Automated cell type annotation and exploration of single-cell signalling dynamics using mass cytometry

Dimitrios Kleftogiannnis[1,2,*], Benedicte Sjo Tislevoll[2], Nisha van der Meer[5] , Inga K. F. Motzfeldt[3] , Monica Hellesøy[3], Stein-Erik Gullaksen[3], Emmanuel Griessinger[5], Oda Fagerholt[2], Andrea Lenartova[4], Yngvar Fløisand[4], Jan Jacob Schuringa[5], Bjørn Tore Gjertsen[2,3] and Inge Jonassen[1,2]

[1] Department of Informatics, Computational Biology Unit, University of Bergen, Bergen, 5020, Norway

[2] Centre for Cancer Biomarkers (CCBIO), Department of Clinical Science, University of Bergen, Bergen, 5021, Norway

[3] Department of Medicine, Hematology Section, Haukeland University Hospital, Helse Bergen HF, Bergen, 5021, Norway

[4] Department of Hematology, Oslo University Hospital, Oslo, 4950, Norway.

[5] Department of Experimental Hematology, University Medical Center Groningen, University of Groningen, Groningen, 9713, The Netherlands

* Correspondence: Dimitrios.Kleftogiannis@uib.no (D.K)

Abstract

Mass cytometry by time-of-flight (CyTOF) allows for in-depth characterisation of cellular heterogeneity in cancer and other diseases. However, computational identification of cell populations from CyTOF, and utilisation of single-cell data for biomarker discoveries is challenging. Here, we deployed a bioinformatics framework that enables automated cell population annotation, and systematic exploration of interactions between signalling proteins in a CyTOF antibody panel. As a case study, we applied the developed framework to analyse a cohort of 45 leukemia patients. We investigated associations between the cellular composition and clinicopathological and genetic features and reported salient signalling interactions that were sufficient to predict short-term survival. We also demonstrate that targeting cell type-specific signalling interactions in leukemia might improve existing patient risk stratification methods. Our work introduces a more general analysis framework for CyTOF data with many applications in future personalised medicine projects.

Keywords

Single-cell analysis, CyTOF, Machine Learning, Predictive biomarkers, Cancer research, Leukemia

Introduction

Mass cytometry by time-of-flight (CyTOF) is a powerful technology that allows for simultaneous quantification of over 30 parameters with single-cell resolution (1-2). The application of CyTOF has revolutionised our understanding about cellular biology and immunology, and several studies have emphasized on the role of cellular phenotypic heterogeneity in human diseases (3-4).

Among the most heterogenous human diseases, Acute Myeloid Leukemia (AML) is an aggressive blood cancer with poor overall survival. In leukemia and other cancer types, although high rates of initial responses to frontline therapies are observed, chemoresistance leading to relapse and metastasis are the most frequent causes of cancer-related death (5-6). Both in cancer and many other diseases, recognising early responders to standard therapies might eliminate potentially harmful side effects, which thereby might improve the quality of patients' life (7). Hence, several risk stratification strategies have been deployed, and many biomarker discovery studies have been presented combining CyTOF data, clinical information, genetic, epigenetic and/or treatment-response data with Artificial Intelligence (AI) algorithms (8-12). However, effective identification of cell populations from CyTOF and cell type-specific analysis of signalling data poses a major challenge.

Typically, CyTOF downstream analyses require accurate identification of cellular phenotypes. For this task a wide range of computational methods have been developed including unsupervised clustering algorithms (13-15) and semi-supervised approaches (16-17). More recently, it has been shown that Linear Discriminant Analysis (LDA) trained with single cells from annotated cell types, achieves high precision (18).

Identification of cell type-specific predictive biomarkers for therapy response or resistance may require mapping cellular phenotypes to signalling states. To quantify signalling responses, approaches such as the Statistical Analysis of Perturbation Responses (SARA) (14), or the Density Resampled Estimate of Mutual Information (DREMI) have been developed (19). However, although the latter approaches are powerful, they have not been fully exploited to identify cell type-specific signalling states that are predictive of survival/relapse.

To mitigate major limitations of CyTOF analyses, here we propose a novel bioinformatics framework that allows for automated cell population annotation and systematic analysis of signalling interactions between proteins in a CyTOF panel. To facilitate improved phenotypic characterisation of individual cells, a K-nearest neighbour (KNN) single-cell classifier was trained and used as "reference" to annotate cell populations from patient samples (20,21). A novel feature engineering approach based on DREMI was utilised to score all pairwise interactions between signalling proteins in the antibody panel.

As a case study we re-analysed a publicly available cohort of 45 diagnostic samples from leukemia patients (22). By applying Machine Learning-based modelling to DREMI scores computed from single-cell populations, we constructed computational models to predict short term survivors. We anticipate that the proposed analysis framework could serve as a paradigm for future personalised medicine projects.

Results

Automated annotation of single-cell populations using a single-cell classifier

Previous studies have shown that supervised learning approaches for cell annotation such as the Linear Discriminant Analysis (LDA), perform equally well to unsupervised clustering algorithms and other approaches that often are difficult to tune (15). However, application of supervised learning for automated cell prediction requires well-characterised set of representative single-cells to be used for training. For this purpose, we first profiled with CyTOF peripheral blood (PBMCs) and bone marrow (in total ~1.2 million cells) from seven healthy donors (HD). We deployed a semi-automated gating strategy using FlowDensity (23) (Supplementary Figure S1) to select individual cells of major myeloid and lymphoid cell types namely B-cells, T-cytotoxic, T-helper, Natural Killer (NK), Granulocytes, Monocytes, Hematopoietic Stem Cells (HSC) and Multipotent Progenitor (MPP) cells based on cell type-defining surface markers available in our antibody panel (24). We visualised the annotated healthy cells with Principal Component Analysis (PCA). Different myeloid and lymphoid cell types were clearly separated in the two-dimensional space considering only the first two principal components capturing ~67% of the total variance (Figure 1a, Supplementary Figure S2). We also visualised the antigen expression of 16 surface phenotypic markers across all annotated cell types from HD (Figure 1b). Clear separation of the annotated myeloid and lymphoid cell types on the two-dimensional space was confirmed using t-distributed stochastic neighbour embedding -- t-SNE (Figure 1 c, d).

The annotated cell types from HD were used to train a single-cell hematopoietic classifier. We compared two alternative machine learning (ML) algorithms namely Linear Discriminant Analysis (LDA) and the K-nearest neighbour (KNN) classifier. The performance was evaluated in a cross-validation setting (using 50% of gated single-cells for training and the remaining 50% for testing), and the training/testing procedures were repeated 1,500 times. In our dataset we observed that KNN outperformed LDA and achieved higher average Precision, Recall and F1-score (Supplementary Figure S3). Upon inspection of the misclassified cases, we confirmed that KNN almost always assigned cells to the correct cell gate (Supplementary Figure S4 a, b). This was also the case irrespective of the selection of parameter K of the KNN classifier (Supplementary Figure S4 c). Collectively, the analysis

conducted on a set of seven HD samples, reliably annotated different healthy hematopoietic cell types with high sensitivity, precision, and F1 score.

Single-cell classification captures the phenotypic heterogeneity of leukemia patients

Leveraging prior knowledge of normal/healthy cell populations to organise cells from leukemia patients into meaningful cell populations enables direct comparisons of these populations in healthy data and within patients (8, 17,20).

To implement this technique in our study, we first projected cells from 45 diagnostic leukemia samples to the PCA space constructed in Figure 1a. Cells from patients with leukemia occupied quite similar PCA-space with a heavy skew towards the $CD34^+$ space (Supplementary Figure S5). Then, we subsampled equal amount of single-cells from patients of the Adverse (n=17) and Favorable (n=13) groups based on European LeukemiaNet (ELN 2017) classification. We observed that the trend towards $CD34^+$ space was greater for the Adverse group of patients in our cohort (Supplementary Figure S6).

We used the KNN single-cell hematopoietic classifier to assign each cell from leukemia patients into the most phenotypically similar cell type from HD. We note here that the KNN-predicted cell type of a leukemic cell is the "closest possible cell type". This is a Transfer Learning approach which does not imply that a leukemic cell is equivalent to a healthy cell (8). Once all cells from patients were classified, we compared the leukemic cellular abundances versus the cellular abundances of healthy controls (Figure 2a). Statistically different distributions were observed for HSC, MPP, Granulocytes and B-cells. Interestingly, the shape of the myeloid distributions and the deviation from the observed median, reflected the vast cellular heterogeneity within leukemia patients.

Next, we utilised information from ELN (2017) classification, Remission status, and Treatment data (Supplementary Figure S7), to explore potential associations within the identified cell populations. By comparing cellular abundance distributions, we found that the Monocytic population was significantly larger in the Favorable group compared to the Adverse. Considering the Remission status, we observed significant differences for both HSC, and Monocytic populations (Supplementary Figure S8). We also compared abundance distributions between mutated and wild-type patients for known hotspot mutations such as DNMT3A, IDH1/2, TET2, SRSF2 and NPM1 identified by Illumina Trusight targeted sequencing (22). Differences were observed for B-cells between DNMT3A wild-type/mutated patients, as well as for HSC, MPP, and T-cytotoxic cells between NPM1 wild-type/mutated patients (Supplementary Figure S9). Lower abundance of HSC/MPP populations for the NPM1-mutated patients, might be associated with the fact that NPM1-mutated cases are often $CD34^-$, and thus additional markers are required to truly quantity these cell populations.

Considering karyotype, differences in the cellular abundance distributions were observed for HSC, and T-cytotoxic cells between patients with normal and complex karyotypes. Interestingly, patients with FLT3-ITD mutations presented significantly lower abundance of T-cytotoxic and T-helper cells, whereas patients harbouring inv(16) translocation showed higher Monocytic, and lower Granulocytic abundances compared to patients without translocation (Figure 2b and Supplementary Figure S10). However, due to the small cohort size the statistical power is limited. Thus, the results must be generalised with caution, as after stratification and comparisons between groups, some distributions included only 5-8 data points.

To investigate possible associations with patients' survival, the frequency of cells in each population was used as input for multivariate Cox proportional hazard modelling. Within our cohort, patients with higher T-helper abundance at time of diagnosis showed survival benefit compared to patients with lower T-helper abundance. (Figure 2c) However, univariate Kaplan-Meier survival analysis did not confirm the same results (log rank test p-value=0.15). Taking together, the deployed ML-based framework allowed prediction of lymphoid and myeloid cell types opening avenues for more advanced cell type-specific downstream analyses via ML approaches.

A novel feature engineering approach to characterise signalling dynamics of cell populations

Once classified, single-cell populations could be used to minutely analyse cell type-specific signalling states. However, analysing single-cell phospho-signalling data with conventional approaches (e.g., mean/median/quantile expression values) is challenging. Technical noise, natural stochasticity, and many other factors collectively have confounding effects on the analysis of signalling markers.

To alleviate this problem, Conditional Density Resampled Estimate of Mutual Information known as DREMI (19) has been successfully applied to CyTOF data to quantify single-cell relationships between pairs of signalling proteins (see Methods).

To build upon the developed single-cell classifier, we used DREMI to engineer cell type-specific features suitable for ML modelling. As a proof-of-principle, in our cohort we engineered 182 features from 14 proteins considering both signalling directions (e.g. X → Y and Y → X), and across seven of the studied cell types. The NK population was discarded from this analysis due to low overall abundance (less than 500 cells) in more than 50% of patients.

We utilised 5-year survival information to stratify leukemia patients into two groups (Supplementary Figure S7) that we hereafter refer to as "Long Term Survivors" (LTS, n=15), with survival equal/greater than 60 months after diagnosis, and "Short Term Survivors" (STS, n=30) with survival shorter than 60 months after diagnosis. The LTS group had median age of 49 years, whereas

the STS group had median age of 63 years. In addition, 12 out of 30 STS patients were transplanted, versus 10 out of 15 patients of the LTS group (Fisher's exact test p-value=0.12).

Figure 3 summarises the average DREMI scores for LTS and STS patients across the studies cell types. It became evident that ~40% of DREMI feature values were close to zero indicating almost no interaction between the respective proteins. Using unsupervised hierarchical clustering we found that DREMI scores were highly cell type-specific, and the same cell type from different patients clustered together irrespective of the patients' survival status. The exception was HSC and MPP populations, where DREMI profiles were more variable, and clustered separately.

Application of machine learning enables prediction of patients' survival

Based on the previous observation we hypothesized that cell type-specific DREMI scores at the time of diagnosis might be sufficient to predict short-term survival. We deployed ML to investigate this hypothesis, and here as an example we present a biomarker discovery application based on CyTOF signalling data.

