Cytometry, Part A

EM/GANN: A multivariate computational strategy for auto-characterizing relationships between cellular and clinical phenotypes and predicting disease progression time using high-dimensional flow cytometry data

Journal:	*Cytometry: Part A*
Manuscript ID:	Draft
Wiley - Manuscript type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Tong, Dong; Nottingham Trent University, The John van Geest Cancer Research Centre Ball, Graham; Nottingham Trent University, The John van Geest Cancer Research Centre Pockley, A; Nottingham Trent University, The John van Geest Cancer Research Centre
Key Words:	FlowCAP, imbalance, cluster analysis, expectation maximization, genetic algorithm-neural network, feature identification, multidimensional, survival time

EM/GANN: A multivariate computational strategy for auto-characterizing relationships between cellular and clinical phenotypes and predicting disease progression time using high-dimensional flow cytometry data

Dong Ling Tong*, Graham R. Ball, A. Graham Pockley

The John van Geest Cancer Research Centre, Nottingham Trent University, Clifton Lane, Nottingham, United Kingdom NG11 8NS.

*Correspondence to: Dr Dong Ling Tong, The John van Geest Cancer Research Centre, Nottingham Trent University, Clifton Lane, Nottingham, United Kingdom NG11 8NS. E-mail: dong.tong@ntu.ac.uk

Abstract

The dramatic increase in the complexity of flow cytometric datasets requires the development of new computational based approaches that can maximize the amount of information derived and overcome the limitations of traditional gating strategies. Herein, we present a multivariate computational analysis of the HIV-infected flow cytometry datasets that were provided as part of the FlowCAP-IV Challenge using unsupervised and supervised learning techniques. Out of 383 samples (stimulated and unstimulated), 191 samples were used as a training set (34 individuals whose disease did not progress, and 157 individuals whose disease did progress). Using the results from the training set, the participants in the Challenge were then asked to predict the condition and progression time of the remaining individuals (45 'non-progressors' and 147 'progressors'). To achieve this, we first scaled down data resolution. We then excluded doublet cells from the analysis using Expectation Maximization approaches. We then standardized all samples into histograms and used Genetic Algorithm-Neural Network to extract feature sets from the datasets, the reliability of which were examined using WEKA-implemented classifiers. The selected feature set resulted in a high sensitivity and specificity for the discrimination of progressors and non-progressors in the training set (average True Positive Rate = 1.00 and average False Positive Rate = 0.033). The capacity of the feature set to predict real-time survival time was better when using data from the 'unstimulated' training set (r = 0.825). The p-values and 95% confidence interval log-rank ratios between actual and predicted survival time in the test set were 0.682 and 0.9542±0.24 for the unstimulated dataset, and 0.4451 and 0.9173±0.23 for the stimulated dataset. Our analytic strategy has demonstrated a promising capacity to extract useful information from complex flow cytometry datasets, despite a significance imbalance and variation between the training and test sets.

Key terms

FlowCAP; imbalance; cluster analysis; expectation maximization; genetic algorithm-neural network; feature identification; multidimensional; survival time

INTRODUCTION

Recent advances in instrumentation and data acquisition platforms have strengthened the potential of flow cytometry as a platform for clinical diagnosis and decision-making (1–3), drug discovery and drug development (4–6). However, the ability to achieve this requires high-dimensional datasets and new approaches for the automated analysis of these. In order to stimulate progress in this area, the FlowCAP Consortium (7) has established a series of competitions, the aims of which are to promote the development of computational methods for automatically identifying cell populations in multi-dimensional flow cytometry data and therefore remove the need for 'subjective' manual gating. The recent competition (FlowCAP-IV) provided flow cytometry data on the phenotype of unstimulated peripheral blood mononuclear cells (PBMCs) from HIV infected individuals, and also of PBMCs that had been stimulated *in vitro* with HIV antigens. The aim of the contest was to use data from the provided 'training' samples which came from individuals with known outcomes (i.e. non-progressor and progressor) and the survival time (in days) to identify which features within the cell populations which could be used to predict the time to progression to Acquired Immune Deficiency Syndrome (AIDS) and outcomes for those individuals whose data were provided in the 'test' set.

