Single cell guided deconvolution of bulk transcriptomics recapitulates differentiation stages of acute myeloid leukemia and predicts drug response

E Onur Karakaslar[1,2,3], Jeppe Severens[1,2,3], Elena Sánchez-López[3,4], Peter A van Veelen[4], Mihaela Zlei[5,6], Jacques JM van Dongen[6], Annemarie M. Otte[4], Constantijn JM Halkes[7], Peter van Balen[7], Hendrik Veelken[3,7], Marcel JT Reinders[1,2,3], Marieke Griffioen[7], Erik B van den Akker[1,2,3]

Affiliations

[1]Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands

[2]Pattern Recognition & Bioinformatics, Delft University of Technology, Delft, The Netherlands

[3]Leiden Center for Computational Oncology, Leiden University Medical Center, Leiden, The Netherlands

[4]Center for Proteomics and Metabolomics, Leiden University Medical Center, Leiden, The Netherlands

[5]Department of Flow Cytometry, Medical Laboratory, Regional Institute of Oncology, Iasi, Romania

[6]Department of Immunology, Leiden University Medical Center, Leiden, The Netherlands

[7]Department of Hematology, Leiden University Medical Center, Leiden, The Netherlands

Materials & Correspondence

Erik B. van den Akker, PhD; Department of Biomedical Data Sciences, Leiden University Medical Center, Leiden, The Netherlands; Einthovenweg 20, 2333 ZC, Leiden, The Netherlands; Tel: +31 (0)71 526 85 57; Fax: +31 (0)71 526 82 80; E-mail: e.b.van_den_akker@lumc.nl

Abstract

The diagnostic spectrum for AML patients is increasingly based on genetic abnormalities due to their prognostic and predictive value. However, information on the AML blast phenotype regarding their maturational arrest has started to regain importance due to its predictive power on drug responses. Here, we deconvolute 1350 bulk RNA-seq samples from five independent AML cohorts on a single-cell healthy BM reference and demonstrate that the morphological differentiation stage (FAB classification) could be faithfully reconstituted using estimated cell compositions (ECCs). Moreover, we show that the ECCs reliably predict *ex-vivo* drug resistances as demonstrated for Venetoclax, a *BCL-2* inhibitor, resistance specifically in AML with CD14+ monocyte phenotype. We further validate these predictions using in-house proteomics data by showing that *BCL-2* protein abundance is split into two distinct clusters for NPM1-mutated AML at the extremes of CD14+ monocyte percentages, which could be crucial for the Venetoclax dosing for these patients. Our results suggest that Venetoclax resistance predictions can also be extended to AML without recurrent genetic abnormalities (NOS), and possibly to MDS-related AML and secondary AML. Collectively, we propose a framework for allowing a joint mutation and maturation stage modeling that could be used as a blueprint for testing sensitivity for new agents across the various subtypes of AML.

Introduction

Acute myeloid leukemia (AML) is an aggressive hematological cancer of the myeloid lineage. AML is caused by a combination of relatively few genetic alterations that are predominantly somatically acquired and cooperatively induce a maturation arrest in combination with rapid uncontrolled proliferation of immature myeloid precursor cells. The prognosis of AML is highly dependent on the presence of such recurrent genetic alterations and varies from > 90% cure rates to less than 10% [1]. Therefore, the current WHO classification primarily defines AML subtypes according to the presence of eleven recurrent genetic aberrations (RGA) changes and an added heterogeneous umbrella subtype composed of a highly diverse set of relatively rare RGAs [3–6]. Only AML cases that lack any detected RGA are characterized according to their maturation stage according to the French-American-British (FAB) classification [2].

Recently, the maturation and differentiation stage of AML blasts has gained importance due to a striking association with sensitivity or resistance to new drugs [7–9]. AML differentiation stage can be assessed by diagnostic flow cytometry more objectively than by morphological examination alone. However, flow cytometry is limited to a relatively low number of differentiation markers. Gene expression profiling by bulk RNA-sequencing (RNA-seq) is an attractive alternative technique as it allows both calling of genetic aberrations and estimation of cell subsets, i.e. estimated cell composition (ECC) from the same sample[10–12] . Reported attempts to estimate ECCs by deconvoluting bulk AML samples utilizing single-cell RNA-seq as an *in-silico* reference, were mostly focused on detection of survival differences without performing thorough validation of the ECCs and/or used leukemic samples as reference[13–15], which prevents assessing whether ECCs are tissue- or sample- specific.

Here, we perform deconvolution of 1350 AML transcriptomic samples via a healthy single-cell reference while validating our findings with our in-house flow cytometry data. Of note, we demonstrate that the ECCs recapitulate the entire FAB landscape (M0-M7). Then, using these ECCs we predict *ex-vivo* drug resistance data from literature and show the agreement of these results at protein level with the help of in-house proteomics data in AML patients, for whom we also had acquired gene expression data. To conclude, we hereby propose a transcription-based single-cell guided deconvolution framework to assess the drug

effectiveness to different maturational stages of AML. We also provide our framework as a CRAN R package available at https://github.com/eonurk/seAMLess.

Results

Deconvolution pipeline recapitulates healthy and malignant hematopoiesis

We initially created a healthy bone marrow (BM) reference atlas from single-cell transcriptomics data (**Figure 1A**). For this purpose, we integrated data from 69,101 cells covering 439 genes from three publicly available datasets (two full-transcriptome and one targeted) from two studies[16,17] (**Figure 1B**, Methods). As expected, T-cells and B-cells formed separate clusters in the UMAP plot, whereas the myeloid lineage cells clustered together. To differentiate early myelopoiesis, we distinguished more than 3,349 hematopoietic stem cells (HSC) and 1,432 erythro-myeloid progenitors (EMP), 2,939 lymphoid multipotent progenitors (LMPP), as well as 3,508 granulocytes-monocytes progenitors (GMP) (**Figure 1C**). Inter-individual variability, possibly due to age differences, mostly affected T- and B- cells and did not influence overall clustering (**Figure S1A**). Homogeneous distribution of the studies on the UMAP plot showed successful integration of the different datasets (**Figure S1B**).

We next performed *in silico* experiments to validate our deconvolution set-up. We first used the healthy BM reference to create *in silico* mixed bulk samples with known cell compositions. Pseudobulk profiles were simulated with one abundant cell type (80% of the cells) with the remaining cells being a mix created by random selection (Methods). Then, MuSiC[12] with the default settings was used to deconvolute the simulated pseudobulk profiles into their respective cell types (**Figure S1C**). To validate on an independent dataset, we also simulated pseudobulks from 40,000 healthy bone marrow cells of the human cell atlas (HCA)[18] (**Figure S1D**) and performed deconvolution via MuSiC. All simulated pseudobulk profiles were successfully deconvoluted into their respective cell types (matching to the most abundant cell type) apart from EMP (**Figure 1D**). For this profile, the annotations were shared mostly among EMP (14%) as well as HSC (25%) and Early Erythrocytes (23%). This discrepancy might be explained by the transcriptional similarity of these cell types (**Figure 1B**).

