Received: 21 November 2022 | Revised: 7 August 2023 | Accepted: 18 September 2023

DOI: 10.1002/cyto.a.24805

ORIGINAL ARTICLE

MIFlowCyt

ISAC INTERNATIONAL SOCIETY FOR ADVANCEMENT OF CYTOMETRY

CYTOMETRY Journal of Quantitative Cell Science PART A

FCM marker importance for MRD assessment in T-cell acute lymphoblastic leukemia: An AIEOP-BFM-ALL-FLOW study group report

Florian Kowarsch[1] | Margarita Maurer-Granofszky[2,3] | Lisa Weijler[1] | Matthias Wödlinger[1,2] | Michael Reiter[1,2] | Angela Schumich[2] | Tamar Feuerstein[4] | Simona Sala[5] | Michaela Nováková[6] | Giovanni Faggin[7] | Giuseppe Gaipa[5] | Ondrej Hrusak[6] | Barbara Buldini[7,8] | Michael N. Dworzak[2,3]

[1]Computer Vision Lab, Faculty of Informatics, Technical University of Vienna, Vienna, Austria

[2]Immunological Diagnostics, St. Anna Children's Cancer Research Institute (CCRI), Vienna, Austria

[3]Labdia Labordiagnostik GmbH, Vienna, Austria

[4]The Rina Zaizov Division of Pediatric Hematology-Oncology, Schneider's Children's Medical Center, Petah Tikva, Israel

[5]M. Tettamanti Foundation Research Center, Department of Pediatrics, University of Milano-Bicocca, Monza, Italy

[6]Department of Pediatric Haematology and Oncology, University Hospital Motol, Prague, Czech Republic

[7]Pediatric Hematology, Oncology and Stem Cell Transplant Division, Maternal and Child Health Department, University of Padova, Padova, Italy

[8]Advanced Diagnostics and Target Discovery in ALL, Fondazione istituto di Ricerca pediatrica Città della Speranza, Padova, Italy

Correspondence
Michael N. Dworzak, Immunological Diagnostics, St. Anna Children's Cancer Research Institute (CCRI), Vienna 1090, Austria.
Email: dworzak@stanna.at

Funding information
Vienna Business Agency

Abstract

T-lineage acute lymphoblastic leukemia (T-ALL) accounts for about 15% of pediatric and about 25% of adult ALL cases. Minimal/measurable residual disease (MRD) assessed by flow cytometry (FCM) is an important prognostic indicator for risk stratification. In order to assess the MRD a limited number of antibodies directed against the most discriminative antigens must be selected. We propose a pipeline for evaluating the influence of different markers for cell population classification in FCM data. We use linear support vector machine, fitted to each sample individually to avoid issues with patient and laboratory variations. The best separating hyperplane direction as well as the influence of omitting specific markers is considered. Ninety-one bone marrow samples of 43 pediatric T-ALL patients from five reference laboratories were analyzed by FCM regarding marker importance for blast cell identification using combinations of eight different markers. For all laboratories, CD48 and CD99 were among the top three markers with strongest contribution to the optimal hyperplane, measured by median separating hyperplane coefficient size for all samples per center and time point (diagnosis, Day 15, Day 33). Based on the available limited set tested (CD3, CD4, CD5, CD7, CD8, CD45, CD48, CD99), our findings prove that CD48 and CD99 are useful markers for MRD monitoring in T-ALL. The proposed pipeline can be applied for evaluation of other marker combinations in the future.

KEYWORDS
CD48, CD99, feature importance, flow cytometry, minimal residual disease, support vector machine, T-lineage acute lymphoblastic leukemia

1 | INTRODUCTION

Acute lymphoblastic leukemia (ALL) is the most common pediatric cancer, with T-cell acute lymphoblastic leukemia (T-ALL) comprising 15% of all newly diagnosed ALL cases in pediatric patients. The event-free survival rates (EFS) have been steadily increasing and now exceed 85% in many contemporary clinical trials [1–3]. The success can be partly attributed to the development of a risk stratification approach to identify patients at high relapse risk. Once relapsed, the survival rate drops below 25% [4]. Hence, there is an urgent need to identify patients at high risk for relapse. In contrast to B-cell precursor acute lymphoblastic leukemia (BCP-ALL), new biological insights in T-ALL

have scarcely been incorporated into current protocols, and risk stratification relies mainly on response to treatment. Measurable residual disease (MRD, defined as the residual leukemic burden at specific time points during intensive therapy, is the single most powerful high-risk factor in acute leukemia, including T-ALL [5, 6]. Nowadays, real-time quantitative polymerase chain reaction-based methodology (PCR-MRD), as well as multicolor flow cytometry (FCM-MRD), are used for MRD assessment in the clinical setting [5]. FCM-MRD is based on the discrimination of normal and leukemic cells due to aberrant expression of specific antigens on T-ALL blasts (leukemia-associated immunophenotype, LAIP). Since immature T-cells are normally not present in peripheral blood (PB) or bone marrow (BM), the leukemic cells can be discriminated from the mature T-cells and Natural killer cells (NK-cells) [7]. Several leukemia-associated immunophenotype (LAIP)s for mostly immature blasts have already been described, such as the dual positive or negative expression of CD4 and CD8 [8, 9] or their isolated single expression, expression of CD34 [9, 10], aberrantly dim/negative or heterogeneous expression of CD45 [9, 11, 12], co-expression of CD56 [9, 13–15] or overexpression of CD7 [9, 16, 17]. In addition, T-ALL blasts display a 7.7 times higher median CD99 expression than normal T-cells from within the same sample [18] and in a study from the Children's Oncology Group (COG), reduced expression of CD48 has been recognized useful for FCM-MRD assessment in T-ALL [7, 19]. We, therefore, aimed to assess the contribution and relevance of these markers for discrimination of T-ALL blasts.

1.1 | Problem description

Often, only a limited number of antigens are used for MRD assessment in T-ALL, and their expression (levels) may vary during the course of therapy. In particular, immature antigens (such as TdT and CD99), which distinguish between leukemic and mature cells at diagnosis, are often affected by such expression changes [18, 20]. Especially T-ALLs that show expression of surface CD3 can often be particularly challenging with respect to MRD assessment. Therefore, the identification of new antigens that complement those currently in use is highly desirable.

1.2 | Research questions and contribution

In this work, we assess the relevance of eight markers for MRD assessment (Table S1) in pediatric T-ALL patients, based on those markers used by large international consortia in clinical practice for T-ALL MRD [7, 21]. We focus on identifying which markers provide crucial information for the separation of normal and leukemic cells, and the impact of omitting certain markers. Additionally, we examine the influence of sampling time point and EGIL stage [22] on marker relevance, and explore the gain of using CD45 in the presence of CD48. Our contribution includes the importance analysis of these markers and the proposal of a pipeline for evaluating cell population classification in FCM samples, addressing cross-patient and cross-laboratory variations, and accounting for class imbalance in MRD blast classification.

2 | RELATED WORK

Hoffmann et al. [23] used logistic regression to assess marker combinations for diagnosing B-cell chronic lymphocytic leukemia (B-CLL) using FCM samples. This method cannot be applied to assess feature importance for MRD determination, as MRD requires classifying individual events within FCM samples.

EuroFlow [24] developed unsupervised and supervised methods to linearly project FCM samples into a 2D space using principal component analysis [25] and canonical correlation analysis [24, 26], respectively. These methods assess marker importance by evaluating the contribution of individual markers to the linear projection. However, these approaches are only suitable for consistently standardized FCM databases, which is not the case for our data.

