 1.2.14/17,18 |
| FeaLect         |                | LASSO classification                                                  |        |
| JKJG            | JAVA source-code | 1D Gaussian and logistic regression                                   | 1.2.15/- |
| PBSC            | C source-code  | Multi-dimensional clustering and cross sample population matching     | 1.2.16/13 |
|                 |                | using a relative distance order                                       |        |
| PRAMS           | R source-code  | 2D Clustering and logistic regression                                  | 1.2.17/- |
| PramSpheres and CIHC | Pseudocode          | Genetic algorithm and gradient boosting                              | 1.2.18/- |
| RandomSpheres   | Pseudocode      | Hypersphere–based Monte Carlo optimization                            | 1.2.18/- |
| SPADE, BCB      | MATLAB, Cytoscape, R/BioC | Density-based sampling, kmeans clustering, and minimum spanning      | 1.2.19/23 |
|                 |                | trees                                                                  |        |
| SPCA+GLM        | Pseudocode      | 1D probability binning and principal component analysis               | 1.2.20/- |
| SWIFT           | MATLAB source-code | SWIFT clustering and support vector machines                         | 1.2.21/ |
| team21          | Python source-code | 1D relative entropies                                                 | 1.2.22/- |

\(^a\) See Supplementary Table 3 for algorithm contact information.

\(^b\) See Supplementary Note XX for more details about each program.

\(^c\) Supplementary Note 1 section (SN) and reference (Ref) citation []
Table 2

Summary of results for the cell identification challenges.

