

# 1 Computational flow cytometry immunophenotyping at diagnosis is 2 unable to predict relapse in childhood B-cell Acute Lymphoblastic 3 Leukemia

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

9 <sup>1</sup>Department of Mathematics, University of Cádiz, 11510 Puerto Real, Spain

10 <sup>2</sup>Biomedical Research and Innovation Institute of Cádiz (INiBICA), Puerta del Mar University Hospital, 11009 Cádiz, Spain

11 <sup>3</sup>Department of Paediatric Hematology and Oncology, Jerez Hospital, 11407 Jerez de la Frontera, Spain

12 <sup>4</sup>Department of Hematology, Virgen del Rocío University Hospital, Instituto de Biomedicina de Sevilla (IBIS)/CSIC, Universidad  
13 de Sevilla, 41013 Sevilla, Spain

14 <sup>5</sup>Oncohematology Unit, Niño Jesús University Children's Hospital, 28009 Madrid, Spain

15 <sup>6</sup>Foundation for Biomedical Research Niño Jesús University Children's Hospital, 28009 Madrid, Spain

16 <sup>7</sup>Health Research Institute La Princesa, 28009 Madrid, Spain

17 <sup>8</sup>Immunology Service, Clinical University Hospital Virgen de la Arrixaca, 30120 Murcia, Spain

18 <sup>9</sup>Instituto Murciano de Investigación Sanitaria (IMIB), University of Murcia, 30120 Murcia, Spain

19 <sup>10</sup>Department of Pediatric Hematology and Oncology, Clinical University Hospital Virgen de la Arrixaca, 30120 Murcia, Spain

20 <sup>11</sup>Mathematical Oncology Laboratory (MOLAB), Department of Mathematics, Instituto de Matemática Aplicada a la Ciencia y la  
21 Ingeniería, Universidad de Castilla-La Mancha, Ciudad Real, Spain.

22 \*Correspondence: [alvaro.martinezrubio@uca.es](mailto:alvaro.martinezrubio@uca.es)

23

## 24 SUMMARY

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

## 38 INTRODUCTION

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

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

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

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

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

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

## 96 **RESULTS**

### 97 **Patient cohort is representative of childhood BCP-ALL population**

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

### 109 **Normalization and merging allows integration of multi-center, multi-sample flow 110 cytometry data**

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

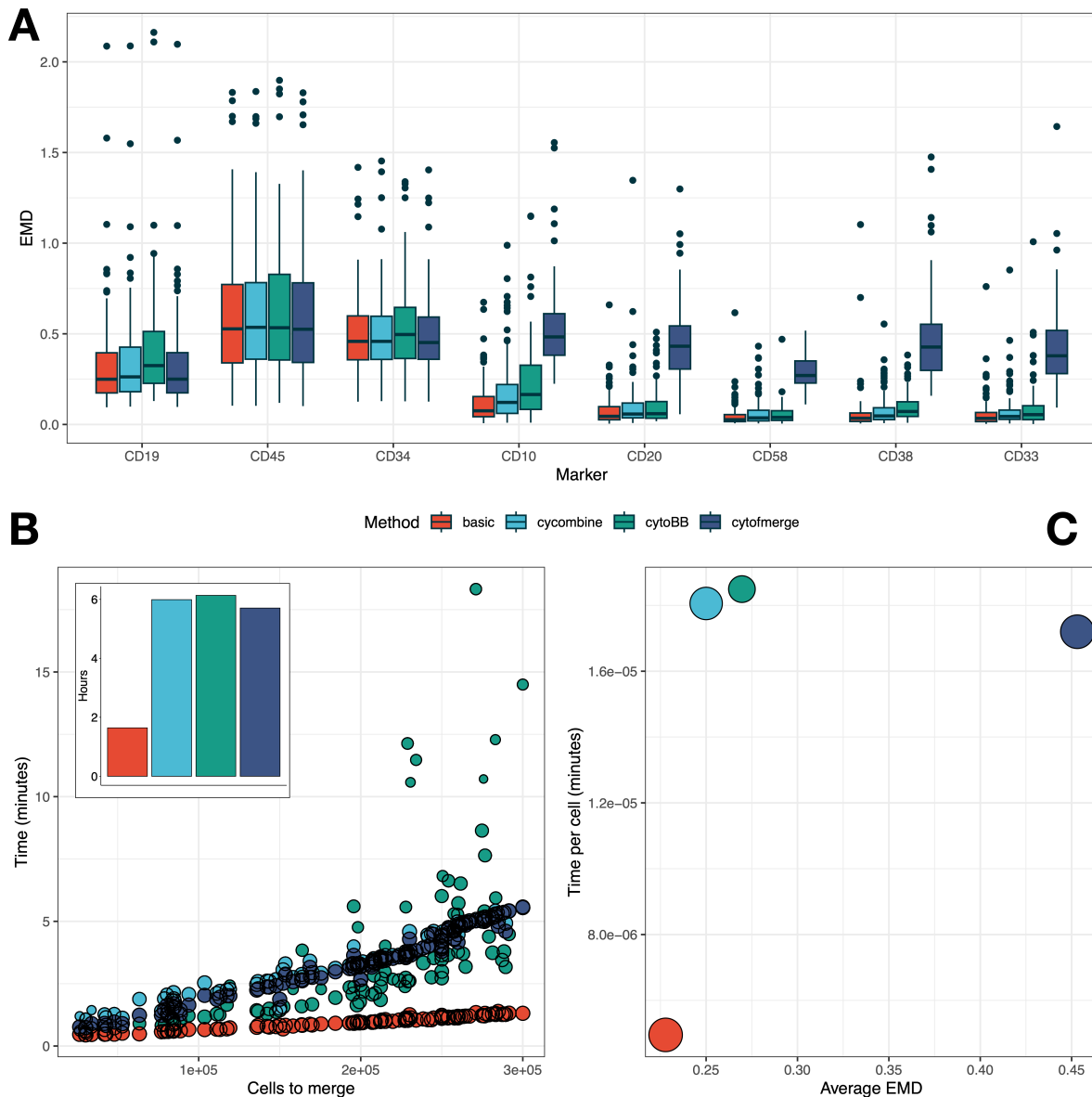
119 These sources of inter-center and inter-aliquot heterogeneity were addressed here  
120 by means of a modified min-max transformation and a quantile normalization step  
121 (Figure S2, see 'Methods'). As for the combination of several FC files into a single  
122 file, various methods have already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