The problem of predicting STS from LTS patients was formulated as a binary classification task. Figure 4a summarises the flowchart of the deployed framework. After single-cell classification and cell type-specific feature engineering, DREMI features (equally denoted scores) were ranked using Minimum Redundancy Maximum Relevance (mRMR) criterion (25). To train models and make predictions, we applied LASSO with ranked features. The proposed machine learning procedure was repeated multiple times, for top-K features starting with K=5,10,15 and up to 182. To ensure reproducibility the training/testing step was repeated 500 times. During each iteration, the samples (n=45) were divided randomly into two disjoint sets with replacement, one for training (50% of all available instances) and one for testing (remaining 50% of all available instances). In total, 126,000 individual LASSO models were trained and tested (i.e., 7 cell types x 36 ranked feature sets x 500 repetitions)

We assessed the performance of the developed cell type-specific models on the testing sets using standard performance metrics (see Methods). We also detected the subsets of DREMI scores that maximised classification performance (26). An overview of the classification performance across different cell types is depicted in Supplementary Figure S11. In summary (Figure 4b), LASSO predicted survival using cell type-specific signalling data with relatively high performance (overall average greater than 0.7). The MPP cell type achieved the best results (overall average above 0.8) using top-5 ranked features namely the DREMI scores of pCREB-pSTAT5, CyclinB1-pCREB, pP38-pAxl, pSTAT5-pP38 and pP38-pSTAT5 proteins (Figure 4c).

Upon inspection of the DREMI scores, we observed that the LTS group showed higher values compared to STS, and the distributions were different based on Wilcoxon rank sum test. Based on this observation we hypothesised that the predictions could be used for survival analysis using other state-of-the-art statistical approaches. To do so, we focused on MPP cells, and stratified the samples (n=36 after discarding patients with less than 500 cells) to low and high groups of DREMI scores. We utilised Kaplan-Meier univariate survival analysis and confirmed survival benefit for patients in our cohort with higher pCREB-pSTAT5 and pP38-pSTAT5 DREMI scores (Figure 5 a-c). Multivariate Cox proportional hazard modelling confirmed our findings only for pCREB-pSTAT5 DREMI score (Figure 5d).

Validation of machine learning-based findings using an external cohort of leukemia patients

To further validate our previous findings, we utilised publicly available CyTOF samples from Levine et al. (14). This dataset focuses on pediatric AML and although does not match our cohort characteristics, it explores a broader spectrum of AML cases. We utilised pre-treatment data from 13 patients with available survival information. To investigate similar MPP-like cells as defined in our Supplementary Figure S1, we downloaded manually gated $CD34^+$ cells, that were further divided on the basis of CD38 expression. On the refined set of gated cells, we computed DREMI scores for pCREB-pSTAT5, pSTAT5-pP38 and pP38-pSTAT5. Figure 5e shows that the selected DREMI scores were higher on average for the survivors, although the results did not reach significance, as the statistical power was limited due to the small cohort size (i.e., seven patients reported still alive, and six not).

DREMI signalling dynamics define distinct survival subgroups via unsupervised clustering

We investigated whether the identified DREMI scores of MPP cells could be used to stratify patients into different subgroups of survival. For this purpose, we utilised unsupervised clustering based on the ConsensusClusterPlus algorithm (27) and we clustered the patients based on selected MPP-derived DREMI scores. After tuning the number of clusters, we detected five clusters, of which two clusters contained only one patient each. Figure 6a shows a t-sne projection of the patient data. Association between clusters and % MPP cellular abundance was not detected, indicating that DREMI scores were not influenced by the size of MPP population.

We performed Kaplan Meier analysis for clusters C1, C2 and C3. Cluster C2 showed the worst survival and grouped together patients with median survival of 8 months after initial diagnosis (cohort median survival is 15 months). One possible explanation might link the identified DREMI scores of patients in C2, with increased heterogeneity within their MPP population. It has been reported that increased DREMI scores are in principle higher within homogeneous cell populations compared to heterogeneous cell populations considering the same pair of signalling proteins (19). This in turn might

reflect more complex clonal/sub-clonal architectures for patients of C2 at time of diagnosis, which could explain the lower values of DREMI scores compared to other patients. Following this rational, we sought to associate genetic information with signalling states of the identified patient subgroups. Focusing on C2 cluster (Figure 6c), we could not link specific mutations to the signalling profiles of co-clustered patients. Interestingly, only one patient out of 14 in C2 harboured NPM1 mutation that was in combination with FLT3-ITD mutation. However, the result for NPM1, and all other mutations presented in Figure 6c, did not reach statistically significant levels compared to the other clusters using Fisher's exact test.

In summary, the computational framework presented here allowed for unbiased stratification of patients, by targeting single-cell interactions between phospho-proteins of MPP cells. And this provides an enhanced stimulus for the development of more effective patient stratification methods based on single-cell omics. Importantly, our results could be exploited in the future, to investigate clonal heterogeneity within MPP or other cell populations. Our framework adds a new dimension to the analysis of cell phenotypes from CyTOF, with the possibility to link signalling states with genomic aberrations and possibly the degree of clonal heterogeneity.

DREMI scores synergise with existing patient risk stratification methods

Current patient stratification methods in leukemia utilise combinations of clinical data with genetic information at the time of diagnosis. Such methods guide clinical decisions, and thus it is of great importance to optimise existing risk stratification approaches. The survival and ELN (2017) classification results for 31 patients (i.e., data from 14 patients were removed due to non-available ELN information, or DREMI quantification for MPPs) in our cohort is depicted in Figure 7a, b. For those patients, previous analyses focusing on MPP cells identified salient DREMI scores namely pCREB-pSTAT5 and pP38-pSTAT5 that were associated with survival benefit. Here, we investigated whether it could be possible to improve the existing ELN scheme by adding signalling information from single-cell omics. To do so, patients of the Favorable and Intermediate/Adverse groups (Intermediate and Adverse groups were merged due to the small sample size) were subdivided given the levels of pCREB-pSTAT5 and pP38-pSTAT5 DREMI scores. Interestingly, patients of the Favorable ELN group that exhibited high levels of pCREB-pSTAT5 and pP38-pSTAT5 scores achieved the highest survival probability. Low DREMI scores were associated with poor survival irrespective of the Favorable ELN status. Our results suggest that single-cell signalling omics could complement the ELN criteria paving the way for improved risk stratification at time of diagnosis.

Discussion

In this study we deployed a bioinformatics framework which allows for automated cell type annotation and systematic exploration of interactions between signalling proteins in a CyTOF antibody panel.

To fully exploit single-cell signalling data, we introduced a novel feature engineering technique based on DREMI scores. Importantly, the proposed feature engineering technique does not rely on pre-defined signalling network topologies that are frequently altered in disease conditions. Instead, it can be applied to all possible pairwise interactions (e.g., edges in a signalling network topology) of the available signalling proteins in the antibody panel. Our approach provides a unique opportunity to discover novel disease-related protein associations, which could be predictive of the disease outcome.

We tested the deployed method using a cohort of 45 leukemia patients. Single-cell classification combined with DREMI-based feature engineering fuelled the development of predictive models of survival using LASSO regression. Within our cohort, modelling with MPP-like cells achieved the highest classification performance. Our experimentation with different variables revealed that MPP-derived DREMI scores between pCREB-pSTAT5 and pP38-pSTAT5 proteins, could predict short-term survival with high level of success. Findings made by LASSO were tested independently using Cox proportional hazards modelling and Kaplan-Meier survival analysis. Similar trends were also observed using an independent cohort of 13 pediatric AML cases. Importantly, the tools and the analysis paradigm presented here, could be useful in the future to investigate mechanisms of therapy resistance and possibly identify vulnerabilities that could serve as novel therapeutic targets.

Focusing on technical aspects, the effectiveness of the developed single-cell classifier strongly depends on the quality of reference samples. Similar to other studies, our experimentation convincingly demonstrates that it is feasible to use reference cell types (e.g., from healthy donors) as "anchors" for automated cell type annotation in disease conditions. In the future, more work is required towards improving the performance of automated cell type annotation. Combining state-of-the-art clustering algorithms, gating strategies and ranked nearest projections using multimodal single-cell data would be one possible direction to explore (20).

There is also room for improvements targeting the ML modelling example presented in our study. Similar to other studies, the limited size cohort restricts the generalisation of findings. Thus, more work is required to minimise overfitting, and to assess the validity of findings in larger cohorts of patients. For example, it will be interesting to extend the proposed methodology to larger cohorts of leukemia patients with compatible CyTOF antibody panels, and/or to compare our performance with results derived from other independent ML methods. However, we note the primary scope of this work is the cell annotation and feature engineering components for CyTOF data, and not the ML-

based modelling of leukemia patients. It is also worth noting that the developed framework, although it has been originally designed and tested to the analysis of leukemia patients, from which 32 are AML cases, it can in principle be applied to different diseases as well. In addition, with different antibody panels (e.g., surface phenotypic markers and/or signalling phosphoproteins), we anticipate that the proposed computational framework could easily be extended to investigate different cell sub-populations and signalling proteins (28).

Focusing on data interpretation aspects, the development of predictive models could benefit from integration of additional single-cell epigenomic, or transcriptomic data. This extension could also improve our understanding of biological mechanisms that control survival of cells, and it might be informative for signalling pathways (i.e., signalling relationship), the effect of genetic aberrations (e.g., CNV or SNVs) and their underlying developmental and proliferation states. In addition, developing predictive models of relapse using data collected after therapy initiation seems a powerful option that we could complement the ML-based results described in ref. (22).

Collectively, our study streamlined the application of ML on CyTOF data to annotate cell populations and engineer features for biomarker discoveries. We believe that the utilisation and exploration of single-cell data, as described here, could provide a more general bioinformatics framework with many applications in future personalised medicine projects.

Methods

Patients' material and healthy donor samples

The primary data presented in this study are publicly available from ref. (22). Briefly, all patients were diagnosed and treated at Haukeland University Hospital, Bergen, Norway and Oslo university Hospital, Rikshospitalet, Norway. The study was performed in accordance with the Declaration of Helsinki, and all samples were collected following written informed consent. All patients were above 16 years of age. The biobank and the clinical protocols were approved by the ethical committee at the University of Bergen (Ethical approval REK Vest 2012/2245, 2012/2247), Norwegian Medicines Agency and The Data Inspectorate. All 45 patient samples were collected from peripheral blood at time of diagnosis. The patient cohort included: 38 AML patients, two acute promyelocytic leukemia (APL) patients, one myelodysplastic syndrome (MDS) patient, three B-cell acute lymphoblastic leukemia (B-ALL) patients and one biphenotypic AML patient. More details about the patient cohort characteristics can be found in the original publication of this dataset and a summary is presented in Supplementary Figure S7. For the development of the single-cell-classifier, four peripheral blood samples and three bone marrow samples from healthy donors were collected and analysed.

Mass cytometry experiments and data pre-processing

The antibody panel used in this study was originally designed and validated from (22). The panel enables classification of cells into cell types by including antibodies for selected surface markers – and to be able to map the status and relationships of important signalling proteins. Here, we exploited subsets of the available markers – to enable predicting survival – by looking at active signalling interactions within particular cell types (or proxies thereof) in diagnostic samples. Specifically, we selected sixteen surface phenotypic markers to define eight major cell types in HD and leukemia patients. This set included antibodies: CD123, CD14, CD16, CD20, CD3, CD33, CD34, CD38, CD4, CD45, CD56, CD64, CD66b, CD7, CD8 and HLA-DR). In addition, we selected 14 intracellular markers were studied to measure intracellular activation/phosphorylation of signalling networks namely pNFkB, pErk, pSTAT1, pSTAT3, pSTAT5, pP38, pHistone3, pRB, CASP3, pAxl, pAkt, pS6, CyclinB1, pCREB. Data pre-processing and normalisation of all CyTOF samples is described in (22). The Helios mass cytometer (Fluidigm) was used for profiling, and all experiments were conducted at the Flow Cytometry Core Facility of the University of Bergen.

Semi-automated gating strategy

A gating strategy was implemented (Supplementary Figure S1), on the full set of cells from healthy donors (~1.2 million cells) using the flowDensity software (23). FlowDensity, is a supervised clustering algorithm based on density estimations that is designed to overcome the tedious work of setting static gates. All healthy single cells were first gated on the basis of CD45 to identify myeloid and lymphoid compartments. Cells from the lymphoid compartment were subsequently gated on the basis of CD3 and CD20 markers to identify representative B and T cells for annotations. Note, that during this gating procedure we selected only a fraction of all available cells within each compartment to improve quality of labeled instances for supervised learning based on experts' knowledge. Selected cells with CD3-high expression were subsequently gated on the basis of CD4, CD7 and CD8a markers to identify representative cells of T-cytotoxic and T-helper subcategories. Selected cells of CD3-low and CD20-low expression were further gated on the basis of CD56, and use of CD7 enabled detection of representative Natural Killer (NK) cells.