The analysis of the data in this competition presented 3 major challenges. Firstly, the data were highly imbalanced: the ratio of non-progressors to progressors was 1:4.6, and the mean survival time of patients varied widely (maximum = 5855 days, minimum = -10 days, mean 947.36 days). Secondly, the number of events that were captured in the non-progressor group ranged between 40131 and 1000000 (mean = 388598, standard deviation = 287653 and median = 335049, unstimulated and stimulated dataset combined) and events captured in the progressor group ranged between 12668 and 1000000 (mean = 281518, standard deviation = 239968 and median = 183671, unstimulated and stimulated dataset combined). Thirdly, the

data exhibited high complexity, in that the cell profiles within the progressor and non-progressor groups were irregular and the data was of a high resolution: the data has 262144 possible channels (equivalent to 18 bit ADCs). Standard computational techniques which assume equality of sample distribution and uniformity of cell distribution, are therefore not very effective for analyzing this dataset. Furthermore, a data pre-processing step which lowers the resolution of the data in order to improve the processing time of the algorithm and the analysis of the histograms, but without compromising the original characteristics of the data is required.

Taking into these challenges into consideration, we proposed a multivariate analysis strategy which would utilize the learning power of unsupervised and supervised learning principles to identify features within the histograms that have the capacity to predict / identify outcome and progression. In order to achieve this, we first scaled down the resolution of the data and used an unsupervised learning method called Expectation Maximization (EM) to remove doublet cells from analyzed cell populations. We then standardized the remaining live, singlet cells into histograms and applied a supervised learning method called Genetic Algorithm-Neural Network (GANN) to identify features that can differentiate non-progressor from progressor and predict the survival time of all patients. We term our method as EM/GANN since it makes use of 2 computing techniques to interpret flow cytometric data: (a) doublet removal using EM clustering and (b) identification of feature sets using GANN.

The use of EM for the analysis of flow cytometry data is not new. The EM algorithm has been commonly used, in conjunction with statistical mixture models, to discover cell populations in samples. Several examples include optimized t-mixture model (8), the skew t-mixture model in FLAME (9) and also in the mixture modeling approach which has been proposed by Boedigheimer and Ferbas (10). However, we did not use EM to optimize the GANN algorithm, rather we used the EM algorithm alone to discover singlet cells in the

population and the GANN algorithm to identify feature sets that can describe the difference between non-progressors and progressors, and the survival time of the patients. GANN is the bespoke algorithm which has been designed for handling large-scale datasets (11).

This study did not analyze the flow cytometry data using the more typically applied gating and hierarchy-based approaches. As a consequence, rather than reporting on features on a 'population' basis (i.e. positive and negative populations for each marker or combination of markers), we examined the capacity of every feature within cell population (i.e. every channel in the histogram plot) to identify features that can (1) discriminate between non-progressors and progressors, and (2) predict the survival time of HIV-infected patients. We will therefore report the identified feature on the basis of channel numbers – for example, IL2_5 represents the channel number 5 in histogram for IL2 expression.

MATERIALS AND METHODS

HIV dataset

The provided dataset contains 17 parameter flow cytometry data (including 3 scatter parameters, a combined VIVID/CD14 parameter, and a Time parameter) derived from 383 peripheral blood mononuclear cell (PBMC) samples that had been stimulated with HIV antigens and their corresponding unstimulated controls. Of the 383 samples, 191 (34 non-progressors and 157 progressors) were used as training samples and the remaining 192 were test samples (45 non-progressors and 147 progressors). Clinical outcomes of the test samples were not provided at the onset of the Challenge. The FlowCAP Consortium suggested the use of specific parameters for doublet removal (FSC-A and FSC-H) and for the identification and gating of live lymphocytes (VIVID/CD14 and SSC). After manually screened all 191 training samples for possible doublet cells and lymphocytes, we found that there is a significant variation in the cell profiles for samples within similar disease groups.

Figure 1 shows 2D-scatterplots based on 4 randomly selected samples from the progressor group which have been generated using combinations of FSC-H, FSC-A, SSC-A and VIVID/CD14. These profiles were not consistent in terms of sample shape, number of cells detected and cell signals. It was also apparent that the Time parameter more consistently identified live, singlet cells than FSC-H. Therefore, we used FSC-A, SSC-A and Time to identify live, singlet cells from cell population.