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

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



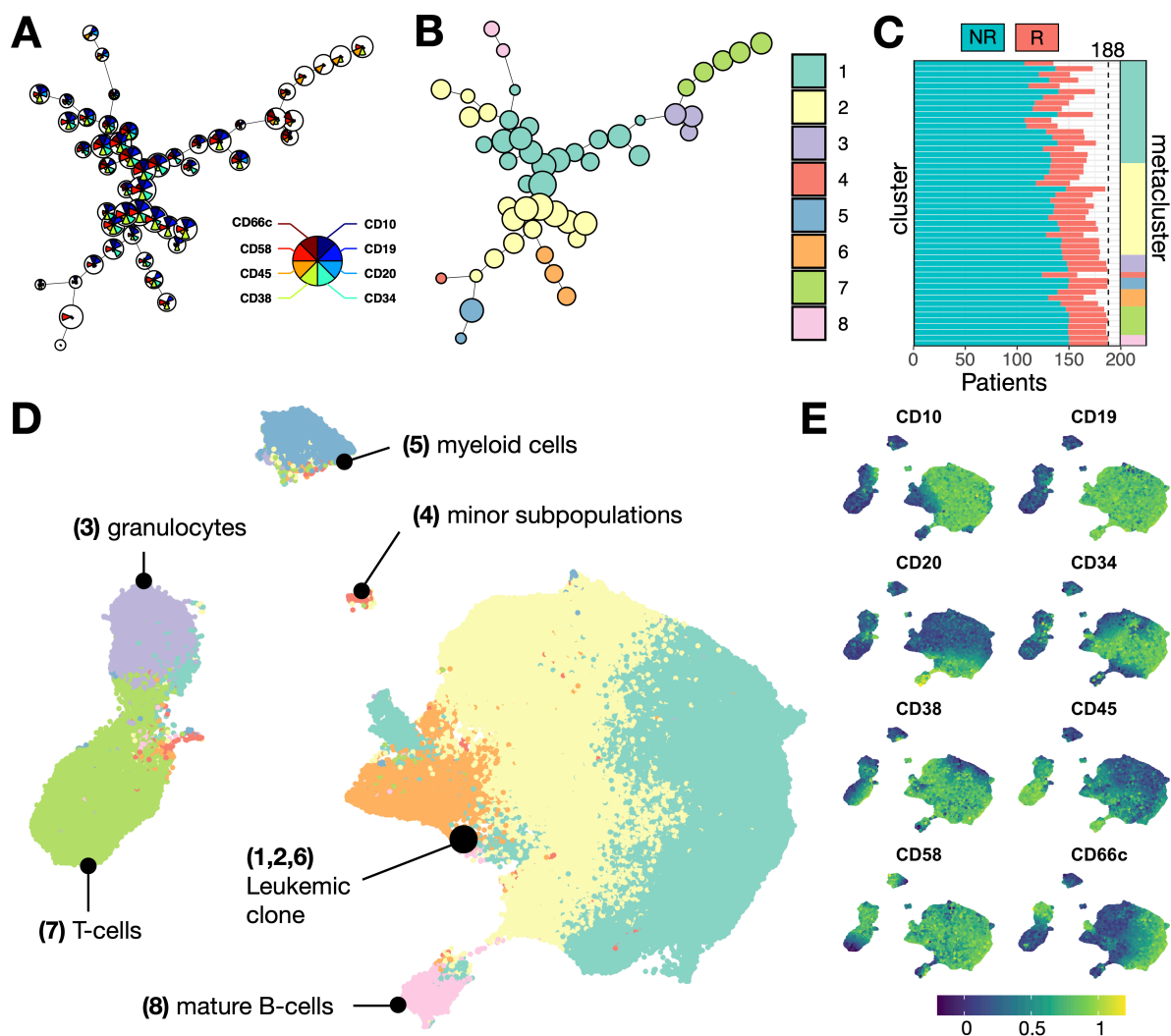
**Figure 1. Comparison of file merging methods.** **A.** Boxplots summarizing the distributions of Earth's Mover Distance (EMD) for each marker. The box includes median (horizontal line) and interquartile range (IQR). **B.** Dots represents the runtime for each patient, with the x-axis displaying the 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).



## 140 **Clustering and dimensionality reduction techniques reveal common structure** 141 **and subpopulations across patients**

142 After preprocessing, file merging and patient selection, the final set of FC markers  
143 included B-cell markers CD19, CD10 and CD20; pan-leukocyte markers CD45 and  
144 CD38; hematopoietic stem cell marker CD34 and myeloid markers CD58 and CD66c.  
145 The next step was to visualize the structure of the bone marrow of all patients. Cell  
146 subpopulations can be obtained by means of clustering techniques, which replace the  
147 traditional manual analysis or ‘gating’<sup>26</sup>. Here, we pooled all the files together and  
148 clustered via FlowSOM<sup>27</sup>. This algorithm produces a low dimensional visualization of  
149 the structure of the data in two steps. Firstly, it clusters on a higher resolution, which  
150 we manually set to 50 clusters (the influence of this number of clusters on the results  
151 will be explored later). Secondly, it obtains an optimal lower number of metaclusters  
152 by aggregating with consensus clustering. It then creates a minimum spanning tree  
153 visualization in which each cluster is represented by a node, and similar clusters are  
154 linked. This is shown in Figure 2. Marker expression per cluster is shown in Figure  
155 2A, while Figure 2B represents the metacluster to which each cluster belongs. Each  
156 metacluster is identified with a cell subpopulation that can be manually annotated. The  
157 number of patients that contribute to each cluster is shown in Figure 2C, split in relapse  
158 (R) and non-relapse patients (NR).