Linear support vector machine (LSVM) has been previously used for FCM sample analysis, such as the approach proposed by Qiu [27] that achieved high accuracy in rare cell identification. This approach fits LSVM to each sample individually to account for inter-laboratory differences (batch effects).

Guyon et al. [28] employed support vector machine (SVM) to evaluate gene importance in cancer classification. Our measurement of FCM marker importance is based on the same idea, where the weights assigned to the hyperplane in the SVM model reflect feature importance for classification.

3 | METHODS

We operate on single FCM samples, described by a set of events $E \in \mathbb{R}^{N \times M}$. N defines the number of events ($50 - 500 \times 10^5$) and M denotes the number of markers. For each sample, a ground truth vector t of length N exists, which defines each event's class membership (blast or non-blast). We aim to assess the feature importance of the M markers for separating the events corresponding to their class in t. SVM is utilized for this goal.

3.1 | Feature importance with SVMs

In brief, SVMs aim to separate two groups of observations by searching for a direction in the feature space that maximizes the distance between the closest observations of both groups [29]. (For further details on the SVM, the reader is referred to 6.1.1 in the supplementary information.) Since linear SVM discriminates along the direction of greatest margin between the closest observations of both classes, the coefficients of the hyperplane can be utilized to gauge the importance of each feature [28]. The larger β_i the more relevant is feature i in the decision function. Only LSVM directly indicate the feature importance in the coefficient vectors [30].

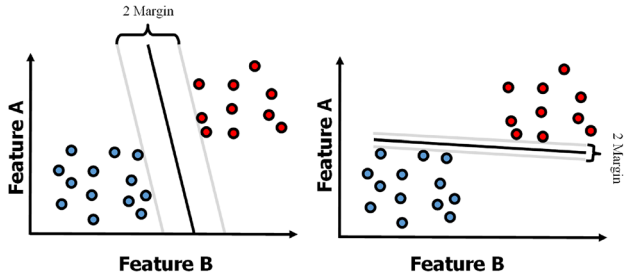

FIGURE 1 Two possible separating hyperplanes of a support vector machine that highlight the importance of feature selection when assessing feature importance. The left plot depicts the separating hyperplane with the widest margin. The right plot depicts another separating hyperplane. Feature (A) and (B) are both capable of separating the classes correctly, though Feature (B) is considered a strong separator due to the wide margin on the left plot in contrast to the margin on the right plot. [Color figure can be viewed at wileyonlinelibrary.com]

Figure 1 demonstrates how feature importance can be inferred from the hyperplane direction. This approach only considers each feature's influence on the direction of the widest separation between classes and may not account for other directions that rely on other features. For instance, in Figure 1, Feature A separates well, but has a low hyperplane coefficient in the left plot, since it only barely contribute to the separation direction with widest margin. This example highlights the importance of considering different combinations of features to assess the feature importance based on the hyperplane coefficients.

3.1.1 | Choosing linear SVM for blast cells separation assessment

Asking for linear separability can be an unnecessary restriction in many machine learning problems, but can be beneficial in certain problems. For the reasons mentioned below, LSVM was chosen in this study to assess the significance of the marker:

1. The direction of the SVM's separating hyperplane in the feature space can be directly used to assess the importance of individual features. The hyperplane coefficient vector can be interpreted as the gradient of the quadratic cost function [28] and therefore indicates directions of greatest sensitivity to data perpetuations. In addition, projections of the individual observations on the linear hyperplane can visually express the separability of a sample in an interpretable way. In contrast to random forest (RF) [31], where the feature importance is based on the impurity decrease within each tree [32] and therefore can be more difficult to interpret.
2. We aim to assess the usefulness of the given markers for blast identification. It is sufficient to measure the marker importance for each sample individually. We do not require the model to generalize from one FCM sample to others (which is especially difficult due to patient and laboratory variations). Nevertheless, the model's capacity must be controlled, since high-capacity models are prone to overfit a FCM sample, resulting in an unreliable estimation of the actual significance of the features. Following structural risk minimization (SRM) [33] we, therefore, opt for a linear model as it allows to fit each sample individual while still obtaining reliable importance scores.
3. Detecting MRD is a classification problem with high imbalance. In our dataset, some samples have only about 30 blasts, which leads to a scarcity of data points. Consequently, low-capacity models like linear models are preferred for evaluating feature importance, despite the possibility of a true decision boundary of higher-order. A low-capacity model allows us to reliable estimate the feature importance on individual samples and account for the limited number of blast cells.
4. LSVM is considered to be robust to outliers since it is only sensitive to the data points closest to the decision boundary. In contrast, logistic regression, another linear classification method, considers all data points and can be stronger influenced from outlying data points (See 7.1.2 Relation to logistic regression in Reference [34]).

3.2 | Assess separation

As we aim to identify a small amount of blast cells in the whole population, an imbalanced binary classification problem is faced. Therefore, the evaluation focuses on the precision, recall, and F1-score to assess the separation of the blast cells. The precision

$$\begin{aligned}\text{Precision} &= \frac{\text{true positives}}{\text{true positives} + \text{false positives}} \\ &= \frac{\text{correctly identified blast cells}}{\text{correctly identified blast cells} + \text{wrongly classified non-blast cells}},\end{aligned} \tag{1}$$

defines, in the context of blast cell classification, the proportion of correctly identified blast cells among all identified blast cells. The Recall

$$\begin{aligned}\text{Recall} &= \frac{\text{true positives}}{\text{true positives} + \text{false negatives}} \\ &= \frac{\text{correctly identified blast cells}}{\text{correctly identified blast cells} + \text{not identified blast cells}},\end{aligned} \tag{2}$$

provides the ratio of correctly identified blast cells to all blast cells. A recall score of one indicates that all blasts have been identified. However, this also holds if a classifier simply predicts all cells as blast cells. Therefore, F1-score

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \tag{3}$$

is used to seek a balance between recall and precision.

4 | EXPERIMENTS

In this section, the used datasets and the experiment setup are described as well as the results are presented.

4.1 | Datasets

We collected 91 leukemic blast-positive samples from 43 pediatric T-ALL patients recruited by five distinct national reference laboratories. Sampling and research were approved by local Ethics Committees, and informed consent was obtained from patients or patient's parents or legal guardians according to the Declaration of Helsinki. The samples were submitted for FCM diagnostics in pediatric leukemias and MRD monitoring and were collected at the time of diagnosis (Day 0) as well as during or after induction therapy (Days 15 and 33) of a Berlin-Frankfurt-Münster (BFM)-type protocol (Table S4 for the number of samples per day and EGIL stage). Local operators annotated all samples with respect to MRD (see Figures S1 and S2 for more information about the distribution of the amount of MRD in our data and see Tables S1–S3 for additional insight on the used clones and fluorochromes.)

All samples have been compensated and in principle manual gating was performed as follows: Debris was excluded via side scatter (SSC) and forward scatter (FSC) gating and singlets were selected using FSC-A/FSC-H gating. Nucleated cells were selected based on Syto41 positivity (if included in the panel). The best separating marker combination was used to define MRD within the CD7 positive compartment (CD7+). We constructed the largest set of common markers among all samples from all laboratories. These markers are referred in the following as default marker set and consist of CD3, CD5, CD7, CD45, CD48, CD99. As described in Table S1, an extended marker set is defined, which includes the additional markers CD4 and CD8. Samples with the extended marker set were available from only three of the five laboratories. See Section 6.2.1 for further information on the individual datasets as well as Table S3 for information on the used antibody clones.

