Computational flow cytometry immunophenotyping at diagnosis is unable to predict relapse in childhood B-cell Acute Lymphoblastic Leukemia

Álvaro Martínez-Rubio[1,2,*], Salvador Chulián[1,2], Ana Niño-López[1,2], Rocío Picón-González[1,2], Juan F. Rodríguez Gutiérrez[3], Eva Gálvez de la Villa[3], Teresa Caballero Velázquez[4], Águeda Molinos Quintana[4], Ana Castillo Robleda[5,6], Manuel Ramírez Orellana[5,6,7], María Victoria Martínez Sánchez[8,9], Alfredo Minguela Puras[8,9], José Luis Fuster Soler[9,10], Cristina Blázquez Goñi[2,4], Víctor M. Pérez-García[11], and María Rosa[1,2]

[1]Department of Mathematics, University of Cádiz, 11510 Puerto Real, Spain
[2]Biomedical Research and Innovation Institute of Cádiz (INiBICA), Puerta del Mar University Hospital, 11009 Cádiz, Spain
[3]Department of Paediatric Hematology and Oncology, Jerez Hospital, 11407 Jerez de la Frontera, Spain
[4]Department of Hematology, Virgen del Rocío University Hospital, Instituto de Biomedicina de Sevilla (IBIS)/CSIC, Universidad de Sevilla, 41013 Sevilla, Spain
[5]Oncohematology Unit, Niño Jesús University Children's Hospital, 28009 Madrid, Spain
[6]Foundation for Biomedical Research Niño Jesús University Children's Hospital, 28009 Madrid, Spain
[7]Health Research Institute La Princesa, 28009 Madrid, Spain
[8]Immunology Service, Clinical University Hospital Virgen de la Arrixaca, 30120 Murcia, Spain
[9]Instituto Murciano de Investigación Sanitaria (IMIB), University of Murcia, 30120 Murcia, Spain
[10]Department of Pediatric Hematology and Oncology, Clinical University Hospital Virgen de la Arrixaca, 30120 Murcia, Spain
[11]Mathematical Oncology Laboratory (MOLAB), Departament of Mathematics, Instituto de Matemática Aplicada a la Ciencia y la Ingeniería, Universidad de Castilla-La Mancha, Ciudad Real, Spain.
[*]Correspondence: alvaro.martinezrubio@uca.es

SUMMARY

B-cell Acute Lymphoblastic Leukemia is the most prevalent form of childhood cancer, with approximately 15% of patients undergoing relapse after initial treatment. Further advancements depend on novel therapies and more precise risk stratification criteria. In the context of computational flow cytometry and machine learning, this paper aims to explore the potential prognostic value of flow cytometry data at diagnosis, a relatively unexplored direction for relapse prediction in this disease. To this end, we collected a dataset of 252 patients from three hospitals and implemented a comprehensive pipeline for multicenter data integration, feature extraction, and patient classification, comparing the results with existing algorithms from the literature. The analysis revealed no significant differences in immunophenotypic patterns between relapse and non-relapse patients and suggests the need for alternative approaches to handle flow cytometry data in relapse prediction.

INTRODUCTION

B-cell progenitor Acute Lymphoblastic Leukemia (BCP-ALL) stands as the most prevalent pediatric cancer, impacting approximately 40,000 children globally each year. Recent clinical trials report survival rates exceeding 90%[1]. However, the remaining 15% experience relapse or refractory disease, with this subset facing a significantly worse prognosis[2]. The advancements in overall survival over the past

decades can be attributed to the implementation of intensive multi-agent chemotherapy regimens tailored to specific risk groups. These groups are identified through cytomorphology, molecular biology, cytogenetics, and immunology[3]. Despite these strides, the latest data suggests that improvements in overall survival will not be reached by further adjusting regimes or incorporating novel chemotherapeutic agents. Instead, hopes for finally achieving a manageable disease lie in immunotherapies for relapsed patients and refined risk stratification criteria at diagnosis[4]. New strategies are therefore necessary to identify and select patients unresponsive to standard chemotherapy and who are at a heightened risk of relapse, given the inaccuracies of current risk allocation schemes[5].

Quantitation of minimal residual disease levels early during therapy, either by flow cytometry (FC) or by clonospecific qPCR, has been consistently reported as a major prognostic factor[6,7]. Despite the fact that FC generates an extensive dataset of single-cell information, it is currently not utilized in risk stratification. In other words, the immunophenotype of the leukemic clone at diagnosis lacks prognostic value. Several factors impede the comprehensive exploitation of this type of data. One of them is the inherent challenge of managing high-dimensional data, especially in the clinical setting[8]. Another reason is the difficulty in gathering a sufficiently large retrospective cohort of patients. Indeed, the lack of prognostic value means that they are less frequently published than other clinical and pathologic information and therefore stored more casually. Lastly, despite ongoing efforts to standardize instruments and protocols[9,10], differences in adherence to standards, cytometer settings, and calibration continue to pose significant challenges for multicenter data integration[11].

The recent emergence of computational flow cytometry[12] has paved the way for automated and more thorough analyses of this type of data. This interdisciplinary field brings together flow cytometry with modern pattern recognition and statistical techniques for data processing and analysis. In combination with machine learning, these techniques can be applied for survival or relapse prediction, sample classification, or subpopulation detection[13]. Surprisingly, there is a notable lack of applications of these tools in the context of BCP-ALL, with only a few published works. For instance, a study by Reiter et al.[14] gathered a dataset of 337 bone marrow samples and employed supervised machine learning to automate minimal residual disease assessment on day +15. Good et al.[15] compiled data from 54 patients and developed a classifier that organized cells based on developmental stage and achieved a high accuracy in relapse prediction[15]. Two additional preliminary works from our group complete this landscape[16,17], one based on percentile differences of marker expression and the other on topological data analysis.

In this work, we set out to fill this gap and determine whether standard flow cytometry panels at the time of diagnosis contain prognostic information. To this end we collected the largest database of FC data of children with BCP-ALL for a computational analysis yet. We integrated tools from computational flow cytometry for data preprocessing and normalization and designed a comprehensive pipeline for

feature extraction and classification. We identified cellular subpopulations across the cohort of patients and we assessed the prognostic value of cell abundance and marker expression with a variety of metrics. We additionally contrasted and confirmed our results with other algorithms for biomarker discovery already presented in the literature. Contrary to our initial hypothesis, our results dismiss the utility of differential expression and distribution-based feature engineering for FC-based classification. We conclude the study by offering insights into the absence of discernible differences between relapse and non-relapse patients and proposing potential avenues for further exploration in this line of research.

RESULTS

Patient cohort is representative of childhood BCP-ALL population

We collected data from 252 patients from three hospitals, diagnosed between 2011 and 2022. Risk stratification criteria, treatment protocols, and outcomes are detailed in the 'Methods' section. Table S1 shows their clinicopathologic characteristics. The full cohort presents a relapse rate of 17,5%, in line with recent world-wide reports[18]. Most patients present a common immunophenotype and belong to the intermediate risk group. The frequency of genetic alterations is also within common ranges reported in European countries[19]. After preprocessing and filtering (see 'Methods' and Figure S1), 188 patients were retained for analysis. Their clinicopathologic characteristics are shown in Table 1. The only relevant differences with respect to the full cohort are a lower proportion of high-risk patients (2.7% VS 4.0%) and a higher percentage of relapse patients (20.2% VS 17.5%), still within reported ranges.

Normalization and merging allows integration of multi-center, multi-sample flow cytometry data

The cornerstone of the study is FC data at diagnosis. The joint analysis of multicenter data presents several challenges that needed to be addressed prior to the classification part of the study. Although FC panels for BCP-ALL are now standardized[10], we needed to account for differences arising from the use of different cytometers, changes in machine calibration with time and other batch effects. Furthermore, due to the maximum number of fluorochromes that can be used in a single experiment, each patient's sample is split in different tubes or aliquots that needed to be integrated if all protein markers were to be analysed together.

These sources of inter-center and inter-aliquot heterogeneity were addressed here by means of a modified min-max transformation and a quantile normalization step (Figure S2, see 'Methods'). As for the combination of several FC files into a single file, various methods have already been developed, relying mostly on nearest neighbor imputation and clustering-based imputation. In order to choose the most suitable method we used the Earth Mover Distance (EMD) to compare the distribution of a marker in the original tube versus the imputed file[20], following a recent review on

	Dataset 1 (HVR) (N=46)	Dataset 2 (HVA) (N=47)	Dataset 3 (HNJ) (N=95)	Total (N=188)
Sex - no. (%)				
Male	27 (58.7)	24 (51.1)	44 (46.3)	95 (50.5)
Female	19 (41.3)	23 (48.9)	51 (53.7)	93 (49.5)
Age at diagnosis - yr				
Median	3	5	4	4
Range	0 - 13	0 - 15	0 - 16	0 - 16
Long term status -no. (%)				
Relapse	11 (23.9)	4 (8.5)	23 (24.2)	38 (20.2)
No relapse	35 (76.1)	43 (91.5)	72 (75.8)	150 (79.8)
Immunophenotype - no. (%)				
Common	29 (63.0)	36 (76.6)	88 (92.6)	153 (81.4)
Pre-B	14 (30.4)	9 (19.1)	4 (4.2)	27 (14.4)
Pro-B	2 (4.3)	2 (4.3)	3 (3.2)	7 (3.7)
Mixed	1 (2.2)	0 (0)	0 (0)	1 (0.5)
Bone Marrow blasts at diagnosis - %				
Median	80.4	78.8	85.0	81.7
Range	10.0 - 96.3	25.6 - 95.0	30.0 - 99.0	10.0 - 99.0
Leukocytes - cell/nL				
Median	8.29	7.16	11.07	8.61
Range	1.61 - 214.21	0.54 - 336.19	0.21 - 294.0	0.21 - 336.19
Central Nervous System involvement - no. (%)				
Yes	2 (4.3)	2 (4.3)	10 (10.5)	14 (7.4)
No	44 (95.7)	45 (95.7)	85 (89.5)	174 (92.6)
Risk at diagnosis - no. (%)				
High	1 (2.2)	3 (6.4)	1 (1.1)	5 (2.7)
Intermediate	20 (43.5)	24 (51.1)	76 (80.0)	120 (63.9)
Low	25 (54.3)	20 (42.5)	18 (18.9)	63 (33.4)
Karyotype - no. (%)				
High hyperdiploidy (>50)	12 (26.2)	2 (4.2)	12 (12.6)	26 (13.8)
Hyperdiploidy (47-50)	3 (6.5)	1 (2.1)	10 (10.5)	14 (7.4)
Normal (46)	16 (34.8)	7 (14.9)	40 (42.1)	63 (33.5)
Hypodiploidy (40-45)	2 (4.3)	0 (0)	5 (5.3)	7 (3.7)
Low hypodiploidy (<40)	1 (2.2)	0 (0)	0 (0)	1 (0.6)
No metaphases	11 (24.0)	6 (12.8)	26 (27.4)	43 (22.9)
No information	1 (2.2)	31 (66.0)	2 (2.1)	34 (18.1)
Chromosomic alterations - no. (%)				
ETV6/RUNX1 t(12;21)	7 (15.2)	10 (21.3)	24 (25.2)	41 (21.8)
TCF3/PBX1 t(1;19)	1 (2.2)	1 (2.1)	4 (4.2)	6 (3.2)
MLL rearrangement	4 (8.7)	1 (2.1)	1 (1.1)	6 (3.2)
BCR/ABL1 t(9;22)	0 (0)	0 (0)	2 (2.1)	2 (1.1)
No alterations	32 (69.6)	34 (72.3)	63 (66.3)	129 (68.6)
No information	2 (4.3)	1 (2.1)	1 (1.1)	4 (2.1)

Table 1. Summary of clinicopathologic characteristics of patients retained for analysis. HVR = Virgen del Rocío Hospital, HVA = Virgen de la Arrixaca Hospital, HNJ = Niño Jesus Hopital.

the topic[21]. We compared the basic approach[22] (direct nearest neighbor imputation) with the algorithms cytoBackBone[23] (non-ambiguous nearest neighbor imputation), CYTOFmerge[24] (median of 50 nearest neighbor imputation) and cyCombine[25] (imputation by drawing from probability density estimates). Figure 1A shows the EMD of all patients for each method and each marker. Ideally, the merged marker would display the same distribution as the actual measurements. Figures 1B and 1C show the tradeoff between merging quality, number of cells per aliquots and runtime. The conclusion was that the basic approach (direct nearest neighbor imputation)

performed better and faster than the other methodologies, preserving the maximum number of cells. CyCombine and cytoBackBone performed similarly, with longer computation time associated with the removal of ambiguous cells. The conclusions were the same across hospitals (Figures S3 and S4). In the light of this result, we chose to continue the analysis with the basic approach and repeat it with the cyCombine method in order to confirm the stability of the results.