Cells from the myeloid compartment were first gated on the basis of CD34 and CD38 markers to discriminate major myeloid compartments namely: CD34-low/CD38-low for Granulocytes and CD34-low/CD38-high for Monocytes. Healthy cells of CD34-high/CD38-high expression was defined as Multipotent Progenitor cells, and in cases of patients with leukemia the notation describes MPP-like blasts. Healthy cells of the CD34-high/CD38-low were defined as Hematopoietic Stem Cells (HSC), and in cases of patients with leukemia notation describes HSC-like blasts. We also note that since we are

analysing cells from HD samples, and based on the deployed sets of rules, HSC and MPP cells were quite rare resulting only into ~120 single cells per category.

Single-cell classifier with LDA and KNN

The selected cells from seven donors were gated into eight cell types describing normal hematopoiesis. A consolidated expression profile of each cell type was compiled using the expression of 16 surface phenotypic markers as shown in Figure 1b. Annotated cells from the aforementioned gating strategy, were used to train a single-cell classifier to predict cell types in unseen CyTOF data. In our cohort we tested two alternative approaches namely LDA and KNN.

Given an unseen cell named c1, a single-cell classifier makes a prediction by assigning c1 to cell type Ci (i.e., in our case i $\in$ [HSC, MPP, Granulocytes, Monocytes, NK, B-cells, T-cytotoxic, T-helper]). LDA considers the posterior probability of c1 to be maximized, across the studied cell types. KNN on the other hand, computes the distance of c1 to k nearest neighbors, and applies a majority voting technique to assign cell types. Selected annotated healthy cells were used to benchmark LDA and KNN based on 2-fold cross validation. A set of 582,010 annotated cells was used to produce random training (50%, n= 291,005) and testing (50%, n= 291,005) sets with replacement, and the procedure was repeated 1,500 times to ensure reproducibility. The original publication by Abdelaal et al. applied a posterior probability threshold to reduce misclassification error of LDA. However, in our study we did not apply this technique. To fine-tune parameter K of the KNN classifier, we performed optimizations with different values of K using 20% of the training data. After comparing the performance of LDA and KNN, we decided to use KNN, and a final hematopoietic classifier was developed using the full set of 582,010 annotated healthy cells.

This classifier was used as a reference to annotate cells from leukemia patients. The expression profiles of the same 16 surface phenotypic markers were used as input. Each cell from patients with leukemia was assigned to its predicted cell type as defined from HD cells.

Feature engineering using DREMI scores

A simple approach to quantify signalling responses of cell types could average the intensity values of phosphoproteins in the antibody panel, and generate feature vectors. However, this approach collapses the rich single-cell CyTOF data into a single point (e.g., mean, median, 75th quantile), and neglects the shape of the entire single-cell distribution of the phosphoprotein intensity (14). In addition, technical noise, natural stochasticity, and many other factors collectively have confounding effects on the analysis of signalling data using single point statistics (19).

To mitigate these problems, several information theory methods (e.g., SARA, DREMI) have been developed. Such approaches consider the full shape of single-cell distributions, and provide better understanding on how cells process and transmit signals in a network. Here, we utilized DREMI to engineer cell type-specific feature vectors suitable for ML-based modelling. A key conceptual advantage of DREMI compared to other approaches, is that DREMI computes mutual information based on resampled conditional density estimations (19,29). Given proteins X and Y in a network/pathway with the edge X→Y, DREMI quantifies how the state of Y varies with different states of X, which is different from describing the density of cells states X and Y in a conventional scatter plot or in a joint density plot. In other words, DREMI reveals how protein Y changes as a function of the activity of protein X, providing a quantitate score to characterise the "strength" of this signalling relationship.

Generating feature vectors using DREMI, could target few proteins that are known to interact in a signalling cascade, or in an agnostic manner by evaluating all possible pairwise combinations of proteins in the antibody panel. The latter approach was applied in our study, considering both signalling protein directions (e.g. X → Y and Y → X), and across the studied cell types. We applied a minimum number of n=500 cells for DREMI estimation per cell type, and thus we excluded cases with low cellular abundance. DREMI was implemented originally in MATLAB, but the codes were re-written in R and incorporated within the developed computational framework.

Cell type-specific ML modelling

Cell type-specific ML models of survival were developed using the cohort of 45 diagnostic samples (n=15 LTS and n=30 STS). ML predictions were derived using the LASSO technique, implemented via the glmnet R package. To identify smaller subsets of predictive features, LASSO was combined with mRMR feature rankings. We note that the feature selection problem in bioinformatics is very well studied and it is well documented that feature selection is a strongly 'data-dependent' process (26). Different feature selection methods have been proposed, including heuristic approaches that have higher algorithmic complexity and exponential computational cost. In contrast, here we used LASSO for its ability to penalize features and reduce overfitting, in combination with mRMR a powerful yet simple feature filtering method.

In summary the developed ML framework works as follows: First, single-cells from patients with leukemia, were annotated using the single-cell hematopoietic classifier that was developed using gated healthy cells. Next, each cell type was processed separately, and cell type-specific features were generated using all possible pairwise DREMI scores of proteins in the antibody panel. The features were ranked with mRMR, and LASSO was trained multiple time, each time taking as input top-K ranked

features. We used K=5, 10, and up to 182, as it provides good compromise between execution time and high-resolution feature searching. Random splitting of the data, to training (50%) and testing (50%) subsets with replacement was performed, and the procedure was repeated 500 times to account for reproducibility. The performance for every cell type and top-K ranked input subset was recorded in the testing set. For performance evaluation we assessed the Area Under the Curve (AUC), Sensitivity, Specificity, Precision/Positive Predictive Value (PPV) and F1-score (F1), with known formulas that can be found elsewhere. At the end, this approach allowed us to identify cell type-specific top-K ranked DREMI scores, achieving the highest (on average) classification performance. Individual samples with less than 500 cells in a particular cell type, were discarded for ML modelling, as DREMI was not possible to be computed. A cell type was completely discarded from LASSO modelling, if more than 50% of samples (n=22) had less than 500 cells.

Statistical analysis and modelling of survival

Data analysis was implemented in the R package. To test statistically significant differences between abundance distributions of different groups, we used Mann-Whitney U test. Patients were stratified to low and high groups of DREMI using the median DREMI score per signalling relationship, estimated from the full cohort of patients.

Kaplan-Meier analysis was used to estimate survival rates. Differences between different groups of survival were assessed using log-rank test. The Cox proportional hazards model was used to investigate the association between the survival time of patients and one or more predicted relationships between signalling proteins. The assumptions of Cox proportional hazards modeling were testing using statistical tests and graphical diagnostics based on the scaled Schoenfeld residuals.

For unsupervised clustering the ConsensusClusterPlus algorithm was used. ConsensusClusterPlus provides quantitative and visual stability evidence for estimating the number of unsupervised classes in a dataset. The software generates Empirical cumulative distribution function (CDF) plots to find the number of clusters at which the distribution reaches an approximate maximum, which indicates a maximum stability and after which divisions are equivalent to random picks rather than true cluster structure.

Acknowledgements

The authors thank nurses and physicians in Bergen and Oslo, the pateint's organization Blodkreftforeningen for following and supporting the project, personnel at the Flow Cytometry Core Facility at the University of Bergen and personnel at the Gjertsen- and Enserink labs for their skillful

technical assistance. A special thanks to the participating patients and their families who suffered from leukemia and nonetheless gave the gift of participation so that others might benefit.

Funding: This work was supported by grants provided by EU ERA PerMed AML_PM (RCN Grant no. 298842), the Norwegian Cancer Society with Solveig & Ole Lunds Legacy, Øyvinn Mølbach-Petersens Fund for Clinical Research (grant nos. 303445 and 190175), Western Norway Health Authorities (grant no. 912308 and F-12148).

Author Contribution

B.S.T, M.H, O.F, I.K.F.M and S.E.G designed the experiments and performed the experiments; D.K developed the algorithm and analysed the data; B.S.T, A.L, Y.F, and B.T.G collected samples and provided clinical input; B.S.T, N.V.D.M, E.G, J.J.S and B.T.G provided intellectual input and help edit the manuscript; D.K conceived the idea and led the project; D.K, and I.J wrote the manuscript.

Declaration of interests

The authors declare they have no competing interests.

Consent for publication

Patients have consented that mass cytometry data with no personal identifiers can be used for publication in an anonymized format.

Code and data availability

The data analysis framework has been implemented using the R language. The codes along with a tutorial is available in GitHub https://github.com/dkleftogi/singleCellClassification. The CyTOF datasets presented in this study are available at https://flowrepository.org/ with ID: FR-FCM-Z5YT, and more details can be provided upon reasonable request.

References

1. Bendall, S.C., Davis, K.L., Amir, E.D., Tadmor, M.D., Simonds, E.F., Chen, T.J., Shenfeld, D.K., Nolan, G.P., and Pe'er, D. (2014). Single-Cell Trajectory Detection Uncovers Progression and Regulatory Coordination in Human B cell Development. Cell 157, 714–725.
2. Spitzer, M.H., and Nolan, G.P. (2016). Mass Cytometry: Single Cells, Many Features. Cell 165, 780–791.
3. Baharlou, H., Canete, N.P., Cunningham, A.L., Harman, A.N., and Patrick, E. (2019). Mass Cytometry Imaging for the Study of Human Diseases—Applications and Data Analysis Strategies. Frontiers in Immunology 10.
4. Welch, J.S., Ley, T.J., Link, D.C., Miller, C.A., Larson, D.E., Koboldt, D.C., Wartman, L.D., Lamprecht, T.L., Liu, F., Xia, J., et al. (2012). The origin and evolution of mutations in acute myeloid leukemia. Cell 150, 264–278.

5. Papaemmanuil, E., Gerstung, M., Bullinger, L., Gaidzik, V.I., Paschka, P., Roberts, N.D., Potter, N.E., Heuser, M., Thol, F., Bolli, N., et al. (2016). Genomic Classification and Prognosis in Acute Myeloid Leukemia. New England Journal of Medicine 374, 2209–2221.
6. Skanderup, A.J., and DasGupta, R. (2021). Targeting the developmental origins of cancer. Nature Cancer 2, 256–257.
7. Carter, J.L., Hege, K., Yang, J., Kalpage, H.A., Su, Y., Edwards, H., Hüttemann, M., Taub, J.W., and Ge, Y. (2020). Targeting multiple signaling pathways: the new approach to acute myeloid leukemia therapy. Signal Transduction and Targeted Therapy *5*, 1–29.
8. 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, 474–483.
9. Georgopoulou, D., Callari, M., Rueda, O.M., Shea, A., Martin, A., Giovannetti, A., Qosaj, F., Dariush, A., Chin, S.-F., Carnevalli, L.S., et al. (2021). Landscapes of cellular phenotypic diversity in breast cancer xenografts and their impact on drug response. Nature Communications 12, 1998.
10. 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. Sci Rep 7.
11. Liu, K., Gao, R., Wu, H., Wang, Z., and Han, G. (2021). Single-cell analysis reveals metastatic cell heterogeneity in clear cell renal cell carcinoma. J Cell Mol Med 25, 4260–4274.
12. Miles, L.A., Bowman, R.L., Merlinsky, T.R., Csete, I.S., Ooi, A.T., Durruthy-Durruthy, R., Bowman, M., Famulare, C., Patel, M.A., Mendez, P., et al. (2020). Single-cell mutation analysis of clonal evolution in myeloid malignancies. Nature 587, 477–482.
13. Stanley, N., Stelzer, I.A., Tsai, A.S., Fallahzadeh, R., Ganio, E., Becker, M., Phongpreecha, T., Nassar, H., Ghaemi, S., Maric, I., et al. (2020). VoPo leverages cellular heterogeneity for predictive modeling of single-cell data. Nature Communications 11, 3738.
14. Levine, J.H., Simonds, E.F., Bendall, S.C., Davis, K.L., Amir, E.D., Tadmor, M.D., Litvin, O., Fienberg, H.G., Jager, A., Zunder, E.R., et al. (2015). Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell 162, 184–197.
15. Liu, X., Song, W., Wong, B.Y., Zhang, T., Yu, S., Lin, G.N., and Ding, X. (2019). A comparison framework and guideline of clustering methods for mass cytometry data. Genome Biology *20*, 297.
16. Lee, H.-C., Kosoy, R., Becker, C.E., Dudley, J.T., and Kidd, B.A. (2017). Automated cell type discovery and classification through knowledge transfer. Bioinformatics 33, 1689–1695.
17. Alpert, A., Nahman, O., Starosvetsky, E., Hayun, M., Curiel, T.J., Ofran, Y., and Shen-Orr, S.S. (2022). Alignment of single-cell trajectories by tuMap enables high-resolution quantitative comparison of cancer samples. Cell Systems 13, 71-82.e8.
18. Abdelaal, T., Unen, V. van, Höllt, T., Koning, F., Reinders, M.J.T., and Mahfouz, A. (2019). Predicting Cell Populations in Single Cell Mass Cytometry Data. Cytometry Part A 95, 769–781.
19. Krishnaswamy, S., Spitzer, M.H., Mingueneau, M., Bendall, S.C., Litvin, O., Stone, E., Pe'er, D., and Nolan, G.P. (2014). Conditional density-based analysis of T cell signaling in single-cell data. Science 346.
20. Triana, S., Vonficht, D., Jopp-Saile, L., Raffel, S., Lutz, R., Leonce, D., Antes, M., Hernández-Malmierca, P., Ordoñez-Rueda, D., Ramasz, B., et al. (2021). Single-cell proteo-genomic reference maps of the hematopoietic system enable the purification and massive profiling of precisely defined cell states. Nat Immunol 22, 1577–1589.
21. van Unen, V., Höllt, T., Pezzotti, N., Li, N., Reinders, M.J.T., Eisemann, E., Koning, F., Vilanova, A., and Lelieveldt, B.P.F. (2017). Visual analysis of mass cytometry data by hierarchical stochastic neighbour embedding reveals rare cell types. Nat Commun *8*, 1740.
22. Tislevoll, B., Hellesøy, M., Fagerholt, O., Gullaksen, S.-E., Srivastava, A., Birkeland, E., Kleftogiannis, D., Ayuda-Durán, P., Piechaczyk, L., Tadele, D., et al. (2023). Early response evaluation by single cell signaling profiling in acute myeloid leukemia. Nat Commun 14, 115.
23. Malek, M., Taghiyar, M.J., Chong, L., Finak, G., Gottardo, R., and Brinkman, R.R. (2015). flowDensity: reproducing manual gating of flow cytometry data by automated density-based cell population identification. Bioinformatics *31*, 606–607.
24. Laurenti, E., and Göttgens, B. (2018). From haematopoietic stem cells to complex differentiation landscapes. Nature *553*, 418–426.
25. Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. IEEE Trans Pattern Anal Mach Intell 27, 1226–1238.