Next, we analyzed our in-house diagnostic flow cytometry data (EuroFlow[19] panels; see Methods) for 22 AML samples with matched bulk RNA-seq data. An overview of mean fluorescence intensity (MFI) values

for these samples' abnormal cells after staining with antibodies for 33 markers distributed over 7 tubes (1 Orientation + 6 AML assignment tubes) is shown in **Figure S1E** (Supplementary Table S1). In line with our expectations, the most abundant cell types for all samples were in the myeloid lineage for both flow cytometry analyses and estimated cell compositions (ECCs) (**Figure 1E**, Supplementary Table S4). To investigate whether monocytic AML can be accurately distinguished from AML with more stem cell-like phenotypes, we plotted CD14+ monocyte percentages as determined by deconvolution against MFI values of antibodies of monocytic markers (CD11b, CD64, IREM2, and CD14) on all BM cells without gating, and observed statistically significant correlations for 3 markers (CD11b, CD64, IREM2), with CD64b being most significant ($R^2 = 0.43$, $P < 0.001$) (**Figure S1F**). Also, percentages of AML cells assigned to the monocytic subset by EuroFlow panels and ECCs showed statistically significant correlations ($R^2 = 0.64$, $P < 0.001$) (**Figure 1F**).

Lastly, we downloaded publicly available TARGET AML and ALL (B-ALL and T-ALL) data and deconvoluted these samples (n=719) with the healthy BM reference (**Figure 1G,** Supplementary Table S1) to show that different leukemic phenotypes could be captured by deconvolution. The estimates of cell type abundances in the three immune lineages (myeloid, B- and T-lymphoid) were visualized using a ternary plot. As expected, each type of acute leukemia was closer to their respective cell type of origin (ECCs), most compact being B-ALL (yellow samples on the right corner), showing the ECCs' ability to capture the patients' immune phenotypes at major cell type levels. Together, these benchmarking and validation analyses demonstrated that given a healthy single cell BM transcriptomic atlas, the cell type proportions can be recapitulated faithfully via deconvolution from bulk transcriptomic leukemia cases.

Deconvolution of bulk AML transcriptomics reveals the dominating immune fraction

To investigate heterogeneity in cell composition of AML, we next applied our framework to deconvolute five independent bulk transcriptomic AML studies, i.e. TCGA-LAML[20] (n=151), BEAT-AML[21] (n=460), TARGET-AML[22] (n=187), LEUCEGENE[23] (n=452) and our cohort LUMC[10] (n=100) totaling in 1350 samples from 1,267 patients (**Figure S2A**). The results are shown in **Figure 2A** as a heatmap where each sample was decomposed into the 22 cell types from the healthy BM reference. We also added information on patients' clinical blast counts, FAB[2], WHO 2016[4], and ELN 2017[6] classes for an overarching picture of AML landscape, and we also calculated the stemness score[24], which is a gene expression signature for patient prognosis trained on engraftment capacity of AML in immunodeficient mice (Supplementary Table S2). Each sample was assigned

to the most abundant cell type as determined by deconvolution (AML phenotype) and arranged according to their fractions within each phenotype. The data showed a clear dominance of one immune cell type for most of the cases, clearly indicating that maturational arrests and lineage skewing are leukemic properties that can be readily assessed using transcriptome sequencing. The majority of AML cases was typically estimated to be dominated by myeloid cells (**Figure 2A, S2A, S2B, S3D**). As notable exceptions, a few pediatric AML samples from the TARGET cohort showed ECC profiles dominated by T-cells or B-cells. This effect is most likely caused by prominent infiltration of the bone marrow by these types of lymphocytes besides the AML blast population and therefore precludes the assignment of the AML blasts to a distinct myeloid cell type. As previously reported these cases can nevertheless be considered as a special subgroup of pediatric AML, and specifically T-cell dominated patients were recognized for poor survival[25]. Furthermore, in line with previous reports, patients with acute promyelocytic leukemia (APL), classified as AML-M3 by FAB, were correctly assigned to have a cell composition dominated by granulocyte-monocyte progenitors (GMPs). Besides APL, there are also AML cases dominated by GMP and large groups of AML assigned as monocytic AML or AML with an earlier HSC or EMP phenotype, again demonstrating that bulk transcriptomics can be used to capture the stage of arrest of AML during hematopoiesis by deconvolution.

Deconvolution of bulk AML transcriptomics agrees upon stemness score

To comprehensively visualize changes in ECCs in relation to metadata, ECCs of AML samples were visualized using a UMAP plot. As expected, samples with similar ECCs clustered, as shown after annotating the samples for their most abundant cell type (**Figure 2B** – left panel). When superimposing the deconvoluted percentages of 6 major cell types on the same UMAP plot (**Figure 2B** – right panels), gradual shifts in cell composition became apparent, in particular, APL samples populated the extreme extension of the GMP cluster. Moreover, a gradient towards monocytic outgrowth and conversely to a high stemness was also observed. Comparisons with the clinical blast percentage and previously published stemness score showed an overall agreement with ECCs, HSCs and EMPs having a higher score than CD14+ Monocytes and GMPs (**Figure S3A, S3B**). Late Erythrocyte also had high stemness score, albeit with large variation. By dividing AML into cases with high and low stemness scores, we observed similar compositional changes with abundance of HSCs and EMPs in AML with high stemness scores in all cohorts. Late erythrocytes, however, were not consistently enriched in all cohorts. Furthermore, we noticed that in TARGET, T cells are abundant in AML with high stemness scores

(**Figure S3C**). This raises the question whether the stemness score, which has been trained on adult AML, is useful to predict prognosis of pediatric patients particularly considering that large subtypes of adult AML such as APL and AML with mutated NPM1 are rare in pediatric AML. The consistent enrichment of HSCs and EMPs in AML with high stemness scores in all cohorts, however, further supports correct determination of the cell composition of AML by deconvoluting bulk transcriptomics.

Estimated cell composition recapitulates FAB classes

Clear associations were also observed between cell type assignments by ECC and FAB classification status (left panel of **Figure 2C**). Furthermore, distinct distributions of cell type-defining gene expression profiles became evident for each FAB AML type by plotting continuous values of deconvolutions rather than categorical assignments: M0 (minimally differentiated AML) cases had high levels of the HSC-defining signature, whereas these levels decreased and EMP- and GMP signatures appeared in M1 (AML without maturation) and M2 (AML with maturation). As expected, M3 (APL) cases were almost completely dominated by the GMP signature, while M4 (Acute Myelomonocytic Leukemia) and M5 (Acute monoblastic/monocytic leukemia) samples resembled CD14+ monocyte cells and GMP. The small groups of M6 (Acute Erythroleukemia) and M7 (Acute Megakaryoblastic Leukemia) AML were dominated by the signatures of Late Erythrocytes and Megakaryocyte Progenitors, respectively. These data demonstrated and confirmed that our single cell guided deconvolution strategy successfully captures the maturational arrest of AML cells at different differentiation stages of hematopoiesis.

Estimated cell composition captures genetic subtype-specific resistances to various drugs

To explore whether ECCs convey information on drug resistances, *ex vivo* drug response data of 122 small molecule inhibitors provided as area under the curves (AUC) for 363 AML samples from BEAT were downloaded. A higher AUC indicates that cancer cells are relatively resistant since higher drug concentrations are needed to induce cell death. To predict drug resistance for AML samples with calculated ECCs, we trained random forest (RF) models per drug via leave-one-out cross validation (LOOCV) setting and then calculated the Spearman ρ values of these models' predictions (**Figure S4A**; Supplementary Table S5-S7). We also stratified these results according to WHO classifications (**Figure 3A, S4B**, Supplementary Table S8).