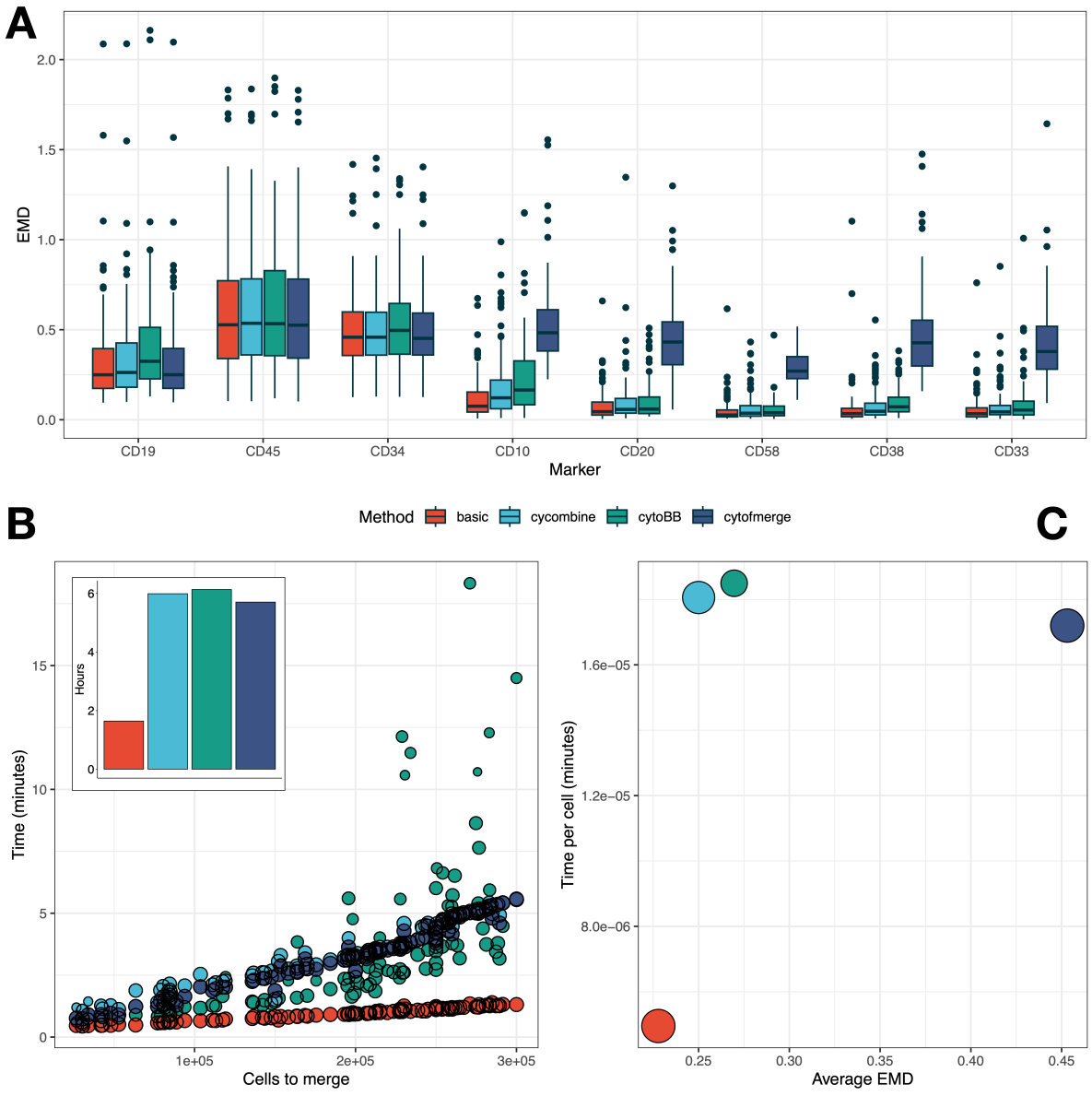

Figure 1. Comparison of file merging methods. A. Boxplots summarizing the distributions of Earth's Mover Distance (EMD) for each marker. The box includes median (horizontal line) and interquartile range (IQR). **B.** Dots represents the runtime for each patient, with the x-axis displaying the number of cells per patient. Marker size represents the ratio between the number of cells to merge and the number of cells in the resulting file. Inset displays accumulated computation time for the complete cohort. **C.** Comparison between average EMD and runtime per cell. Marker size represents the ratio between the number of cells to merge and the number of cells in the resulting file, averaging all patients. Patients analysed here belong to hospital HNJ. Similar results are obtained for hospitals HVA and HVR (Figures S3 and S4).

Clustering and dimensionality reduction techniques reveal common structure and subpopulations across patients

After preprocessing, file merging and patient selection, the final set of FC markers included B-cell markers CD19, CD10 and CD20; pan-leukocyte markers CD45 and CD38; hematopoietic stem cell marker CD34 and myeloid markers CD58 and CD66c. The next step was to visualize the structure of the bone marrow of all patients. Cell subpopulations can be obtained by means of clustering techniques, which replace the traditional manual analysis or 'gating'[26]. Here, we pooled all the files together and clustered via FlowSOM[27]. This algorithm produces a low dimensional visualization of the structure of the data in two steps. Firstly, it clusters on a higher resolution, which we manually set to 50 clusters (the influence of this number of clusters on the results will be explored later). Secondly, it obtains an optimal lower number of metaclusters by aggregating with consensus clustering. It then creates a minimum spanning tree visualization in which each cluster is represented by a node, and similar clusters are linked. This is shown in Figure 2. Marker expression per cluster is shown in Figure 2A, while Figure 2B represents the metacluster to which each cluster belongs. Each metacluster is identified with a cell subpopulation that can be manually annotated. The number of patients that contribute to each cluster is shown in Figure 2C, split in relapse (R) and non-relapse patients (NR).

To visualize the clustering information on a single-cell level we used UMAP. This dimensionality reduction technique computes a two-dimensional representation that preserves the structure of the cell subpopulations[28]. The result is shown in Figure 2D. Each cell is colored according to FlowSOM metacluster. For comparison, the marker expression of each region of the UMAP embedding is shown in Figure 2E. We note that both FlowSOM and UMAP yield a similar structure, as shown by the proximity of the different metaclusters and UMAP regions. FlowSOM obtained an optimal number of 8 metaclusters. There were two main metaclusters (1 and 2) that comprised most of the CD19+ cells and that we identified with the leukemic clone. These are immature B-cells with intermediate expression of CD45 and heterogeneous expression of CD34 and CD38. The two metaclusters were distinguished by relative expression of CD66c. We also assigned metacluster 6 to the leukemic cell population, distinguished from the other two by a negative expression of CD10. These metaclusters contained the majority of cells since the bone marrow of BCP-ALL patients at diagnosis are almost fully invaded. Metacluster 8, with a high expression of CD45 and CD20, represents healthy, mature B-cells. The remaining metaclusters represent other bone marrow cell types, including T-cells with high expression of CD45 (metacluster 7) and myeloid subpopulations (metaclusters 3, 4 and 5). While these subpopulations are seldom considered in B-cell malignancies, here we also explored them for prognostic value. With respect to the robustness and generality of these results, we note that most of the clusters contained more than 80% of the patients. When considering the metacluster scale, virtually all patients contribute to all cell subpopulations, with the exception of the minor myeloid subpopulations (metacluster 4). This confirmed that all patients adhere to the global structure described in this section.