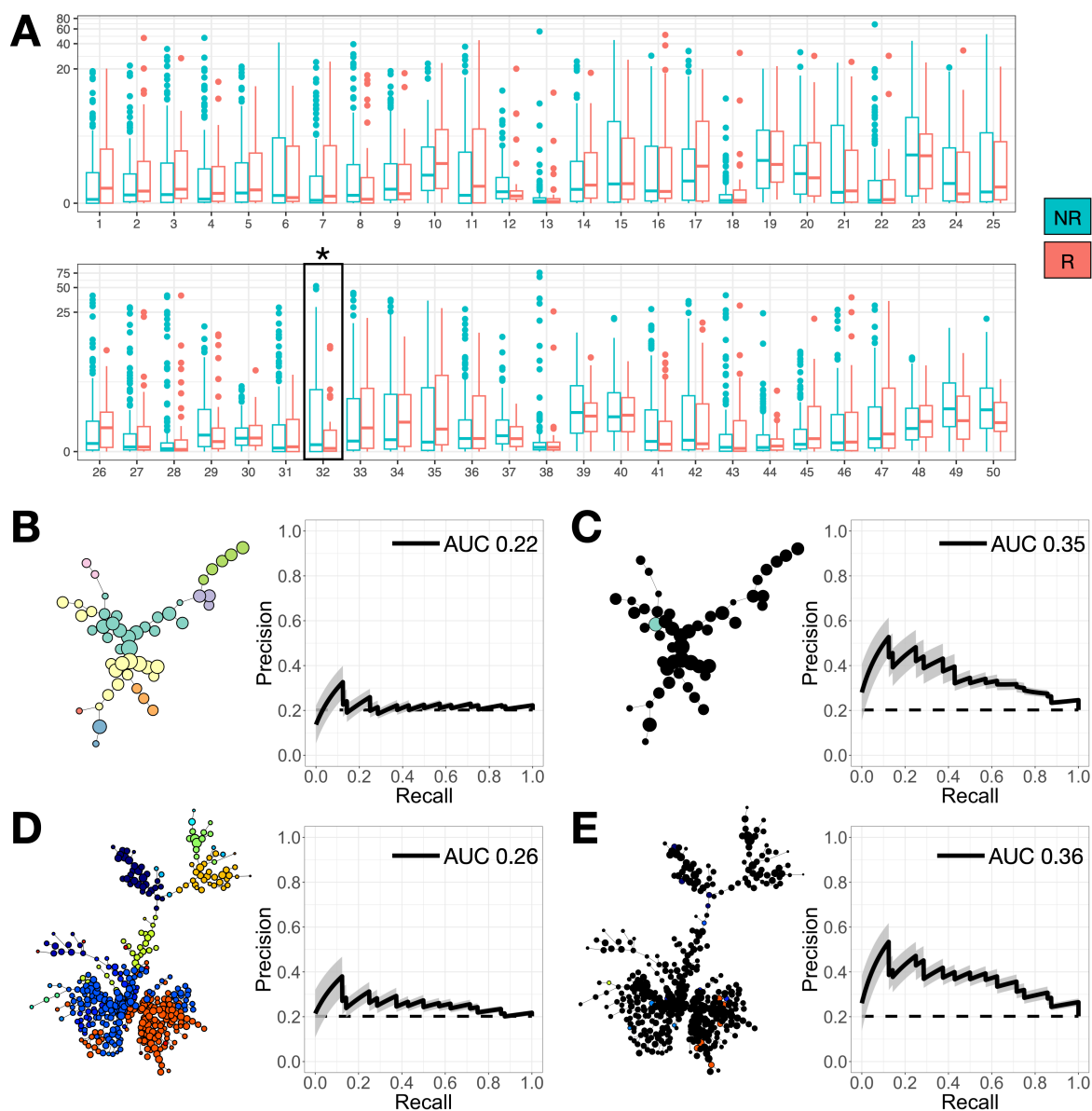
159 To visualize the clustering information on a single-cell level we used UMAP. This  
160 dimensionality reduction technique computes a two-dimensional representation that  
161 preserves the structure of the cell subpopulations<sup>28</sup>. The result is shown in Figure 2D.  
162 Each cell is colored according to FlowSOM metacluster. For comparison, the marker  
163 expression of each region of the UMAP embedding is shown in Figure 2E. We note  
164 that both FlowSOM and UMAP yield a similar structure, as shown by the proximity of  
165 the different metaclusters and UMAP regions. FlowSOM obtained an optimal number  
166 of 8 metaclusters. There were two main metaclusters (1 and 2) that comprised most  
167 of the CD19+ cells and that we identified with the leukemic clone. These are immature  
168 B-cells with intermediate expression of CD45 and heterogeneous expression of CD34  
169 and CD38. The two metaclusters were distinguished by relative expression of CD66c.  
170 We also assigned metacluster 6 to the leukemic cell population, distinguished from  
171 the other two by a negative expression of CD10. These metaclusters contained the  
172 majority of cells since the bone marrow of BCP-ALL patients at diagnosis are almost  
173 fully invaded. Metacluster 8, with a high expression of CD45 and CD20, represents  
174 healthy, mature B-cells. The remaining metaclusters represent other bone marrow  
175 cell types, including T-cells with high expression of CD45 (metacluster 7) and myeloid  
176 subpopulations (metaclusters 3, 4 and 5). While these subpopulations are seldom  
177 considered in B-cell malignancies, here we also explored them for prognostic value.  
178 With respect to the robustness and generality of these results, we note that most of the  
179 clusters contained more than 80% of the patients. When considering the metacluster  
180 scale, virtually all patients contribute to all cell subpopulations, with the exception of the  
181 minor myeloid subpopulations (metacluster 4). This confirmed that all patients adhere  
182 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.

### 183 Individual patient cells per cluster are unable to predict relapse

184 Figure 2C shows that only a number of clusters contain a proportion of relapse  
 185 patients above the baseline 20%, without any particular cluster being dominated by  
 186 either relapse or non-relapse patients. To investigate the predictive power of cell  
 187 abundance per cluster, however, we had to check not only the number of patients but  
 188 also how many cells each patient contributed with. The idea was to test if relapse  
 189 patients tended to participate more in a subset of clusters, or if instead all patients  
 190 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.