4.2 | Setup

To account for differences in the data due to antibody clone choice, used FCM, or gating strategy of different operators we process each sample separately. By avoiding to combine events from different samples we ensure that we measure how good specific markers serve as separators only in relation to the other markers in the given sample. This sample-wise processing also ensures that samples with a low number of events are not disregarded, as each sample equally contributes to the obtained results. In addition, individual processing of each sample followed allows parallelization and of cache the calculations of different samples, before the aggregated results are obtained. We conducted the following experimental setup, in which, for each FCM sample, the following steps are performed (as depicted in Figure 2):

1. To investigate the effects of different markers on the separability, different subsets of the data are considered. k combinations of markers are constructed from the initial set of m markers. In each combination, 1–3 markers of the initial set are left out. For instance, one subset can contain all markers except CD48 to measure the impact of omitting CD48.
2. All events outside the CD7+ gate are omitted, since we are only interested in feature importance related to T-ALL specific events and not related to discriminant more general cell groups as performed by the initial gates. The operators provide the CD7+ gate in the ground truth.
3. The remaining events are fitted k times (for each of the k marker combinations) with a linear SVM.

From these computations the following information is obtained:

- For each sample, the direction of the separating SVM hyperplane is retrieved to infer the importance of each marker for this separation. As shown in Equation 4 for each marker i of all p markers the normalized parameter value $\widehat{\beta}_i$ is computed. $\widehat{\beta}_i \in [0,1]$ therefore expresses the influence of the ith marker for the separation. Such that $\widehat{\beta}_i = 0$ indicates that the ith marker had no influence on the

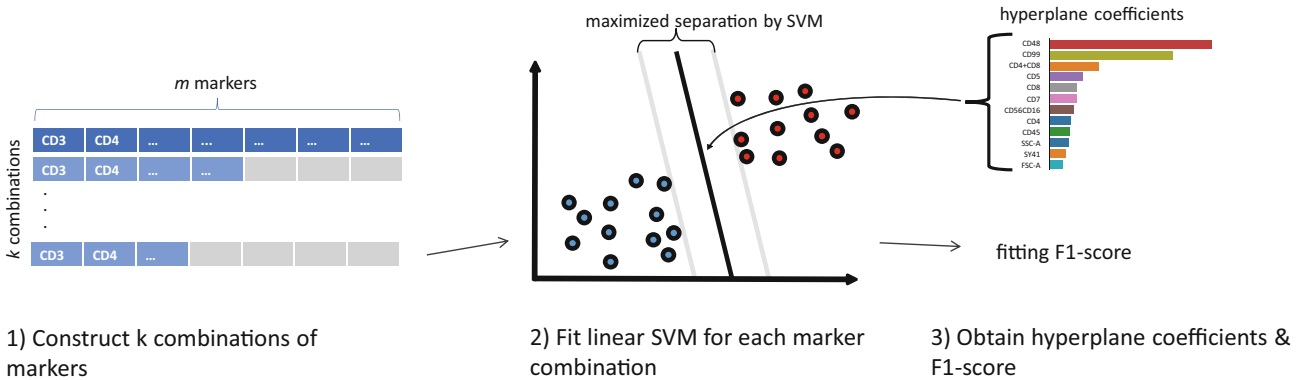

FIGURE 2 The processing pipeline of flow cytometry sample. k marker combinations are constructed, and for each combination, a linear support vector machine (SVM) is fitted to the data. The obtained hyperplane coefficient and the fitting F1-score are analyzed in comparison to the other samples. [Color figure can be viewed at wileyonlinelibrary.com]

separation and that $\widehat{\beta}_i = 1$ describes that the ith marker entirely separates the data.

$$\widehat{\beta}_i = \frac{|\beta_i|}{\sum_{j=1}^{p} |\beta_j|}, \tag{4}$$

- For each fit, the F1-score is obtained, which serves as a metric to rate the fit quality. It, therefore, allows measuring the impact of leaving out one marker e.g. CD48 on the sample fit quality.

4.3 | Results

The results are considered from three different perspectives: First, the parameters of the separating SVM hyperplanes are evaluated. Then the impact of omitting markers on the fitting quality is investigated, followed by the visualizations of the separability of the hyperplanes on a single sample level.

4.3.1 | Hyperplane direction

Figure 3 depicts the normalized parameters of the separating linear SVM hyperplane of each sample using the default marker set. The markers are ordered by their median parameter size. For all five datasets, CD48 is among the top two parameters with the largest median coefficient. This means that it strongly contributes to the direction of maximized class separation. CD99 is among the top two parameters in Vienna, Padova, and Petah Tikva datasets and third place in the Monza and Prague dataset.

Figure 4 shows the impact of omitting markers on the hyperplane parameters. The top three normed hyperplane parameters are compared among the k different combinations of omitted markers. Omitting CD48 clearly increases the parameter size of CD45 while omitting CD45 seems not to affect the parameter size of CD48. If CD99 and CD48 are removed from the data, CD45 contributes the strongest to the separating direction. Removing all three markers (CD45, CD48, CD99) reveals CD3 and CD5 as the markers with the strongest contribution among the default marker set.

4.3.2 | Fitting quality

To measure the impact of individual markers on the separation quality, we measure the F1-scores of the fitted hyperplane for each FCM sample. Figure 4 shows the fit F1-scores of all samples compared among the k different marker combinations. The violin plots are ordered by the median F1-score. The best median F1-score is achieved when all markers of the default marker set are present. The worst median F1-score is obtained when CD48, CD99, and CD45 are omitted. The exact values can be derived from Table S5.

4.3.3 | Extended marker set

As stated in Table S1 three of the five datasets also share two additional markers CD4 and CD8, which forms the extended marker set. We reevaluated all experiments on this extended marker set and compared the results to the default marker set. Figure S3 depicts the parameters of the separating hyperplane of each sample using the extended marker set. The additional markers CD4 and CD8 are considered individual and combined. CD48 is among the top two parameters with the largest median coefficient for all three datasets. Figure S5 shows the fit F1-scores of the samples using the extended marker set. Nearly the same order of the violin plots as in Figure S4 is present. Only the violin plots of "w/o CD45" and "w/o CD5" changed their position. Including CD4 and CD8 generally did not change the result's main takeaways, neither as simple markers nor in combination.

4.3.4 | Influence of acquisition time point

Since the obtained data contains measurements from several acquisition time-points the results were analyzed regarding dependence on the acquisition day. In Figure 5 the normalized hyperplane coefficient values of all samples separated by acquisition day using both marker sets are displayed. Regardless of the acquisition time-point CD48 and CD99 are among the top two markers. The fit F1-scores divided per sampling time point is depicted in Figure 6. In general, marker omission reduces the fit F1-scores less on Day 0, than on Days 15 and 33, indicating that blast event separation is easier at diagnosis. For both marker sets, throughout all analyzed time points, omitting CD48, CD99, and CD45 show the strongest decline in median fit F1-score.

4.3.5 | Influence of maturation stage

Flow MRD can be influenced by stage of maturation. We categorized the samples into 3 categories according to EGIL stage [22] (T-I-II, T-III, T-IV) and analyzed the results regarding dependence on the EGIL stage. Due to the small sample size, data from EGIL T-I and T-II were combined in the study. In Figure S6 the normed hyperplane coefficient values of all samples separated by EGIL stage using both marker sets are displayed. Regardless of the stage CD48 and CD99 are among the top two markers.