Drug responses to the EGF-R inhibitor Erlotinib (Spearman ρ = 0.376 and the mTOR pathway inhibitor Rapamycin (Spearman ρ = 0.368) showed high correlations with AML with mutated NPM1 (n = 77, R^2 = 0.29, P= $3.7x10^{-7}$) and inv(3) (n = 6, R^2 = 0.91, P= $3.2x10^{-3}$), respectively. Notably, responses to Flavopiridol (Spearman ρ = 0.325), a CDK kinase inhibitor, showed strong correlations with ECC within the group of AML-NOS (n = 77, R^2 = 0.31, P= $3.7x10^{-7}$) while showing higher resistance towards AML with more stem cell like cell phenotypes (**Figure 3A, S5A, S5B**). The strongest correlation (Spearman ρ = 0.509), however, was observed for Venetoclax (ABT-199), a BCL2 inhibitor (**Figure S5C**). Annotation with primary diagnosis from WHO classes (**Figure 3B**) revealed that PML-RARA-carrying APL samples (n = 8) are dominated by GMPs are sensitive for Venetoclax, whereas AML with CBFB-MYH11 (n = 7) had the best fit (R^2 = 0.78, P = 0.009) for Venetoclax resistance, although the small sample size of CBFB-MYH11 cases prohibits drawing robust conclusions. Larger groups of AML, however, such as NPM1 mutated cases (n = 55, R^2 = 0.33, P < 0.001), and AML-NOS (n = 42, R^2 = 0.17, P = 0.007), together accounting for approximately 58% of AML, also showed a clear trend for Venetoclax resistance (**Figure 3B**). Annotating drug responses with predicted resistances by deconvolution (**Figure S5C**) revealed that CD14+ Monocytic AML are relatively resistant to Venetoclax. This resistance of monocytic AML was also clear by overlaying the AUC for Venetoclax onto the UMAP plot (**Figure 3C**) or boxplots (**Figure S5D**) (one-way ANOVA p = $6.31x10^{-13}$; Supplementary Table S9). Also, within the group of NPM1 mutated samples, which is the largest class of AML, CD14+ Monocyte dominated cases were most resistant to Venetoclax (**Figure S5E, S5F**; one-way ANOVA p = $7.6x10^{-4}$; Supplementary Table S10). Briefly, these findings suggest that information on the ECCs of AML might yield therapeutic implications even within one genetic subtype.

Estimated CD14+ Monocyte percentages predict Venetoclax resistance better than *BCL-2* mRNA expression

Since Venetoclax is targeting the anti-apoptotic *BCL-2* protein, we next checked *BCL-2* mRNA expression levels in AML and overlaid the CD14+ Monocyte percentages for these cases (**Figure 3D**). The data showed that low BCL2 expression indeed correlated with strong resistance to Venetoclax. However, a few samples were resistant despite relatively high *BCL-2* expression (samples at upper right corner in **Figure 3D**). As the majority of these cases had a CD14+ Monocyte phenotype, we univariately associated *BCL-2* expression and CD14+ Monocyte percentage to investigate which factor best explains the Venetoclax response (**Figure 3E**).

We also associated metadata such as reported blast percentages, ELN status and primary diagnosis to determine their effect sizes and significance (Supplementary Table S11). This analysis demonstrated that both CD14+ Monocyte percentages (adjusted P = $1.7x10^{-18}$) and *BCL-2* mRNA expression (adjusted P = $4.8x10^{-14}$) were associated significantly with the Venetoclax response. Next, we created a multivariate model to investigate whether the significance of CD14+ Monocyte percentages diminishes along with the presence of *BCL-2* expression in the same model (**Figure 3F**, Multivariate Venetoclax Tab in Supplementary Table 12). In this model, CD14+ Monocyte percentages remained significantly associated with Venetoclax resistance (P = $1.92x10^{-5}$), while *BCL-2* expression was below the significance threshold of 0.01 (P = 0.015). In conclusion, these results indicate that cellular composition is a more robust marker than *BCL-2* mRNA expression to predict Venetoclax resistance, specifically for AML from NOS and NPM1 mutated patients.

Estimated CD14+ Monocyte percentages associates with BCL-2 protein abundance

Next, we compared the effects of *BCL-2* gene expression and CD14+ Monocyte percentages with *BCL-2* protein abundance within each sample. For this purpose, we used our in-house produced proteomics data after batch correction (**Figure S6A, S6B**) for matched AML cases (n = 39; LUMC) to correlate the abundance of Apoptosis regulator *BCL-2* protein vs gene expression (**Figure S6C**) and CD14+ Monocyte percentages (**Figure S6D**). The data showed that *BCL-2* gene expression and CD14+ Monocyte percentages both correlated with *BCL-2* protein abundance to a similar extent (R^2 = 0.45, P < 0.001). We also overlaid plots for *BCL-2* expression and CD14+ Monocyte percentages with information on patients' genetic abnormalities (**Figure 4A**). As expected, based on their low AUC for Venetoclax response (**Figure 3B**), two AML cases with PML-RARA had high *BCL-2* protein expression (n=2, dark green). We also confirmed the strong variability in AUC for Venetoclax response within the group of AML with mutated NPM1 (n=18, green) by showing two distinct groups at the extremes of CD14+ Monocyte percentages which correlated with *BCL-2* protein expression (**Figure 4B**). Within the group of AML patients with mutated NPM1, CD14+ Monocyte percentage associated stronger with both *BCL-2* protein abundance (R^2 = 0.56, P < 0.001) than *BCL-2* gene expression (R^2 = 0.52) (**Figure 4C, 4D**). In conclusion, the data demonstrate the relevance of our deconvolution approach on bulk RNA-Seq to separate AML, especially those with mutated NPM1, with high and low monocyte percentages to predict the patient's response to Venetoclax.

Discussion

In this work, we utilized single cell guided deconvolution to decompose bulk transcriptomics data from five independent AML cohorts and show that the obtained estimated cell compositions (ECCs) faithfully reconstitute the FAB landscape (Mo-M7). Moreover, using same-sample flow cytometry, we were able to validate our deconvolution framework. Hence, following our previous work using deep transcriptomics to call various types of leukemia-defining genetic aberrations[10,26], our current findings further underpin the power of transcriptomic-based approaches as a comprehensive and versatile platform for AML diagnosis. We next illustrated the potential use of our deconvolution framework for precision medicine applications, by correlating the estimated ECC to the results of an *ex vivo* drug resistance screening of 122 small-molecule inhibitors in the BEAT-AML study. For the *BCL-2* inhibitor Venetoclax we show that higher levels of the estimated ECC subset 'CD14+ Monocyte' correspond to a higher resistance, and intriguingly, that estimated CD14+ Monocyte levels is a better explanatory variable of resistance to Venetoclax than *BCL-2* expression alone. Nevertheless, using same-sample in-house proteomics data in 39 patients, we show that the estimated CD14+ Monocyte levels accurately mark BCL protein expression, and that for *NPM1*-mutated patients the presence or absence of a CD14 monocytic outgrowth corresponds with a distinct NPM1 protein abundance.