26. Kleftogiannis, D., Ashoor, H., and Bajic, V.B. (2018). TELS: A Novel Computational Framework for Identifying Motif Signatures of Transcribed Enhancers. Genomics Proteomics Bioinformatics 16, 332–341.
27. Wilkerson, M.D., and Hayes, D.N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics 26, 1572–1573.
28. de Boer, B., Prick, J., Pruis, M.G., Keane, P., Imperato, M.R., Jaques, J., Brouwers-Vos, A.Z., Hogeling, S.M., Woolthuis, C.M., Nijk, M.T., et al. (2018). Prospective Isolation and Characterization of Genetically and Functionally Distinct AML Subclones. Cancer Cell 34, 674-689.e8.
29. Qin, P., Pang, Y., Hou, W., Fu, R., Zhang, Y., Wang, X., Meng, G., Liu, Q., Zhu, X., Hong, N., et al. (2021). Integrated decoding hematopoiesis and leukemogenesis using single-cell sequencing and its medical implication. Cell Discovery 7, 1–17.

Main Figures

Figure 1: Phenotypic characterisation of hematopoietic cells using CyTOF. a) PCA analysis of 30,000 randomly selected cells from seven healthy donors. The cells were gated using a semi-automated gating strategy based on flowDensity (Supplementary Figure S1). The principal components were constructed using 16 surface phenotypic markers. In total, the first two components explained ~67% of the variance. For each cell population, we report the centroids of the population, which is the mean X and Y coordinate position. b) Distribution of antigen expression in gated cells from seven healthy donors. First a set of 582,010 cells were selected and annotated using gating rules, then they were used for training and testing models. The raw expression values of 16 markers were first arsinh-transformed. For visualization purposes, the arsinh-transformed values were scaled between 0 and 1 using low (1%) and high (99%) percentiles as boundaries. The x-axis shows the density of the scaled expression. c), d) Visualization of the annotated myeloid and lymphoid healthy cell populations via t-SNE**.** Single-cells were gated based on the same gating strategy presented in Supplementary Figure S1, and were projected into the 2D space using the same set of 16 surface phenotypic markers.

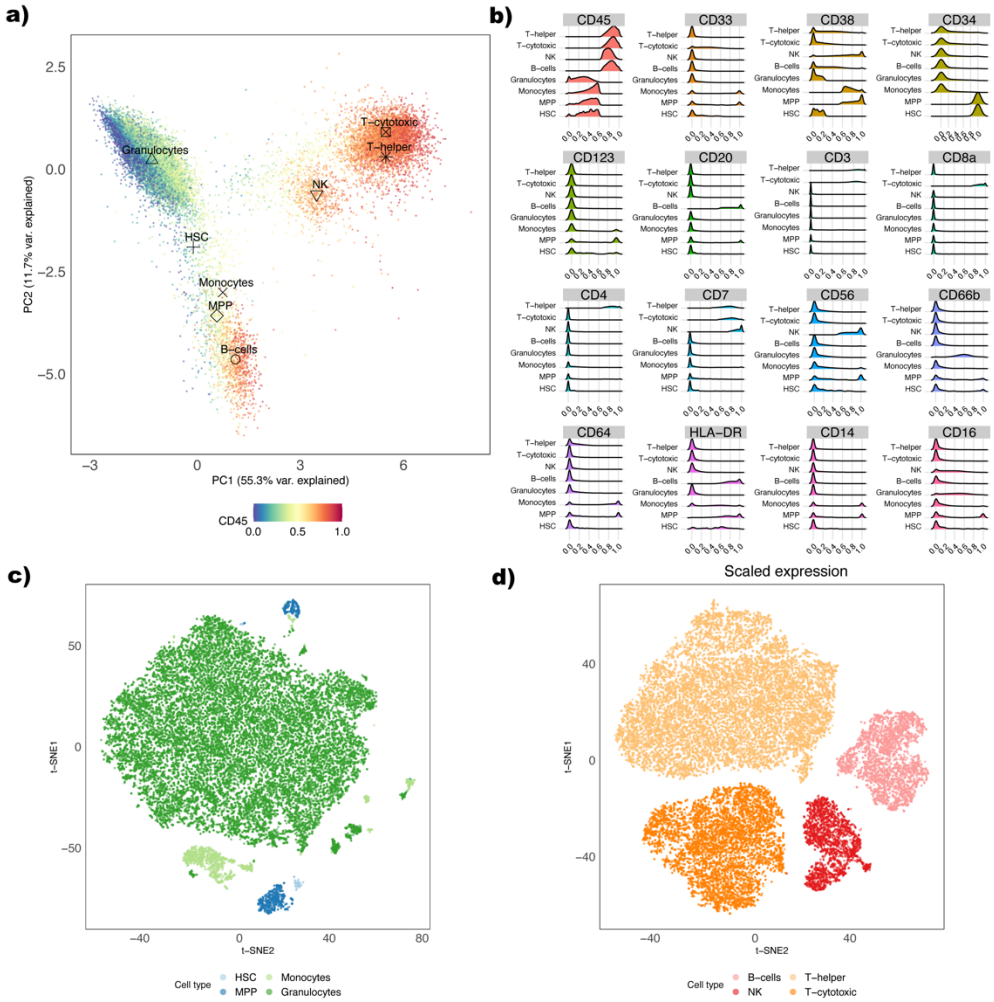


Figure 2: Single-cell classification of cells from leukemia patients allows for cellular abundance comparisons with healthy control. a) Percentages (%) of annotated cells after single-cell classification at time of diagnosis. The boxplots show the cellular abundance of 45 patients and seven healthy donors. b) Percentages (%) of annotated T-cytotoxic and T-helper cells. The boxplots show the cellular abundance of patients with FLT3-ITD mutation present (n=8) and absent (n=37). c) Percentages (%) of annotated Monocytes and Granulocytes. The boxplots show the cellular abundance of patients with inv16 translocation present (n=6) and absent (n=35). c) Cox proportional hazards modelling using as input the cellular abundance (%) measured at time of diagnosis. Statistically significant differences were tested using Mann-Whitney U test. The levels of significance are: * for p<0.05, ** for p<0.01 and *** for p<0.001, ns is reported for non-significant results.

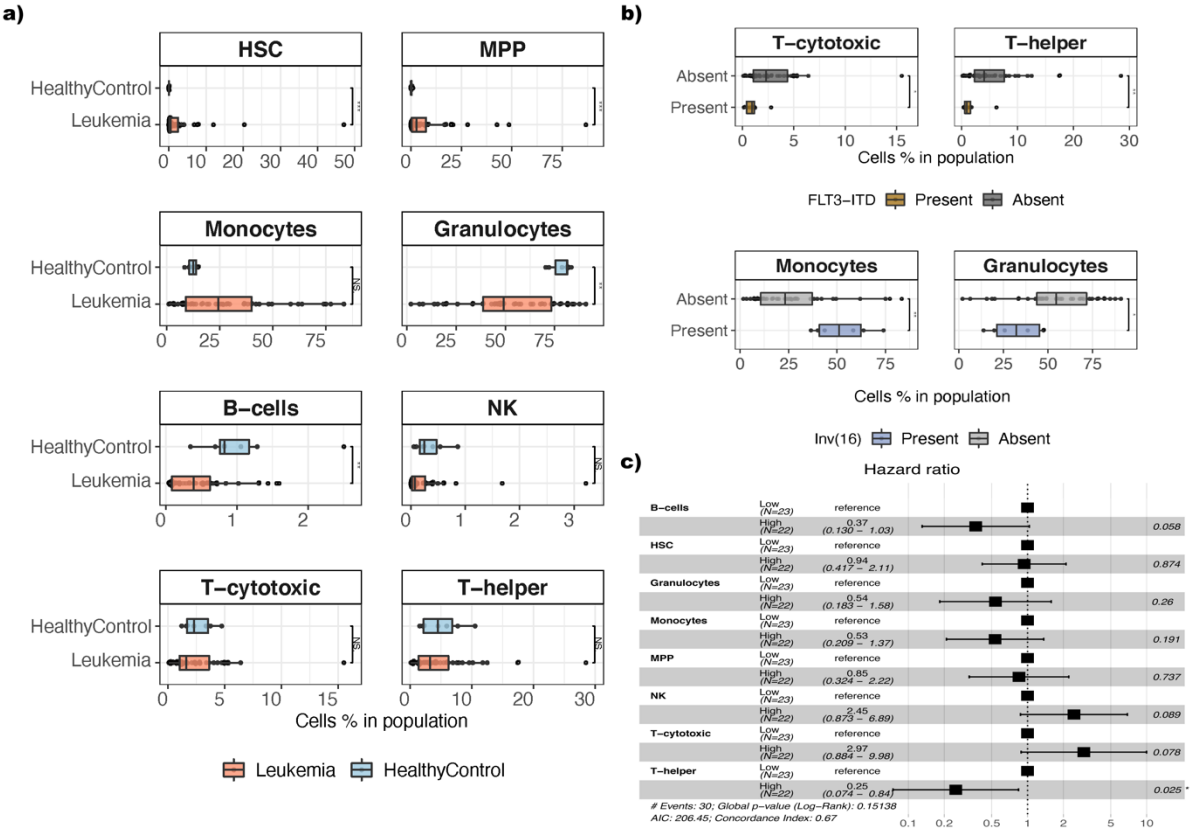


Figure 3: DREMI signalling profiles are highly cell type-specific. Heatmap showing cell type-specific DREMI scores for LTS and STS patients. For each cell type and group of survival the average DREMI score was computed, and used for hierarchical clustering on the rows and columns. The x-axis shows 184 signalling interactions between proteins.