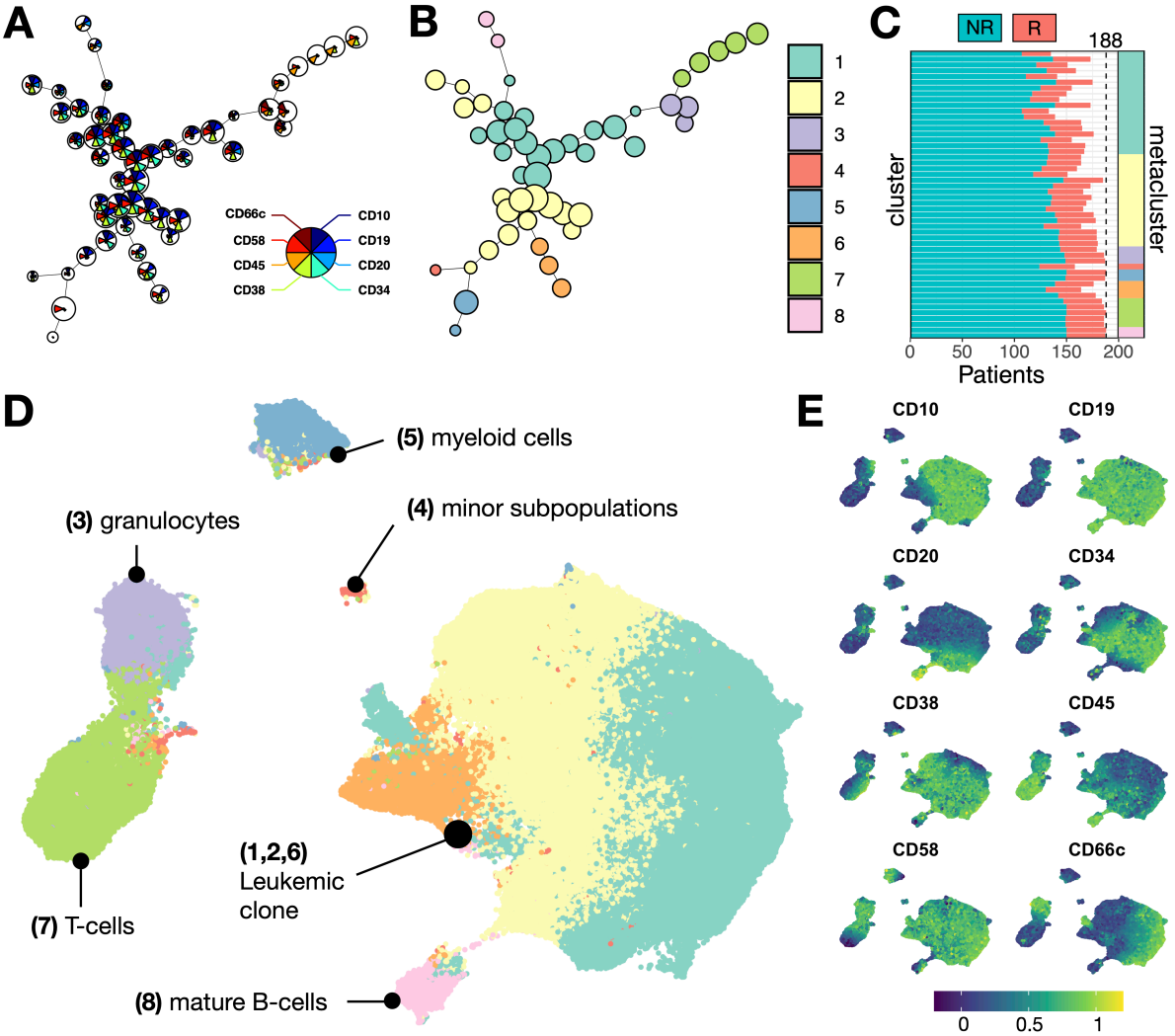

Figure 2. Clustering and visualization of flow cytometry data. A. Minimum spanning tree generated by FlowSOM. Each node represents a cluster, and similar clusters are linked. Pie plot represent the relative expression of protein markers within each cluster. **B.** Minimum spanning tree generated by FlowSOM. Color denotes the metacluster to which each cluster belongs. **C.** Number of patients per cluster colored according to outcome (R=relapse, NR=non-relapse). Clusters are sorted according to the metacluster they belong to (vertical bar on the left). Vertical dashed line represents the maximum amount of patients. **D.** Single-cell UMAP embedding. Each cell is colored according to the FlowSOM metacluster it belongs to. **E.** Single-cell UMAP embedding. Each cell is colored according to relative marker expression.

Individual patient cells per cluster are unable to predict relapse

Figure 2C shows that only a number of clusters contain a proportion of relapse patients above the baseline 20%, without any particular cluster being dominated by either relapse or non-relapse patients. To investigate the predictive power of cell abundance per cluster, however, we had to check not only the number of patients but also how many cells each patient contributed with. The idea was to test if relapse patients tended to participate more in a subset of clusters, or if instead all patients contributed equally. To do so, we calculated the percentage of cells per cluster for

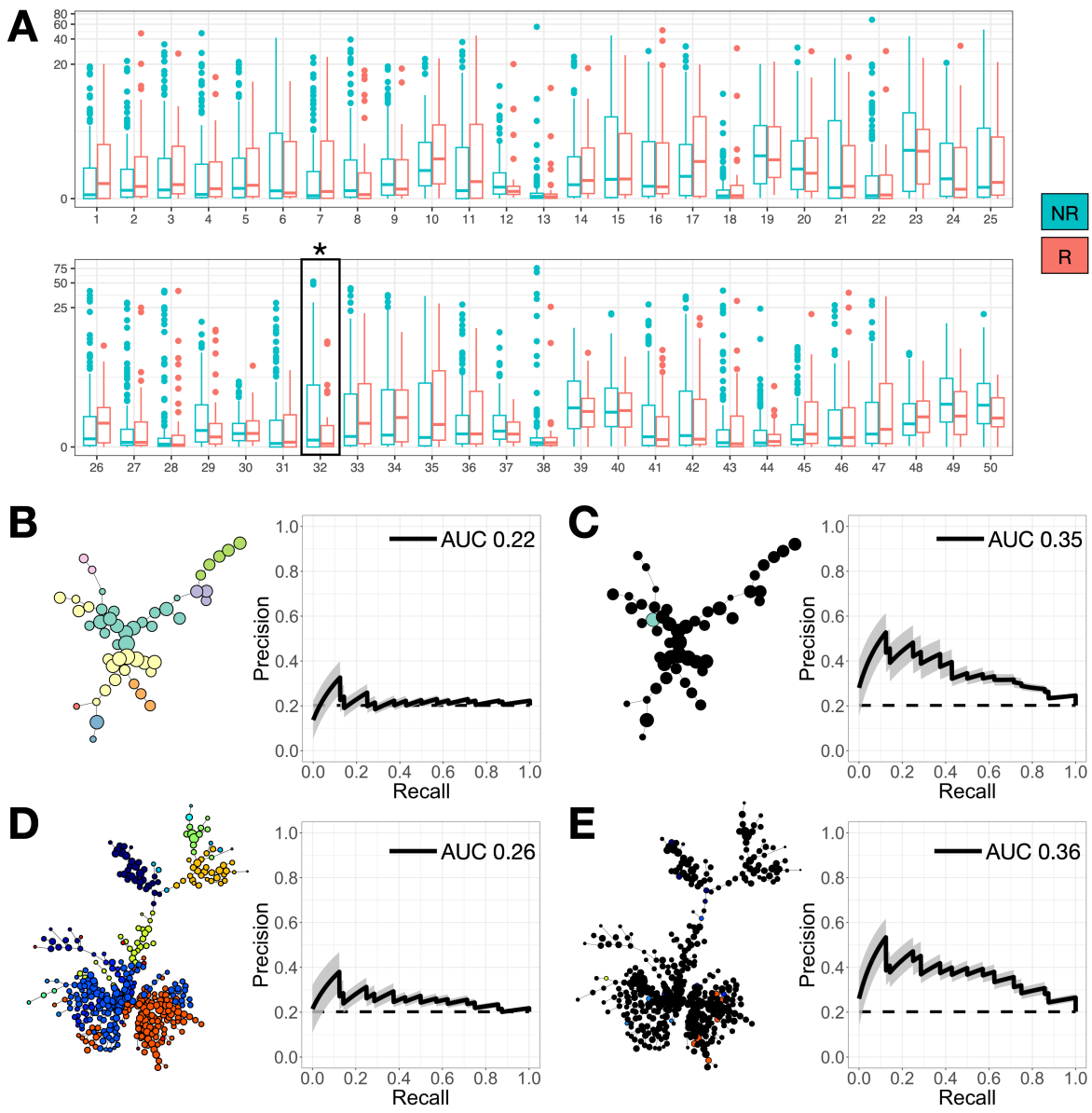

Figure 3. Results of abundance-based classification. A. Comparison of cell percentage per cluster between relapse (R) and non-relapse (NR) patients. Boxplot includes median and IQR. The scale has been transformed with an inverse hyperbolic sine for clarity. Black box and asterisk denote clusters with significant differences in cell abundance (two-sided Kolmogorov-Smirnov test). **B.** Classification results in terms of Area Under the Precision Recall Curve using information from all clusters. The shaded region represents the standard deviation of 10 repetitions of the classification routine. Horizontal dashed line represents the baseline precision. **C.** Classification results in terms of Area Under the Precision Recall Curve using information from clusters with significant differences in cells per cluster (shown in color in the minimum spanning tree). **D.** Classification results in terms of Area Under the Precision Recall Curve using information from all clusters, for a new FlowSOM clustering with 400 clusters. **E.** Classification results in terms of Area Under the Precision Recall Curve using information from clusters with significant differences in cells per cluster (shown in color in the minimum spanning tree), for the new FlowSOM clustering with 400 clusters.

every patient, and the results are shown in Figure 3A. Only one cluster (cluster 32 in metacluster 1) exhibited statistical significance ($p<0.05$), as determined by a two-sided Kolmogorov-Smirnov test.

This, however, was insufficient to conclude the lack of predictive power of the

number of cells per cluster. Indeed, although each cluster individually did not present clear differences, non-linear interactions between all clusters could create a region in which relapse patients are more clearly distinguished. We tested this by building a classifier for relapse prediction that uses cells per cluster as input. We implemented a nested cross-validation scheme and included four supervised machine learning algorithms: Naive Bayes, Random Forest, K-Nearest-Neighbors and linear Support Vector Machine. For robustness, we repeated the classification 10 times. More details about the classification routine can be consulted in the 'Methods' section. The average Precision-Recall curve obtained is shown in Figure 3B. We used the Area Under the Precision-Recall Curve (AUCPR) to summarize the result. This is equivalent to the average precision of the classifier and can be interpreted as the probability that a predicted relapse is a true relapse. Its value was close to the baseline precision, which is the proportion of relapse patients in our dataset (0.202, Table 1). This means that the features used for classification had no prognostic value. We repeated the classification but using only the cluster in which significant differences were found (Figure 3C). We obtained a higher precision compared to using all information, although still close to the baseline classifier. To explore the possibility of finding more relevant prognostic information, we repeated the clustering with 400 FlowSOM clusters. When using all clusters, the classification results were almost identical to the 50 cluster case (Figure 3D). We finally repeated the classification using only the significant clusters (Figure 3E). This scenario reported the highest performance, but still far from a significant enhancement compared to the baseline classifier. We finally assessed the reliability of these results by performing stability and overfitting checks (Figure S5).

Relative marker expression is similar between relapse and non-relapse patients across cell subpopulations

Following the assessment of the prognostic significance of cell abundance, we turned to marker expression within each cell subpopulation. The distributions depicted in Figure 4A portray the aggregated marker expression for relapse and non-relapse patients within each metacluster. The patient-specific distributions that contribute to these aggregated distributions are shown in Figure S6. Most markers across the majority of metaclusters did not exhibit noteworthy disparities between the relapse and non-relapse groups. The exceptions are metacluster 4, which showed underexression of CD10 and overexpression of CD20 in relapse patients, as well as more general differences in markers CD38, CD45 and CD58; and metacluster 6, which displayed differences in CD10 and CD34. These metaclusters are associated with minor subpopulations and part of the leukemic clone respectively. It remained to be seen that the significance of these disparities were reproducible at an individual-patient level, rather than being confined to the population level. Following the rationale of the previous section, we aimed to test whether individual patients' marker expression could predict relapse.