As Venetoclax was approved by the US Food and Drug Administration (FDA) in October 2020 for its use in clinics for untreated AML patients [27], other researchers also experimented with it and also observed a similar phenomenon that we had observed. For instance, Kuusanmaki *et al.* reported that monocytic differentiation of AML reduced sensitivity to Venetoclax ex vivo[8]. Similarly, Pei *et al.* reported that the different monocytic subclones in vivo created resistance to Venetoclax treatment[9], and also recently, White *et al.* showed that *BCL-2* inhibitor resistance could be predicted via the genes associated with monocytes. Recently, using the flow cytometry data from untreated MDS patients Ganan-Gomez *et al.* showed that after relapse and becoming secondary AML (sAML) patients, those patients with less maturated cell types (EMP) before treatment had a faster complete remission (CR) and longer relapse free status compared to more matured cell type (GMP) with Venetoclax treatment, supporting our hypothesis that ECCs could also predict Venetoclax resistance for MDS and sAML patients[28]. Collectively, these studies suggest that Venetoclax has different resistances at different maturational stages, and especially higher resistance for patients with CD14+

Monocyte dominated phenotype as we have shown in this manuscript, and we provide an open-source framework, *seAMLess*, for replicating our results or applying it to other clinically relevant datasets.

Our findings may potentially have important implications on drug use especially for genetically uncharacterized patients (AML, NOS) currently accounting for ~40% of all AML as well as other well-characterized patients such as NPM1 mutated samples. For instance, one interesting case report was from Korea [29], an elderly AML patient was treated with Venetoclax along with Azacitidine, and even though the patient had a favorable NPM1 status this treatment lead into tumor lysis syndrome (TLS) and patient ended up in ER 12 h into the treatment. Although the manuscript does not indicate the FAB status, AML cells expressed CD33, CD117 and myeloperoxidase, demonstrating a more stem cell than monocytic phenotype, thus one possible hypothesis could be the patient having a more sensitive cell types (e.g., EMP, GMP) to Venetoclax and thus causing immediate apoptosis of these cells and leading to TLS.

One limitation of our deconvolution strategy is that it cannot distinguish the cancerous cell types as it uses a healthy bone marrow as a reference. The reason we wanted to emphasize this is, for instance in **Figure 4D** there is a relapsed NPM1 patient with 41% of CD14+ Monocytes, as deconvolution percentage might not be exact, this patient could potentially have a healthy monocytic population but rather the reason for his relapse could be due to a different cell type, possibly 17% EMP as it is the second most abundant cell type for this patient. However, we have two rationales behind not using a single cell reference with cancerous cells. First, without mutation calling for all cells, one cannot be sure whether a cell is cancerous or not. Strategies like predicting cancer cells based on their transcriptional similarity of cells with mutation calling, proposed by Van Galen *et al.*[16], adds another level of ambiguity to already not perfect deconvolution pipelines. Secondly and more importantly, heterogeneity of AML causes further sub-clustering within individual AML cases (e.g. UMAP plots of Triana *et al.*, 2021), therefore creating a not well-characterized cell type signature but rather patient specific clusters[17]. Additionally, using healthy subsets as reference allows our framework to provide more interpretable and intuitive results for clinicians and doctors, as it reports immune phenotypes and percentages on contrary to score-based prognostic values[15,24,26]. To summarize, we believe our proposed pipeline could be a blueprint for assessing new drugs' resistances on different cell types of AML and along

with our framework, they may provide better insights for clinicians and help paving the way into precision medicine in AML.

Methodology

Creating the healthy BM single-cell reference

We downloaded three different healthy BM datasets from two different studies, namely Van Galen et. al.[16] (full-transcriptome, n=6,915), and Triana et. al.[17] (full-transcriptome, n = 13,165; 462 targeted mRNA, n=49,057). Then, all cell labels were uplifted up via *Seurat* package[30] (v4.0.3) default query annotation pipeline to match with the Triana's full-transcriptome cell labels as it had the most recent and detailed labels. Next, *Seurat*'s integration pipeline with *CCA* was run and the cells that were labeled as doublets/multiplets were removed from the down-stream analyses, and this yielded a healthy BM atlas of 69,130 cells in total, covering 439 genes. Also, we used the 40,000 cell subset of HCA provided by *SeuratObject* package (v4.0.2) for the validation analyses (**Figure 1C, 1D**).

Different schemes of creating pseudobulks

To create a cancerous-like pseudobulk profile from healthy BM reference and HCA subset, we selected a total of 1,000 cells for each profile, majority (80%) of them coming from a cell-type (over-abundant) and the rest of the cells were distributed according to inverse proportion of the numbers of cells for the remaining cell types. To achieve this, first integrated counts were exponentiated to make them non-log scale and *slice_sample* function from *dplyr* package (v1.0.7) was used with a replacement option. Then, these non-log scale cell counts were summed up to create the pseudobulk profiles. For individual-based pseudobulk, *AggregateExpression* from *Seurat* package was used.

Flow Cytometry data

AML cases were stained with fluorescent antibodies and analyzed by flow cytometry for diagnosis, prognosis, and disease monitoring of AML in the diagnostic laboratory of the department of Hematology in the Leiden

University Medical Center. The flow cytometric test has been developed and performed according to EuroFlow standard operating procedures (www.euroflow.org)[19]. EuroFlow antibody combinations have been tested against references databases of normal cells from healthy individuals and allow multidimensional identification and distinction of aberrant cells from normal cell populations. The flow cytometric test includes 8 tubes with different antibody combinations, i.e. one ALOT tube (acute leukemia orientation tube) and 7 AML tubes (AML1-7). The ALOT tube contains antibodies against CD3, CD45, MPO, CD79, CD19 and CD7. The AML tubes contain antibodies against CD16, CD13, CD34, CD117, CD11b, CD10, HLA-DR and CD45 (AML-1), CD35, CD64, CD34, CD117, CD300e/IREM2, CD14, HLA-DR and CD45 (AML-2), CD36, CD105, CD34, CD117, CD33, CD71, HLA-DR and CD45 (AML-3), NuTdT, CD56, CD34, CD117, CD7, CD19, HLA-DR and CD45 (AML-4), CD15, NG2, CD34, CD117, CD22, CD38, HLA-DR and CD45 (AML-5), CD42a and CD61, CD203c, CD34, CD117, CD123, CD4, HLA-DR and CD45 (AML-6). AML tube 7 contains antibodies against CD41, CD25, CD34, CD117, CD42b, CD9, HLA-DR and CD45, but this tube has not been used to stain AML cases analyzed in this study.

Obtaining bulk transcriptomic data and their preprocessing

We have downloaded the non-normalized count matrices (htseq-counts) and the meta files of the four discovery cohorts (TCGA-LAML, BEAT-AML, TARGET-AML and TARGET-ALL) from https://portal.gdc.cancer.gov. For LEUCEGENE, count data was downloaded from their dedicated site (https://data.leucegene.iric.ca/) along with their provided meta data. All meta/count data were pre-processed using *R* (v4.1.0). For the meta data, genomic aberration labels were relabeled to the main AML WHO 2016 classes, non-AML samples were removed from the down-stream analyses, ELN-classes were relabeled according to ELN 2017 recommendations. For the count data, ERCC spike-ins and mitochondrial genes were removed, and the count matrix was then sorted according genes standard deviation in order to remove the duplicated genes that had less variation thus providing less information, and lastly the gene ensembl ids were converted to gene symbols. Before converting ensembl ids into gene symbols, the stemness score for each patient was calculated via count-per-million (cpm) normalized libraries, and these libraries were normalized using *cpm* function from *edgeR* package[31] (v.3.34.1).