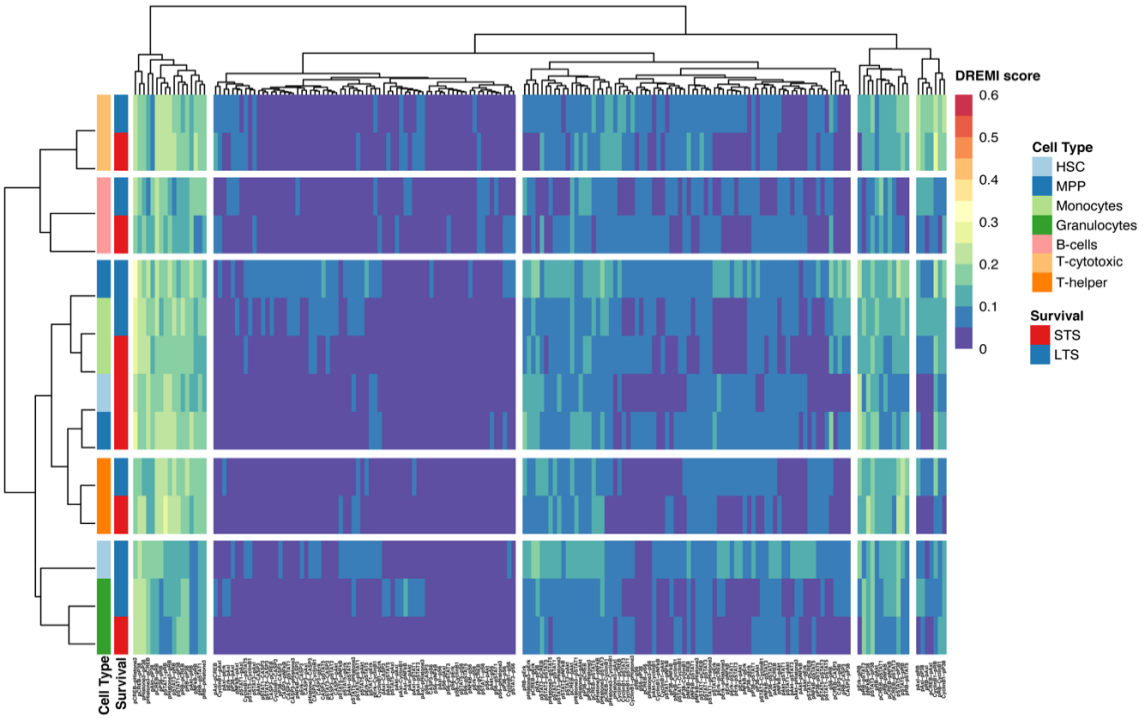


Figure 4: Example of Machine Learning-based modelling to identify cell type-specific predictors of survival. a) The flowchart of the developed ML method. Single-cell data from 45 individuals were annotated using single-cell classification. Cell type-specific features (n=182) were computed using DREMI. Features were ranked using mRMR, and LASSO was used for training/testing of models of survival with feature selection. b) The average classification performance achieved for top-5 ranked features using different metrics across and across the studied cell types. c) Top-5 selected features computed from MPP cells. The boxplots show the DREMI scores of LTS and STS patients. Statistically significant differences were tested using Mann-Whitney U test. The levels of significance are: * for p<0.05, ** for p<0.01 and *** for p<0.001, ns is reported for non-significant results.

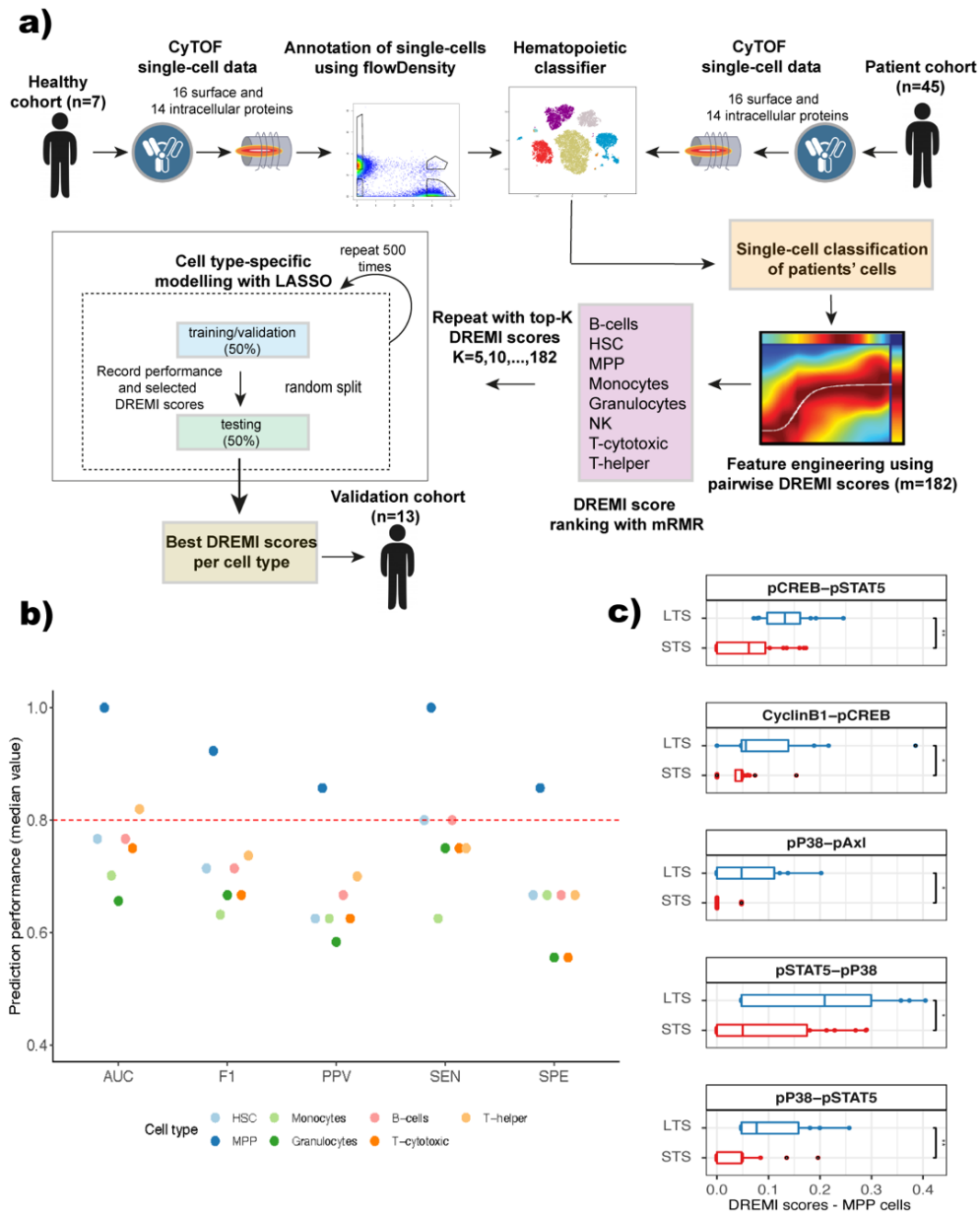


Figure 5: Survival analysis using DREMI-derived scores from MPP cells. a-c) Kaplan-Meier analysis of survival using pCREB→pSTAT5 and pSTAT5→pP38 and pP38→pSTAT5 DREMI scores from MPP cells. d) Cox proportional hazards modelling reveals that pCREB→pSTAT5 measured at time of diagnosis from MPP cells is predictive of survival. A subset of 36 patients with MPP abundance greater than 500 cells was used for this analysis, and the patients were stratified to high and low groups based on their median DREMI scores. Log-rank test was used to compute the p values. e) Visualizing the DREMI score values of pCREB→pSTAT5 and pSTAT5→pP38 and pP38→pSTAT5 using 13 AML patients from Levine et al. Statistically significant differences were not detected using Mann-Whitney U test.

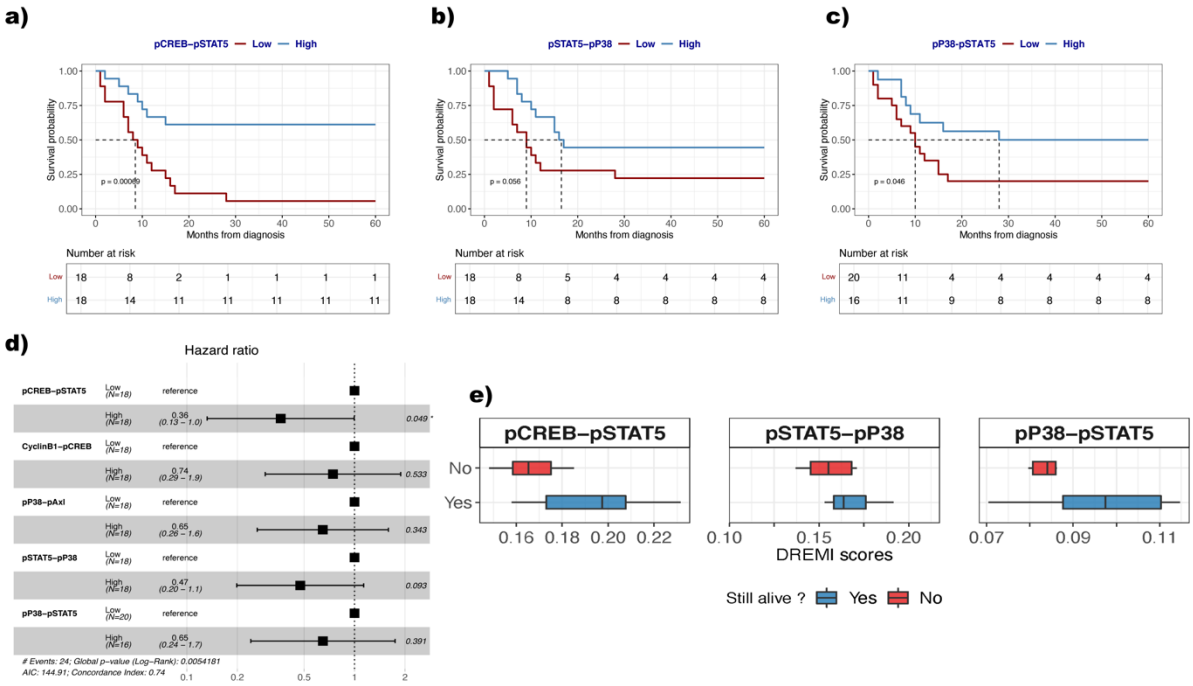


Figure 6: DREMI scores from MPP cells cluster patients into groups of different survival. Unsupervised clustering of patients using MPP-derived DREMI scores identify five distinct clusters. We show a t-sne projection of patients into the 2D space. The size of data points indicates the (%) abundance of MPP cells. b) Kaplan-Meier survival analysis of patients' cluster C2 versus all other patients. Log-rank test was used to compute the p value. c) Heatmap showing the MPP-derived DREMI scores of all patients included in the analysis. The heatmap is annotated using clinical and genetic information.