To address this, we summarized the marker expression distributions of each

patient within every metacluster using the median and the first four statistical moments: Mean, standard deviation, skewness and kurtosis (Statistical moments of order 1 to 4). This procedure produced five distinct datasets, each corresponding to a specific metric, where each row encapsulated a patient's marker distribution summary. Additionally, we constructed a combined dataset with all features to explore whether a combination of metrics would yield more informative results. Furthermore, we created a final dataset comprising exclusively those features displaying statistically significant differences, as done in the previous section (Kolmogorov Smirnov test, α = 0.05). We used the classification routine to check the

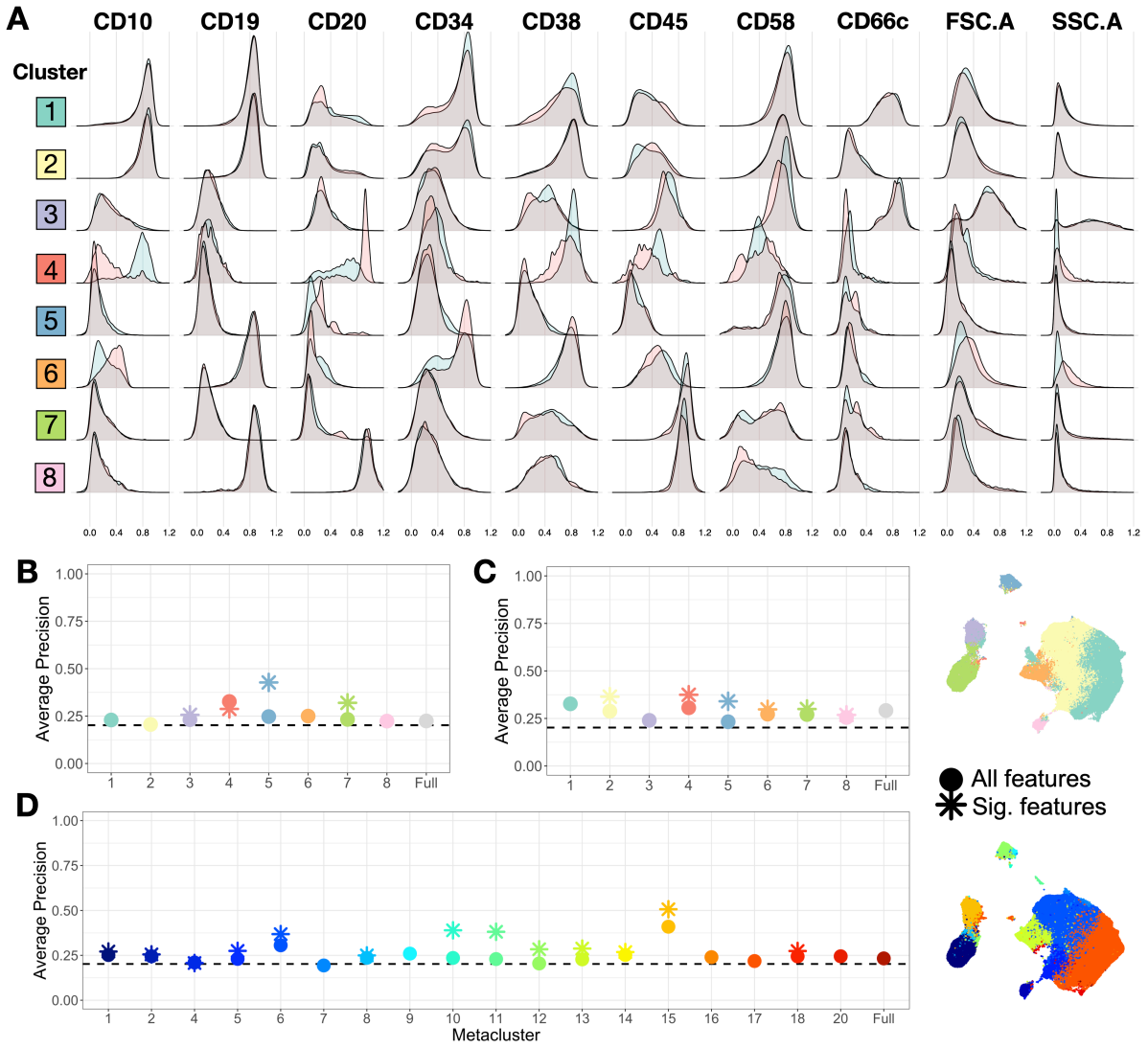

Figure 4. Expression-based classification. A. Aggregated marker expression of relapse and non-relapse patients per metacluster. **B.** Classification results in terms of AUCPR. Black dashed line represents baseline precision. Color denotes metacluster. Circles represents the average precision obtained when using all the distribution metrics together to train the classifier. Asterisk represents the same average precision when using only those features with significant differences according to a Kolmogorov-Smirnov statistical test (α = 0.05). **C.** Classification results in terms of AUCPR for a subset of patients (N=158) with an increased number of markers. **D.** Classification results in terms of AUCPR for a higher resolution clustering (400 FlowSOM clusters, 20 metaclusters).

predictive power of each dataset. A summary of the workflow followed in this section is shown in Figure S7. The results are shown in Figure 4B. For each metacluster, we show the average precision (equivalent to AUCPR) obtained by using all the information versus only the features with significant differences, as done in the previous section with cell abundance. We also show the same information for the full cohort, without segregating by metaclusters. The results for individual metrics (median, mean, standard deviation, skewness and kurtosis) are shown in Figure S8. The conclusion is straightforward: the information contained in marker expression distribution lacks predictive capacity, given that the majority of AUCPRs marginally exceeded the baseline precision. The best result was a precision of 42% when using only the significant features in metacluster 5, which correspond to myeloid cells. Notably, the leukemic clone metaclusters (1, 2 and 6) contained no prognostic information.

We explored the possibility that the observed outcome could stem from an insufficiency of detail in the information under examination for each patient. To address this, we repeated the analysis incorporating two alterations. Firstly, we expanded the set of markers selected for analysis. However, this came at a cost: the patient count diminished from 188 to 158. Employing the same clusters identified in the original analysis, we searched for differences within the new markers (IgM, cyTDT, cyMPO, cyCD3, CD13, CD22, CD3, CD33), but the improvement in the predictive power of the routine was negligible: we obtained an increment of 5% of precision on average (Figure 4C). In subsequent investigation, we studied whether the challenge laid not in marker quantity but rather in the size of the clusters. Using the initial set of 8 markers, we reconsidered the 400-cluster outcome from the previous section, which resulted in 20 metaclusters. Two of the metaclusters (3 and 19) lacked enough patients to reliably estimate performance. For the remaining metaclusters, the improvement was also unremarkable, especially in clusters associated with the leukemic clone (red and blue colors) (Figure 4D). The best result was an average precision of 50%, slightly superior than the best precision in the default analysis but in a different subpopulation. The reliability of this set of results was also assessed as in the previous section by calculating the stability of the classifier (Figure S8).

To conclude, we performed two additional analyses. We first considered whether the preprocessing of the data could be responsible for the lack of predictive information. To explore this, we replicated the analysis using the cyCombine algorithm for file matching (see ‘Methods’), and our findings concurred with the conclusions detailed earlier (Figure S9). Lastly, we considered only those patients which were initially diagnosed as intermediate risk, to check if the more intensive treatment received by high risk patients could bias the results. This resulted in a reduced cohort of 119 patients. The results were also similar to the above (Figure S10). Hence, irrespective of treatment received, preprocessing technique employed, number of markers considered, cluster size and distribution metric, marker expression of FC data at diagnosis failed to predict relapse.

Biomarker discovery algorithms from the literature support the findings of the main analysis

We contrasted our findings with other algorithms from the literature designed for biomarker discovery and outcome prediction. A description of their functionality and implementation can be found in the 'Methods' section. The results for each of them are shown in Figure 5. The first example is Cydar[29], which is designed for differential abundance discovery. The clusters (hyperspheres in Cydar terminology) with a sufficient number of cells are projected onto the UMAP embedding employed in the previous sections (Figure 2). Those hyperspheres with significant differences in abundance (according to a lasso-regularized logistic regression) are plotted with wider radius and colored according to the fold change in abundance between both group of patients (Figure 5A). To check the predictive power of such hyperspheres, we extracted the number of cells per patient and hypersphere and ran the classification routine previously described, with results similar to the best models in the previous section (Figure 5B). The difference here is that due to the lower size of the clusters (hyperspheres), there are less patients per cluster (Figure 5C), which makes the results less generalizable. The second example is Citrus[30]. The results for both abundance and median expression (Figure 5D) indicate the lack of predictive information, regardless of regularization threshold. In both cases the null classifier (no features, leftmost regularization threshold) was the best classifier, with an error of 20.2%. This number is the proportion of relapse patients in our dataset, which means the algorithm was classifying all patients as non-relapse. Further, the False Discovery Rate shows all the characteristics of a classifier unable to discriminate[31]. The Citrus algorithm also reports the clusters with potential predictive capacity (Figure S11A-B). The third example is cellCNN[32], which uses a convolutional neural network. We complemented it with a nested loop that allowed us to provide two conclusions: First, the lack of a inner validation routine makes the algorithm more prone to overfitting, as we see in the comparison between the accuracies of the inner and outer loops (Figure 5E). Second, the performance in terms of AUCPR did not improve previous tests (Figure 5F). This algorithm also reports the characteristics of the most significant cells, which are included in Figure S11C-D. Finally, we tested Diffcyt[26] on the metaclusters that were already obtained by FlowSOM (Figure 2). This algorithm showed no significant differences in cell abundance per metacluster (Figure 5G) but it did detect significant differences in expression (Figure 5H). On closer inspection, we noticed that those significant features were the ones that displayed differences in aggregated marker expression (Figure 4A). We already showed how this sum of distributions does not necessarily entail that individual patients follow the same trend (see Figure S7) and that a classifier could still be unable to properly predict relapse, as shown in Figures 4B-C.

The conclusion of this section is that other algorithms that aim for the same goal as this study and follow a comparable methodology are also unable to detect differences between relapse and non-relapse patients. This applies to analyses centered on both cell abundance and marker expression.