5 | DISCUSSION

Across laboratories, different panels are utilized for MRD assessment in T-ALL, which makes it difficult to compare results among laboratories. Since immature T-cells are normally not present in PB or BM a combination of immaturity markers with mature T-cell markers can be used for T-ALL MRD detection [35, 36]. However, the used panels often rely on a relatively small number of antigens, of which not all are stable during therapy [7, 18]. Phenotypic shifts can occur, which

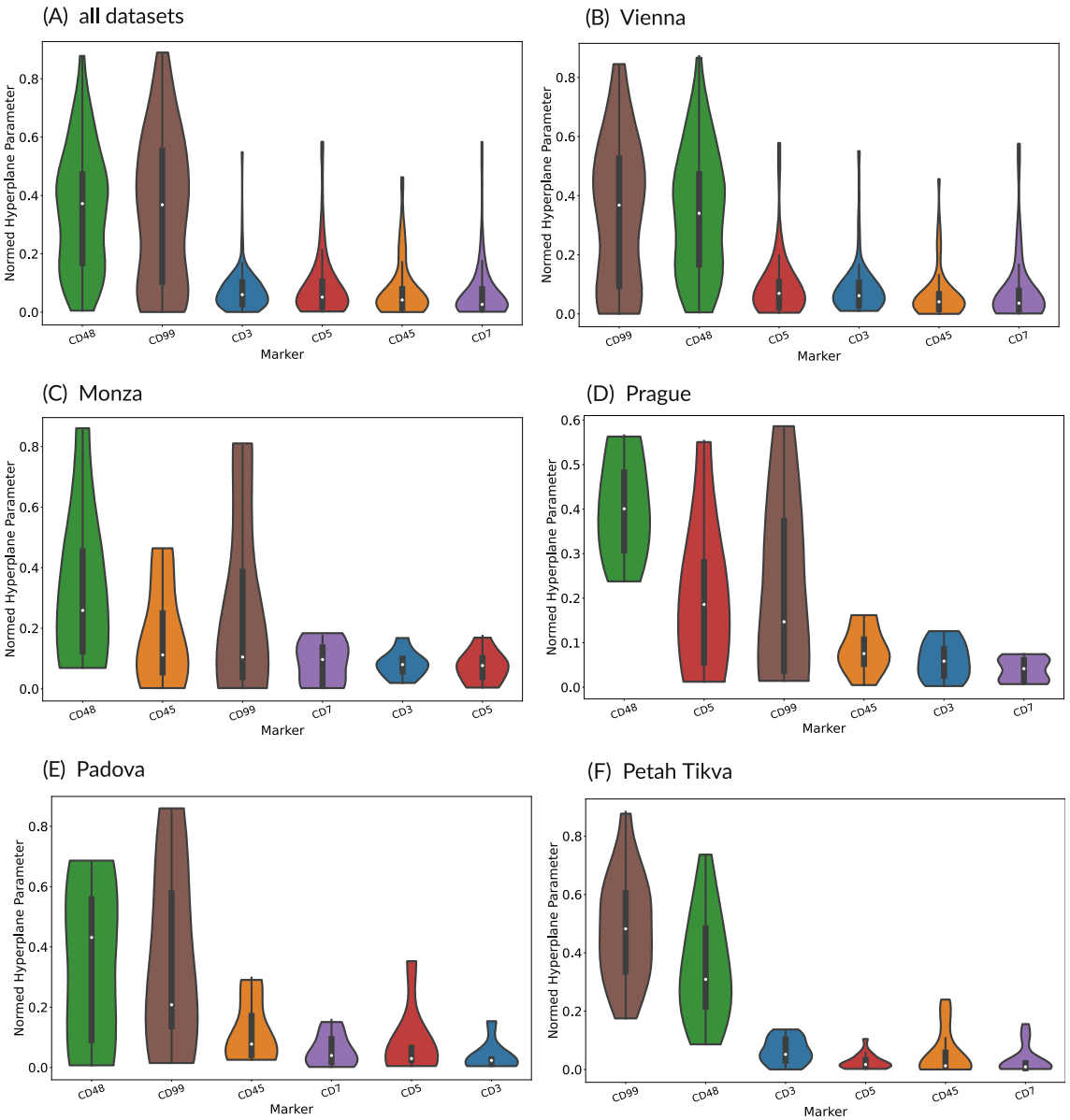

FIGURE 3 CD48 is always among the markers with top two largest hyperplane parameters. Higher values indicate a stronger contribution to the direction of maximized class separation. The Figure shows violin plots with boxplots of the normalized hyperplane parameters ordered by median using the default marker set. (A) All 91 samples. (B) The 51 samples of the Vienna dataset. (C) The 8 samples of the Monza dataset. (D) The 8 samples of the Prague dataset. (E) The 7 samples of the Padova dataset. (F) The 17 samples of the Petah Tikva dataset (see Figure S3 for the same plot on the extended marker set). [Color figure can be viewed at wileyonlinelibrary.com]

urges the use of broader panels to avoid false negative MRD results [20, 36]. While the occurrence of phenotypic shifts during leukemia treatment is well described for B-cell acute lymphoblastic leukemia (B-ALL) [37, 38], less literature exists on early phenotypic shifts in T-ALL [18, 39]: Roshal et al. [20] reports that markers of immaturity such as TdT and CD99 decline on leukemic cells during therapy. In contrast, Janeliūnienė et al. [39] observed that the TdT and CD99 mean fluorescence intensity (MFI) of blast cells fluctuate but are always higher than in normal T-cells. Several markers and their combination, including TdT, CD99, CD7, CD4 & CD8, CD3, CD5, and CD45 have been established as suitable for MRD assessment in T-ALL [9, 20].

The deployment of TdT is technically complicated by requiring intracellular staining and is, therefore, less often used compared to surface-only approaches [9, 40, 41]. In addition, TdT is not applicable for all T-ALL patients for MRD assessment [18, 42].

Dworzak et al. [18] discovered that CD99 has in median 7.7 times higher expression on T-ALL blasts than on normal T-cells from within the same sample. Enein et al. [36] showed that CD99 was expressed in all T-ALL patients with a higher median expression level compared to age and sex-matched healthy control patients, making it therefore especially attractive for MRD assessment.

The T-cell marker CD7 is stable across therapy. It appears at the earliest stage of T-cell development and is almost always expressed

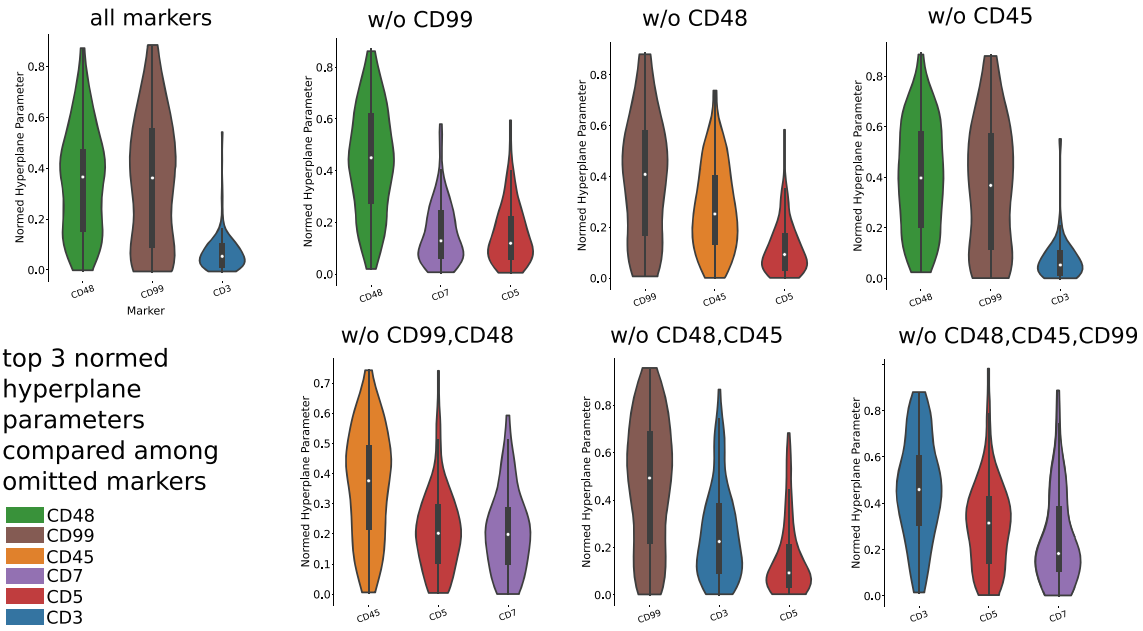

FIGURE 4 If not omitted, CD48 or CD99 have the highest median hyperplane coefficient values. Omitting CD48 leads to an increase of CD45 coefficient size. The figures shows violin plots with boxplots of the top three hyperplane coefficient values using the default marker set on the samples of all five centers. Each plot corresponds to one combination of omitted markers (e.g., "w/o CD99" depicts the top three highest hyperplane coefficient values without using CD99 for fitting the support vector machine). [Color figure can be viewed at wileyonlinelibrary.com]