|                   | GvHD | DLBCL | F-measure $^d$ | HSCT | WNV | ND | Mean | Runtime | Rank Score |
|-------------------|------|-------|----------------|------|-----|----|------|---------|------------|
| **Challenge 1: Completely Automated** |      |       |                |      |     |    |      |         |            |
| ADICyt            | 0.81 (0.72, 0.88) | 0.93 (0.91, 0.95) | 0.93 (0.90, 0.96) | 0.86 (0.84, 0.87) | 0.92 (0.92, 0.93) | 0.89 | 04:50:37 | 52        |
| flowMeans         | 0.88 (0.82, 0.93) | 0.92 (0.89, 0.95) | 0.92 (0.90, 0.94) | 0.88 (0.86, 0.90) | 0.85 (0.76, 0.92) | 0.89 | 00:02:18 | 49        |
| FLOCK             | 0.84 (0.76, 0.90) | 0.88 (0.85, 0.91) | 0.86 (0.83, 0.89) | 0.83 (0.80, 0.86) | 0.91 (0.89, 0.92) | 0.86 | 00:00:20 | 45        |
| FLAME             | 0.85 (0.77, 0.91) | 0.91 (0.88, 0.93) | 0.94 (0.92, 0.95) | 0.80 (0.76, 0.84) | 0.90 (0.89, 0.90) | 0.88 | 00:04:20 | 44        |
| SamSPECTRAL       | 0.87 (0.81, 0.93) | 0.86 (0.82, 0.90) | 0.85 (0.82, 0.88) | 0.75 (0.60, 0.85) | 0.92 (0.92, 0.93) | 0.85 | 00:03:51 | 39        |
| MM&PCA            | 0.84 (0.74, 0.93) | 0.85 (0.82, 0.88) | 0.91 (0.88, 0.94) | 0.64 (0.51, 0.71) | 0.76 (0.75, 0.77) | 0.80 | 00:00:03 | 29        |
| FlowVB            | 0.85 (0.79, 0.91) | 0.87 (0.85, 0.90) | 0.75 (0.70, 0.79) | 0.81 (0.78, 0.83) | 0.85 (0.84, 0.86) | 0.82 | 00:38:49 | 28        |
| MM                | 0.83 (0.74, 0.91) | 0.90 (0.87, 0.92) | 0.73 (0.66, 0.80) | 0.69 (0.60, 0.75) | 0.75 (0.74, 0.76) | 0.78 | 00:00:10 | 28        |
| flowClust/Merge   | 0.69 (0.55, 0.79) | 0.84 (0.81, 0.86) | 0.81 (0.77, 0.85) | 0.77 (0.74, 0.79) | 0.73 (0.58, 0.85) | 0.77 | 02:12:00 | 24        |
| L2kmeans          | 0.64 (0.57, 0.72) | 0.79 (0.74, 0.83) | 0.70 (0.65, 0.75) | 0.78 (0.75, 0.81) | 0.81 (0.80, 0.82) | 0.74 | 00:08:03 | 20        |
| CDP               | 0.52 (0.46, 0.58) | 0.87 (0.85, 0.90) | 0.50 (0.48, 0.52) | 0.71 (0.68, 0.75) | 0.88 (0.86, 0.90) | 0.70 | 00:00:57 | 19        |
| SWIFT             | 0.63 (0.56, 0.70) | 0.67 (0.62, 0.71) | 0.59 (0.53, 0.62) | 0.69 (0.64, 0.74) | 0.87 (0.86, 0.88) | 0.69 | 01:14:50 | 15        |
| Ensemble Clustering | 0.88 | 0.94 | 0.97 | 0.88 | 0.94 | 0.92 | - | 64     |
| **Challenge 2: Manually Tuned** |      |       |                |      |     |    |      |         |            |
| ADICyt            | 0.81 (0.71, 0.89) | 0.93 (0.91, 0.95) | 0.93 (0.90, 0.96) | 0.86 (0.84, 0.87) | 0.92 (0.92, 0.93) | 0.89 | 04:50:37 | 34        |
| SamSPECTRAL       | 0.87 (0.79, 0.94) | 0.92 (0.89, 0.94) | 0.90 (0.86, 0.93) | 0.85 (0.83, 0.88) | 0.91 (0.91, 0.92) | 0.89 | 00:06:47 | 31        |
| FLOCK             | 0.84 (0.76, 0.90) | 0.88 (0.85, 0.91) | 0.86 (0.83, 0.89) | 0.84 (0.82, 0.86) | 0.89 (0.87, 0.91) | 0.86 | 00:00:15 | 23        |
| FLAME             | 0.81 (0.75, 0.87) | 0.87 (0.84, 0.90) | 0.87 (0.82, 0.90) | 0.84 (0.83, 0.85) | 0.87 (0.86, 0.87) | 0.85 | 00:04:20 | 23        |
| SamSPECTRAL-FK    | 0.87 (0.80, 0.94) | 0.85 (0.81, 0.89) | 0.90 (0.86, 0.92) | 0.76 (0.71, 0.81) | 0.92 (0.91, 0.93) | 0.86 | 00:04:25 | 23        |
| CDP               | 0.74 (0.67, 0.80) | 0.89 (0.86, 0.91) | 0.90 (0.88, 0.92) | 0.75 (0.71, 0.78) | 0.86 (0.85, 0.88) | 0.83 | 00:00:18 | 19        |