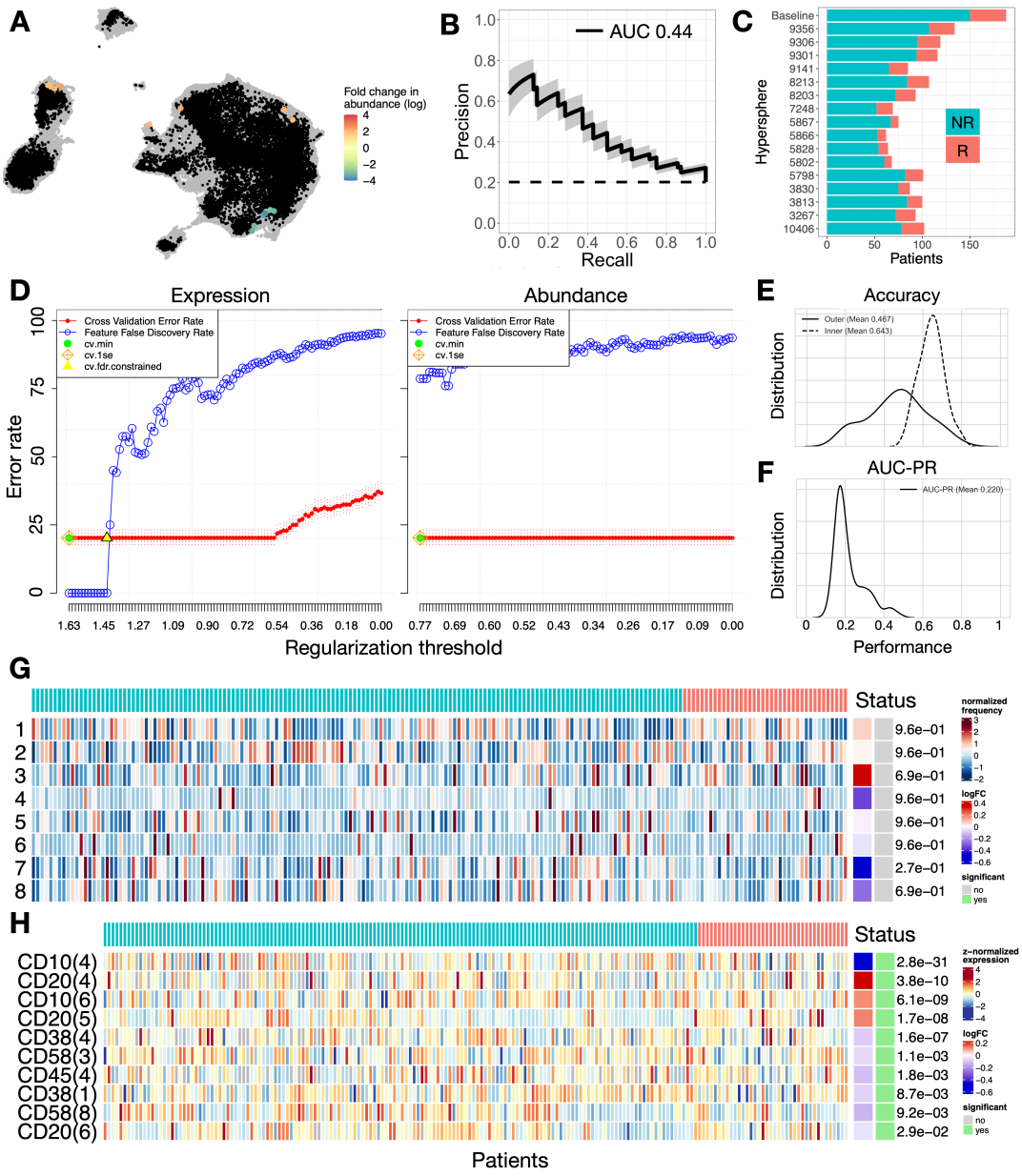

Figure 5. Results from other biomarker discovery algorithms. **A.** Cydar hyperspheres (black) projected on UMAP embedding from Figure 2C (gray). Significant hyperspheres are colored according to fold change in abundance. **B.** Classification results from the cell abundance of the significant hyperspheres. Interpretation is as in Figure 3B-E. **C.** Number of patients in significant cydar hyperspheres split in relapse (R) and non-relapse (NR). Top row displays the reference of 188 patients (38 relapses). **D.** Citrus results for median expression (left) and abundance (right). Represented are cross-validation error rate (Red) and false discovery rate (blue). Green dot represents the error rate of the best model according to the minimum cross-validation error rate. Orange rhomboid represents the error rate of the best model according to the one standard deviation criterion. Yellow triangle represents the best model according to the lowest compatible false discovery rate. **E.** Comparison between the accuracy in the outer and inner loops of the CellCNN algorithm. **F.** AUCPR curve in the outer loop of the CellCNN algorithm. **G.** Diffcyt differential abundance test. Each row contains the individual patient cell percentage in a metacluster (1 to 8). The algorithm includes the fold change between status (relapse R in red vs non-relapse NR in blue) and the statistical significance of the results (gray vs green) **H.** Diffcyt differential expression test. Row annotation includes the marker and the metacluster in which significant differences were found.

DISCUSSION

Approximately 15% of children diagnosed with BCP-ALL will suffer relapse or refractory disease, and the prognosis for this subgroup is significantly worsened. Despite advancements in therapy through refined chemotherapy regimens, the potential for further therapeutic success appears rooted in alternative treatments or more precise risk assessment upon diagnosis. This underscores the importance of enhancing our capability to anticipate disease progression at the individual patient level. In this investigation, we have compiled an extensive FC database for childhood BCP-ALL. A total of 252 patients from three hospitals participated in the study, with 188 patients advancing to the computational analysis phase. The objective of the study was to examine whether patients experiencing relapse exhibit distinctive patterns within their FC data at diagnosis. In other words, the goal was to test if FC data at diagnosis has prognostic value with regards to long-term response.

To fulfill this objective, we preprocessed and normalized the data and we carried out a file merging step in order to integrate the different aliquots of each patient into a single file, after comparing the performance of different imputation methods. We concluded that direct nearest neighbor imputation was the most distribution-preserving algorithm. We hypothesize, however, that this may be only applicable to the kind of data considered in this study (in terms of markers included and type of distributions). This is clearer after noting the differences with the conclusions reached in a recent review on imputation methods[21]. Without being exhaustive, the presence of a dense and homogeneous clone could make the data more suitable for merging algorithms of one kind, whereas more balanced or heterogeneous bone marrow distributions would benefit from other algorithms. The preprocessing step, the normalization, and the previous clustering step can also impact the values and range of the metrics employed to measure distribution differences. We therefore recommend repeating this assessment when dealing with a different disease or high-dimensional data of other kind.

The selected patients were then pooled together and clustered with FlowSOM. We visually examined the data structure through its UMAP embedding, revealing minimal disparities between relapse and non-relapse patients. We extracted cell abundance per patient at the cluster level and summarized marker expression at the metacluster level by means of the first four statistical moments of the expression distribution: Mean, standard deviation, skewness, and kurtosis. We also computed the median of the distribution, a classical FC metric. All these features were input into a nested cross-validation scheme, which aimed to identify the optimal classifier for each dataset and assess its performance on unseen data. The performance of such classifiers served as an indicator of the prognostic value of the dataset.

The outcome of the primary analysis directly contradicts the initial hypothesis: FC data obtained at diagnosis does not appear to harbor information relevant to the prediction of relapse. Cell abundance per cluster is unable to predict relapse, even when increasing the number of clusters and when using only the ones with significant

differences between relapse and non-relapse groups. Likewise, no distribution metric is able to significantly improve the baseline precision. Considering all metrics together in a single dataset or retaining only the ones with significant differences also failed to improve outcomes. We further increased the number of clusters and the number of markers, the latter with a reduction in the number of patients from 188 to 158, and we also repeated the analysis considering a different file merging algorithm, to test the possibility that the preprocessing routine masked differences in abundance or expression. Finally, we restricted the analysis to intermediate risk patients, to account for the possible confounding effect of the more intensive treatment received by high risk patients. The conclusion remained unaltered across all these studies. The most precise classifier was found when increasing the number of clusters and using only the metrics with significant differences between groups. This classifier achieved an average precision of 0.507. This indicates that, within this particular classifier, the likelihood of a predicted relapse corresponding to an actual relapse stands at 50.7%.

The pipeline followed in this study was designed to encompass and extend previously published algorithms by offering a more comprehensive characterization of marker expression distributions and employing non-linear classifiers with a more rigorous resampling scheme. Despite these advancements, we verified the outcomes against other open-source algorithms. We specifically assessed Cydar[29], Citrus[30], Diffcyt[26], and CellCNN[32], which are among the most frequently referenced algorithms for discovery analysis in FC. Cydar, Citrus, and Diffcyt incorporate tests for differential abundance. Cydar identified several clusters exhibiting significant differences in abundance, with a performance akin to the classifiers obtained in the primary analysis. As a drawback, those clusters only contained a subset of the full cohort of patients. Citrus and Diffcyt failed to identify differences bearing prognostic value. These two algorithms additionally include tests for differential expression. Citrus identified three features, but the classifier's performance proved inferior to the null model. In the case of Diffcyt, the identified features held significance at a population level but struggled to consistently discern individual patients. Finally, the outcomes from cellCNN mirrored those of the other classifiers, with performance marginally surpassing the baseline classifier. The aforementioned findings further underscore the established conclusion that the metrics used to characterize the distributions of surface markers fail to differentiate between patients who experience relapse and those who do not.

The initial hypothesis of this study rested on the premise that the leukemic clone in relapsing patients differs from that of successfully treated individuals, and that such distinctions manifest in the immunophenotype and could then be captured through FC measurements. The negative outcome we have obtained in this study offers room for diverse interpretations. It is possible that the immunophenotype of relapsing patients does not exhibit distinctive characteristics. While genetic differences are known to play a fundamental role in the origin and potentially the relapse of leukemia[33,34], these differences may not necessarily translate to variations in marker expression distributions. Rather, they may only be found through genomics, transcriptomics or

metabolomics. In this line, recent research has demonstrated the feasibility of predicting relapse in infants with MLL-rearranged ALL by single-cell transcriptomics[35]. It remains essential to conduct further investigations to ascertain the predictive potential of a comprehensive panel of mutations for the broader population. Alternatively, immunophenotypic disparities might emerge post-therapy. Such a scenario could be attributed to chemotherapy-induced bottleneck selection, which has been shown to impact the phenotype more significantly than genotype[36]. This could be probed by revisiting this study with FC data from a later time point, although this approach would deviate from the initial goal of refining risk stratification at diagnosis.

With respect to the conditions of this study, it is also feasible that immunophenotypic distinctions exist but are only discernible within small cell subpopulations. Such differences might elude detection even with high-resolution clustering if the number of cells per patient is not increased. This hypothesis could be explored by imposing stricter limitations on the number of cells per patient, although this would inevitably reduce the total number of patients in the study. Another potential consideration is that immunophenotypic disparities manifest in markers beyond the ones routinely assessed in clinical practice. Evaluating this notion would require prospective studies. Finally, it can be the case that immunophenotypic disparities exist but are obscured by the extensive preprocessing and normalization required to integrate data from multiple centers. No immediate alternative exists until the clinical adoption of next-generation cytometers that can measure a larger number of markers simultaneously and are more amenable to standardization.

Despite the scope and scale of this study, as well as the evidence gathered in support of the negative conclusion, there are still alternative ways of exploring the potential prognostic value of FC, a line of research that is still relatively unexplored for this particular disease. Indeed, a number of works employ machine learning techniques to answer questions relative to BCP-ALL, but applications for relapse prediction from FC data at diagnosis are still uncommon. For instance, Pan et al.[37] utilized clinical data from a cohort of 336 patients to predict relapse. However, this study lacked FC data and incorporated response variables (such as MRD at days 15 and 33), thereby limiting its applicability to the diagnosis phase. A similar predictive framework based on clinical features was presented by Mahmood et al.[38]. Moving closer to the objectives of the present study, Good et al.[15] gathered mass cytometry data at diagnosis to achieve a relapse prediction AUC of 0.85 using an elastic net model. However, their database only encompassed 54 patients, and the validation was confined to a single train-validation split, thereby hampering direct comparability with our results. Similar constraints apply to an earlier work by our own group that included 56 patients to identify differences in expression[16]. Finally, we recently published a framework that uses topological data analysis for feature extraction and includes a classifier that reached high accuracy and AUC with an increased number of patients (N = 96)[17]. This study meets the criterion of moving beyond the conventional feature engineering in FC and the preliminary results encourage the search for differences in immunophenotype of relapsing patients by means of more

complex methods.

To sum up, we have performed a machine learning-based relapse classification study involving 252 patients diagnosed with childhood BCP-ALL. A detailed characterization of immunophenotype and different cluster resolutions have been unable to distinguish relapse from non-relapse patients, and other algorithms from the literature exhibited similar outcomes. We conclude that different characterizations of FC data are required to uncover its potential prognostic value, pending the availability of high-dimensional omics data at diagnosis and more advanced cytometers that circumvent some of the challenges found throughout our study.

METHODS

Study population

252 patients and three different spanish hospitals participated in this study. We collected data from 116 patients from Hospital Niño Jesús, Madrid (HNJ), diagnosed between January 2013 and January 2022; 80 patients from Hospital Virgen de la Arrixaca, Murcia (HVA), diagnosed between May 2011 and July 2022; and 56 patients from Hospital Virgen del Rocío, Sevilla (HVR), diagnosed between January 2012 and July 2021. 207 patients had long-term remission and 44 patients relapsed. All patients are in the age range 0-19. We dropped those which continued treatment at another institution or that had not reached 1 year of follow up, with 211 patients finally proceeding to the main analysis (Figure S1). The data collected included FC files from bone marrow samples at diagnosis and additional clinical information: Age, sex, phenotype, risk group, CNS involvement, absolute lymphocyte count (ALC), immunophenotype and genetic information (karyotype and chromosomal translocations). Informed consent was obtained from the parents or legal guardians according to the Helsinki Declaration.