Data scaling

Given $x \in \{1, 2, ..., i\}$ is the cell events in which $\{1, 2, ...i\}$ is the relative intensity value and channel number of events, s is the scaling factor and R_d is the reduced resolution channels. A simple equation is given as follows:

$$\text{Scale}_{\text{d}}(x) = \frac{x_i}{s}, \quad (1)$$

where $s = \frac{\max(x)}{R_d}$.

In this study, we used 1024 channels (equivalent to 10 bit ADCs).

Doublet removal using EM

EM is an effective iterative procedure to produce Maximum Likelihood (ML) estimates of unknown parameters when the outcomes of these parameters are clumped together (12,13). The EM algorithm contains 2 iterative steps: an expectation step (E-step) and a maximization step (M-step). The E-step computes an estimate for the unknown variables using the observed variables and current estimate of the parameters. The M-step provides new estimates of the parameters.

Given that Y is a set of observed data of flow cytometry parameters SSC-A, FSC-A and Time; X is a set of unknown latent values from Y that can describe outcomes x (singlet cells in this study) and θ is a set of density values to describe likelihood of x. The probability density function of x is $f(x|\theta)$. To optimize (maximize) log-likelihood $\ln f(x|\theta)$, EM iteratively find (E-step) and maximize (M-step) lower boundaries of $f(x|\theta)$ based on Y and current estimate of θ_k with the following equations:

$$\text{E-step: } g_k(\theta) = E\left[\ln f(x|\theta)|Y, \theta_k\right] \quad (2)$$

$$\text{M-step: } \theta_{k+1} = \arg\max_{\theta} g_k(\theta) \quad (3)$$

The E-step and the M-step will be iteratively performed until convergence, i.e. when $\|\theta_k - \theta_{k-1}\| < \gamma$ and γ is the defined threshold for convergence.

In our study, the simple EM algorithm from the WEKA data mining suite (14), version 3.6 was used to cluster singlet cells from the cell populations. The default parameter

setting was used, except that the number of clusters was set between 3 and 4, depending on the variability of background noise in the population. In the default setting, a 10 fold cross-validation procedure and $\gamma = 0.0001$ were applied.

Histogram transformation

Cumulative histograms based on the actual number of channel values for the gated cell populations were created. This yielded 2047 bins ranging from -1023 to +1023 for each fluorochrome marker in the data (2047 bins x 12 markers = 24564 binned features). The dimension of the data is then reduced by simply removing bins with more than 50% of NULL values, leaving a total of 1578 and 1644 features in the training sets (stimulated and unstimulated).

Feature selection using GANN

GANN is the bespoke hybrid method of a genetic algorithm (GA) and an artificial neural network (ANN) which co-evolve together to find optimal set of features that, given a basic ANN classifier (i.e. feedforward neural network), will effectively classify the data (15). GA is tasked with finding the optimum ANN with a feature set that will correctly discriminate the classes and ANN is tasked to optimize the fitness values of GA.

Given $N \in \{1,2,...,s\}$ is the total number of samples in the dataset, s_{ik} is the sample i in class k, T_{ik} is the target output of the sample i in class k, A_{ik} is the activation value of ANN, C_k is the centroid value of class k and O_{ik} is the final prediction value generated by ANN, fitness of GA is iteratively examined by ANNs in the following equation:

$$\text{fitness} = \sum_{i=1}^{N}\sum_{k=1}^{C} s_{ik}\,, \tag{3}$$

where $s_{ik} = T_{ik} - \sqrt{(A_{ik} - C_{ik})^2} \begin{cases} \geq f(x), O_{ik} = T_{ik} \\ < f(x), O_{ik} \neq T_{ik} \end{cases}$ and $C_k = \frac{1}{s_k} \sum_{s \in k} A_{sk}$.

For network activation function, $f(x)$, given w is the network weight, i is the input node to the hidden node j and b is the bias nodes, $f(x)$ can be expressed as:

$$f(x_j) = \left[\frac{2}{1 + \exp(-x_j)} - 1 \right] + b_j \quad (4)$$

The remaining parameter was set as follows: Total repetitions = 1000, Population size = 100, Fitness evaluation = 40000, Input node = 10 and Hidden node = 5.

Classification using cost-sensitive classifier

Considering the degree of imbalance in the data, cost-sensitive classifiers (CSC), which allow the control of True Positive Rate (TPR) and False Positive Rate (FPR) were used. The WEKA-implemented CSC Nearest Neighbor (CSC-IBk) was used to analyze the data. As the misclassification cost is dependent on the underlying probability of the classifiers (16), experiments on different sets of misclassification costs were performed until a minimal FPR is achieved.