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

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



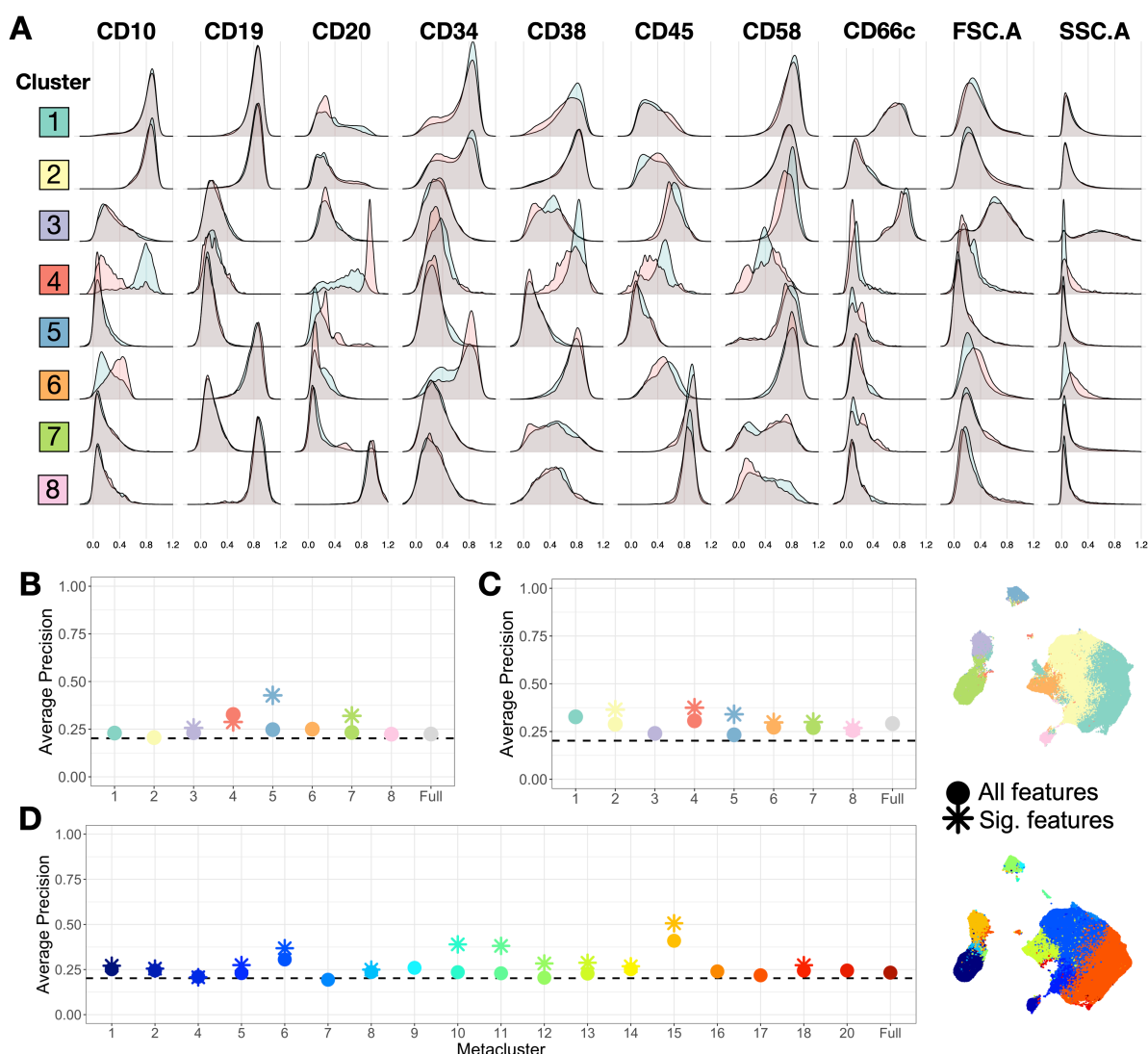
195 number of cells per cluster. Indeed, although each cluster individually did not present  
196 clear differences, non-linear interactions between all clusters could create a region in  
197 which relapse patients are more clearly distinguished. We tested this by building a  
198 classifier for relapse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

### 219 **Relative marker expression is similar between relapse and non-relapse patients** 220 **across cell subpopulations**

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

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

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

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

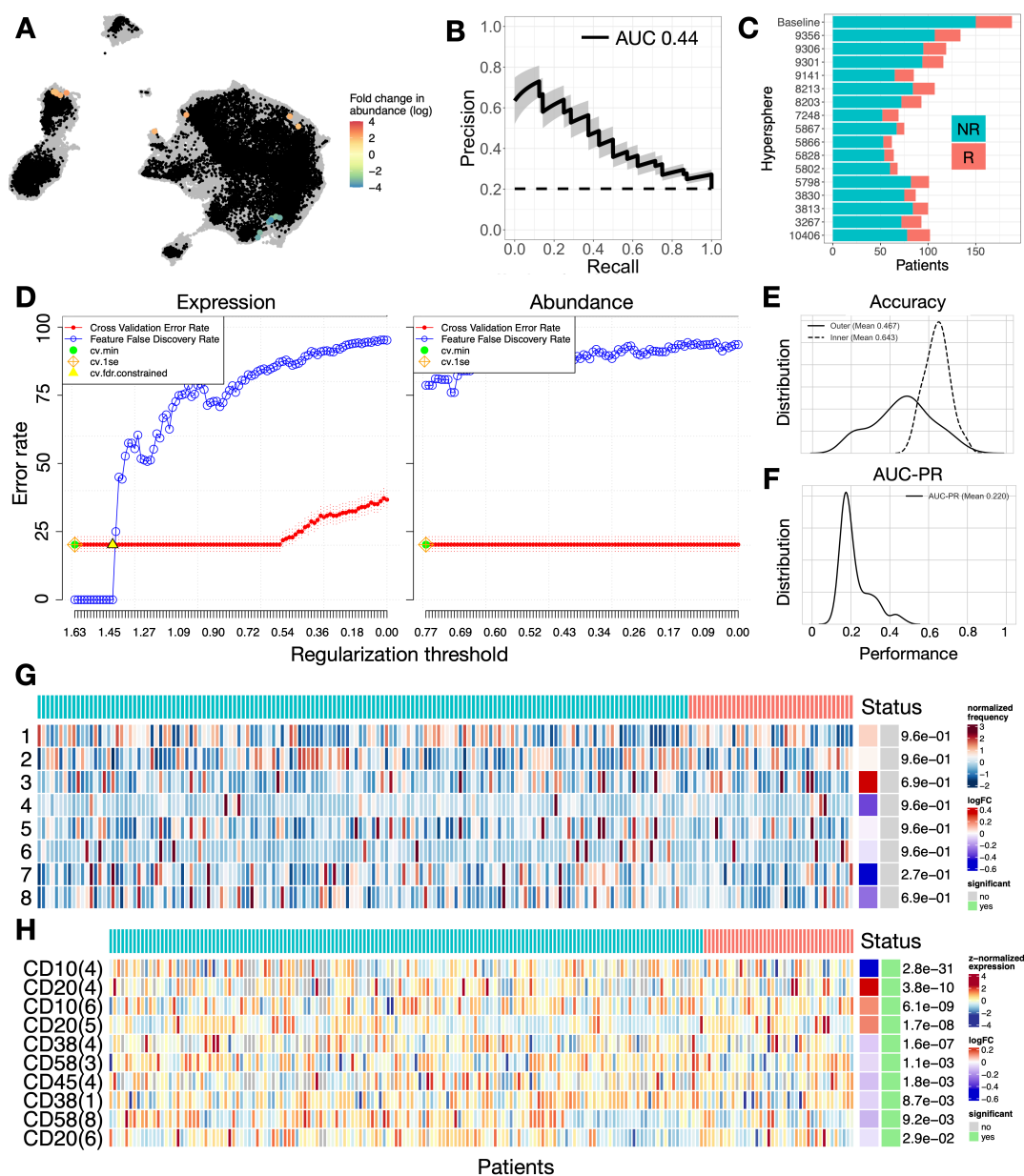
259 We explored the possibility that the observed outcome could stem from an  
260 insufficiency of detail in the information under examination for each patient. To  
261 address this, we repeated the analysis incorporating two alterations. Firstly, we  
262 expanded the set of markers selected for analysis.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