Our 100 AML samples (LUMC) deposited to EGA with accession number EGAS00001003096 and they are accessible upon request. QC benchmark analyses for these samples were done in our previous paper[10].

Therefore, we ran default *HT-SEQ* pipeline (v0.11.2) with paired-end option aligning fastq files to *hg38* to obtain the count matrix. All above mentioned preprocessing steps (filtering, gene name conversion) were also conducted for these samples as well before deconvolution.

Deconvolution pipeline

To deconvolute the simulated pseudobulks and bulk RNA-seq AML patients, we used *MuSiC*[12] package (v0.1.1) as it benchmarked highly and consistently across different cell types at various settings[11] and had an easy-to-use open-source (GPL-3) implementation (https://github.com/xuranw/MuSiC). We used non-log scaled count values as inputs and set the normalization option to false. Patients were assigned to the groups (e.g., GMP, CD14+ Monocyte etc.) according to their most abundant deconvoluted ECCs. In the heatmap (**Figure 2A**), patients were re-ordered according to their ECCs within each assignment.

UMAP of estimated cell compositions

First, to obtain reproducible results with umap plots, we set a seed to 2 as UMAP procedure involves random initialization. Then, we ran *umap* function with default parameters from *umap* package (v0.2.7) and used the first two reduced dimensions to create the plots. All related figures were plotted using *ggplot2* package (v3.3.5).

Drug resistance predictions via Random Forest

BEAT-AML has drug resistance data for 122 small-molecule inhibitors, we downloaded these from their manuscript (Supplementary Table S5). Then, each drug response was min-max normalized, then matched with their available RNA-seq samples. Next, drug resistances were predicted with the deconvoluted ECCs. Random forest algorithm from *randomForest* package (v.4.6-14) with default parameters was used for the predictions. Each sample within each drug was predicted at leave-one-out cross-validation settings. Then, for each drug, Spearman ρ values were calculated between predicted and actual drug resistance values. To stratify the drugs according to their primary diagnosis of WHO classification, samples within each diagnosis are selected and then each drugs predicted, and normalized drug resistance were associated, then the correlation and p-values were calculated using *summary* function in base R.

Venetoclax Association Analysis

First, each attribute was associated to standardized (min-max normalized) Venetoclax resistance from BEAT-AML study at univariate settings using *lm* function in R environment (v4.1.0) (Supplementary Table S11). Then, p-values and coefficients were calculated using *summary* function and then p-values were multiple hypothesis corrected using Benjamini-Hochberg procedure. For multivariate models, FAB classification, and ECC levels (except CD14+ Monocytes) were excluded as only 76 out of 460 samples of BEAT-AML had FAB classifications and as other ECCs are not independent of CD14+ Monocyte percentages (as the question is whether CD14+ Monocyte levels are independently predictive of Venetoclax resistance given *BCL-2* expression in the same model). Again, p-values and coefficients were calculated with *summary* function and plotted in a volcano plot (Figure 3F, Supplementary Table S12).

Proteomics Sample preparation

Cell lysis, digestion and TMT labeling was performed as described in Paula and et al[32]. Cell lysis was performed using 5% SDS lysis buffer (100 mM Tris-HCl pH7.6) and 5 U benzonase nuclease (Thermo Scientific) with incubation at 95 °C for 4 minutes. Protein concentration was determined using Pierce BCA Gold protein assay (Thermo Fisher Scientific). 100 µg protein of each sample was then reduced with 5 mM TCEP. Reduced disulfide bonds were alkylated using 15 mM iodoacetamide. Excess iodoacetamide was quenched using 10 mM DTT. Protein lysates were precipitated using chloroform/methanol; resulting pellets were re-solubilized in 40 mM HEPES pH 8.4 and digested using TPCK treated trypsin (1:12.5 enzyme/protein ratio) overnight at 37 °C. Peptide concentration was then determined using Pierce BCA Gold protein assay.

The different samples, and reference samples, were arranged into five TMTpro 16plex sets. The peptides were labeled with TMTpro Label Reagents (Thermo Fisher Scientific) in a 1:4 ratio by mass (peptides/TMT reagents), total volume was 35 µL, for 1 h at RT. Excess TMT reagent was quenched with 5 µL 6% hydroxylamine for 15 min at RT. Samples corresponding to a TMT set were pooled and lyophilized.

Each TMT sample (80 ug) was fractionated by high pH reverse phase chromatrography on a Zorbax RRHD Eclipse Plus C18 2.1x150 mm 1.8-micron column, at 800 ul/min using an Agilent1200 binary HPLC system, equipped with a UV detector. The mobile phases were 10 mM Ambic pH 8.4 (A) and 10 mM Ambic/Acetonitrile 20/80 pH 8.4 (B). The gradient was from 2% to 90%B in 35 minutes. 20 fractions were

collected in a circular fashion, i.e., collection per vial for 20 sec before moving to the next collection vial. After collection in the last vial collection is continued in the first vial. Fractions were subsequently freeze dried.

Mass spectrometry

TMT-labeled peptide fractions were dissolved in water/formic acid (100/0.1 v/v) and analyzed by on-line C18 nanoHPLC MS/MS with a system consisting of an Ultimate3000nano gradient HPLC system (Thermo, Bremen, Germany), and an Exploris480 mass spectrometer (Thermo) as in Rossi et al[33]. Fractions were injected onto a cartridge precolumn (300 µm × 5 mm, C18 PepMap, 5 um, 100 A, and eluted via a homemade analytical nano-HPLC column (50 cm × 75 µm; Reprosil-Pur C18-AQ 1.9 um, 120 A (Dr. Maisch, Ammerbuch, Germany)). Solvent A was water/formic acid 100/0.1 (v/v). The gradient was run from 2% to 40% solvent B (20/80/0.1 water/acetonitrile/formic acid (FA) v/v) in 120 min. The nano-HPLC column was drawn to a tip of ~10 µm and acted as the electrospray needle of the MS source. The mass spectrometer was operated in data-dependent MS/MS mode with a cycle time of 3 seconds, with the HCD collision energy at 36 V and recording of the MS2 spectrum in the orbitrap, with a quadrupole isolation width of 1.2 Da. In the master scan (MS1) the resolution was 120,000, the scan range 350-1200, at standard AGC target @maximum fill time of 50 ms. A lock mass correction on the background ion m/z=445.12003 was used. Precursors were dynamically excluded after n=1 with an exclusion duration of 45 s, and with a precursor range of 20 ppm. Charge states 2-5 were included. For MS2 the first mass was set to 110 Da, and the MS2 scan resolution was 45,000 at an AGC target of 200% @maximum fill time of 60 ms.