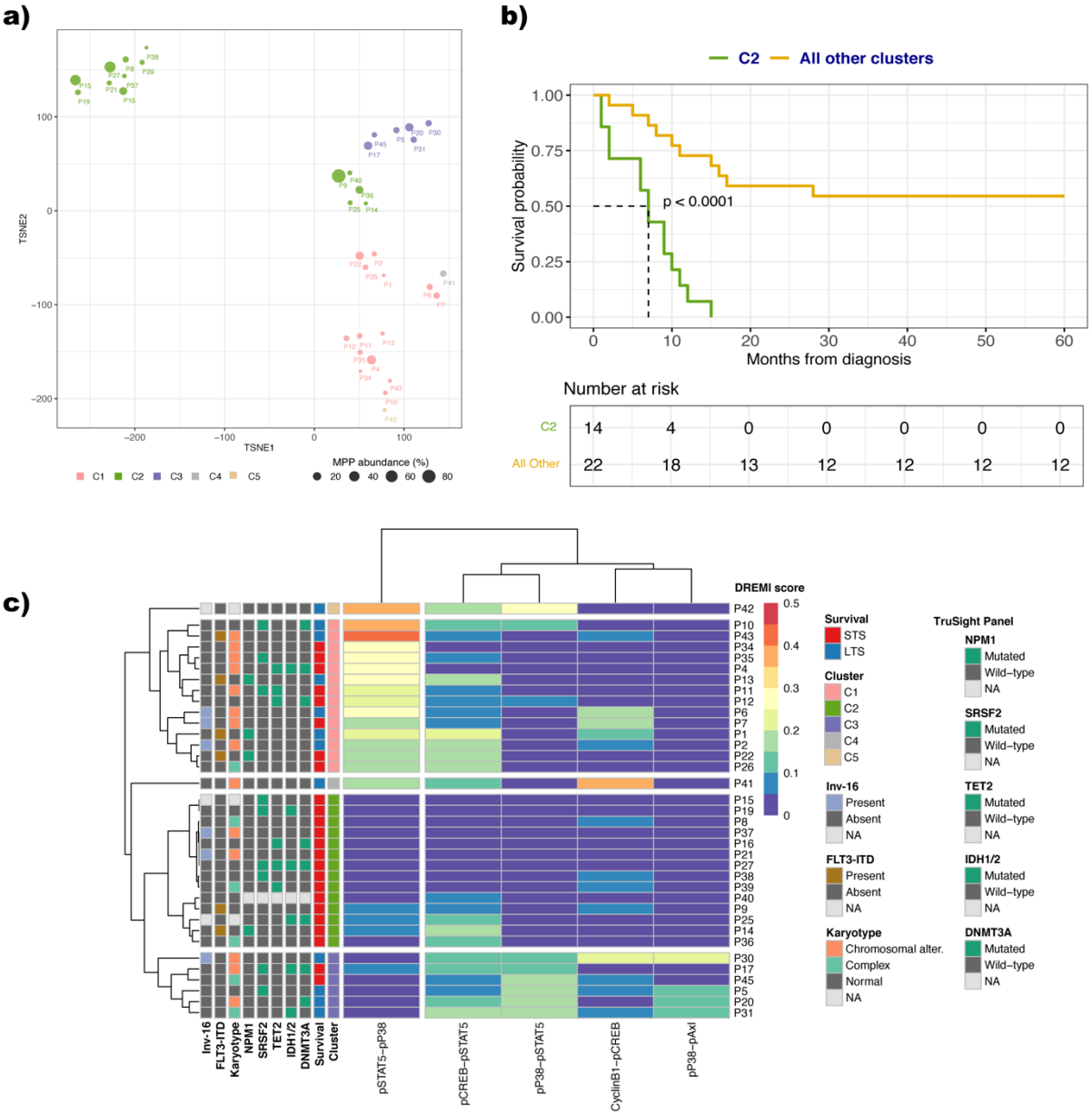
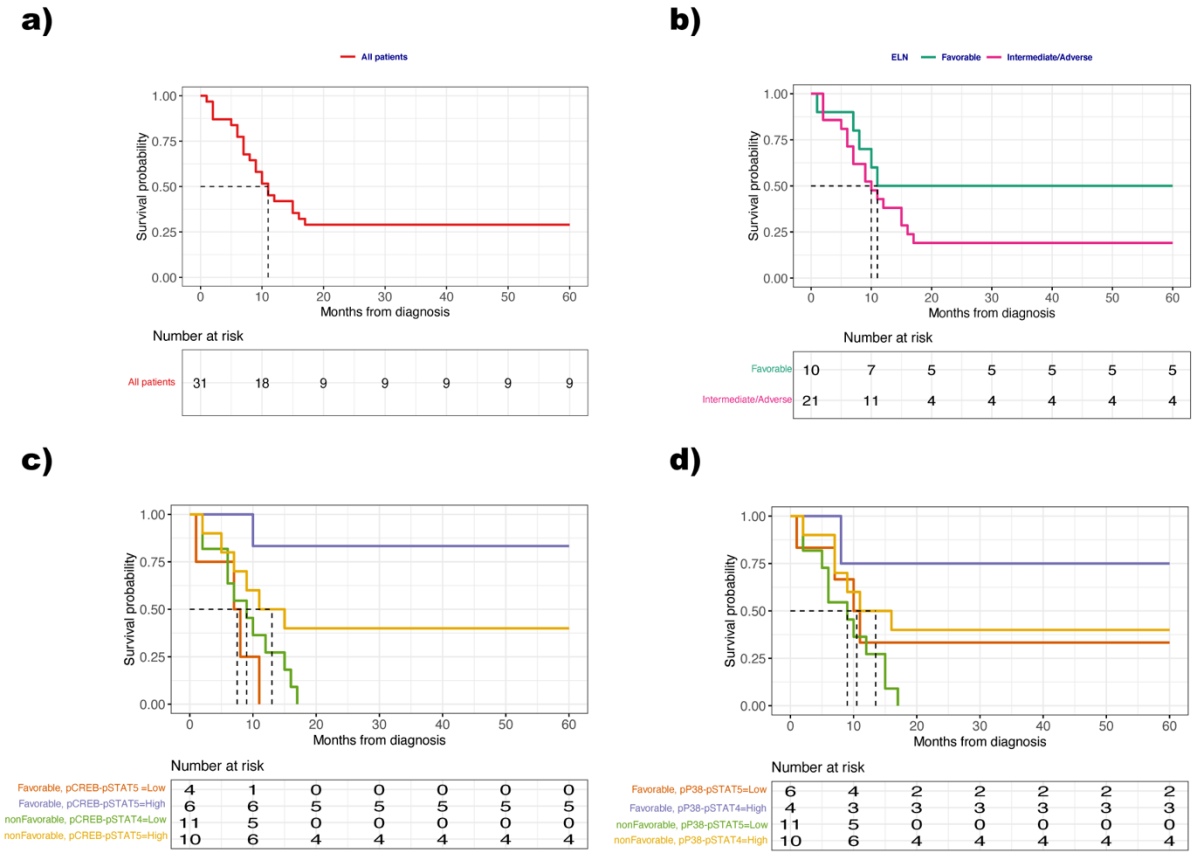


Figure 7: DREMI scores synergise with ELN risk stratification criteria. a) The survival probability of all patients with available ELN information and DREMI-derived features from MPP cells. b) The survival probability of ELN Favorable patients versus the survival probability of ELN Adverse and Intermediate patients combined (p=0.16). c) Kaplan-Meier survival analysis of patients using MPP-derived pCREB-pSTAT5 DREMI score and ELN stratification (p=0.0034). d) Kaplan-Meier survival analysis of patients using MPP-derived pP38-pSTAT5 DREMI score and ELN stratification (p=0.06). Log-rank test was used to compute the p values.



Supplementary Figures

Figure S1: Overview of the deployed gating strategy. Gating strategy to identify hematopoietic cell types from seven healthy donors. In total, ~1.2 million cells were analysed using flowDensity. From them, 582,010 cells were selected to describe expression profiles of B-cells, T-cells (Cytotoxic and Helper), Natural Killer cells, Multi Potent Progenitor (MPP), Hematopoietic Stem Cells (HSC), Granulocytes and Monocyte cells. The selected healthy cells were used for the rest of the analysis to construct the reference of hematopoietic cell types, and to train the single-cell hematopoietic classifier.

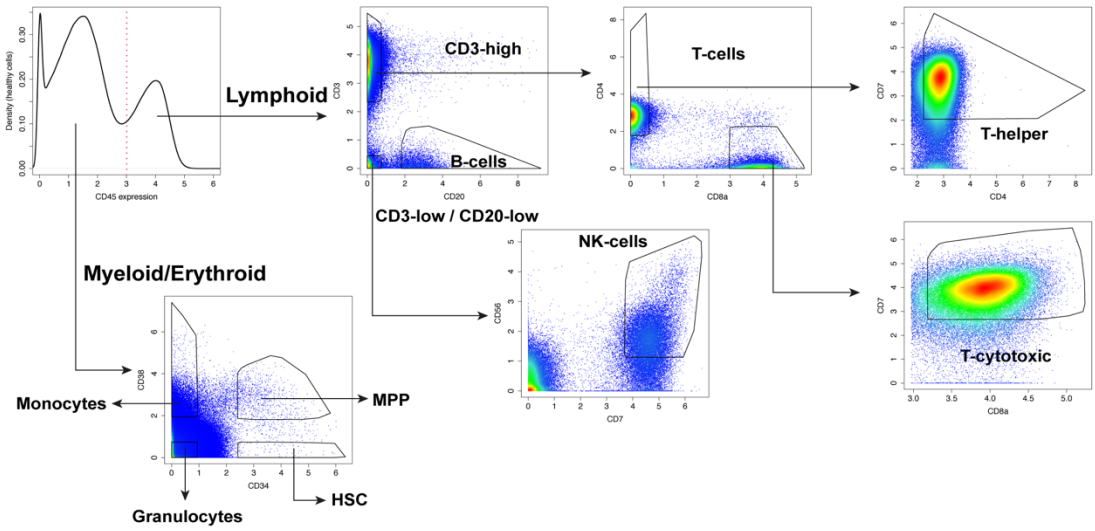


Figure S2: Visualization of the selected healthy cell populations via Principal Component Analysis (PCA). a) We use the same PCA space shown in Figure 1a. b-d) Projected cells colored based on the normalised expression of CD38, CD33, CD66b and markers.

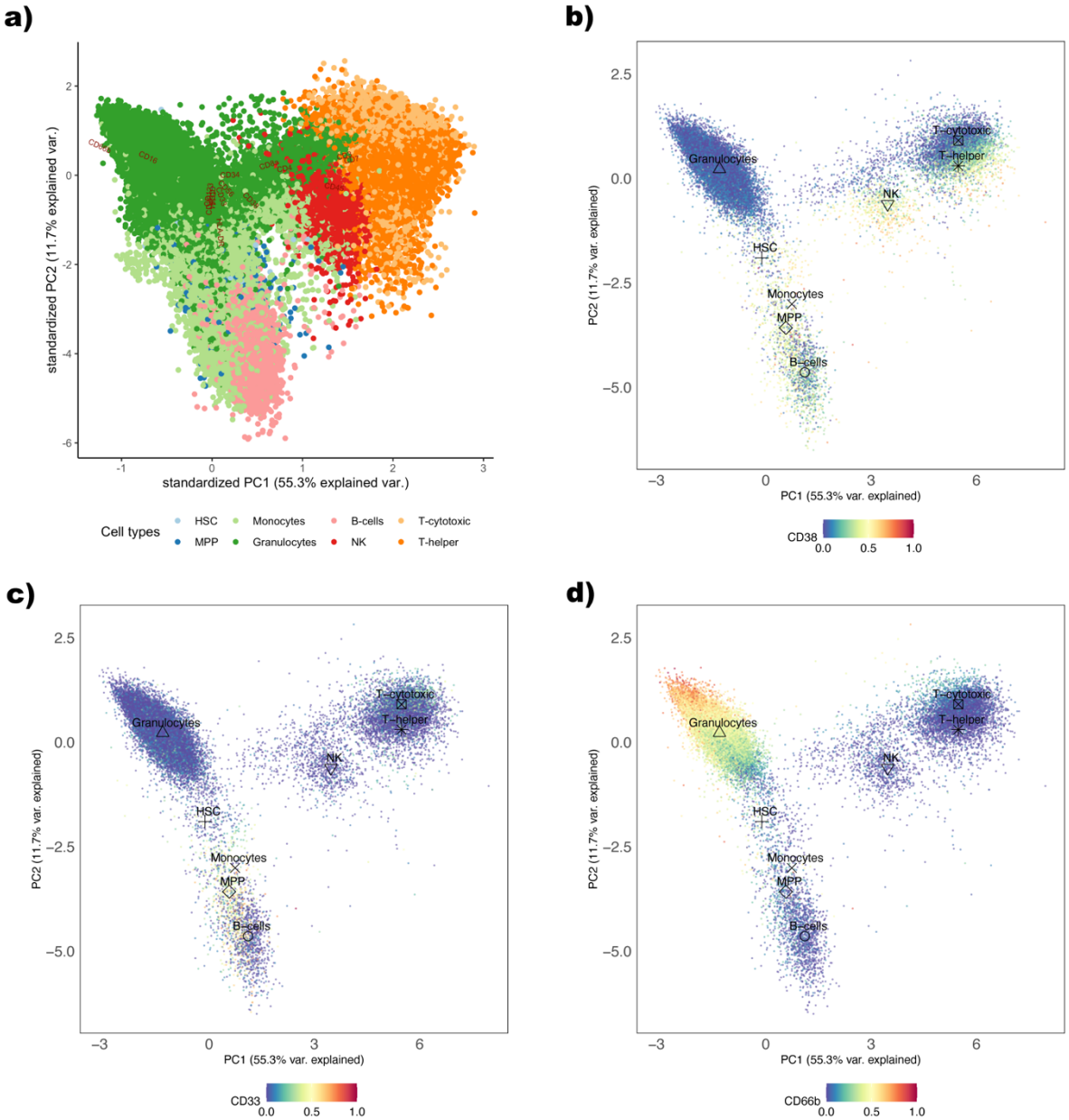
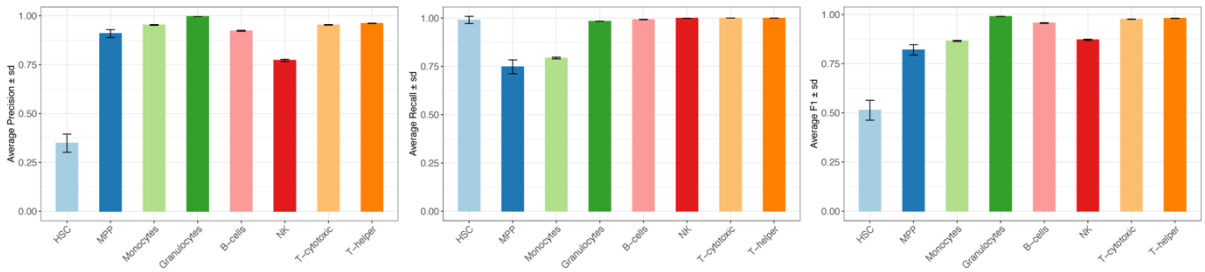


Figure S3: Comparison of LDA and KNN methods for single-cell classification using our HD dataset. The classification performance was computed using the hold-out experimental setup and repeated 1,500 times with random training/testing subsets. Bar plots show average Precision Recall and F1 score ± std. a) Performance achieved for LDA; and b) Performance achieved for KNN using K=5.

a) LDA



b) KNN

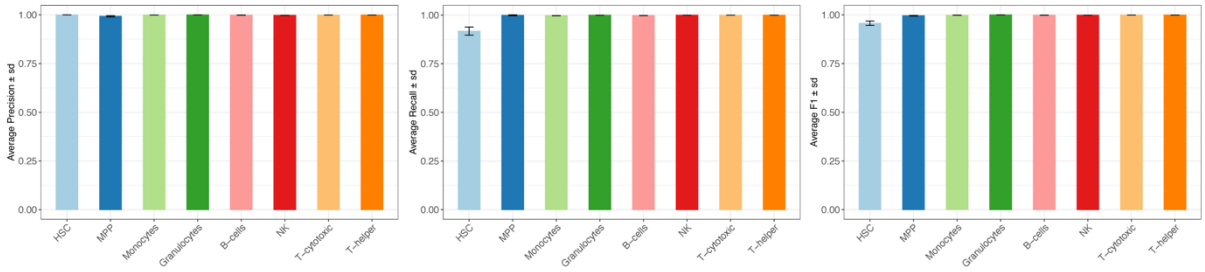


Figure S4: Investigating misclassified cells from LDA and KNN methods using HD data. a-b) We show the number of misclassified single-cells from LDA and KNN. Each subplot corresponds to a cell gate that is considered groud-truth. The y-axis shows the average % fraction of cells by LDA and KNN, and the x-axis shows the predicted cell gates. c) Tuning of parameter K for KNN method. We used 20% of the training data for optimizations, and we show the average F1 achieved across 1,500 experiments.