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

## 289 **Biomarker discovery algorithms from the literature support the findings of the** 290 **main analysis**

291 We contrasted our findings with other algorithms from the literature designed for  
292 biomarker discovery and outcome prediction. A description of their functionality and  
293 implementation can be found in the 'Methods' section. The results for each of them  
294 are shown in Figure 5. The first example is Cydar<sup>29</sup>, which is designed for differential  
295 abundance discovery. The clusters (hyperspheres in Cydar terminology) with a  
296 sufficient number of cells are projected onto the UMAP embedding employed in the  
297 previous sections (Figure 2). Those hyperspheres with significant differences in  
298 abundance (according to a lasso-regularized logistic regression) are plotted with  
299 wider radius and colored according to the fold change in abundance between both  
300 group of patients (Figure 5A). To check the predictive power of such hyperspheres,  
301 we extracted the number of cells per patient and hypersphere and ran the  
302 classification routine previously described, with results similar to the best models in  
303 the previous section (Figure 5B). The difference here is that due to the lower size of  
304 the clusters (hyperspheres), there are less patients per cluster (Figure 5C), which  
305 makes the results less generalizable. The second example is Citrus<sup>30</sup>. The results for  
306 both abundance and median expression (Figure 5D) indicate the lack of predictive  
307 information, regardless of regularization threshold. In both cases the null classifier  
308 (no features, leftmost regularization threshold) was the best classifier, with an error of  
309 20.2%. This number is the proportion of relapse patients in our dataset, which means  
310 the algorithm was classifying all patients as non-relapse. Further, the False Discovery  
311 Rate shows all the characteristics of a classifier unable to discriminate<sup>31</sup>. The Citrus  
312 algorithm also reports the clusters with potential predictive capacity (Figure S11A-B).  
313 The third example is cellCNN<sup>32</sup>, which uses a convolutional neural network. We  
314 complemented it with a nested loop that allowed us to provide two conclusions: First,  
315 the lack of an inner validation routine makes the algorithm more prone to overfitting, as  
316 we see in the comparison between the accuracies of the inner and outer loops  
317 (Figure 5E). Second, the performance in terms of AUCPR did not improve previous  
318 tests (Figure 5F). This algorithm also reports the characteristics of the most  
319 significant cells, which are included in Figure S11C-D. Finally, we tested Diffcyt<sup>26</sup> on  
320 the metaclusters that were already obtained by FlowSOM (Figure 2). This algorithm  
321 showed no significant differences in cell abundance per metacluster (Figure 5G) but it  
322 did detect significant differences in expression (Figure 5H). On closer inspection, we  
323 noticed that those significant features were the ones that displayed differences in  
324 aggregated marker expression (Figure 4A). We already showed how this sum of  
325 distributions does not necessarily entail that individual patients follow the same trend  
326 (see Figure S7) and that a classifier could still be unable to properly predict relapse,  
327 as shown in Figures 4B-C.

328 The conclusion of this section is that other algorithms that aim for the same goal as  
329 this study and follow a comparable methodology are also unable to detect differences  
330 between relapse and non-relapse patients. This applies to analyses centered on both  
331 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 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.



## 332 DISCUSSION

333 Approximately 15% of children diagnosed with BCP-ALL will suffer relapse or  
334 refractory disease, and the prognosis for this subgroup is significantly worsened.  
335 Despite advancements in therapy through refined chemotherapy regimens, the  
336 potential for further therapeutic success appears rooted in alternative treatments or  
337 more precise risk assessment upon diagnosis. This underscores the importance of  
338 enhancing our capability to anticipate disease progression at the individual patient  
339 level. In this investigation, we have compiled an extensive FC database for childhood  
340 BCP-ALL. A total of 252 patients from three hospitals participated in the study, with  
341 188 patients advancing to the computational analysis phase. The objective of the  
342 study was to examine whether patients experiencing relapse exhibit distinctive  
343 patterns within their FC data at diagnosis. In other words, the goal was to test if FC  
344 data at diagnosis has prognostic value with regards to long-term response.

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

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

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

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

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

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

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

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

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

460 complex methods.

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

## 469 **METHODS**

### 470 **Study population**

471 252 patients and three different spanish hospitals participated in this study. We  
472 collected data from 116 patients from Hospital Niño Jesús, Madrid (HNJ), diagnosed  
473 between January 2013 and January 2022; 80 patients from Hospital Virgen de la  
474 Arrixaca, Murcia (HVA), diagnosed between May 2011 and July 2022; and 56 patients  
475 from Hospital Virgen del Rocío, Sevilla (HVR), diagnosed between January 2012 and  
476 July 2021. 207 patients had long-term remission and 44 patients relapsed. All  
477 patients are in the age range 0-19. We dropped those which continued treatment at  
478 another institution or that had not reached 1 year of follow up, with 211 patients finally  
479 proceeding to the main analysis (Figure S1). The data collected included FC files  
480 from bone marrow samples at diagnosis and additional clinical information: Age, sex,  
481 phenotype, risk group, CNS involvement, absolute lymphocyte count (ALC),  
482 immunophenotype and genetic information (karyotype and chromosomal  
483 translocations). Informed consent was obtained from the parents or legal guardians  
484 according to the Helsinki Declaration.

### 485 **Treatment**

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

### 495 **Risk stratification**

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

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

## 505 **Patient outcome**

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

## 511 **Flow cytometry data**

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

## 517 **Preprocessing of flow cytometry data**

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

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

533 Finally, we had to consider the issue of backbone markers displaying inter-aliquot  
534 differences. Some causes of this variability are staining problems, acquisition errors  
535 and other batch effects. To account for this source of heterogeneity we first sampled  
536 10000 cells from each tube and then performed quantile normalization, a technique



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

## 541 **File merging**

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

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

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

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

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

## 596 **Patient selection**

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

## 606 **Flow cytometry visualization**

607 FlowSOM was run with parameters  $x = 5$ ,  $y = 10$  and  $\max K = 20$ . We consistently  
608 obtained an optimal number of 8 metaclusters. For UMAP, we subset 1000 cells from  
609 each patient and pool the subset files to obtain the embedding of the bone marrow of  
610 all patients. After a visual exploration, we selected UMAP hyperparameters  $\min\_dist =$   
611  $0.1$ ,  $n\_neighbors = 50$  and the rest with default values (Figure S12).

## 612 **Feature extraction**

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

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

## 627 **Classification**

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

642 Without any previous knowledge about the characteristics of this region and given  
643 that it can be quite different depending on which metric we are considering, we could  
644 not say a priori which classification model was best for this task, nor which  
645 hyperparameters of such model were optimal. For this reason, the classification  
646 routine had to include some form of internal validation to make this decision based on  
647 the data. We did this by means of nested cross-validation<sup>44,45</sup>. This approach  
648 consists of two cross-validation loops, an outer loop and an inner loop. The inner loop  
649 is used to find the best model and its hyperparameters, and the outer loop is used to  
650 get an estimate of performance in unseen data. For the inner loop we performed  
651 9-fold cross-validation repeated 20 times to get a more robust estimate, and for the  
652 outer loop we performed 5-fold cross-validation, repeated 10 times. This resampling  
653 scheme implied that each inner fold contained 16 patients on average, with 2 of them  
654 belonging to the relapse group.