| flowClust/Merge   | 0.69 (0.53, 0.78) | 0.87 (0.85, 0.90) | 0.96 (0.94, 0.97) | 0.77 (0.75, 0.79) | 0.88 (0.81, 0.91) | 0.83 | 02:12:00 | 18        |
| NMF-cur-HDR       | 0.76 (0.69, 0.82) | 0.84 (0.83, 0.86) | 0.70 (0.67, 0.74) | 0.81 (0.77, 0.84) | 0.83 (0.83, 0.84) | 0.79 | 01:39:42 | 13        |
| Ensemble Clustering | 0.87 | 0.94 | 0.98 | 0.87 | 0.92 | 0.91 | - | 41     |
| **Challenge 3: Assignment of Cells to Populations with Pre-defined Number of Populations** |      |       |                |      |     |    |      |         |            |
| ADICyt            | 0.91 (0.84, 0.96) | 0.96 (0.94, 0.97) | 0.98 (0.97, 0.99) | 0.95 | 00:10:49 | 26.2 |        |            |
| Algorithm          | GvHD       | DLBCL      | F-measure<sup>a</sup> | HSCF       | WNV        | ND         | Mean | Runtime <br>hh:mm:ss<sup>b</sup> | Rank Score<sup>c</sup> |
|--------------------|------------|------------|------------------------|------------|------------|------------|------|----------------------------------|------------------------|
| SamSPECTRAL        | 0.85 (0.75, 0.93) | 0.93 (0.91, 0.95) | 0.97 (0.95, 0.98) |          |            |            | 0.92 | 00:02:30                          | 26.2                   |
| flowMeans          | 0.91 (0.84, 0.96) | 0.94 (0.91, 0.96) | 0.95 (0.93, 0.96) |          |            |            | 0.93 | 00:00:01                          | 23.4                   |
| TCLUST             | 0.93 (0.91, 0.96) | 0.93 (0.91, 0.95) | 0.93 (0.90, 0.95) |          |            |            | 0.93 | 00:00:40                          | 23.4                   |
| FLOCK              | 0.86 (0.79, 0.93) | 0.92 (0.89,0.94) | 0.97 (0.95, 0.98) |          |            |            | 0.92 | 00:00:02                          | 22.2                   |
| CDP                | 0.85 (0.77, 0.92) | 0.92 (0.89, 0.94) | 0.76 (0.72, 0.81) |          |            |            | 0.84 | 00:00:21                          | 16.9                   |
| flowCluster/Merge  | 0.88 (0.82, 0.93) | 0.90 (0.86, 0.94) | 0.83 (0.79, 0.88) |          |            |            | 0.87 | 00:49:24                          | 15.9                   |
| FLAME              | 0.85 (0.79, 0.91) | 0.90 (0.86, 0.93) | 0.86 (0.82, 0.91) |          |            |            | 0.87 | 00:03:20                          | 15.9                   |
| SWIFT              | 0.90 (0.84, 0.95) | 0.00 (0.00, 0.00) | 0.88 (0.84, 0.92) |          |            |            | 0.59 | 00:01:37                          | 11.9                   |
| flowKoh            | 0.85 (0.80, 0.90) | 0.85 (0.82, 0.88) | 0.87 (0.84, 0.91) |          |            |            | 0.86 | 00:00:42                          | 9.5                    |
| NMF                | 0.74 (0.69, 0.78) | 0.84 (0.80, 0.88) | 0.80 (0.76, 0.84) |          |            |            | 0.79 | 00:01:00                          | 7.5                    |
| Ensemble Clustering| 0.95        | 0.97        | 0.98                  |          |            |            | 0.97 | -                                 | 35.0                   |