Proteomics data processing and down-stream analysis

In a post-analysis process, raw data were first converted to peak lists using Proteome Discoverer version 2.4 (Thermo Electron), and submitted to the Uniprot database (Homo sapiens, 20596 entries), using Mascot v. 2.2.07 (www.matrixscience.com) for protein identification. Mascot searches were performed with 10 ppm and 0.02 Da deviation for precursor and fragment mass, respectively, and the enzyme trypsin was specified. Up to two missed cleavages were allowed. Methionine oxidation and acetyl on protein N-terminus were set as variable modifications. Carbamidomethyl on Cys and TMTpro on Lys and N-terminus were set as fixed modifications. Protein and peptide FDR were set to 1%. Normalization was on the total peptide amount. The 5 TMT-16plex analyses were normalized to each other by the bridge samples.

First, the abundance data is log-cpm transformed to stabilize variance among samples and then to ensure dealing with the technical batch effects, we ran *removeBatchEffect* function from *limma* package (v3.48.3) providing batch information. Then, we overlaid the transformed protein abundances onto a PCA plot to observe the clearance of batch related clustering. Then, the cell lines (positive-control) and controls (negative-control) were removed from the down-stream analyses, selecting only primary AML (n=39). Next, the transformed BCL-2 protein abundance was associated with the BCL-2 expression from LUMC cohort (**Figure 4A**) and CD14+ Monocyte percentage within each sample (**Figure 4B**). R^2 and p-values were calculated using *stat_poly_eq* function from *ggpmisc* (v0.4.5) package and lines were drawn with *geom_smooth* function from ggplot2 package (v3.3.5) via method option set to linear model.

Data analysis and reproducibility

All analyses performed in this study are provided with the following link: https://github.com/eonurk/lumc-sc-aml. We have also implemented an GPL-3 licensed R package at https://github.com/eonurk/seAMLess, which deconvolutes a given bulk RNA-seq count matrix to 22 cell types from our single-cell reference and predicts drug resistances via RF algorithm.

Data Sharing Statement

For original data, please contact: e.b.van_den_akker@lumc.nl

Acknowledgements

Authors want to thank Daniele Bizzarri for his fruitful critics on the project. This study was funded by a strategic investment of the Leiden University Medical Center, embedded within the Leiden Oncology Center, and executed within the Leiden Center for Computational Oncology. EvdA was funded by a personal grant of the Dutch Research Council (NWO; VENI: 09150161810095). The funding bodies had no role in the study design, the collection, analysis, and interpretation of data, the writing of the manuscript, and the decision to submit the manuscript for publication. Authors declare no competing interests.

Authorship and conflict-of-interest statements

Conceptualization: MJR, EvdA, MG

Resources: EvdA

Methodology: EOK, EvdA, PvA, AMO

Investigation: EOK, JS, MG, EvdA, PvA

Visualization: EOK

Funding acquisition: EvdA

Project administration: EvdA

Supervision: MJR, MG, EvdA

Writing – original draft: EOK, MG, EvdA

Writing – review & editing: All authors

Authors declare no conflict of interest.

Figure Captions

Figure 1 – Schematic of the study and compositional validation

A The overview of the study. Integrated cells for healthy bone marrow (BM) dataset were collected from two studies[16,17]. Five independent bulk transcriptomic AML cohort (TCGA-LAML, BEAT-AML, TARGET-AML, LEUCEGENE and LUMC) samples (n=1,350) were deconvoluted. Drug resistance data (n=122) from BEAT-AML were predicted using ECCs. Additional ALL samples (n=532) from TARGET study were used for validating deconvolution framework. Proteomics (n=39) and flow cytometry samples (n=22) from LUMC cohort were used for further validation of the framework. **B** UMAP plot of integrated healthy BM (n=69,101) cells. Annotations are lifted via Azimuth framework. T and B cell subset were merged, resulting into 22 cell types (abbreviations; GMP: Granulocyte Monocyte Progenitors, LMPP: Lymphoid Primed Multipotent Progenitors, NK: Natural Killer, EMP: Erythroid Megakaryocyte Progenitor, pDC: Plasmacytoid Dendritic Cell, CLP: Common Lymphoid Progenitor, HSC: Hematopoietic Stem Cell, BaEoMA: Basophil Eosinophil Mast Progenitor, Prog Mk: Progenitor Megakaryocyte, pre-mDC: Precursor Myeloid Dendritic Cell, ASDC: AXL+ Dendritic Cell). **C** Barplot for cells per cell type of the healthy BM. X-axis is in log10 scale. **D** Heatmap showing the fraction of deconvoluted cell types for simulated pseudobulks from HCA subset. Each row indicates a simulated pseudobulk with an overabundant cell type (80%) and adds up to 1, and each column is a cell type from the healthy BM reference. **E** Cellular composition of the same samples via flow gating of orientation

tube (ALOT) and ECCs (RNA-seq). Abnormal assignment within myeloid cells is determined by flow gating, thus only exists for flow cytometry. **F** Percentage of monocytic subsets measured by flow cytometry and ECCs. **G** Ternary plot showing the ECCs of TARGET AML and ALL cohorts (n=719), each dot represents a sample and corners indicate a major cell type and colors indicate the primary diagnosis of each leukemic sample (Supplementary Table S1).

Figure S1 – Single-cell integration evaluation & flow markers

A Healthy BM reference overlaid on UMAP separated by samples **B** and by different studies. **C** Similar to Figure 1D; heatmap showing the deconvolution results of simulated pseudobulks from healthy BM reference via an over-abundant cell type (80%). **D** UMAP for HCA subset with the cell type annotations, which were lifted via Azimuth framework. **E** Dot plot of MFI values of all EuroFlow markers for all samples, size of the dot represents the abnormal cell percentages per sample assigned with EuroFlow analyses. **F** MFI values of four EuroFlow monocytic markers (CD11b, CD64, IREM2 and CD14) and CD14+ Monocyte %. Colors indicate the abnormal percentage of cells by flow cytometry assignments.

Figure 2 – Capturing differentiation stages with ECCs

A Heatmap showing the deconvoluted fractions for of five AML cohorts (TCGA-LAML, BEAT-AML, TARGET-AML, LEUCEGENE and LUMC). Each column represents a sample, which is deconvoluted into 22 cell types (bottom part with blue) via healthy single cell BM reference. Top part of the annotations shows the provided meta data (FAB, WHO and ELN classifications, origin, sex and reported blast percentages) for these cohorts. Each patient was assigned to its most abundant deconvoluted cell type (AML phenotype) and samples within each assignment were sorted according to the assigned phenotype. **B** UMAP plot of deconvolution levels annotated via most abundant cell type as the left panel; and on the right, ECCs of 6 out of 22 cell types were shown in continuous scale (HSC, CD14+ Monocytes, Late Erythrocytes, GMP, EMP and cDC). **C** Similar to UMAP plot in (B) but colored with FAB classifications (left panel), and the fractions of cell types for each of these FAB classifications (right panel).

Figure S2 – Overview of AML cohorts

A Detailed table for sample characteristics for age, sex, FAB, ELN, and primary diagnosis (WHO) for each cohort. Missing data is labeled as 'Unknown' for each attribute. *AML, NOS patients represents %39 of WHO classes. **B** Ternary plots like Figure 1G for all five AML cohorts.