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

660 curve, which is more suitable for problems with unbalanced data<sup>46</sup>. Hyperparameter  
661 estimation and model selection were thus performed together<sup>47</sup>.

662 For every dataset, the nested cross-validation routine produces 50 performance  
663 estimates (AUC-PR) and identifies 50 ‘best models’ (obtained from the 10 repetitions  
664 of 5-fold cross-validation in the outer loop). We summarized the 50 AUC-PR values by  
665 calculating their average, and the 50 best models by using a measure of heterogeneity  
666 as a surrogate of the stability of the routine. This stability measure is assigned a value of  
667 1 if the same model is consistently selected in all outer folds, and 0 if the four models are  
668 equally frequent. It is important to note that while this measure can indicate instability  
669 or unsuccessful optimization, it is possible for two models to perform nearly equally  
670 well in identifying the best boundary, making them equally suitable for the task at hand.  
671 Thus, it’s essential to consider the degree by which the top model has been selected  
672 and its associated level of performance.

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

## 678 **Comparison with other algorithms**

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

- 694 • Cydar<sup>29</sup> identifies differentially abundant cell populations between groups. It was  
695 originally proposed for mass cytometry data but can be extended to any  
696 multidimensional dataset. It clusters cells into hyperspheres, extracts cell  
697 abundance and tests for significant differences by means of a negative binomial  
698 generalized linear model, controlling for the spatial false discovery rate. In this  
699 study we subsampled 1000 cells from each patient, clustered with scaling factor  
700 0.2, removed hyperspheres with average counts below 5 and applied the QL  
701 framework to test for significant differences. After correcting for multiple testing

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

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

## 730 **Software**

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



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

## 748 **Hardware**

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

## 754 **Data and code availability**

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

## 759 **SUPPLEMENTAL INFORMATION**

760 Supplemental information can be found online at XXXX.

## 761 **ACKNOWLEDGMENTS**

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

## 775 AUTHOR CONTRIBUTIONS

776 Conceptualization, V.M.P.G., M.R., and C.B.G.; Data curation, Á.M.R., R.P.G., A.N.L.,  
777 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  
778 C.B.G.; Formal analysis, Á.M.R.; Funding acquisition, M.R. and V.M.P.G.;  
779 Investigation, Á.M.R., R.P.G., A.N.L. and S.C.; Methodology, Á.M.R.; Project  
780 administration, C.B.G., V.M.P.G. and M.R.; Resources, T.C.V., Á.M.Q., A.C.R.,  
781 M.R.O., M.V.M.S., A.M.P., J.L.F.S. and C.B.G.; Software, Á.M.R.; Supervision,  
782 V.M.P.G. and M.R.; Writing—original draft, Á.M.R.; Writing—review & editing, Á.M.R.,  
783 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.,  
784 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  
785 published version of the manuscript.

## 786 DECLARATION OF INTERESTS

787 The authors declare no conflicts of interest.

## 788 GENERATIVE AI USAGE

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

## 793 References

- 794 1. Pui, C.-H., Yang, J. J., Hunger, S. P., Pieters, R., Schrappe, M., Biondi, A., Vora, A., Baruchel, A.,  
795 Silverman, L. B., Schmiegelow, K., et al. (2015). Childhood acute lymphoblastic leukemia: Progress  
796 through collaboration. *Journal of Clinical Oncology* 33, 2938. [https://doi.org/10.1200/jco.2014.59.](https://doi.org/10.1200/jco.2014.59.1636)  
797 [1636](https://doi.org/10.1200/jco.2014.59.1636).
- 798 2. Ceppi, F., Cazzaniga, G., Colombini, A., Biondi, A., and Conter, V. (2015). Risk factors for relapse in  
799 childhood acute lymphoblastic leukemia: prediction and prevention. *Expert Review of Hematology*  
800 *8*, 57–70. <https://doi.org/10.1586/17474086.2015.978281>.
- 801 3. Schultz, K. R., Pullen, D. J., Sather, H. N., Shuster, J. J., Devidas, M., Borowitz, M. J., Carroll, A. J.,  
802 Heerema, N. A., Rubnitz, J. E., Loh, M. L., et al. (2007). Risk-and response-based classification of  
803 childhood B-precursor acute lymphoblastic leukemia: A combined analysis of prognostic markers  
804 from the Pediatric Oncology Group (POG) and Children's Cancer Group (CCG). *Blood* 109, 926–  
805 935. <https://doi.org/10.1182/blood-2006-01-024729>.
- 806 4. Talleur, A. C., Pui, C.-H., and Karol, S. E. (2023). What Is Next in Pediatric B-Cell Precursor Acute  
807 Lymphoblastic Leukemia. *Lymphatics* 1, 34–44. <https://doi.org/10.3390/lymphatics1010005>.
- 808 5. Teachey, D. T. and Hunger, S. P. (2013). Predicting relapse risk in childhood acute lymphoblastic  
809 leukaemia. *British Journal of Haematology* 162, 606–620. <https://doi.org/10.1111/bjh.12442>.