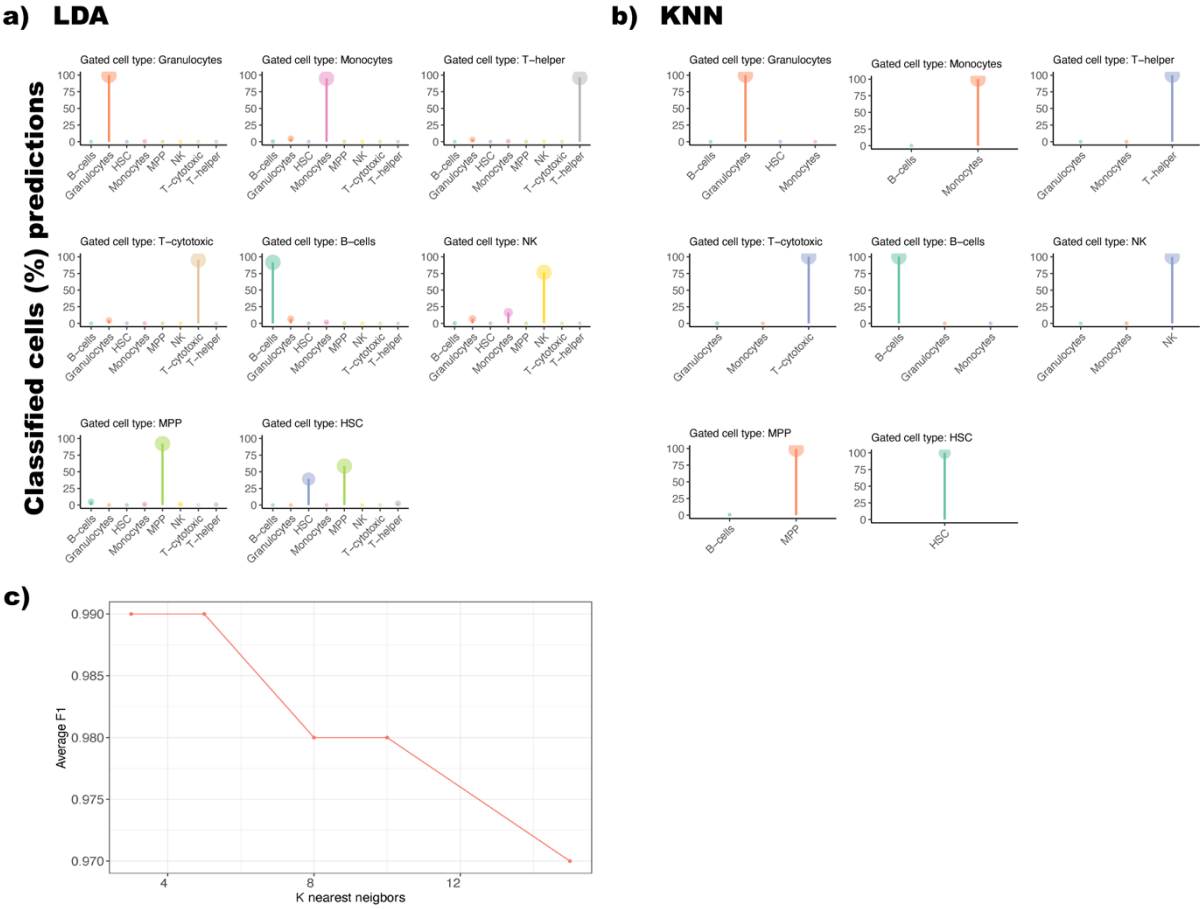


Figure S5: Analysis of single-cells from patients with leukemia. PCA analysis of 30,000 randomly selected single-cells from 45 patients irrespective of clinicopathological classification. The cells have been projected into the PCA space constructed using healthy reference cell types from Figure 1a. For comparisons purposes the PCA space has been colored using normalized expression levels of CD45, CD33, CD34 and CD66b markers. The centroids of the gated healthy cell types from Figure 1a are highlighted for comparisons.

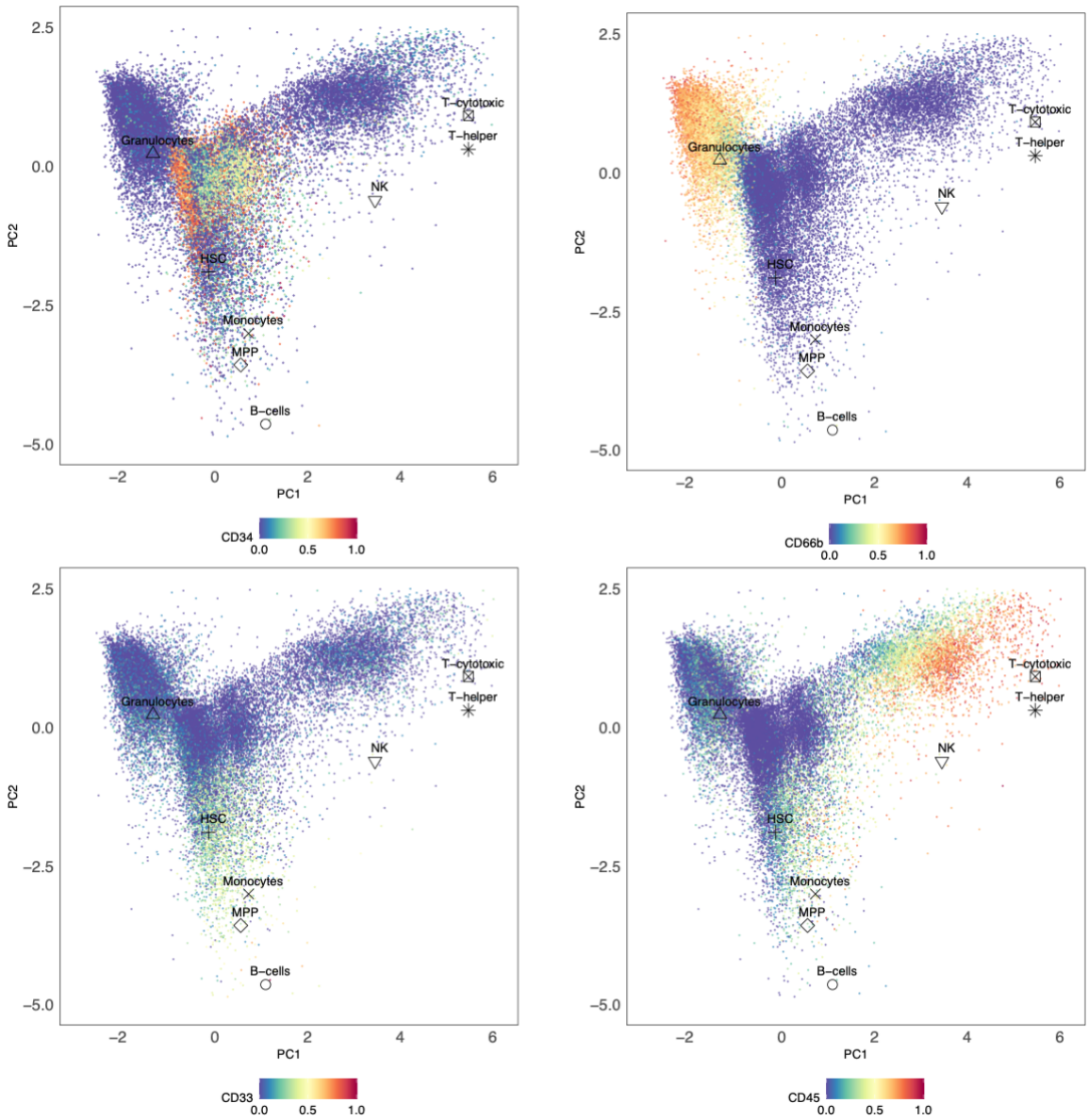


Figure S6: Analysis of single-cells from patients with leukemia considering ELN (2017) patient stratification. PCA analysis of 30,000 randomly selected single-cells from patients of ELN Adverse a) and ELN Favorable groups b). The cells have been projected into the PCA space constructed using healthy reference cell types from Figure 1a. For comparisons purposes the PCA space has been colored using normalized expression levels of surface markers CD34, CD66b, CD123. We also show signalling markers pErk, pSTAT1 and pNFkB, to highlight the complexity of analysing single-cell signalling data for cell types. The centroids of the gated healthy cell types from Figure 1a are highlighted for comparisons.

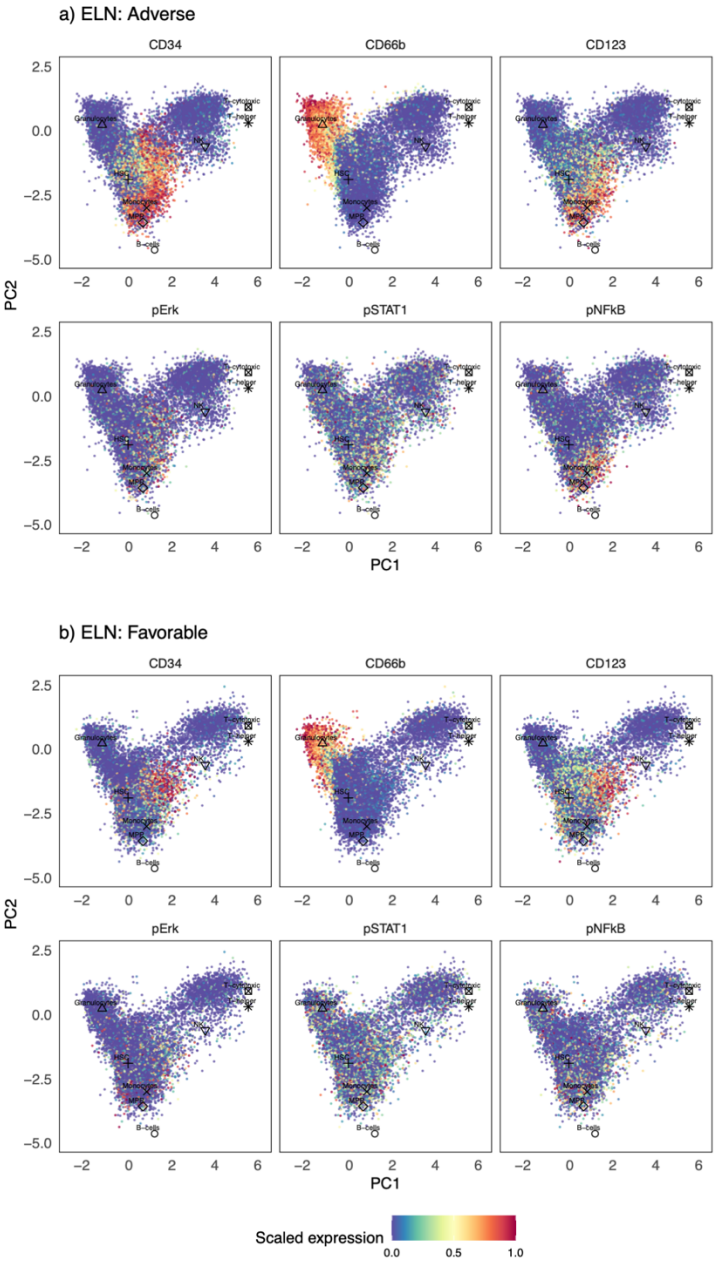


Figure S7: Patient cohort characteristics. a) We show the proportion (%) of different treatments for Short Term Survivors (STS; ST=17, ST+Len=4, Other=8, NA=1) and Long Term Survivors (LTS; ST=6, ST+Len=6, Other=3) patients. b) We show the proportion (%) of ELN (2017) Favorable (8 vs. 5), Intermediate (7 vs. 2) and Adverse (12 vs. 5) in STS and LTS groups respectively. Six patients did not have ELN information available. c) We show the proportion (%) of patients with complete remission (CR) (10 vs. 10), non-complete remission (nonCR) (10 vs. 1), as well as the fraction of aplastic (6 vs. 1) cases for STS and LTS groups. Seven patients did not have Remission information available. d) We show the proportion (%) of male (20 vs. 9) and female (10 vs. 6) patients in STS and LTS groups.