Survival time prediction using additive regression method

Considering the variability in the survival time of the patients: mean survival time of 947.36 days in the training set (maximum 5855 days, minimum -10 days, median 722 days) and mean survival time of 966.61 days in the test set (maximum 4925 days, minimum -8 days, median 745 days), a Gaussian-based additive regression method which is less sensitive to the data dimension and data skewness, was used. The WEKA-implemented Additive Regression model with Gaussian-based classifier was therefore used to predict the survival time of the patients.

RESULTS

Doublet removal

Figure 2 shows the gating results generated by EM in order to illustrate its elasticity for processing flow cytometry data. The plot of the original samples was depicted in Figure 1. Using SSC-A, FSC-A and Time, most of the doublets and unwanted cells were eliminated from the subsequent analysis, thereby leaving singlet and "lymphocyte-like" cell populations.

Combining all 383 samples (stimulated and unstimulated), more singlet cells were found in the non-progressor group, with average ratio of 0.52:0.48. The breakdown ratio of singlet cells identified in the non-progressor group in each dataset are as follow: 0.51:0.49 in the stimulated dataset (mean = 198091.7 in the non-progressor group and mean = 192174.5 in the progressor group), and 0.53:0.47 in the unstimulated dataset (mean = 208097.5 in the non-progressor group and 187743.6 in the progressor group). Considering the degree of imbalance in the data (average ratio of 1 non-progressor sample to 3.9 progressor samples) and number of events captured in each group, EM is able to cluster singlet cells from such cell populations.

Feature sets differentiating non-progressor from progressor

For the stimulated dataset, a total of 5 features (IL2_5, IL2_6, IL2_7, IL2_8 and IL2_9) were identified as being able to discriminate between non-progressors and progressors. For the unstimulated dataset, 7 features from 4 separate markers (CD154_5, CD154_6, CD45RO_1023, CCR7_9, IL2_6, IL2_7 and IL2_8) were deemed to be significant. Table 1 shows the summary results for the non-progressor group, and the misclassification cost for the group is presented in Figure 3.

All of the 34 samples from the non-progressor group in the training set were perfectly classified. An False Positive Rate (FPR) lower than 20% confirmed the predictive capacity of

the identified features. Although a slight increase in FPR (~6% increment) and a reduction in the True Positive Rate (TPR) suggested that the identified features are still significant for describing progressor samples in the test samples, these features were no longer effective for predicting non-progressor samples . This could be due to possible variations in some of the features that are associated with non-progression between the training and test sets (Table 2). Significant variations in IL2_5 and IL2_6 in the stimulated samples and IL2_6 in the unstimulated samples in the training and test sets were detected, thereby indicating that these features might be associated with certain (training set) patients in the data. This led to the low TPR in the non-progressor group in the test set.

Feature sets to predict survival time of HIV patients

A total of 6 features from 4 markers (IL2_1, CD3_1, CD3_2, CD8_9, CD107A_6 and CD107A_7) and 7 features from CD45RO marker (CD45RO_16, CD45RO_17, CD45RO_18, CD45RO_19, CD45RO_22, CD45RO_23 and CD45RO_26) in unstimulated and stimulated datasets exhibited a strong capacity to predict real-time survival time of patients in the training set (r = 0.535 in the stimulated dataset and r = 0.825 in the unstimulated dataset) (Table 3). The predictive capacity of these feature sets was significantly lower in the test sets (r = 0.149, stimulated and r = 0.173. unstimulated). This is not too surprising given that our regression model was trained using actual progression times of patients.

The capacity of these feature sets to predict patient survival times was evaluated using a log-rank (Mantel-Cox) test which was based on the actual and the predicted progression time for stimulated and unstimulated datasets was calculated and the survival plot based on these dataset is presented in Figure 4. The p-values and 95% confidence interval log-rank ratios between actual and predicted survival time for stimulated data are 0.2956 and

0.8944±0.21 in the training set and 0.4451 and 0.9173±0.23 in the test set. The p-values and 95% confidence interval log-rank ratios between actual and predicted survival time for unstimulated data are 0.7859 and 0.9702±0.24 in the training set and 0.682 and 0.9542±0.24 in the test set. High p-values show that there is no significant different between the actual survival curve and the predicted survival curve in these datasets, and demonstrates that the selected feature sets are able to provide proximity prediction to the actual results. Interestingly, the feature set identified from the unstimulated dataset has better prediction accuracy than the feature set selected from the stimulated dataset.