- 810 6. Basso, G., Veltroni, M., Valsecchi, M. G., Dworzak, M. N., Ratei, R., Silvestri, D., Benetello, A.,  
811 Buldini, B., Maglia, O., Maserà, G., et al. (2009). Risk of relapse of childhood acute lymphoblastic  
812 leukemia is predicted by flow cytometric measurement of residual disease on day 15 bone marrow.  
813 *Journal of Clinical Oncology* 27, 5168–5174. <https://doi.org/10.1200/jco.2008.20.8934>.
- 814 7. Dongen, J. J. van, Velden, V. H. van der, Brüggemann, M., and Orfao, A. (2015). Minimal residual  
815 disease diagnostics in acute lymphoblastic leukemia: need for sensitive, fast, and standardized  
816 technologies. *Blood, The Journal of the American Society of Hematology* 125, 3996–4009.
- 817 8. Pedreira, C. E., Costa, E. S., Lecrevisse, Q., van Dongen, J. J., and Orfao, A. (2013). Overview  
818 of clinical flow cytometry data analysis: recent advances and future challenges. *Trends in*  
819 *Biotechnology* 31, 415–425. <https://doi.org/10.1016/j.tibtech.2013.04.008>.
- 820 9. Kalina, T., Flores-Montero, J., Van Der Velden, V., Martín-Ayuso, M., Böttcher, S., Ritgen, M.,  
821 Almeida, J., Lhermitte, L., Asnafi, V., Mendonça, A., et al. (2012). EuroFlow standardization of  
822 flow cytometer instrument settings and immunophenotyping protocols. *Leukemia* 26, 1986–2010.  
823 <https://doi.org/https://doi.org/10.1038/leu.2012.122>.
- 824 10. Van Dongen, J., Lhermitte, L., Böttcher, S., Almeida, J., Van Der Velden, V., Flores-Montero, J.,  
825 Rawstron, A., Asnafi, V., Lecrevisse, Q., Lucio, P., et al. (2012). EuroFlow antibody panels for  
826 standardized n-dimensional flow cytometric immunophenotyping of normal, reactive and malignant  
827 leukocytes. *Leukemia* 26, 1908–1975. <https://doi.org/https://doi.org/10.1038/leu.2012.120>.
- 828 11. Duetz, C., Bachas, C., Westers, T. M., and van de Loosdrecht, A. A. (2020). Computational analysis  
829 of flow cytometry data in hematological malignancies: future clinical practice? *Current Opinion in*  
830 *Oncology* 32, 162–169. <https://doi.org/10.1097/cco.0000000000000607>.
- 831 12. Saeys, Y., Van Gassen, S., and Lambrecht, B. N. (2016). Computational flow cytometry: helping  
832 to make sense of high-dimensional immunology data. *Nature Reviews Immunology* 16, 449. <https://doi.org/10.1038/nri.2016.56>.
- 833
- 834 13. Robinson, J. P., Rajwa, B., Patsekina, V., and Davisson, V. J. (2012). Computational analysis of  
835 high-throughput flow cytometry data. *Expert Opinion on Drug Discovery* 7, 679–693. <https://doi.org/10.1517/17460441.2012.693475>.
- 836
- 837 14. Reiter, M., Diem, M., Schumich, A., Maurer-Granofszky, M., Karawajew, L., Rossi, G. J., Ratei, R.,  
838 Groeneveld-Krentz, S., and Sajaroff, O. E. (2019). Automated flow cytometric mrd assessment in  
839 childhood acute b- lymphoblastic leukemia using supervised machine learning. *Cytometry Part A*  
840 95, 966–975. <https://doi.org/10.1002/cyto.a.23852>.
- 841 15. Good, Z., Sarno, J., Jager, A., Samusik, N., Aghaeepour, N., Simonds, E. F., White, L., Lacayo,  
842 N. J., Fantl, W. J., Fazio, G., et al. (2018). Single-cell developmental classification of b cell precursor  
843 acute lymphoblastic leukemia at diagnosis reveals predictors of relapse. *Nature Medicine* 24, 474.  
844 <https://doi.org/10.1038/nm.4505>.
- 845 16. Chulián, S., Martínez-Rubio, Á., Pérez-García, V. M., Rosa, M., Blázquez Goñi, C., Rodríguez  
846 Gutiérrez, J. F., Hermosín-Ramos, L., Molinos Quintana, Á., Caballero-Velázquez, T., Ramírez-  
847 Orellana, M., et al. (2020). High-dimensional analysis of single-cell flow cytometry data predicts  
848 relapse in childhood acute lymphoblastic leukaemia. *Cancers* 13, 17. <https://doi.org/10.3390/cancers13010017>.
- 849
- 850 17. Chulián, S., Stolz, B. J., Martínez-Rubio, Á., Blázquez Goñi, C., Rodríguez Gutiérrez, J. F.,  
851 Caballero Velázquez, T., Molinos Quintana, Á., Ramírez Orellana, M., Castillo Robleda, A.,  
852 Fuster Soler, J. L., et al. (2023). The shape of cancer relapse: Topological data analysis predicts  
853 recurrence in paediatric acute lymphoblastic leukaemia. *PLoS Computational Biology* 19,  
854 e1011329. <https://doi.org/10.1371/journal.pcbi.1011329>.

- 855 18. Pui, C.-H., Yang, J. J., Bhakta, N., and Rodriguez-Galindo, C. (2018). Global efforts toward the cure  
856 of childhood acute lymphoblastic leukaemia. *The Lancet Child & Adolescent Health* 2, 440–454.  
857 [https://doi.org/10.1016/s2352-4642\(18\)30066-x](https://doi.org/10.1016/s2352-4642(18)30066-x).
- 858 19. Agarwal, M., Seth, R., and Chatterjee, T. (2021). Recent advances in molecular diagnosis and  
859 prognosis of childhood B cell lineage acute lymphoblastic leukemia (B-ALL). *Indian Journal of*  
860 *Hematology and Blood Transfusion* 37, 10–20. <https://doi.org/10.1007/s12288-020-01295-8>.
- 861 20. Orlova, D. Y., Zimmerman, N., Meehan, S., Meehan, C., Waters, J., Ghosn, E. E., Filatenkov, A.,  
862 Kolyagin, G. A., Gernez, Y., Tsuda, S., et al. (2016). Earth mover's distance (EMD): a true metric  
863 for comparing biomarker expression levels in cell populations. *PLoS One* 11, e0151859. <https://doi.org/10.1371/journal.pone.0151859>.
- 865 21. Mocking, T., Duetz, C., van Kuijk, B., Westers, T., Cloos, J., and Bachas, C. (2023). Merging and  
866 imputation of flow cytometry data: a critical assessment. *Cytometry Part A*. <https://doi.org/10.1002/cyto.a.24774>.
- 868 22. Pedreira, C. E., Costa, E. S., Barrena, S., Lecrevisse, Q., Almeida, J., van Dongen, J. J., and Orfao,  
869 A. (2008). Generation of flow cytometry data files with a potentially infinite number of dimensions.  
870 *Cytometry Part A* 73, 834–846. <https://doi.org/10.1002/cyto.a.20608>.
- 871 23. Leite Pereira, A., Lambotte, O., Le Grand, R., Cosma, A., and Tchitchek, N. (2019). CytoBackBone:  
872 an algorithm for merging of phenotypic information from different cytometric profiles. *Bioinformatics*  
873 35, 4187–4189. <https://doi.org/10.1093/bioinformatics/btz212>.
- 874 24. Abdelaal, T., Höllt, T., van Unen, V., Lelieveldt, B. P., Koning, F., Reinders, M. J., and Mahfouz,  
875 A. (2019). CyTOFmerge: integrating mass cytometry data across multiple panels. *Bioinformatics*  
876 35, 4063–4071. <https://doi.org/10.1093/bioinformatics/btz180>.
- 877 25. Pedersen, C. B., Dam, S. H., Barnkob, M. B., Leipold, M. D., Purroy, N., Rassenti, L. Z., Kipps, T. J.,  
878 Nguyen, J., Lederer, J. A., Gohil, S. H., et al. (2022). cyCombine allows for robust integration of  
879 single-cell cytometry datasets within and across technologies. *Nature Communications* 13, 1698.  
880 <https://doi.org/10.1038/s41467-022-29383-5>.
- 881 26. Weber, L. M., Nowicka, M., Soneson, C., and Robinson, M. D. (2019). diffcyt: Differential discovery  
882 in high-dimensional cytometry via high-resolution clustering. *Communications Biology* 2, 183. <https://doi.org/10.1038/s42003-019-0415-5>.
- 884 27. Van Gassen, S., Callebaut, B., Van Helden, M. J., Lambrecht, B. N., Demeester, P., Dhaene, T.,  
885 and Saeys, Y. (2015). FlowSOM: Using self-organizing maps for visualization and interpretation of  
886 cytometry data. *Cytometry Part A* 87, 636–645. <https://doi.org/10.1002/cyto.a.22625>.
- 887 28. Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W., Ng, L. G., Ginhoux, F., and  
888 Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP.  
889 *Nature Biotechnology* 37, 38–44. <https://doi.org/10.1038/nbt.4314>.
- 890 29. Lun, A. T., Richard, A. C., and Marioni, J. C. (2017). Testing for differential abundance in mass  
891 cytometry data. *Nature Methods* 14, 707–709. <https://doi.org/10.1038/nmeth.4295>.
- 892 30. Bruggner, R. V., Bodenmiller, B., Dill, D. L., Tibshirani, R. J., and Nolan, G. P. (2014). Automated  
893 identification of stratifying signatures in cellular subpopulations. *Proceedings of the National*  
894 *Academy of Sciences* 111, E2770–E2777. <https://doi.org/10.1073/pnas.1408792111>.
- 895 31. Polikowsky, H. G. and Drake, K. A. (2019). Supervised machine learning with CITRUS for single  
896 cell biomarker discovery. *Mass Cytometry: Methods and Protocols*, 309–332. [https://doi.org/10.1007/978-1-4939-9454-0\\_20](https://doi.org/10.1007/978-1-4939-9454-0_20).