Figure S3 – Stemness score and meta data on UMAP

A Box plots of blast percentage vs AML phenotype and **B** Stemness score vs AML phenotype. Boxplot were sorted according to median values. **C** Compositional changes of each cohort (except LEUCEGENE as it lacks entire list of genes to calculate stemness score) for low and high stemness score (split by median). Violin plots show the deconvoluted cellular percentages for 22 cell type and heatmap on the right panel summarizes the changes. A linear model is fit to calculate the slope and significance per cell type vs. stemness, and a red slope indicates more of that cell type with higher stemness score, and blue is the vice versa. Asterix indicates a statistically significant trend ($P < 0.05$). **D** UMAP plots of deconvoluted percentages (ECCs) annotated via age, study, reported blast level, sex, ELN and stemness score.

Figure 3 – *Ex-vivo* drug response predictions

A Correlation of predicted AUC and normalized AUC of *ex-vivo* drug resistances split by WHO classes (top 10 drugs). Bigger bubble corresponds to higher correlation, and colors indicate statistical significance. Only groups with >5 samples were fit linear models thus R^2 and p values are shown. **B** Random Forest predictions of Venetoclax resistance via ECCs with LOOCV ($\rho = 0.509$; Supplementary Table - BEAT AUC RF Predictions). Each dot represents a patient. Higher resistance (AUC) means more cells are viable after administering the drug. **C** Samples with normalized Venetoclax resistance data are shown on the UMAP plot of ECCs, along with phenotype map from (**Figure 2B**). Patients without available data were shown as gray dots for the sake of completeness. **D** Venetoclax resistance vs *BCL-2* expression; also, CD14+ Monocyte percentages for these patients were annotated. Expression data was count-per-million (cpm) normalized. **E** Univariate analyses of different attributes (i.e., CD14+ Monocyte %, reported blast percentages etc.) vs normalized Venetoclax resistance. P values were corrected using Benjamini-Hochberg procedure and the names of the attributes are shown for FDR > 0.01. **F** Similar to **E**; A multivariate model is associated with different attributes, and the names of the attributes with P> 0.01 is shown. Full list of attributes' p values and estimates are provided in Supplementary Table for both **E** and **F**.

Figure S4 – Summary plots for *ex-vivo* drug response predictions

A Spearman ρ values for the predictions of the drug resistance of 122 small molecule inhibitors from BEAT-AML (see Supplementary Table S7). For each drug, the resistance per sample is predicted via random forest algorithm using ECCs at leave-one-out cross validation setting. **B** Dot plot showing the stratified associations with predicted resistance values. Colors of the dot show the significance and sizes indicate the correlation strength. Groups with more than 5 samples were included in the plot.

Figure S5 – Venetoclax response on different WHO classes

A Flavopiridol resistance predictions with random forest annotated with patients' AML phenotype. **B** Flavopiridol resistance predictions are faceted and highlighted by WHO classes, classes with >5 samples were fit linear models thus R^2 and p values are shown. **C** Venetoclax resistance predictions with random forest annotated with patients' AML phenotype. **D** Venetoclax resistance split by AML phenotypes. All drugs' one-way anova p values are provided in Supplementary Table S9 **E** UMAP plot showing Venetoclax resistance only for NPM1 mutated samples **F** Venetoclax resistance split by AML phenotype only for NPM1 mutated samples.

Figure 4 – Proteomics data and ECC relations

A *BCL-2* expression vs protein abundance (n = 39) annotated with genetic abnormalities ($R^2 = 0.45$, $P < 0.001$). Each dot represents a sample. **B** CD14+ Monocytes percentage vs *BCL-2* protein abundance annotated with genetic abnormalities ($R^2 = 0.45$, $P < 0.001$). **C** Similar to A; but only limited to NPM1 mutated samples (n = 18) and annotated with patients' AML phenotypes (most abundant cell type) ($R^2 = 0.52$, $P < 0.001$) **D** Similar to B; but only limited to NPM1 mutated samples (n = 18) and annotated with patients' AML phenotypes (most abundant cell type) ($R^2 = 0.56$, $P < 0.001$).

Figure S6 – QC plots & BCL-2 expression for proteomics

A All-proteomics data (n = 39) without batch correction and **B** batch corrected (with *limma*) log-cpm normalized proteomics data. All abundances are log-cpm normalized. **C** *BCL-2* expression vs protein abundance annotated with patients' AML phenotypes ($R^2 = 0.45$, $P < 0.001$). **D** CD14+ Monocytes percentage vs *BCL-2* protein abundance annotated with patients' AML phenotypes ($R^2 = 0.45$, $P < 0.001$).

Supplementary Tables

Tab 1 TARGET deconvolutions

Deconvolutions of TARGET AML and ALL cohorts

Tab 2 Deconvolution Results

Estimated cell compositions for 22 cell types along with meta data of 1,350 samples

Tab 3 EuroFlow MFI values

EuroFlow MFI values of abnormal populations of LUMC samples

Tab 4 EuroFlow cell type percentages

EuroFlow gating results for different cell types (abnormals are not annotated) of LUMC samples

Tab 5 BEAT AUC

Min-max normalized AUC values for 122 small-molecule inhibitors from BEAT-AML

Tab 6 BEAT AUC RF Predictions

Predictions of each drug resistance with random forest at LOOCV

Tab 7 Drug Spearman Rho

Spearman rho values for drug resistance predictions of 122 small-molecule inhibitors

Tab 8 Drug Associations stratified with WHO

Predicted drug and normalized resistance associations within each primary diagnosis

Tab 9 Anova - Drug vs Cell Types (All)

One-way anova test results with sample phenotype assignments for drug resistances

Tab 10 Anova - Drug vs Cell Types (NPM1)

One-way anova test results with sample phenotype assignments for drug resistances (NPM1 only)

Tab 11 Univariate Venetoclax

The results of linear model associations of Venetoclax resistance to different attributes

Tab 12 Multivariate Venetoclax

The results of linear model association of Venetoclax resistance with different attributes

References

1. Cancer Facts & Figures 2022. American Cancer Society; .

2. Bennett JM, Catovsky D, Daniel M-T, et al. Proposals for the Classification of the Acute Leukaemias French-American-British (FAB) Co-operative Group. *Br. J. Haematol.* 1976;33(4):451–458.

3. Vardiman JW, Thiele J, Arber DA, et al. The 2008 revision of the World Health Organization (WHO) classification of myeloid neoplasms and acute leukemia: rationale and important changes. *Blood*. 2009;114(5):937–951.

4. Arber DA, Orazi A, Hasserjian R, et al. The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood*. 2016;127(20):2391–2405.

5. Khoury JD, Solary E, Abla O, et al. The 5th edition of the World Health Organization Classification of Haematolymphoid Tumours: Myeloid and Histiocytic/Dendritic Neoplasms. *Leukemia*. 2022;36(7):1703–1719.

6. Döhner H, Estey E, Grimwade D, et al. Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood*. 2017;129(4):424–447.

7. Canaani J, Beohou E, Labopin M, et al. Impact of FAB classification on predicting outcome in acute myeloid leukemia, not otherwise specified, patients undergoing allogeneic stem cell transplantation in CR1: An analysis of 1690 patients from the acute leukemia working party of EBMT: CANAANI ET AL. *Am. J. Hematol.* 2017;92(4):344–350.