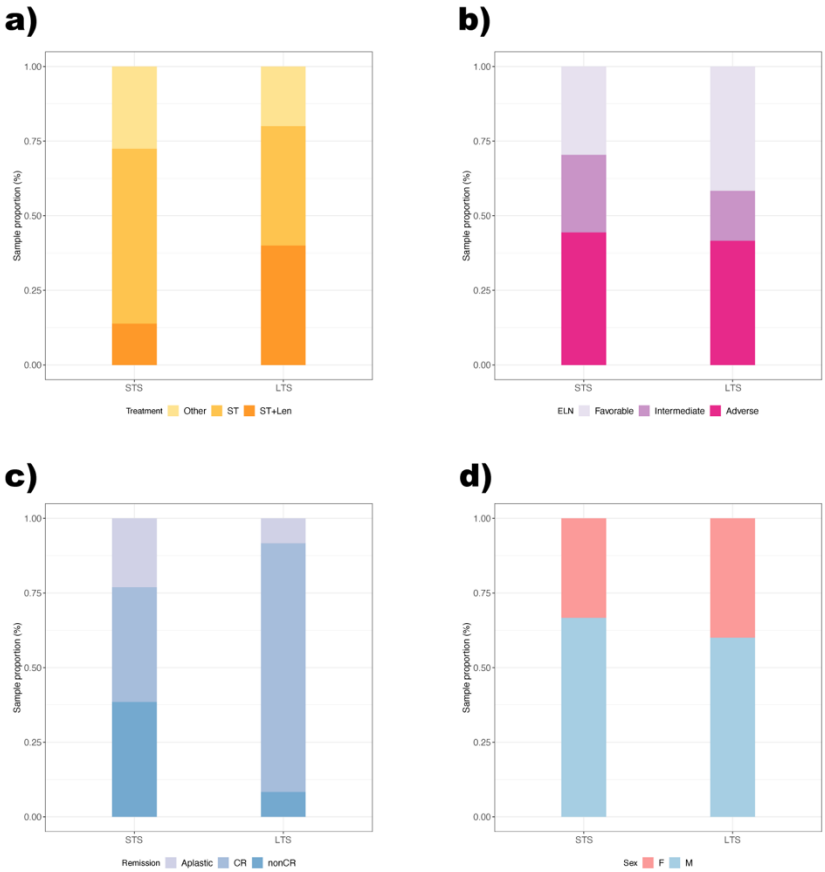


Figure S8: Single-cell classification of cells from leukemia patients allows for cellular abundance comparisons using clinical info. a) Percentages (%) of annotated cells after single-cell classification at time of diagnosis. The boxplots show the cellular abundance of 45 patients stratified based on ELN (2017) group. b) Percentages (%) of annotated cells after single-cell classification at time of diagnosis. The boxplots show the cellular abundance of 45 patients stratified based on Remission group. c) Percentages (%) of annotated cells after single-cell classification at time of diagnosis. The boxplots show the cellular abundance of 45 patients stratified based on Treatment group. Statistically significant differences were tested using Mann-Whitney U test. The levels of significance are: * for p<0.05, ** for p<0.01 and *** for p<0.001, ns is reported for non-significant results.

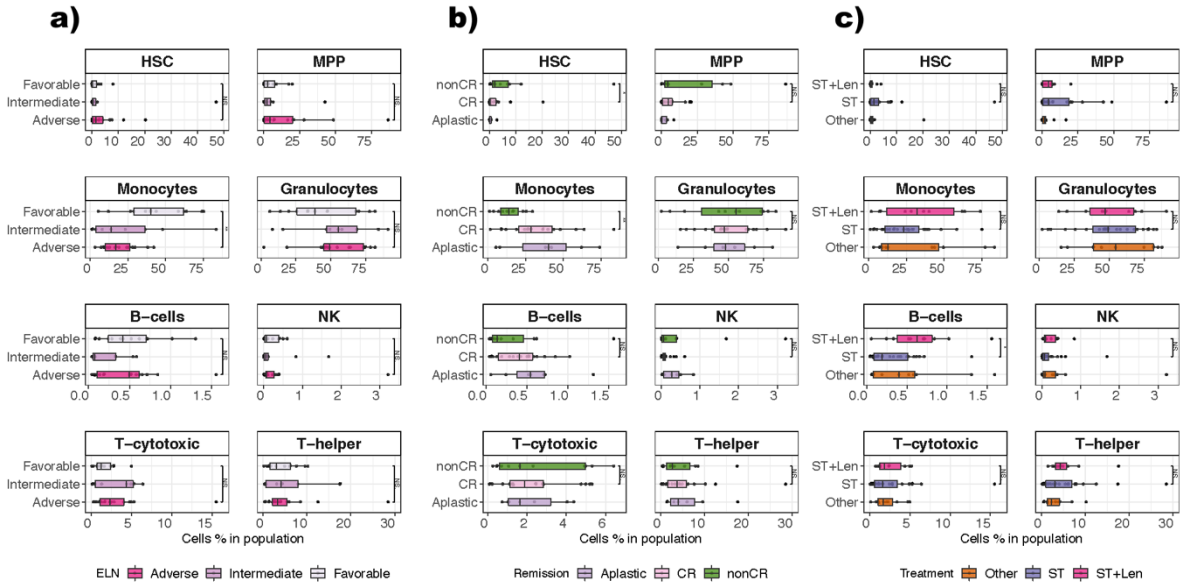


Figure S9: Single-cell classification of cells from leukemia patients allows for cellular abundance comparisons using mutational information from sequencing data. a) Percentages (%) of annotated cells after single-cell classification at time of diagnosis. The boxplots show the cellular abundance of 45 patients stratified based on DNMT3A mutational status. b) Percentages (%) of annotated cells after single-cell classification at time of diagnosis. The boxplots show the cellular abundance of 45 patients stratified based on NPM1 mutational status. Statistically significant differences were tested using Mann-Whitney U test. The levels of significance are: * for p<0.05, ** for p<0.01 and *** for p<0.001, ns is reported for non-significant results.

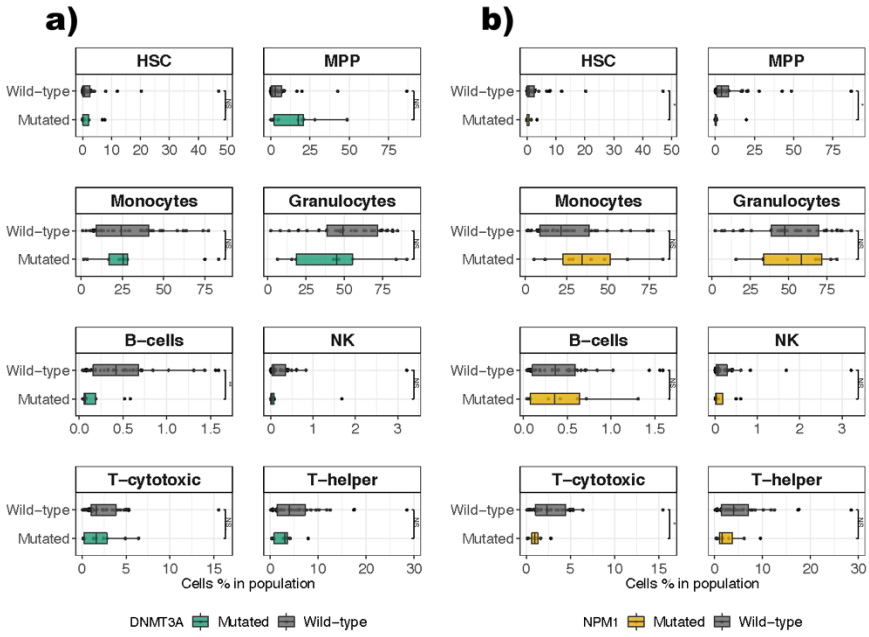


Figure S10: Single-cell classification of cells from leukemia patients allows for cellular abundance comparisons using genetic information. a) Percentages (%) of annotated cells after single-cell classification at time of diagnosis. The boxplots show the cellular abundance of 45 patients stratified based on karyotype status. b) Percentages (%) of annotated cells after single-cell classification at time of diagnosis. The boxplots show the cellular abundance of 45 patients stratified based on presence/absence of FLT3-IDT mutations. c) Percentages (%) of annotated cells after single-cell classification at time of diagnosis. The boxplots show the cellular abundance of 45 patients stratified based on presence/absence of inv16 translocation. Statistically significant differences were tested using Mann-Whitney U test. The levels of significance are: * for p<0.05, ** for p<0.01 and *** for p<0.001, ns is reported for non-significant results.

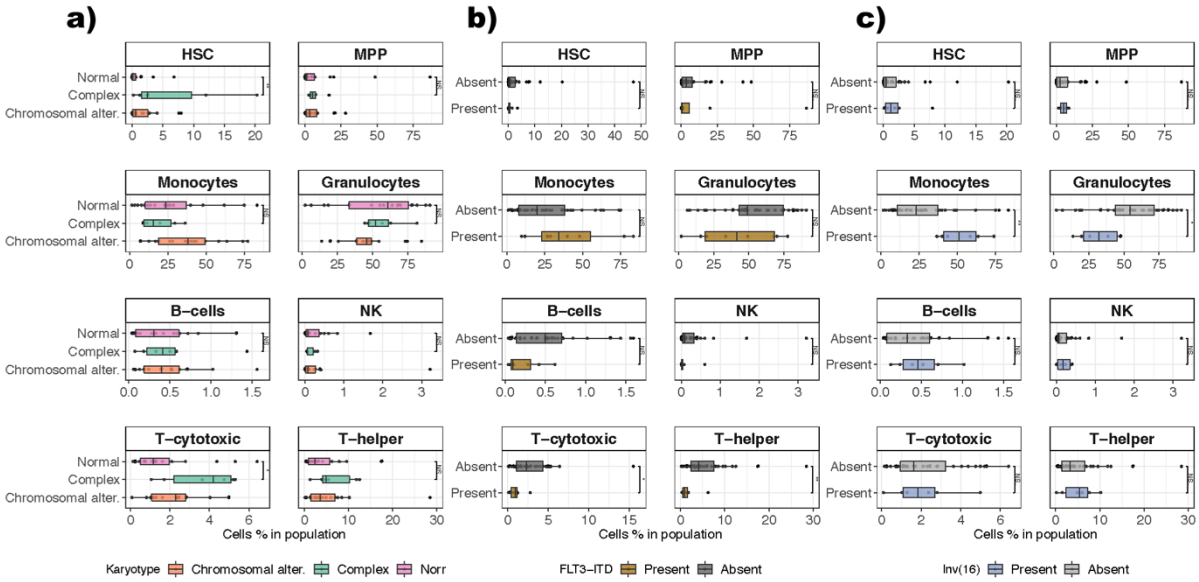
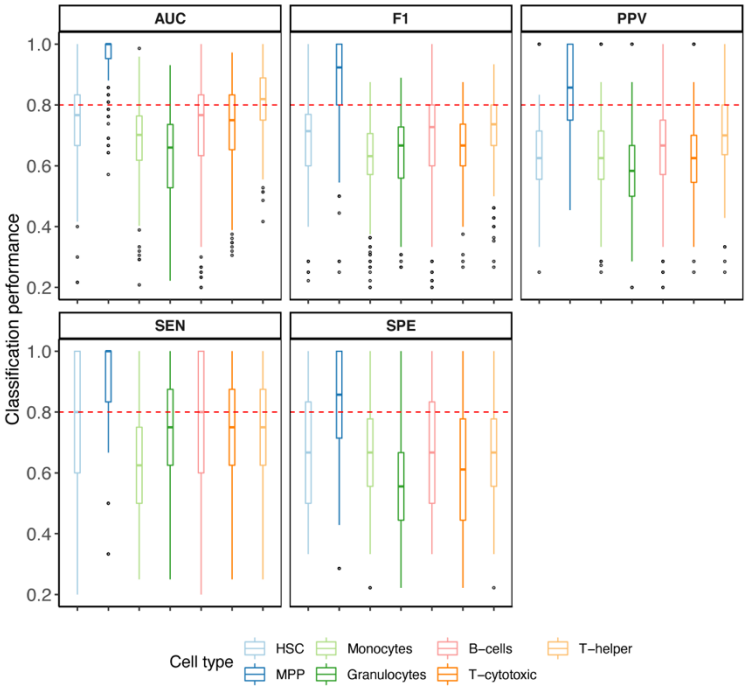



Figure S11: Summary of LASSO classification performance with feature selection. Boxplots showing the classification performance measured using different metrics across the studied cell types, and across different top-K sets of features.