on T-ALL blasts [16, 17]. Therefore in most gating strategies related to MRD in T-ALL, CD7 is used for the initial gating of cells of the T-cell lineage. In addition, CD7 is often overexpressed in T-ALL and, together with CD99, can help to identify MRD. Weak(er) expression of CD45 (compared to the normal lymphocytic counterparts) is a hallmark of ALL cells, including T-ALL blasts [9] although in general B-ALL blasts express lower CD45 than T-ALL blasts. The weak expression of CD45 can be used for blast identification, especially in combination with overexpression of CD99 and/or downregulation of CD5. Nevertheless, blast identification can be difficult in cases with higher CD45 expression as for example in mature T-ALL.

More recently, we implemented CD48 in our panel, an antigen that has been identified as one of the few antigens that showed consistent differential expression between mature T cells and NK cells in comparison to T-ALL blasts [7]. In our experience, the CD48 staining pattern of normal leukocytes is reminiscent of that of CD45. We found, however, that CD48 was, in most cases, superior to CD45 in separating blasts from normal T-cells. Nevertheless, those results can always only be interpreted minding the limited set of available markers (CD3, CD4, CD5, CD7, CD8, CD45, CD48, CD99) in our study.

5.1 | Research questions

In the following, we revisit the research questions as formulated in Subsection 1.2 for further discussion:

1. Which markers and marker combinations allow the best separation between normal and leukemic T-cell? CD48 has in median higher hyperplane coefficient values than any other marker except CD99 (Figure 3). This indicates that CD48 is important for the direction of separating hyperplane that maximizes the margin between both classes. However, CD48 seems to be also important for the general separation performance (Figure S4). Removing CD48 and CD99 from the available data reduces the F1-score (calculated from the per-sample fit). Removing CD48 alone does not show a significant reduction of F1-score but the second greatest reduction when removing a single marker according to the median value. The decrease in overall separation performance (measured by F1-score) when a marker is removed, indicates that this marker includes essential information for class differentiation that no other feature implies. It is, therefore, justifiable to state that based on the given data CD48 benefits the separation of blast cells in T-ALL samples for MRD assessment. Similar to CD48, the results show high importance of CD99 for class distinction. CD99 has the second largest median hyperplane coefficients over all samples (Figure 3). And, although not significant, removing CD99 shows the second-largest median decrease in F1-score among single marker omission (Figure S4).
2. The omission of which markers worsen the separation significantly? Removing CD45, CD48, and CD99 yields the largest reduction in median F1-score, followed by either removing CD45 and CD48 or CD48 and CD99 (Figures S4 and S5). Omitting CD3, CD5 or CD45 alone shows no reduction in median F1-score. The exact median F1-score values can be derived from Table S5.
3. Does the time point of sampling or the EGIL stage [22] influence the relevance of different markers? In line with [18, 20] a decline of CD99 on leukemic cells during therapy occurs. However, similar to the findings in [39] CD99 still retains enough separability on

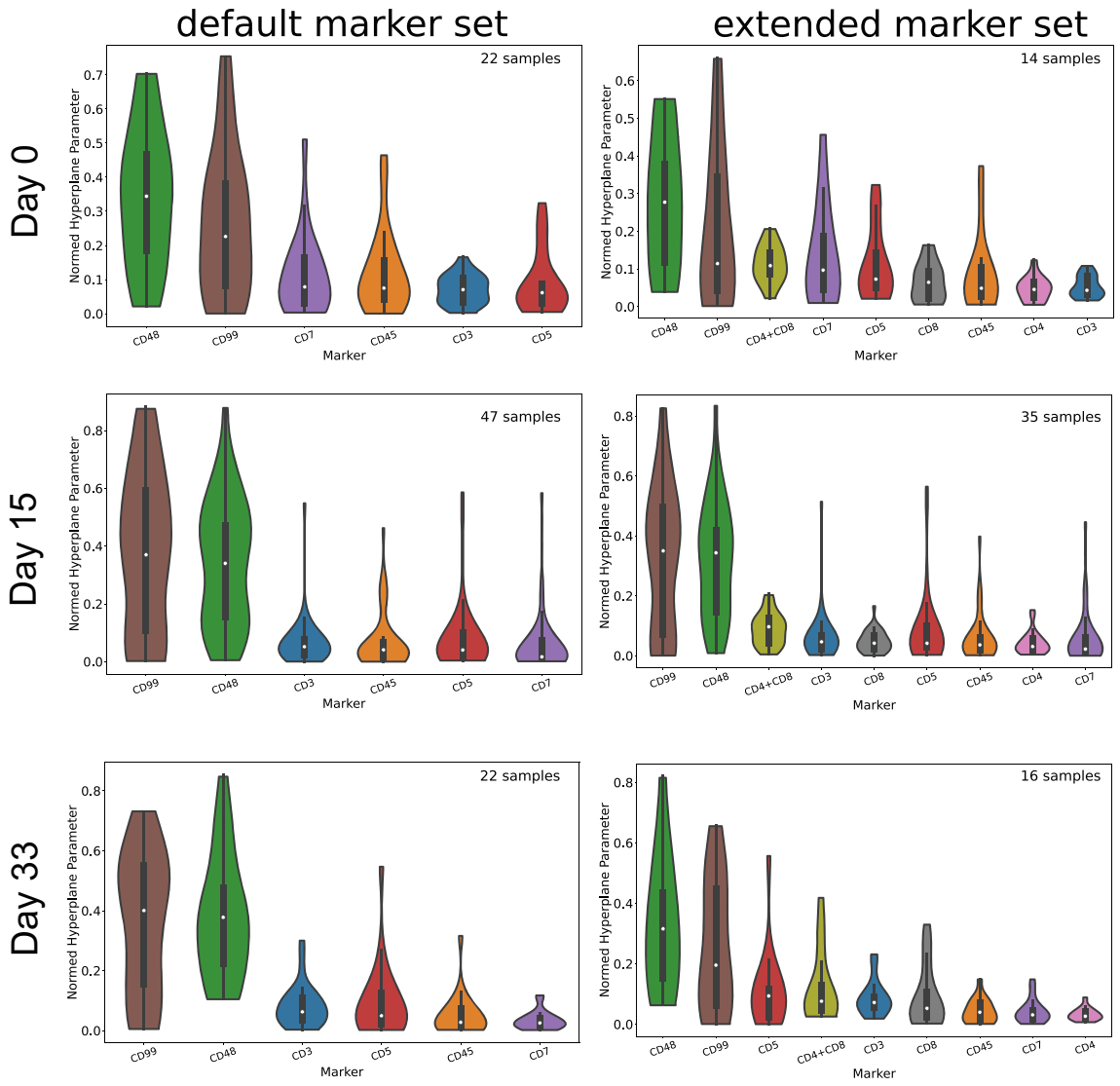

FIGURE 5 CD48 and CD99 are both among the top two markers for all three time points. CD7 shows higher coefficient median values for Day 0 than Day 15, and Day 33. The CD4 + CD8 combination ranks among top four markers at all time points. The Figure depicts violin plots with boxplots of normed hyperplane coefficient values separated by acquisition day (rows) of all 91 samples using the default marker set (left column) and all 65 samples using the extended marker set (right column). [Color figure can be viewed at wileyonlinelibrary.com]