CONCLUSIONS

We proposed a multivariate analysis strategy that assembles unsupervised and supervised learning principles to interpret complex flow cytometry data. We used EM and GANN to identify the capacity of individual features within fluorescence histograms to predict the outcome of patients with HIV (i.e. non-progressor and progressor) and their survival time (in days). The classification models based on these feature sets were trained using 191 training samples and validated using a separate set of data (192 test samples).

Fewer than 7 features (i.e. 5 features from stimulated and 7 features from unstimulated) have been shown to be able to discriminate non-progressors from progressors and to predict survival time of patients (6 features for the stimulated dataset and 7 features for the unstimulated dataset). The better performance of these feature sets in the training set than in the test set is likely due to the significant variation on these features between training and test sets. Somewhat counter-intuitively, features from the unstimulated dataset were better at predicting the survival time than the features identified from the stimulated dataset. In summary, we have developed a new approach for analyzing complex flow cytometry datasets which has the potential to extract features that have the capacity to predict disease progression and survival in patients with HIV infection.

ACKNOWLEDGMENTS

We would like to thank FlowCAP Consortium for the organization of this highly interesting challenge.

CONFLICT OF INTEREST

The authors have no conflict of interest related to the research presented herein.

LITERATURE CITED

1. Laerum OD, Farsund T. Clinical application of flow cytometry: a review. Cytometry 1981;2:1–13. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7023887. Accessed April 18, 2013.

2. Barlogie B, Raber MN, Schumann J, Johnson TS, Drewinko B, Swartzendruber DE, Göhde W, Andreeff M, Freireich EJ. Flow cytometry in clinical cancer research. Cancer Res. 1983;43:3982–3997. Available at: http://www.ncbi.nlm.nih.gov/pubmed/6347364.

3. Jaye DL, Bray RA, Gebel HM, Harris WAC, Waller EK. Translational applications of flow cytometry in clinical practice. J. Immunol. 2012;188:4715–4719. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22556132. Accessed March 31, 2013.

4. Edwards BS, Oprea T, Prossnitz ER, Sklar LA. Flow cytometry for high-throughput, high-content screening. Curr. Opin. Chem. Biol. 2004;8:392–398. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15288249. Accessed March 21, 2013.

5. Krutzik PO, Crane JM, Clutter MR, Nolan GP. High-content single-cell drug screening with phosphospecific flow cytometry. Nat. Chem. Biol. 2008;4:132–142. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18157122. Accessed March 15, 2013.

6. Valet G. Cytomics as a new potential for drug discovery. Drug Discov. Today 2006;11:785–791. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16935745. Accessed April 18, 2013.

7. Aghaeepour N, Finak G, FlowCap Consortium T, DREAM Consortium T, Hoos H, Mosmann TR, Brinkman R, Gottardo R, Scheuermann RH. Critical assessment of automated flow cytometry data analysis techniques. Nat. Methods 2013;10:228–238.

Available at: http://www.nature.com/doifinder/10.1038/nmeth.2365. Accessed February 28, 2013.

8. Lo K, Brinkman RR, Gottardo R. Automated gating of flow cytometry data via robust model-based clustering. Cytom. Part A 2008;73:321–332. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18307272. Accessed March 4, 2013.

9. Pyne S, Hu X, Wang K, Rossin E, Lin T-I, Maier LM, Baecher-Allan C, McLachlan GJ, Tamayo P, Hafler DA, De Jager PL, Mesirov JP. Automated high-dimensional flow cytometric data analysis. Proc. Natl. Acad. Sci. 2009;106:8519–8524. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2682540&tool=pmcentrez&rendertype=abstract.

10. Boedigheimer MJ, Ferbas J. Mixture modeling approach to flow cytometry data. Cytom. Part A 2008;73:421–429. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18383311. Accessed April 10, 2013.

11. Tong DL, Schierz AC. Hybrid genetic algorithm-neural network: feature extraction for unpreprocessed microarray data. Artif. Intell. Med. 2011;53:47–56. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21775110. Accessed March 10, 2013.