Treatment

Treatment was administered according to the Spanish National protocols SEHOP-PETHEMA 2013 and INTERFANT-06 in patients under 1 year old. Older patients from HVR and HVA followed the previous consecutive versions of this protocol (LAL/SEHOP 01 for low risk patients, LAL/SEHOP 96 for intermediate risk patients and LAL/SHOP 05 for high risk patients). These protocols are based on the Berlin–Frankfurt–Munster (BFM) backbone and consists of a four-drug induction phase (IA), followed by induction IB, consolidation, reinduction, and maintenance. High risk patients receive three specific high-risk blocks, three reinduction cycles, and maintenance. The total duration of therapy is 2 years.

Risk stratification

Risk stratification criteria is based on age, lymphocyte count at diagnosis, extramedullary infiltration, cytogenetics and early response to treatment.

SEHOP-PETHEMA 2013 assigns a low risk to patients who meet the following criteria: Age between 1 and 10 years, ALC less than $20 \cdot 10^9$ cells/liter at diagnosis, absence of CNS or testicular infiltration, high hyperdiploidy or presence of t(12;21), absence of t(1;19), no MLL rearrangement, good early response and good response to prednisone. High risk patients verify at least one of the following: presence of t(4;11), hypodiploidy, BCR-ABL rearrangement or poor early and prednisone response. Patients who do not meet either criteria are assigned to intermediate risk[39].

Patient outcome

Patients are assigned to either relapse or non-relapse group. Bone marrow relapse is diagnosed with the same criteria as the initial diagnosis: presence of >25% of leukemic blasts in bone marrow. Extramedullary relapses require a biopsy of the tissue or a sample of cerebrospinal fluid for confirmation. For a patient to be included in the non-relapse group we require at least one year of disease-free survival after treatment.

Flow cytometry data

All data is retrospective. Bone marrow samples have been handled following standard clinical procedures (there is no specific design for this study). Monoclonal fluorochrome-conjugated antibody combinations employed at each hospital are shown in Tables S3-S5. Some patients presented variations from this standard (marker changes, additions or omissions).

Preprocessing of flow cytometry data

Preprocessing encompassed a manual and a computational step. The manual step consisted in checking each aliquot for acquisition errors and removing doublets and debris (Fig. S2A). At this step we required that all aliquots contain CD19 and CD45 markers. For this reason, certain patients (mostly those diagnosed at earlier dates) were excluded from the study (1 from HVA and 7 from HVR). Aliquots with too little cells of with strong batch effects were also removed.

The compensated files were subsequently exported to undergo the computational preprocessing step[40]. This preprocessing involved transforming data with the standard Logicle transform, removing margin events (this is done more efficiently here than manually) and renaming the channels to uniformize marker names across patients. Finally, each marker was normalized to the [0,1] interval by means of a modified max-min transformation: Instead of taking the maximum and minimum values, we took the 99^{th} and 1^{st} quantile respectively, making the normalization more robust to outliers. This transformation implies that we are comparing relative expression of a marker instead of the absolute expression.

Finally, we had to consider the issue of backbone markers displaying inter-aliquot differences. Some causes of this variability are staining problems, acquisition errors and other batch effects. To account for this source of heterogeneity we first sampled 10000 cells from each tube and then performed quantile normalization, a technique

already used in RNA-seq data to make distributions more similar. Instead of normalizing the whole distribution we followed the approach in the cytoNorm algorithm[41]: we performed flowSOM clustering with 5 clusters and then normalized on a per cluster basis (Figure S2B).

File merging

File merging (also file matching, panel merging or imputation) refers to the process of combining all the information from a FC experiment into a single file. The issue arises from the fact that flow cytometers can measure a limited number of colors, i.e. the expression of a limited number of protein cell markers. To obtain information for more markers, the sample is divided in several tubes or aliquots and each tube measures a different set of proteins, while maintaining a subset of them constant (backbone markers). This is enough for manual inspection of the sample but for data analysis the combined file allows for a much deeper analysis. Figure S2C illustrates the starting point and endpoint of this part of the analysis.

Several methods have already been developed for this purpose. Most of them rely on nearest neighbor imputation: Backbone markers are used to find the closest neighbors (cells with the highest surface protein similarity), and the missing information is copied from the respective neighbor. This was first published by Pedreira et. al.[22]. Later works use slightly modified versions that aim to correct artifacts and biases: cytoBackBone[23] includes the concept of acceptable and non-ambiguous nearest neighbors (data is only imputed if a cell's closest neighbor is also the other cell's closest neighbor) and CYTOFmerge[24] used median expression from the closest 50 neighbors instead of the single closest one. A more recent method (cyCombine)[25] follows a different methodology: It finds clusters in the space of backbone markers and then approximates the distribution of the remaining markers using kernel density estimation. The missing information is then imputed using probability draws. This is similar to other approach by Lee et. al.[42], which requires domain knowledge but demonstrated that pre-matching clustering enhances performance and reduces the risk of spurious cell populations appearing in the data. These previous steps improve quality of merging in terms of preserving the original distribution at the expense of removing cells that are too exclusive of one file and that would otherwise impute noise.

In light of these advances, the question arises as to which one is the most suitable method for conducting downstream analysis on a patient dataset. A recent comprehensive review delved into this question[21], using an array of metrics to compare the performance of the different algorithms. They concluded that there is not a clear winner and caution needs to be taken when performing downstream analysis with imputed data. A similar approach was carried out by Perdersen et. al.[25] when demonstrating the cyCombine functionality. The Earth's Mover Distance (EMD) was employed to compare the distribution of a marker in the original tube versus the merged file. This distance, also known as Wasserstein distance, measures the minimum cost required to transform one distribution into another. In the context of

flow cytometry, this cost is associated with moving cells from one marker expression state to another. Lower EMD values indicate a closer match between the original and imputed distributions, suggesting a more accurate imputation process. Its suitability for comparing marker expression distributions in the context of flow cytometry was recently demonstrated[20].

Here, we preprocessed patients from each hospital as described above and imputed the missing values according to the four methods mentioned in the main text: Direct nearest-neighbor imputation (basic), CYTOFmerge, cytoBackBone and cyCombine. For the backbone markers (CD19, CD34 and CD45), since these markers are present in all aliquots, we measured the difference of expression in each aliquot with the expression in the merged file and computed the average to get an upper bound for acceptable inter-aliquot differences. For the remaining markers we computed the EMD between the expression in the merged file and the expression in the specific aliquot in which they were present. Other metrics that were used to assess the performance of the different algorithms were the ratio between cells to merge and final number of cells (some of the algorithms discard cells) and the computation time.

Patient selection

The selection of patients and markers that proceeded to the final study was conducted post-normalization and merging. This is due to the fact that certain aliquots were excluded during these steps, resulting in a variation in the markers available for each patient compared to the preprocessing phase. We first reduced each patient (i.e. each merged file) to 10000 cells, removing patients that did not reach this amount. There exists a tradeoff between the number of markers analyzed and the number of patients. For the main analysis, 10 markers were retained (FSC.A, SSC.A, CD19, CD10, CD20, CD34, CD66, CD58, CD45, CD38) for a total of 188 patients (95 from HNJ, 46 from HVR and 47 from HVA).

Flow cytometry visualization

FlowSOM was run with parameters x = 5, y = 10 and maxK = 20. We consistently obtained an optimal number of 8 metaclusters. For UMAP, we subset 1000 cells from each patient and pool the subset files to obtain the embedding of the bone marrow of all patients. After a visual exploration, we selected UMAP hyperparameters min_dist = 0.1, n_neighbors = 50 and the rest with default values (Figure S12).

Feature extraction

The most common features for analyzing flow and mass cytometry data are abundance (relative or absolute) and expression, measured as the median intensity of a marker (MFI), in general or on a per-cluster basis. This has been the case in most of the studies and methods used for biomarker discovery in FC data applied to leukemia (Table S5). However, a single number might not be enough to characterize

the full marker distribution and thus to discover differences in expression, intensity and immunophenotype. Here, we computed for each cluster not only the abundance and median expression but also the first four moments of the distribution (mean, standard deviation, skewedness, and kurtosis). We created a dataset for each feature and a dataset with all features together, in order to find which characterization is best for detecting differences in expression and to see if the combination of all enhances the predictive capacity. We finally created a dataset with only those features that present significant differences between relapse and non-relapse patients according to a two-tailed Kolmogorov-Smirnov test.

Classification

Most of the published methods for analyzing FC data (Table S5) use linear models to perform moderated tests in order to find significant differences in expression (median intensity). The exception are neural network based algorithms, which do not explicitly perform feature selection but include the FC file as input for the algorithm. The differential expression methodology is standard in transcriptomics analysis, when looking for genes that are overexpressed under given conditions[43]. For the problem and the hypothesis of this study, finding a significantly over- or under-expressed marker might not be enough to distinguish a relapse from a non-relapse patient. In other words, while we would be able to make a statement of the kind “relapse patients on average have a higher expression of marker X”, we would not be able to say whether a new patient belongs in the relapse or non-relapse group. Further, these analyses consider markers individually, but it could be the case that, while there might not be significant differences in MFI of a marker, we could find a region in the space of MFIs that separates both groups of patients.

Without any previous knowledge about the characteristics of this region and given that it can be quite different depending on which metric we are considering, we could not say a priori which classification model was best for this task, nor which hyperparameters of such model were optimal. For this reason, the classification routine had to include some form of internal validation to make this decision based on the data. We did this by means of nested cross-validation[44,45]. This approach consists of two cross-validation loops, an outer loop and an inner loop. The inner loop is used to find the best model and its hyperparameters, and the outer loop is used to get an estimate of performance in unseen data. For the inner loop we performed 9-fold cross-validation repeated 20 times to get a more robust estimate, and for the outer loop we performed 5-fold cross-validation, repeated 10 times. This resampling scheme implied that each inner fold contained 16 patients on average, with 2 of them belonging to the relapse group.

We chose 4 models that are widely used and ensure that different types of boundaries are explored: K-Nearest Neighbors, Naïve Bayes Classifier, Random Forest, and Linear Support Vector Machine. Each time we trained a model we use random grid search to select the optimal hyperparameters (Table S6). The best model was selected based on the one standard deviation rule using the AUCPR

curve, which is more suitable for problems with unbalanced data[46]. Hyperparameter estimation and model selection were thus performed together[47].

For every dataset, the nested cross-validation routine produces 50 performance estimates (AUC-PR) and identifies 50 'best models' (obtained from the 10 repetitions of 5-fold cross-validation in the outer loop). We summarized the 50 AUC-PR values by calculating their average, and the 50 best models by using a measure of heterogeneity as a surrogate of the stability of the routine. This stability measure is assigned a value of 1 if the same model is consistently selected in all outer folds, and 0 if the four models are equally frequent. It is important to note that while this measure can indicate instability or unsuccessful optimization, it is possible for two models to perform nearly equally well in identifying the best boundary, making them equally suitable for the task at hand. Thus, it's essential to consider the degree by which the top model has been selected and its associated level of performance.

To sum up, for each dataset we had a measure of the predictive information it contains (average AUCPR) and a proxy of the reliability of this measure (stability index). These two metrics were employed in conjunction to assess the predictive information across different metrics and metaclusters. Figure S7 provides an overview of the feature extraction and classification steps.