### Challenge 4: Supervised Approaches Trained using Human-Provided Gates

| Algorithm          | GvHD       | DLBCL      | F-measure<sup>a</sup> | HSCF       | WNV        | ND         | Mean | Runtime <br>hh:mm:ss<sup>b</sup> | Rank Score<sup>c</sup> |
|--------------------|------------|------------|------------------------|------------|------------|------------|------|----------------------------------|------------------------|
| Radial SVM         | 0.89 (0.83, 0.95) | 0.84 (0.80,0.87) | 0.98 (0.96, 0.99) | 0.96 (0.94, 0.97) | 0.93 (0.92, 0.94) |          | 0.92 | 00:00:18                          | 21                     |
| flowCluster/Merge  | 0.92 (0.88, 0.95) | 0.92 (0.89, 0.94) | 0.95 (0.92, 0.97) | 0.84 (0.82, 0.86) | 0.89 (0.88, 0.90) |          | 0.90 | 05:31:50                          | 19                     |
| randomForests      | 0.85 (0.78, 0.91) | 0.78 (0.74, 0.83) | 0.81 (0.79, 0.83) | 0.87 (0.84,0.90) | 0.94 (0.92, 0.95) |          | 0.85 | 00:02:06                          | 15                     |
| FLOCK              | 0.82 (0.77,0.87)  | 0.91 (0.89,0.93) | 0.86 (0.76, 0.93) | 0.86 (0.82, 0.89) | 0.86 (0.77, 0.92) |          | 0.86 | 00:00:08                          | 13                     |
| CDP                | 0.78 (0.68, 0.87) | 0.95 (0.93, 0.97) | 0.75 (0.71, 0.78) | 0.86 (0.84,0.88) | 0.83 (0.80, 0.86) |          | 0.83 | 00:00:15                          | 11                     |
| Ensemble Clustering| 0.91        | 0.94        | 0.95                  | 0.92        | 0.94        |            | 0.93 | -                                 | 26                     |