12. Moon TK. The expectation-maximization algorithm. IEEE Signal Process. Mag. 1996;13:47–60. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=543975. Accessed March 12, 2013.

13. Do CB, Batzoglou S. What is the expectation maximization algorithm? Nat. Biotechnol. 2008;26:897–899. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18688245.

14. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: An update. SIGKDD Explor. 2009;11.

15. Tong DL, Mintram R. Genetic Algorithm-Neural Network (GANN): a study of neural network activation functions and depth of genetic algorithm search applied to feature selection. Int. J. Mach. Learn. Cybern. 2010;1:75–87. Available at: http://www.springerlink.com/index/10.1007/s13042-010-0004-x. Accessed March 2, 2013.

16. Schierz AC. Virtual screening of bioassay data. J. Cheminform. 2009;1:21. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2820499&tool=pmcentrez&rendertype=abstract. Accessed August 5, 2014.

Legends to Figures

Figure 1: Scatterplots of original samples, with the intensity based on Time represents with 2-color shades, i.e. blue means low frequency and yellow means high frequency. These plots pose challenges for manual gating, as each sample has to be independently measured and the gate for one sample cannot be used for another sample.

Figure 2: Scatterplots of gated 'live', singlet cells using EM.

Figure 3: Misclassification costs per dataset.

Figure 4: Kaplan Meier curves for stimulated and unstimulated cells. (Top) The predicted progression time in the training set. (Bottom) The predicted progression time in the test set.

Table 1: The TPR and FPR for the non-progressor group.

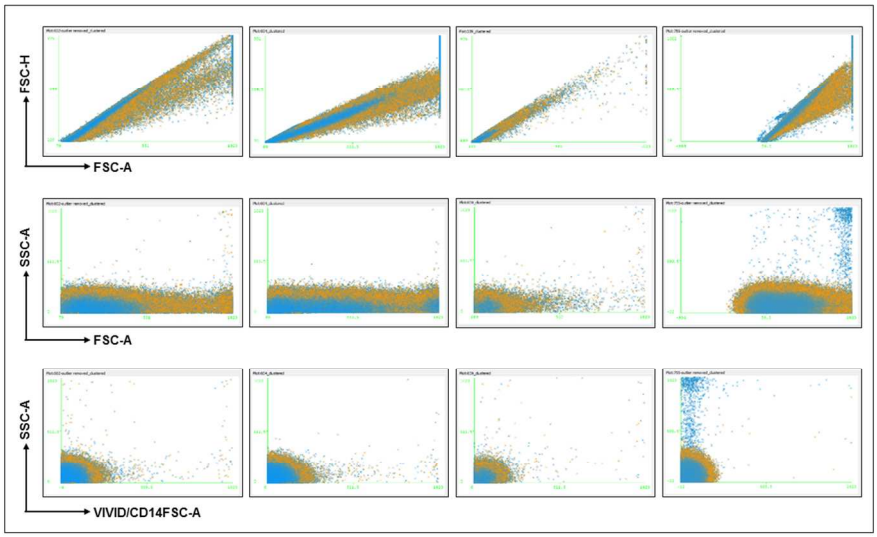
Dataset	Set	TPR (%)	FPR (%)	Accuracy (%)
Stimulated	Training	100.00	17.83	85.34
	Test	37.78	26.53	65.10
Unstimulated	Training	100.00	19.11	84.29
	Test	35.56	22.45	67.71

Table 2: The median intensities of the feature sets and the intensities different between sample sets.