treatment Days 15 and 33 such that it, together with CD48, influences the direction of greatest separation the most, therefore remaining the strongest separators throughout all analyzed time points. While the results in Figure 4 illustrated, that CD3 and CD5 do have the strongest hyperplane coefficients if CD45, CD48, and CD99 are omitted from the default marker set, we see in Figure 6 that the separation in this case leads on Day 33 to poor (≤0.2 median F1 default marker set) median F1 fitting scores. This effect is less distinct on the extended marker set (≤0.5 median F1 extended marker set), which could be an indication that the additional markers (CD4 and CD8) in combination with CD3 and CD5 are beneficial on Day 33 in absence of CD45, CD48, and CD99. The analysis of different EGIL stages in Figure S6 shows no contrast to the results compiled among all samples. Both CD48 and CD99 are among the top two markers over all stages.

4. Since CD45 and CD48 show a very similar expression pattern on normal leukocytes, what is the gain of using CD45 in the presence of CD48 and vice versa? The results highlight a dependence between CD48 and CD45. In Figure 4, which displays the top 3 highest median hyperplane coefficients for different marker combinations, the coefficient size of CD45 increases if CD48 is omitted from the data. If both CD48 and CD99 are removed CD45 has the top largest median hyperplane coefficient size, and this size is significantly greater compared to using all other markers. Removing CD48 harms the F1-score, while removing CD45 does not harm the fitting F1-score if CD48 is used (Figures S4 and S5). This indicates that CD48 includes both the information about the class distinction of CD45 and essential information that CD45 does not have. We conclude that both are useful discriminators that should both be utilized if possible. The emerging growth in deployable panel size on modern cytometers will allow

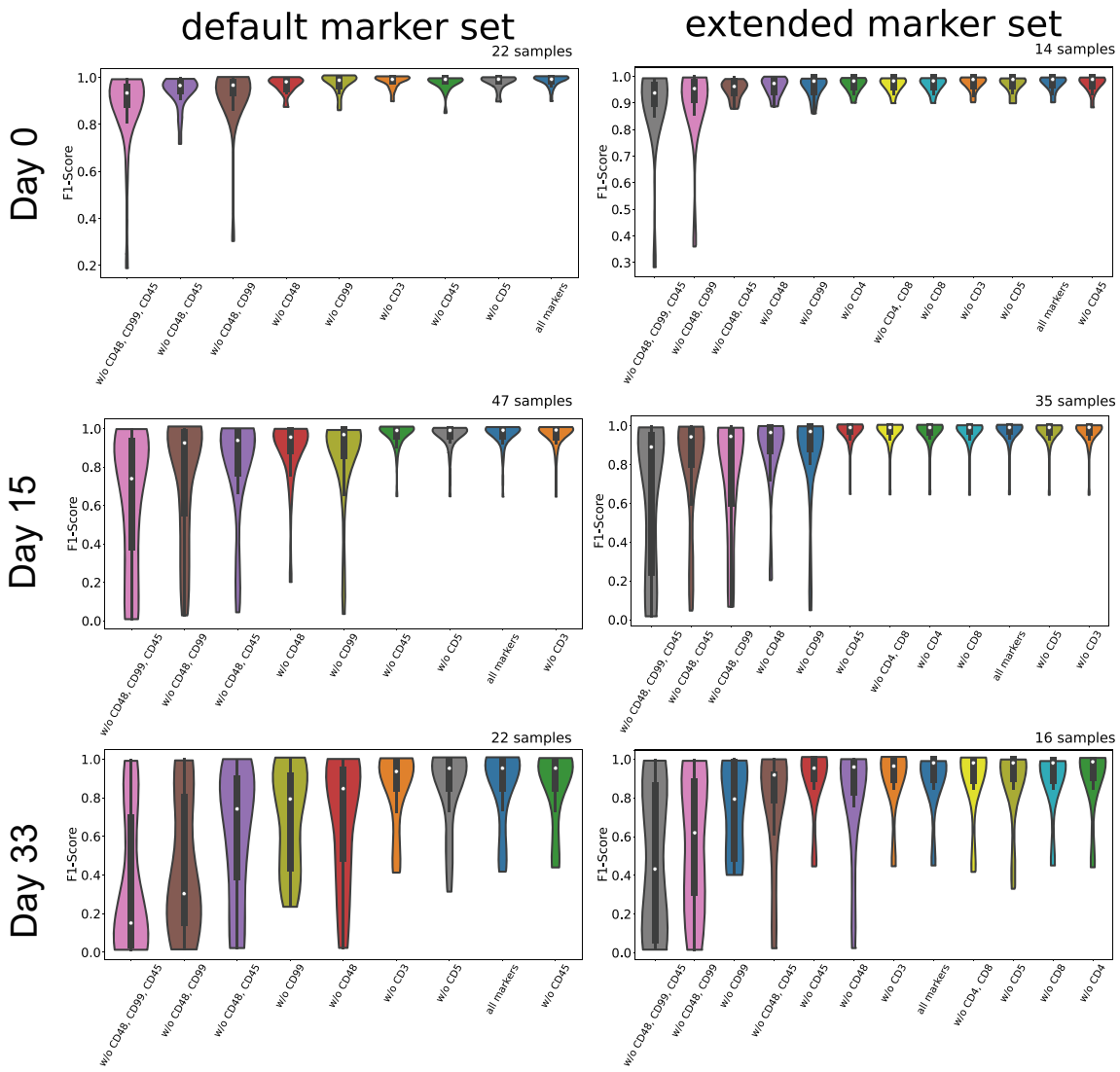

FIGURE 6 In general marker omission reduces the fit F1-scores less on Day 0, than on Days 15 and 33, indicating that blast event separation is easier at diagnosis. The figures shows violin plots with boxplots of the fit F1-scores separated by acquisition day (rows) of all 91 samples using the default marker set (left column) and all 65 samples using the extended marker set (right column). [Color figure can be viewed at wileyonlinelibrary.com]

extending established panels with CD48. However, when a choice must be made, CD48 should be chosen in favor of CD45.

5.2 | Limitations

Since our analysis was conducted under retrospective conditions, the operators could not be inhibited from using, for example, CD48 during manual gating. Therefore, the results may be biased toward the operators' decision to gate the blast populations via specific markers/marker combinations. Future work could further investigate the importance of selected markers without using them in manual gating. The data sets used contain a different number of samples. Furthermore, the centers use varying panel setups with regard to marker composition and fluorochromes. The choice of the fluorochrome is certainly important and may impact the separative power of certain markers. In addition, the study involves different patient cohorts with heterogeneous immunophenotypes. Therefore, we cannot exclude center-specific issues. Given the limited number of samples available, it was unfeasible to perform separate analyses on all the maturation stages as defined by the EGIL classification, resulting in a consolidation of data from EGIL T-I and T-II for the study. In addition, the panel we used for the study does not contain all of the markers that are reported to be used for MRD assessment in T-ALL. However, our panel is based on those used by large international consortia in clinical practice for T-ALL MRD [7, 21] and has already been shown to have a good predictive power and correlates well with outcome [21]. A future prospective study could investigate the importance also for those markers that are less common used in clinical practice for T-ALL MRD.