- 898 32. Arvaniti, E. and Claassen, M. (2017). Sensitive detection of rare disease-associated cell subsets via  
899 representation learning. *Nature Communications* 8, 14825. <https://doi.org/10.1038/ncomms14825>.
- 900 33. Jan, M. and Majeti, R. (2013). Clonal evolution of acute leukemia genomes. *Oncogene* 32, 135–  
901 140. <https://doi.org/10.1038/onc.2012.48>.
- 902 34. Rothenberg-Thurley, M., Amler, S., Goerlich, D., Köhnke, T., Konstandin, N. P., Schneider, S.,  
903 Sauerland, M. C., Herold, T., Hubmann, M., Ksienzyk, B., et al. (2017). Persistence of  
904 pre-leukemic clones during first remission and risk of relapse in acute myeloid leukemia.  
905 *Leukemia*, 1–27. <https://doi.org/10.1038/leu.2017.350>.
- 906 35. Candelli, T., Schneider, P., Garrido Castro, P., Jones, L. A., Bodewes, E., Rockx-Brouwer, D.,  
907 Pieters, R., Holstege, F. C., Margaritis, T., and Stam, R. W. (2022). Identification and  
908 characterization of relapse-initiating cells in MLL-rearranged infant ALL by single-cell  
909 transcriptomics. *Leukemia* 36, 58–67. <https://doi.org/10.1038/s41375-021-01341-y>.
- 910 36. Turati, V. A., Guerra-Assunção, J. A., Potter, N. E., Gupta, R., Ecker, S., Daneviciute, A., Tarabichi,  
911 M., Webster, A. P., Ding, C., May, G., et al. (2021). Chemotherapy induces canalization of cell  
912 state in childhood B-cell precursor acute lymphoblastic leukemia. *Nature Cancer* 2, 835–852. <https://doi.org/10.1038/s43018-021-00219-3>.
- 914 37. Pan, L., Liu, G., Lin, F., Zhong, S., Xia, H., Sun, X., and Liang, H. (2017). Machine learning  
915 applications for prediction of relapse in childhood acute lymphoblastic leukemia. *Scientific*  
916 *Reports* 7, 1–9. <https://doi.org/10.1038/s41598-017-07408-0>.
- 917 38. Mahmood, N., Shahid, S., Bakhshi, T., Riaz, S., Ghufuran, H., and Yaqoob, M. (2020). Identification  
918 of significant risks in pediatric acute lymphoblastic leukemia (ALL) through machine learning (ML)  
919 approach. *Medical & Biological Engineering & Computing*, 1–10. <https://doi.org/10.1007/s11517-020-02245-2>.
- 921 39. Mesegué, M., Alonso-Saladrigues, A., Pérez-Jaume, S., Comes-Escoda, A., Dapena, J. L.,  
922 Faura, A., Conde, N., Catalá, A., Ruiz-Llobet, A., Zapico-Muñiz, E., et al. (2021). Lower incidence  
923 of clinical allergy with PEG-asparaginase upfront versus the sequential use of native *E. coli*  
924 asparaginase followed by PEG-ASP in pediatric patients with acute lymphoblastic leukemia.  
925 *Hematological Oncology* 39, 687–696. <https://doi.org/10.1002/hon.2914/v1/review2>.
- 926 40. O’Neill, K., Aghaeepour, N., Špidlen, J., and Brinkman, R. (2013). Flow cytometry bioinformatics.  
927 *PLoS Computational Biology* 9, e1003365. <https://doi.org/10.1371/journal.pcbi.1003365>.
- 928 41. Van Gassen, S., Gaudilliere, B., Angst, M. S., Saeys, Y., and Aghaeepour, N. (2020). CytoNorm:  
929 a normalization algorithm for cytometry data. *Cytometry Part A* 97, 268–278. <https://doi.org/10.1002/cyto.a.23904>.
- 931 42. Lee, G., Finn, W., and Scott, C. (2011). Statistical file matching of flow cytometry data. *Journal of*  
932 *Biomedical Informatics* 44, 663–676. <https://doi.org/10.1016/j.jbi.2011.03.004>.
- 933 43. Law, C. W., Alhamdoosh, M., Su, S., Smyth, G. K., and Ritchie, M. E. (2016). RNA-seq analysis  
934 is easy as 1-2-3 with limma, Glimma and edgeR. *F1000Research* 5. <https://doi.org/10.12688/f1000research.9005.3>.
- 936 44. Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the*  
937 *Royal Statistical Society: Series B (Methodological)* 36, 111–133. <https://doi.org/10.1111/j.2517-6161.1974.tb00994.x>.
- 939 45. Cawley, G. C. and Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection  
940 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](https://doi.org/10.1016/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>.