Comparison with other algorithms

We already mentioned the existence of other algorithms and studies that aim to predict a clinical outcome from flow cytometry data (table S5). The way they are designed follows a similar pattern: All of them begin from a set of flow cytometry files (one per patient) and cells are clustered with a different algorithm depending on the method. Each cluster is summarized by means of the abundance and the median fluorescence intensity of a marker, and these are in turn used for classification. Generalized linear models are the usual choice, as many algorithms are inspired by RNA or DNA microarray data analysis. The exception to this two-step process are neural networks based algorithms, since feature extraction is performed in the inner layers of the network. The pipeline that we followed here aimed to generalize this 'classical' approach by going beyond the typical characterization of a marker distribution (MFI) and by including a broader and more thorough classification routine. To validate the conclusions of this study, we selected four of the most cited algorithms and compared the results. Below we summarize the characteristics and functionality of the selected algorithms.

- Cydar[29] identifies differentially abundant cell populations between groups. It was originally proposed for mass cytometry data but can be extended to any multidimensional dataset. It clusters cells into hyperspheres, extracts cell abundance and tests for significant differences by means of a negative binomial generalized linear model, controlling for the spatial false discovery rate. In this study we subsampled 1000 cells from each patient, clustered with scaling factor 0.2, removed hyperspheres with average counts below 5 and applied the QL framework to test for significant differences. After correcting for multiple testing

(spatial FDR<0.05), relevant hyperspheres and the respective fold changes in abundance were visualized on the UMAP embedding of the dataset.

- Citrus[30] identifies cell subpopulations associated with a clinical or experimental outcome. It clusters cells in a hierarchical manner, extracts either abundance or median expression and uses regularized supervised learning algorithms to identify clusters of interest. For this method we also subsampled 1000 cells from each patient. We clustered with a minimum cluster size of 5% and 5 folds and tested with the nearest shrunken centroids algorithm (PAMR).
- CellCNN[32] uses a convolutional neural network to detect rare cell subsets associated with disease. As explained above, it bypasses an explicit feature extraction process to go directly from the multicell inputs to the model prediction, drawing inspiration from multiple instance learning. We ran the convolutional neural network with 1000 cells, 1000 subsets, quantile normalization and scaling already performed and the rest of parameters with the default values. The default function performs hyperparameter tuning via a single train-test split. We further included an outer loop (20 repeats of 5-fold cross-validation) to obtain an unbiased estimate of performance, since a single train-test split would make the estimation more prone to bias.
- Diffcyt[26] employs a combination of high-resolution clustering and empirical Bayes moderated tests adapted from transcriptomics to perform differential discovery analyses. It is specifically intended for complex and/or flexible experimental designs. Like Citrus, each cluster is characterized by abundance and median marker expression and these are modelled by statistical methods based on the negative binomial distribution (Bayes estimation and generalized linear models among others). We followed a previously published workflow to run this framework[48]. We reused the FlowSOM clustering obtained in the visualization step of the study and used the edgeR method for differential abundance testing and the limma method for differential expression testing.

Software

Manual preprocessing step was performed by means of FlowJo™ v10.9 Software (BD Life Sciences). The computational step was carried out in RStudio (v2023.06.1+524, Posit team 2023) with the R Statistical Software (v4.2.2, R Core Team 2022), using packages flowCore (v2.12.2, available at Bioconductor) and flowWorkspace (v4.12.1, available at Bioconductor). File matching was also performed in R adapting the code from packages cytoBackBone (https://github.com/tchitchek-lab/CytoBackBone), cyCombine (v0.2.15, available at https://github.com/biosurf/cyCombine) and CYTOFmerge (https://github.com/tabdelaal/CyTOFmerge). Visualization made use of packages FlowSOM (v2.8.0, Bioconductor) and uwot (v0.1.16, available at CRAN). Classification was performed with caret (v6.0-94, CRAN) and rsample (v1.1.1, CRAN) packages. For the other algorithms of the literature, packages Cydar (v1.24.0,

Bioconductor), Citrus (v0.0.8, available at https://github.com/nolanlab/citrus) and Diffcyt (v1.20.0, Bioconductor) were run in R and cellCNN (https://github.com/eiriniar/CellCnn) was run in Python v2.7 (Python Software Foundation https://www.python.org/), all of them making use of the open source code provided at their respective websites.

Hardware

The computational preprocessing, file merging, visualization and feature extraction routines were performed on a 3,4 GHz, 4-core, 16 GB memory iMac machine. The classification routine was run on a 3,2 GHz, 16-core, 96 GB memory Mac Pro machine. Runtime per dataset was 8-9 minutes (running each outer fold in a 31-core parallel cluster).

Data and code availability

The source code and functions used in this article can be consulted at https://github.com/Almr95/Relapse-Prediction. This repository also includes the preprocessed and merged files of the 188 patients selected for the main analysis. The full database of anonymized FC files is available at http://flowrepository.org/id/FR-FCM-Z7A2.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at XXXX.

ACKNOWLEDGMENTS

This work was partially supported by project PDC2022-133520-I00 funded by Ministerio de Ciencia e Innovación/ Agencia Estatal de investigación (doi:10.13039/501100011033) and European Union NextGenerationEU/PRTR; by project PID2022-140451OA-I00 funded by Ministerio de Ciencia e Innovación/Agencia Estatal de investigación (doi:10.13039/501100011033) and ERDF A way of making Europe; and by University of Castilla-La Mancha / ERDF, A way of making Europe (Applied Research Projects) under grant 2022-GRIN-34405. The support of Fundación Española para la Ciencia y la Tecnología (FECYT project PR214), Asociación Pablo Ugarte (APU, Spain) and Junta de Andalucía (Spain) group FQM-201 is also acknowledged. This work was also subsidized in its early stages by a grant for the research and biomedical innovation in the health sciences within the framework of the Integrated Territorial Initiative (ITI) for the province of Cadiz (grant number ITI-0038-2019).

AUTHOR CONTRIBUTIONS

Conceptualization, V.M.P.G., M.R., and C.B.G.; Data curation, Á.M.R., R.P.G., A.N.L., S.C., J.F.R.G., E.G.V., T.C.V., Á.M.Q., A.C.R., M.R.O., M.V.M.S., A.M.P., J.L.F.S. and C.B.G.; Formal analysis, Á.M.R.; Funding acquisition, M.R. and V.M.P.G.; Investigation, Á.M.R., R.P.G., A.N.L. and S.C.; Methodology, Á.M.R.; Project administration, C.B.G., V.M.P.G. and M.R.; Resources, T.C.V., Á.M.Q., A.C.R., M.R.O., M.V.M.S., A.M.P., J.L.F.S. and C.B.G.; Software, Á.M.R.; Supervision, V.M.P.G. and M.R.; Writing—original draft, Á.M.R.; Writing—review & editing, Á.M.R., R.P.G., A.N.L., S.C., J.F.R.G., E.G.V., T.C.V., Á.M.Q., A.C.R., M.R.O., M.V.M.S., A.M.P., J.L.F.S., C.B.G., V.M.P.G. and M.R. All authors have read and agreed to the published version of the manuscript.

DECLARATION OF INTERESTS

The authors declare no conflicts of interest.

GENERATIVE AI USAGE

During the preparation of this work, the authors used chatGPT (powered by OpenAI's language model, GPT-3.5; http://openai.com) in order to improve readability and language of the work. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

1. Pui, C.-H., Yang, J. J., Hunger, S. P., Pieters, R., Schrappe, M., Biondi, A., Vora, A., Baruchel, A., Silverman, L. B., Schmiegelow, K., et al. (2015). Childhood acute lymphoblastic leukemia: Progress through collaboration. Journal of Clinical Oncology *33*, 2938. https://doi.org/10.1200/jco.2014.59.1636.

2. Ceppi, F., Cazzaniga, G., Colombini, A., Biondi, A., and Conter, V. (2015). Risk factors for relapse in childhood acute lymphoblastic leukemia: prediction and prevention. Expert Review of Hematology *8*, 57–70. https://doi.org/10.1586/17474086.2015.978281.

3. Schultz, K. R., Pullen, D. J., Sather, H. N., Shuster, J. J., Devidas, M., Borowitz, M. J., Carroll, A. J., Heerema, N. A., Rubnitz, J. E., Loh, M. L., et al. (2007). Risk-and response-based classification of childhood B-precursor acute lymphoblastic leukemia: A combined analysis of prognostic markers from the Pediatric Oncology Group (POG) and Children's Cancer Group (CCG). Blood *109*, 926–935. https://doi.org/10.1182/blood-2006-01-024729.

4. Talleur, A. C., Pui, C.-H., and Karol, S. E. (2023). What Is Next in Pediatric B-Cell Precursor Acute Lymphoblastic Leukemia. Lymphatics *1*, 34–44. https://doi.org/10.3390/lymphatics1010005.

5. Teachey, D. T. and Hunger, S. P. (2013). Predicting relapse risk in childhood acute lymphoblastic leukaemia. British Journal of Haematology *162*, 606–620. https://doi.org/10.1111/bjh.12442.

6. Basso, G., Veltroni, M., Valsecchi, M. G., Dworzak, M. N., Ratei, R., Silvestri, D., Benetello, A., Buldini, B., Maglia, O., Masera, G., et al. (2009). Risk of relapse of childhood acute lymphoblastic leukemia is predicted by flow cytometric measurement of residual disease on day 15 bone marrow. Journal of Clinical Oncology *27*, 5168–5174. https://doi.org/10.1200/jco.2008.20.8934.

7. Dongen, J. J. van, Velden, V. H. van der, Brüggemann, M., and Orfao, A. (2015). Minimal residual disease diagnostics in acute lymphoblastic leukemia: need for sensitive, fast, and standardized technologies. Blood, The Journal of the American Society of Hematology *125*, 3996–4009.

8. Pedreira, C. E., Costa, E. S., Lecrevisse, Q., van Dongen, J. J., and Orfao, A. (2013). Overview of clinical flow cytometry data analysis: recent advances and future challenges. Trends in Biotechnology *31*, 415–425. https://doi.org/10.1016/j.tibtech.2013.04.008.

9. Kalina, T., Flores-Montero, J., Van Der Velden, V., Martin-Ayuso, M., Böttcher, S., Ritgen, M., Almeida, J., Lhermitte, L., Asnafi, V., Mendonça, A., et al. (2012). EuroFlow standardization of flow cytometer instrument settings and immunophenotyping protocols. Leukemia *26*, 1986–2010. https://doi.org/https://doi.org/10.1038/leu.2012.122.

10. Van Dongen, J., Lhermitte, L., Böttcher, S., Almeida, J., Van Der Velden, V., Flores-Montero, J., Rawstron, A., Asnafi, V., Lecrevisse, Q., Lucio, P., et al. (2012). EuroFlow antibody panels for standardized n-dimensional flow cytometric immunophenotyping of normal, reactive and malignant leukocytes. Leukemia *26*, 1908–1975. https://doi.org/https://doi.org/10.1038/leu.2012.120.

11. Duetz, C., Bachas, C., Westers, T. M., and van de Loosdrecht, A. A. (2020). Computational analysis of flow cytometry data in hematological malignancies: future clinical practice? Current Opinion in Oncology *32*, 162–169. https://doi.org/10.1097/cco.0000000000000607.