5.2.1 | Choice of machine learning pipeline

Our study aims to assess the usefulness of given markers for blast identification. We recognize that generalization is a well-known challenge in FCM data analysis and assuming linear separability introduces bias. Our choice of a linear model based on SRM was motivated by the need to control the model's capacity and prevent overfitting of a particular FCM sample, which could lead to unreliable feature importance scores. However, this does not mean that a linear model is the only viable approach, nor does it imply that generalization is not important. We do not aim to generalize from one FCM sample to others, since the aim is not finding stable decision regions across samples. This does not imply that generalization within a sample is not important for the task at hand. This means that the variance of the hyperparameter coefficients should be small if adapted to several random event subsets of the same sample (See 6.2. Preliminary Evaluations in Supplementary Information).

6 | CONCLUSION

Based on the limited observed markers, our experiments indicate that CD99, CD48, and CD45 are all strong discriminators for blast cells among the CD7+ cell compartment when monitoring MRD in T-ALL. Although CD99 expression may decline during therapy, it retains strong separation capabilities. This conclusion does not differ when analyzing the results per laboratory or maturation stage. The proposed evaluation pipeline is designed for cross-laboratory MRD assessment, accounting for the lack of standardization between individual samples. It can also be used beyond MRD assessment to quantify the importance of FCM markers for event population separation in other contexts.

AUTHOR CONTRIBUTIONS

Florian Kowarsch: Software; visualization; writing – original draft; conceptualization; methodology; validation; writing – review and editing; formal analysis. **Margarita Maurer-Granofszky:** Investigation; validation; writing – review and editing; data curation; supervision; writing – original draft; conceptualization; project administration; methodology; funding acquisition. **Lisa Weijler:** Writing – review and editing; software; visualization. **Matthias Wödlinger:** Writing – review and editing. **Michael Reiter:** Methodology; supervision; project administration; writing – review and editing; funding acquisition; resources; conceptualization. **Angela Schumich:** Data curation; writing – review and editing. **Tamar Feuerstein:** Data curation; writing – review and editing. **Simona Sala:** Data curation; writing – review and editing. **Michaela Nováková:** Data curation; writing – review and editing. **Giovanni Faggin:** Data curation; writing – review and editing. **Giuseppe Gaipa:** Data curation; writing – review and editing. **Ondrej Hrusak:** Data curation; writing – review and editing. **Barbara Buldini:** Data curation; writing – review and editing. **Michael N. Dworzak:** Data curation; supervision; project administration; validation; investigation; conceptualization; writing – review and editing; funding acquisition; methodology; resources.

ACKNOWLEDGMENTS

This study was initiated and promoted by the International BFM Study Group. The authors thank all doctors, nurses, and technicians of the participating study groups for their close collaboration. We thank Dieter Printz (FACS Core Unit, CCRI) for flow-cytometer maintenance and quality control, as well as Daniela Scharner, Susanne Suhendra, Alice Bramböck, and Claudia Mitteregger (CCRI) for excellent technical assistance. We also thank Pamela Scarparo, Elena Varotto (Padova) as well as Oscar Maglia (Monza) for the generation and provision of data. We thank Markus Kaymer and Michael Kapinsky (both from Beckman Coulter Inc.) for kindly providing customized DuraCloneTM tubes for this study as designed by the authors. We thank Fondazione GrandeAle ONLUS for its support in the research on T-ALL in the center of Monza (IT) as well as Fondazione Cariparo for the support of the center in Padova (IT) (through Grant no. 20/12 to B. Buldini). Notably, Beckman Coulter Inc. did not have any influence on study design, data acquisition, and interpretation, or manuscript writing.

CONFLICT OF INTEREST STATEMENT

M.N.D. received payments for invited lectures and travel expenses from Beckman-Coulter. Beckman-Coulter, Exbio, and Becton Dickinson supported the study with respect to privileged material acquisition. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

INFORMED CONSENT

Informed consent was obtained from all patients or patient's parents or legal guardians according to the Declaration of Helsinki.

ORCID

Florian Kowarsch https://orcid.org/0000-0001-7946-6169
Matthias Wödlinger https://orcid.org/0000-0002-3872-7470
Michael Reiter https://orcid.org/0000-0002-8004-6839
Simona Sala https://orcid.org/0000-0001-6357-0375
Michaela Nováková https://orcid.org/0000-0003-2964-5956
Giuseppe Gaipa https://orcid.org/0000-0002-1006-5946

REFERENCES

1. Möricke A, Zimmermann M, Valsecchi MG, Stanulla M, Biondi A, Mann G, et al. Dexamethasone vs prednisone in induction treatment of pediatric ALL: results of the randomized trial AIEOP-BFM ALL 2000. Blood. 2016;127(17):2101–12.
2. Vora A, Goulden N, Wade R, Mitchell C, Hancock J, Hough R, et al. Treatment reduction for children and young adults with low-risk acute lymphoblastic leukaemia defined by minimal residual disease (UKALL 2003): a randomised controlled trial. Lancet Oncol. 2013; 14(3):199–209.
3. Place AE, Stevenson KE, Vrooman LM, Harris MH, Hunt SK, O'Brien JE, et al. Intravenous pegylated asparaginase versus