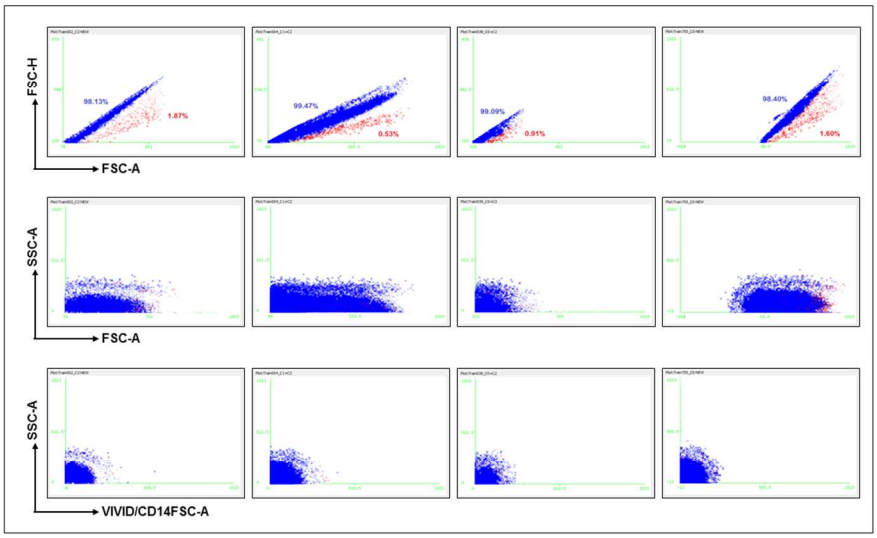
		Non-progressor		Progressor		Intensity difference (ratio)	
Dataset	**Feature set**	**Training**	**Test**	**Training**	**Test**	**Training**	**Test**
Stimulated	IL2_5	6576.5	5323	6632	8775	0.991	0.606
	IL2_6	3726	3860	4600	6463	0.81	0.597
	IL2_7	2170.5	2354	3337	4103	0.65	0.573
	IL2_8	979	1313	2180	2463	0.449	0.533
	IL2_9	380.5	526	1196	1432	0.318	0.367
Unstimulated	CD154_5	6094.5	6247	7564	7520	0.805	0.83
	CD154_6	3098	3537	5011	5141	0.618	0.687
	CD45RO_1023	14	11	40	28	0.35	0.392
	CCR7_9	2581.5	2448	3324	3338	0.776	0.733
	IL2_6	7017.5	5451	6523	7320	1.075	0.744
	IL2_7	4688	4699	5867	5297	0.799	0.887
	IL2_8	2391.5	2584	3980	4006	0.6	0.645

Table 3: Survival analysis results.

Dataset	Set	Mean (Actual)	Std. dev. (Actual)	Mean (Predicted)	Std. dev. (Predicted)	r
Stimulated	Training	947.36	966.96	947.05	430.65	0.535
	Test	933.61	851.00	942.16	489.05	0.149
Unstimulated	Training	947.36	966.96	947.13	721.88	0.825
	Test	933.61	851.00	966.21	545.17	0.173

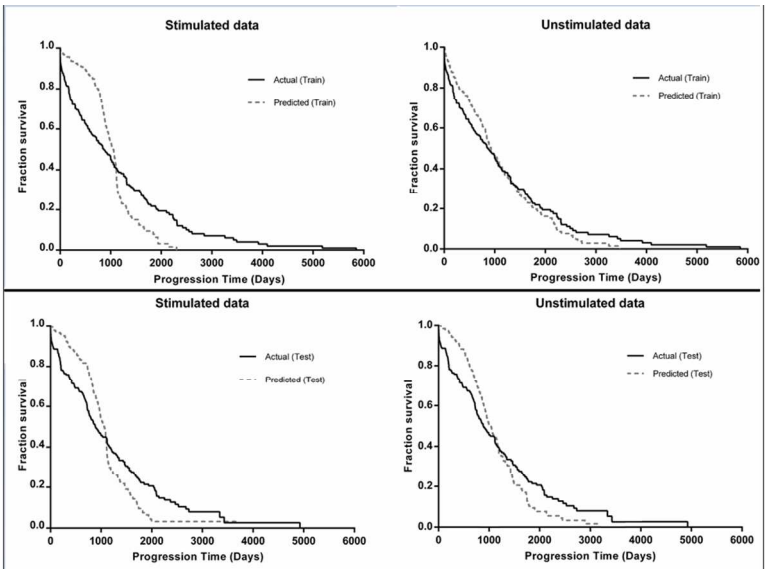


Scatterplots of original samples, with the intensity based on Time represents with 2-color shades, i.e. blue means low frequency and yellow means high frequency. These plots pose challenges for manual gating, as each sample has to be independently measured and the gate for one sample cannot be used for another sample.
338x190mm (96 x 96 DPI)



Scatterplots of gated 'live', singlet cells using EM.
338x190mm (96 x 96 DPI)

Misclassification costs per dataset.
338x190mm (96 x 96 DPI)



Kaplan Meier curves for stimulated and unstimulated cells. (Top) The predicted progression time in the training set. (Bottom) The predicted progression time in the test set.
338x190mm (96 x 96 DPI)