12. Saeys, Y., Van Gassen, S., and Lambrecht, B. N. (2016). Computational flow cytometry: helping to make sense of high-dimensional immunology data. Nature Reviews Immunology *16*, 449. https://doi.org/10.1038/nri.2016.56.

13. Robinson, J. P., Rajwa, B., Patsekin, V., and Davisson, V. J. (2012). Computational analysis of high-throughput flow cytometry data. Expert Opinion on Drug Discovery *7*, 679–693. https://doi.org/10.1517/17460441.2012.693475.

14. Reiter, M., Diem, M., Schumich, A., Maurer-Granofszky, M., Karawajew, L., Rossi, G. J., Ratei, R., Groeneveld-Krentz, S., and Sajaroff, O. E. (2019). Automated flow cytometric mrd assessment in childhood acute b- lymphoblastic leukemia using supervised machine learning. Cytometry Part A *95*, 966–975. https://doi.org/10.1002/cyto.a.23852.

15. Good, Z., Sarno, J., Jager, A., Samusik, N., Aghaeepour, N., Simonds, E. F., White, L., Lacayo, N. J., Fantl, W. J., Fazio, G., et al. (2018). Single-cell developmental classification of b cell precursor acute lymphoblastic leukemia at diagnosis reveals predictors of relapse. Nature Medicine *24*, 474. https://doi.org/10.1038/nm.4505.

16. Chulián, S., Martínez-Rubio, Á., Pérez-García, V. M., Rosa, M., Blázquez Goñi, C., Rodríguez Gutiérrez, J. F., Hermosín-Ramos, L., Molinos Quintana, Á., Caballero-Velázquez, T., Ramírez-Orellana, M., et al. (2020). High-dimensional analysis of single-cell flow cytometry data predicts relapse in childhood acute lymphoblastic leukaemia. Cancers *13*, 17. https://doi.org/10.3390/cancers13010017.

17. Chulián, S., Stolz, B. J., Martínez-Rubio, Á., Blázquez Goñi, C., Rodríguez Gutiérrez, J. F., Caballero Velázquez, T., Molinos Quintana, Á., Ramírez Orellana, M., Castillo Robleda, A., Fuster Soler, J. L., et al. (2023). The shape of cancer relapse: Topological data analysis predicts recurrence in paediatric acute lymphoblastic leukaemia. PLoS Computational Biology *19*, e1011329. https://doi.org/10.1371/journal.pcbi.1011329.

18. Pui, C.-H., Yang, J. J., Bhakta, N., and Rodriguez-Galindo, C. (2018). Global efforts toward the cure of childhood acute lymphoblastic leukaemia. The Lancet Child & Adolescent Health *2*, 440–454. https://doi.org/10.1016/s2352-4642(18)30066-x.

19. Agarwal, M., Seth, R., and Chatterjee, T. (2021). Recent advances in molecular diagnosis and prognosis of childhood B cell lineage acute lymphoblastic leukemia (B-ALL). Indian Journal of Hematology and Blood Transfusion *37*, 10–20. https://doi.org/10.1007/s12288-020-01295-8.

20. Orlova, D. Y., Zimmerman, N., Meehan, S., Meehan, C., Waters, J., Ghosn, E. E., Filatenkov, A., Kolyagin, G. A., Gernez, Y., Tsuda, S., et al. (2016). Earth mover's distance (EMD): a true metric for comparing biomarker expression levels in cell populations. PLoS One *11*, e0151859. https://doi.org/10.1371/journal.pone.0151859.

21. Mocking, T., Duetz, C., van Kuijk, B., Westers, T., Cloos, J., and Bachas, C. (2023). Merging and imputation of flow cytometry data: a critical assessment. Cytometry Part A. https://doi.org/10.1002/cyto.a.24774.

22. Pedreira, C. E., Costa, E. S., Barrena, S., Lecrevisse, Q., Almeida, J., van Dongen, J. J., and Orfao, A. (2008). Generation of flow cytometry data files with a potentially infinite number of dimensions. Cytometry Part A *73*, 834–846. https://doi.org/10.1002/cyto.a.20608.

23. Leite Pereira, A., Lambotte, O., Le Grand, R., Cosma, A., and Tchitchek, N. (2019). CytoBackBone: an algorithm for merging of phenotypic information from different cytometric profiles. Bioinformatics *35*, 4187–4189. https://doi.org/10.1093/bioinformatics/btz212.

24. Abdelaal, T., Höllt, T., van Unen, V., Lelieveldt, B. P., Koning, F., Reinders, M. J., and Mahfouz, A. (2019). CyTOFmerge: integrating mass cytometry data across multiple panels. Bioinformatics *35*, 4063–4071. https://doi.org/10.1093/bioinformatics/btz180.

25. Pedersen, C. B., Dam, S. H., Barnkob, M. B., Leipold, M. D., Purroy, N., Rassenti, L. Z., Kipps, T. J., Nguyen, J., Lederer, J. A., Gohil, S. H., et al. (2022). cyCombine allows for robust integration of single-cell cytometry datasets within and across technologies. Nature Communications *13*, 1698. https://doi.org/10.1038/s41467-022-29383-5.

26. Weber, L. M., Nowicka, M., Soneson, C., and Robinson, M. D. (2019). diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering. Communications Biology *2*, 183. https://doi.org/10.1038/s42003-019-0415-5.

27. Van Gassen, S., Callebaut, B., Van Helden, M. J., Lambrecht, B. N., Demeester, P., Dhaene, T., and Saeys, Y. (2015). FlowSOM: Using self-organizing maps for visualization and interpretation of cytometry data. Cytometry Part A *87*, 636–645. https://doi.org/10.1002/cyto.a.22625.

28. Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W., Ng, L. G., Ginhoux, F., and Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. Nature Biotechnology *37*, 38–44. https://doi.org/10.1038/nbt.4314.

29. Lun, A. T., Richard, A. C., and Marioni, J. C. (2017). Testing for differential abundance in mass cytometry data. Nature Methods *14*, 707–709. https://doi.org/10.1038/nmeth.4295.

30. Bruggner, R. V., Bodenmiller, B., Dill, D. L., Tibshirani, R. J., and Nolan, G. P. (2014). Automated identification of stratifying signatures in cellular subpopulations. Proceedings of the National Academy of Sciences *111*, E2770–E2777. https://doi.org/10.1073/pnas.1408792111.

31. Polikowsky, H. G. and Drake, K. A. (2019). Supervised machine learning with CITRUS for single cell biomarker discovery. Mass Cytometry: Methods and Protocols, 309–332. https://doi.org/10.1007/978-1-4939-9454-0_20.

32. Arvaniti, E. and Claassen, M. (2017). Sensitive detection of rare disease-associated cell subsets via representation learning. Nature Communications *8*, 14825. https://doi.org/10.1038/ncomms14825.

33. Jan, M. and Majeti, R. (2013). Clonal evolution of acute leukemia genomes. Oncogene *32*, 135–140. https://doi.org/10.1038/onc.2012.48.

34. Rothenberg-Thurley, M., Amler, S., Goerlich, D., Köhnke, T., Konstandin, N. P., Schneider, S., Sauerland, M. C., Herold, T., Hubmann, M., Ksienzyk, B., et al. (2017). Persistence of pre-leukemic clones during first remission and risk of relapse in acute myeloid leukemia. Leukemia, 1–27. https://doi.org/10.1038/leu.2017.350.

35. Candelli, T., Schneider, P., Garrido Castro, P., Jones, L. A., Bodewes, E., Rockx-Brouwer, D., Pieters, R., Holstege, F. C., Margaritis, T., and Stam, R. W. (2022). Identification and characterization of relapse-initiating cells in MLL-rearranged infant ALL by single-cell transcriptomics. Leukemia *36*, 58–67. https://doi.org/10.1038/s41375-021-01341-y.

36. Turati, V. A., Guerra-Assunção, J. A., Potter, N. E., Gupta, R., Ecker, S., Daneviciute, A., Tarabichi, M., Webster, A. P., Ding, C., May, G., et al. (2021). Chemotherapy induces canalization of cell state in childhood B-cell precursor acute lymphoblastic leukemia. Nature Cancer *2*, 835–852. https://doi.org/10.1038/s43018-021-00219-3.

37. Pan, L., Liu, G., Lin, F., Zhong, S., Xia, H., Sun, X., and Liang, H. (2017). Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia. Scientific Reports *7*, 1–9. https://doi.org/10.1038/s41598-017-07408-0.

38. Mahmood, N., Shahid, S., Bakhshi, T., Riaz, S., Ghufran, H., and Yaqoob, M. (2020). Identification of significant risks in pediatric acute lymphoblastic leukemia (ALL) through machine learning (ML) approach. Medical & Biological Engineering & Computing, 1–10. https://doi.org/10.1007/s11517-020-02245-2.

39. Mesegué, M., Alonso-Saladrigues, A., Pérez-Jaume, S., Comes-Escoda, A., Dapena, J. L., Faura, A., Conde, N., Catalá, A., Ruiz-Llobet, A., Zapico-Muñiz, E., et al. (2021). Lower incidence of clinical allergy with PEG-asparaginase upfront versus the sequential use of native E. coli asparaginase followed by PEG-ASP in pediatric patients with acute lymphoblastic leukemia. Hematological Oncology *39*, 687–696. https://doi.org/10.1002/hon.2914/v1/review2.

40. O'Neill, K., Aghaeepour, N., Špidlen, J., and Brinkman, R. (2013). Flow cytometry bioinformatics. PLoS Computational Biology *9*, e1003365. https://doi.org/10.1371/journal.pcbi.1003365.

41. Van Gassen, S., Gaudilliere, B., Angst, M. S., Saeys, Y., and Aghaeepour, N. (2020). CytoNorm: a normalization algorithm for cytometry data. Cytometry Part A *97*, 268–278. https://doi.org/10.1002/cyto.a.23904.

42. Lee, G., Finn, W., and Scott, C. (2011). Statistical file matching of flow cytometry data. Journal of Biomedical Informatics *44*, 663–676. https://doi.org/10.1016/j.jbi.2011.03.004.

43. Law, C. W., Alhamdoosh, M., Su, S., Smyth, G. K., and Ritchie, M. E. (2016). RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR. F1000Research *5*. https://doi.org/10.12688/f1000research.9005.3.

44. Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society: Series B (Methodological) *36*, 111–133. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x.

45. Cawley, G. C. and Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. The Journal of Machine Learning Research *11*, 2079–2107.

941 46. Jeni, L. A., Cohn, J. F., and De La Torre, F. (2013). “Facing imbalanced data–recommendations
942 for the use of performance metrics”. *2013 Humaine association conference on affective computing*
943 *and intelligent interaction*. IEEE, 245–251. https://doi.org/10.1109/acii.2013.47.

944 47. Wainer, J. and Cawley, G. (2021). Nested cross-validation when selecting classifiers is overzealous
945 for most practical applications. Expert Systems With Applications *182*, 115222. https://doi.org/10.
946 1016/j.eswa.2021.115222.

947 48. Nowicka, M., Krieg, C., Crowell, H. L., Weber, L. M., Hartmann, F. J., Guglietta, S., Becher, B.,
948 Levesque, M. P., and Robinson, M. D. (2017). CyTOF workflow: differential discovery in
949 high-throughput high-dimensional cytometry datasets. F1000Research *6*.
950 https://doi.org/10.12688/f1000research.11622.4.