intramuscular native Escherichia coli L-asparaginase in newly diagnosed childhood acute lymphoblastic leukaemia (DFCI 05-001): a randomised, open-label phase 3 trial. Lancet Oncol. 2015;16(16):1677–90.
4. Reismüller B, Attarbaschi A, Peters C, Dworzak MN, Pötschger U, Urban C, et al. Long-term outcome of initially homogenously treated and relapsed childhood acute lymphoblastic leukaemia in Austria–a population-based report of the Austrian Berlin-Frankfurt-Münster (BFM) Study Group. Br J Haematol. 2009;144(4):559–70.
5. Modvig S, Madsen H, Siitonen S, Rosthøj S, Tierens A, Juvonen V, et al. Minimal residual disease quantification by flow cytometry provides reliable risk stratification in T-cell acute lymphoblastic leukemia. Leukemia. 2019;33(6):1324–36.
6. Conter V, Valsecchi MG, Buldini B, Parasole R, Locatelli F, Colombini A, et al. Early T-cell precursor acute lymphoblastic leukaemia in children treated in AIEOP centres with AIEOP-BFM protocols: a retrospective analysis. Lancet Haematol. 2016;3(2):e80–6.
7. Wood BL, Levin G, Wilson M, Winter SS, Dunsmore K, Loh ML, et al. High-throughput screening by flow cytometry identifies reduced expression of CD48 as a universal characteristic of T-ALL and a suitable target for minimal residual disease (MRD) detection. Blood. 2011;118(21):2547.
8. Bhandoola A, von Boehmer H, Petrie HT, Zúñiga-Pflücker JC. Commitment and developmental potential of extrathymic and intrathymic T cell precursors: plenty to choose from. Immunity. 2007;26(6):678–89.
9. Tembhare PR, Chatterjee G, Khanka T, Ghogale S, Badrinath Y, Deshpande N, et al. Eleven-marker 10-color flow cytometric assessment of measurable residual disease for T-cell acute lymphoblastic leukemia using an approach of exclusion. Cytometry B Clin Cytom. 2021;100(4):421–33.
10. Porwit-MacDonald A, Björklund E, Lucio P, Van Lochem E, Mazur J, Parreira A, et al. BIOMED-1 concerted action report: flow cytometric characterization of CD7+ cell subsets in normal bone marrow as a basis for the diagnosis and follow-up of T cell acute lymphoblastic leukemia (T-ALL). Leukemia. 2000;14(5):816–25.
11. Nakamura A, Tsurusawa M, Kato A, Taga T, Hatae Y, Miyake M, et al. Prognostic impact of CD45 antigen expression in high-risk, childhood B-cell precursor acute lymphoblastic leukemia: Children's cancer and leukemia study group (CCLSG). Leuk Lymphoma. 2001;42(3):393–8.
12. DiGiuseppe JA, Wood BL. Applications of flow cytometric immunophenotyping in the diagnosis and posttreatment monitoring of B and T lymphoblastic leukemia/lymphoma. Cytometry B Clin Cytom. 2019;96(4):256–65.
13. Azzam HS, Grinberg A, Lui K, Shen H, Shores EW, Love PE. CD5 expression is developmentally regulated by T cell receptor (TCR) signals and TCR avidity. J Exp Med. 1998;188(12):2301–11.
14. Fischer L, Gökbuget N, Schwartz S, Burmeister T, Rieder H, Brüggemann M, et al. CD56 expression in T-cell acute lymphoblastic leukemia is associated with non-thymic phenotype and resistance to induction therapy but no inferior survival after risk-adapted therapy. Haematologica. 2009;94(2):224–9.
15. Fuhrmann S, Schabath R, Möricke A, Zimmermann M, Kunz JB, Kulozik AE, et al. Expression of CD56 defines a distinct subgroup in childhood T-ALL with inferior outcome. Results of the ALL-BFM 2000 trial. Br J Haematol. 2018;183(1):96–103.
16. Arber DA, Orazi A, Hasserjian R, Thiele J, Borowitz MJ, Le Beau MM, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. Blood. 2016;127(20):2391–405.
17. Patel JL, Smith LM, Anderson J, Abromowitch M, Campana D, Jacobsen J, et al. The immunophenotype of T-lymphoblastic lymphoma in children and adolescents: a Children's Oncology Group report. Br J Haematol. 2012;159(4):454–61.
18. Dworzak M, Fröschl G, Printz D, De Zen L, Gaipa G, Ratei R, et al. CD99 expression in T-lineage ALL: implications for flow cytometric detection of minimal residual disease. Leukemia. 2004;18(4):703–8.
19. Gupta S, Devidas M, Loh ML, Raetz EA, Chen S, Wang C, et al. Flow-cytometric vs.-morphologic assessment of remission in childhood acute lymphoblastic leukemia: a report from the Children's Oncology Group (COG). Leukemia. 2018;32(6):1370–9.
20. Roshal M, Fromm JR, Winter S, Dunsmore K, Wood BL. Immaturity associated antigens are lost during induction for T cell lymphoblastic leukemia: implications for minimal residual disease detection. Cytom B: Clinic Cytom. 2010;78(3):139–46.
21. Basso G, Veltroni M, Valsecchi MG, Dworzak M, Ratei R, Silvestri D, et al. Risk of relapse of childhood acute lymphoblastic leukemia is predicted by flow cytometric measurement of residual disease on day 15 bone marrow. J Clin Oncol. 2009;27(31):5168–74.
22. Bene M, Castoldi G, Knapp W, Ludwig WD, Matutes E, Orfao A, et al. Proposals for the immunological classification of acute leukemias. European Group for the Immunological Characterization of Leukemias (EGIL). Leukemia. 1995;9(10):1783–6.
23. Hoffmann J, Rother M, Kaiser U, Thrun MC, Wilhelm C, Gruen A, et al. Determination of CD43 and CD200 surface expression improves accuracy of B-cell lymphoma immunophenotyping. Cytometry B Clin Cytom. 2020;98(6):476–82.
24. Pedreira C, da Costa ES, Lecrevise Q, Grigore G, Fluxá R, Verde J, et al. From big flow cytometry datasets to smart diagnostic strategies: the EuroFlow approach. J Immunol Methods. 2019;475:112631.
25. Costa E, Pedreira CE, Barrena S, Lecrevisse Q, Flores J, Quijano S, 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. 2010;24(11):1927–33.
26. Peltier C, Visalli M, Schlich P. Comparison of canonical variate analysis and principal component analysis on 422 descriptive sensory studies. Food Qual Prefer. 2015;40:326–33.
27. Qiu P. Computational prediction of manually gated rare cells in flow cytometry data. Cytometry A. 2015;87(7):594–602.
28. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Machine Learn. 2002;46(1):389–422.
29. Vapnik V, Chervonenkis A. Theory of pattern recognition, 1974. Russian 1974.
30. Chang YW, Lin CJ. Feature ranking using linear SVM. Causation and prediction challenge PMLR, Maastricht, Netherlands undefined; 2008. p. 53–64.
31. Breiman L. Random forests. Machine Learn. 2001;45:5–32.
32. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn machine learning in python. J Machine Learn Res. 2011;12:2825–30.
33. Vapnik V. Principles of risk minimization for learning theory. Advances in neural information processing systems, Holmdel, New Jersey; 1992. p. 831–8.
34. Bishop CM. Pattern recognition and machine learning. Springer, New York, USA; 2006.
35. Bradstock K, Janossy G, Tidman N, Papageorgiou E, Prentice H, Willoughby M, et al. Immunological monitoring of residual disease in treated thymic acute lymphoblastic leukaemia. Leuk Res. 1981;5(4–5):301–9.
36. Enein AAA, Rahman HAA, Sharkawy NE, Elhamid SA, Abbas S, Abdelfaatah R, et al. Significance of CD99 expression in T-lineage acute lymphoblastic leukemia. Cancer Biomark. 2016;17(2):117–23.
37. Gaipa G, Basso G, Maglia O, Leoni V, Faini A, Cazzaniga G, et al. Drug-induced immunophenotypic modulation in childhood ALL: implications for minimal residual disease detection. Leukemia. 2005;19(1):49–56.
38. Stahnke K, Eckhoff S, Mohr A, Meyer L, Debatin KM. Apoptosis induction in peripheral leukemia cells by remission induction treatment in vivo: selective depletion and apoptosis in a CD34+ subpopulation of leukemia cells. Leukemia. 2003;17(11):2130–9.

39. Janeliūnienė M, Matuzevičienė R, Griškevičius L, Kučinskienė ZA. Monitoring of T-cell acute lymphoblastic leukemia by flow cytometry. Central Europ J Med. 2010;5(6):651–8.
40. Gujral S, Tembhare P, Badrinath Y, Subramanian P, Kumar A, Sehgal K, et al. Intracytoplasmic antigen study by flow cytometry in hematolymphoid neoplasm. Indian J Pathol Microbiol. 2009;52(2):135–44.
41. Illingworth A, Liu L, Rolf N. Current Status of TdT Testing by Flow Cytometry (ICCS Module 15). Sponsored and reviewed by the Quality and Standards Committee of the International Clinical Cytometry Society (ICCS). 2019. 12.
42. Dworzak MN, Frooeschl G, Printz D, Mann G, Poetschger U, Muuehlegger N, et al. Prognostic significance and modalities of flow cytometric minimal residual disease detection in childhood acute lymphoblastic leukemia. Blood. 2002;99(6):1952–8.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Kowarsch F, Maurer-Granofszky M, Weijler L, Wödlinger M, Reiter M, Schumich A, et al. FCM marker importance for MRD assessment in T-cell acute lymphoblastic leukemia: An AIEOP-BFM-ALL-FLOW study group report. Cytometry. 2024;105(1):24–35. https://doi.org/10.1002/cyto.a.24805