<sup>a</sup> In each dataset/challenge, the top algorithm (highest mean F-measure) and the algorithms with overlapping CIs with the top algorithm are bolded (see Online Methods for F-measure calculations).

<sup>b</sup> Algorithms are sorted by rank score within each challenge (see Online Methods for rank score calculations).

<sup>c</sup> Runtime was calculated as time per CPU per sample.
## Table 3

Performance of algorithms in the sample classification challenges on the validation cohort.

| Algorithm           | Recall Challenge 1: HEUvsUE | Precision Challenge 1: HEUvsUE | Accuracy Challenge 1: HEUvsUE | F-measure Challenge 1: HEUvsUE | Recall Challenge 2: AML | Precision Challenge 2: AML | Accuracy Challenge 2: AML | F-measure Challenge 2: AML | Recall Challenge 3: HVTN | Precision Challenge 3: HVTN | Accuracy Challenge 3: HVTN | F-measure Challenge 3: HVTN |
|---------------------|-----------------------------|--------------------------------|-------------------------------|--------------------------------|--------------------------|---------------------------|----------------------------|----------------------------|--------------------------|---------------------------|----------------------------|---------------------------|
| 2DhistsSVM          | 0.50                        | 0.091                          | 0.50                          | 0.15                           | 0.00                     | 0.95                      | 0.99                       | 0.97                       | 1.00                     | 1.00                      | 1.00                       | 1.00                       |
| EMMIXCYTOM          | 0.95                        | 0.95                           | 0.99                          | 0.95                           | 0.95                     | 0.95                      | 0.95                       | 0.95                       | 1.00                     | 1.00                      | 1.00                       | 1.00                       |
| flowBin             | 0.012                       | 0.00                           | 0.45                          | 0.00                           | 0.10                     | 0.30                      | 0.92                       | 0.46                       | 1.00                     | 1.00                      | 1.00                       | 1.00                       |
| flowCore-flowStats | 0.56                        | 0.455                          | 0.55                          | 0.50                           | 1.00                     | 1.00                      | 1.00                       | 1.00                       | 1.00                     | 1.00                      | 1.00                       | 1.00                       |
| flowPeakasvm        | 1.00                        | 1.00                           | 1.00                          | 1.00                           | 1.00                     | 1.00                      | 1.00                       | 1.00                       | 1.00                     | 1.00                      | 1.00                       | 1.00                       |
| flowType            | 0.58                        | 0.636                          | 0.59                          | 0.61                           | 0.95                     | 0.95                      | 0.99                       | 0.95                       | 1.00                     | 1.00                      | 1.00                       | 1.00                       |
| flowType-FeaLect    | 0.33                        | 0.273                          | 0.36                          | 0.30                           | 1.00                     | 1.00                      | 1.00                       | 1.00                       | 1.00                     | 1.00                      | 1.00                       | 1.00                       |
| Kmeanssvm           | 1.00                        | 1.00                           | 1.00                          | 1.00                           | 1.00                     | 1.00                      | 1.00                       | 1.00                       | 1.00                     | 1.00                      | 1.00                       | 1.00                       |
| PBSC                | 0.55                        | 0.545                          | 0.55                          | 0.55                           | 0.75                     | 0.75                      | 0.94                       | 0.75                       | 0.95                     | 0.95                      | 0.95                       | 0.95                       |
| PRAMS               | 1.00                        | 1.00                           | 1.00                          | 1.00                           | 1.00                     | 1.00                      | 1.00                       | 1.00                       | 1.00                     | 1.00                      | 1.00                       | 1.00                       |
| PramSpheres         | 0.36                        | 0.364                          | 0.36                          | 0.36                           | 0.90                     | 0.90                      | 0.90                       | 0.90                       | 1.00                     | 1.00                      | 1.00                       | 1.00                       |
| RandomSpheres       |                             |                                |                               |                                | 0.95                     | 0.95                      | 0.99                       | 0.95                       | 1.00                     | 1.00                      | 1.00                       | 1.00                       |
| SPADE               | 1.00                        | 1.00                           | 1.00                          | 1.00                           | 1.00                     | 1.00                      | 1.00                       | 1.00                       | 1.00                     | 1.00                      | 1.00                       | 1.00                       |
| SWIFT               | 0.67                        | 0.545                          | 0.64                          | 0.60                           | 1.00                     | 1.00                      | 1.00                       | 1.00                       | 1.00                     | 1.00                      | 1.00                       | 1.00                       |
| DREAM               |                             |                                |                               |                                |                         |                           |                           |                           |                         |                           |                           |                           |
| admire-lvq          | 1.00                        | 1.00                           | 1.00                          | 1.00                           |                         |                           |                           |                           |                         |                           |                           |                           |
| bcb                 | 1.00                        | 1.00                           | 1.00                          | 1.00                           |                         |                           |                           |                           |                         |                           |                           |                           |
| biolobe             | 1.00                        | 1.00                           | 1.00                          | 1.00                           |                         |                           |                           |                           |                         |                           |                           |                           |
| cihc                | 1.00                        | 0.95                           | 0.99                          | 0.97                           |                         |                           |                           |                           |                         |                           |                           |                           |
| daltons             | 1.00                        | 1.00                           | 1.00                          | 1.00                           |                         |                           |                           |                           |                         |                           |                           |                           |
| DREAM-A             | 0.95                        | 0.95                           | 0.99                          | 0.95                           |                         |                           |                           |                           |                         |                           |                           |                           |
| DREAM-B             | 1.00                        | 0.85                           | 0.98                          | 0.92                           |                         |                           |                           |                           |                         |                           |                           |                           |
| DREAM-C             | 1.00                        | 0.85                           | 0.98                          | 0.92                           |                         |                           |                           |                           |                         |                           |                           |                           |
| DREAM-D             | 0.95                        | 0.95                           | 0.99                          | 0.95                           |                         |                           |                           |                           |                         |                           |                           |                           |
|        | Recall | Precision | Accuracy | F-measure | Recall | Precision | Accuracy | F-measure |
|--------|--------|-----------|----------|----------|--------|-----------|----------|----------|
| **Challenge 1: HEUvsUE** |        |           |          |          |        |           |          |          |
| fivebyfive | 0.95   | 1.00      | 0.99     | **0.98** |        |           |          |          |
| jkg    | 1.00   | 1.00      | 1.00     | **1.00** |        |           |          |          |
| SPCA+GLM | 0.89   | 0.85      | 0.97     | **0.87** |        |           |          |          |
| team21 | 1.00   | 1.00      | 1.00     | **1.00** |        |           |          |          |
| uqs    | 1.00   | 0.95      | 0.99     | **0.97** |        |           |          |          |

\( ^a \) Not all algorithms were applied in all challenges. Particularly, a large number of algorithms participated through the DREAM project that only included the AML dataset.