8. Kuusanmäki H, Leppä A-M, Pölönen P, et al. Phenotype-based drug screening reveals association between venetoclax response and differentiation stage in acute myeloid leukemia. *Haematologica*. 2020;105(3):708–720.

9. Pei S, Pollyea DA, Gustafson A, et al. Monocytic Subclones Confer Resistance to Venetoclax-Based Therapy in Patients with Acute Myeloid Leukemia. *Cancer Discov.* 2020;10(4):536–551.

10. Arindrarto W, Borràs DM, de Groen RAL, et al. Comprehensive diagnostics of acute myeloid leukemia by whole transcriptome RNA sequencing. *Leukemia*. 2021;35(1):47–61.

11. Avila Cobos F, Alquicira-Hernandez J, Powell JE, Mestdagh P, De Preter K. Benchmarking of cell type deconvolution pipelines for transcriptomics data. *Nat. Commun.* 2020;11(1):5650.

12. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nat. Commun.* 2019;10(1):380.

13. Dai C, Chen M, Wang C, Hao X. Deconvolution of Bulk Gene Expression Profiles with Single-Cell Transcriptomics to Develop a Cell Type Composition-Based Prognostic Model for Acute Myeloid Leukemia. *Front. Cell Dev. Biol.* 2021;9:762260.

14. Li H, Sharma A, Ming W, Sun X, Liu H. A deconvolution method and its application in analyzing the cellular fractions in acute myeloid leukemia samples. *BMC Genomics*. 2020;21(1):652.

15. Zeng AGX, Bansal S, Jin L, et al. A cellular hierarchy framework for understanding heterogeneity and predicting drug response in acute myeloid leukemia. *Nat. Med.* 2022;28(6):1212–1223.

16. van Galen P, Hovestadt V, Wadsworth II MH, et al. Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. *Cell*. 2019;176(6):1265-1281.e24.

17. Triana S, Vonficht D, Jopp-Saile L, et al. Single-cell proteo-genomic reference maps of the hematopoietic system enable the purification and massive profiling of precisely defined cell states. *Nat. Immunol.* 2021;22(12):1577–1589.

18. Hay SB, Ferchen K, Chetal K, Grimes HL, Salomonis N. The Human Cell Atlas bone marrow single-cell interactive web portal. *Exp. Hematol.* 2018;68:51–61.

19. on behalf of the EuroFlow Consortium (EU-FP6, LSHB-CT-2006-018708), van Dongen JJM, Lhermitte L, et al. EuroFlow antibody panels for standardized n-dimensional flow cytometric immunophenotyping of normal, reactive and malignant leukocytes. *Leukemia*. 2012;26(9):1908–1975.

20. The Cancer Genome Atlas Research Network. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *N. Engl. J. Med.* 2013;368(22):2059–2074.

21. Tyner JW, Tognon CE, Bottomly D, et al. Functional genomic landscape of acute myeloid leukaemia. *Nature*. 2018;562(7728):526–531.

22. Farrar JE, Schuback HL, Ries RE, et al. Genomic Profiling of Pediatric Acute Myeloid Leukemia Reveals a Changing Mutational Landscape from Disease Diagnosis to Relapse. *Cancer Res.* 2016;76(8):2197–2205.

23. MacRae T, Sargeant T, Lemieux S, et al. RNA-Seq Reveals Spliceosome and Proteasome Genes as Most Consistent Transcripts in Human Cancer Cells. *PLoS ONE*. 2013;8(9):e72884.

24. Ng SWK, Mitchell A, Kennedy JA, et al. A 17-gene stemness score for rapid determination of risk in acute leukaemia. *Nature*. 2016;540(7633):433–437.

25. Jiang F, Wang X-Y, Wang M-Y, et al. An Immune Checkpoint-Related Gene Signature for Predicting Survival of Pediatric Acute Myeloid Leukemia. *J. Oncol.* 2021;2021:1–14.

26. Docking TR, Parker JDK, Jädersten M, et al. A clinical transcriptome approach to patient stratification and therapy selection in acute myeloid leukemia. *Nat. Commun.* 2021;12(1):2474.

27. FDA grants regular approval to venetoclax in combination for untreated acute myeloid leukemia. U.S. Food and Drug Administration; 2020.

28. Ganan-Gomez I, Yang H, Ma F, et al. Stem cell architecture drives myelodysplastic syndrome progression and predicts response to venetoclax-based therapy. *Nat. Med.* 2022;28(3):557–567.

29. Kim M, Bang H-J, Song G-Y, et al. Venetoclax with Azacitidine Induced Tumor Lysis Syndrome in an Elderly Patient with Acute Myeloid Leukemia: A Case Report. *Electrolytes Blood Press.* 2021;19(2):46.

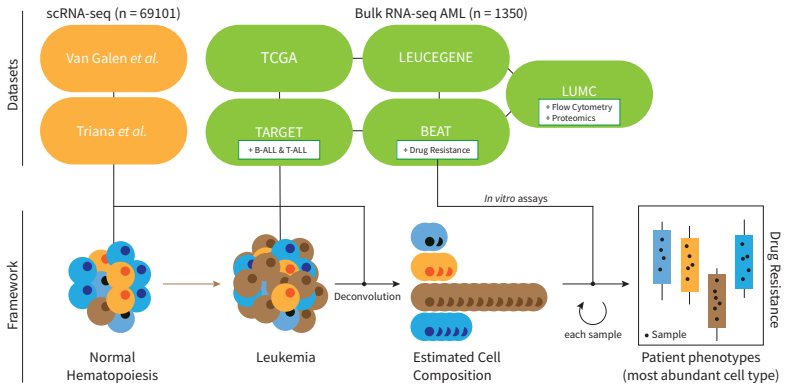
30. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. *Cell.* 2021;184(13):3573-3587.e29.

31. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinforma. Oxf. Engl.* 2010;26(1):139–140.

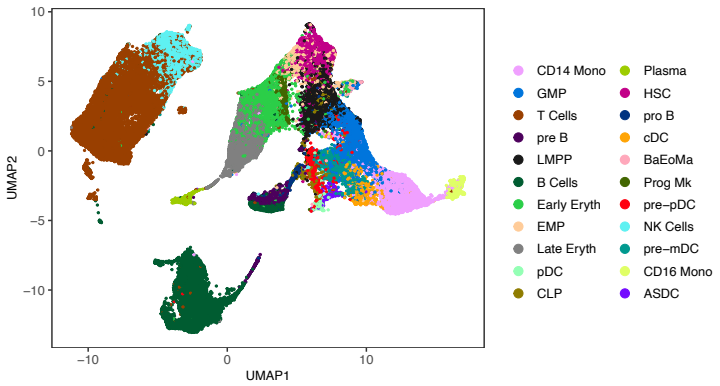
32. Paulo JA, Gygi SP. Nicotine-induced protein expression profiling reveals mutually altered proteins across four human cell lines. *Proteomics*. 2017;17(1–2):.

33. Rossi M, Altea-Manzano P, Demicco M, et al. PHGDH heterogeneity potentiates cancer cell dissemination and metastasis. *Nature*. 2022;605(7911):747–753.

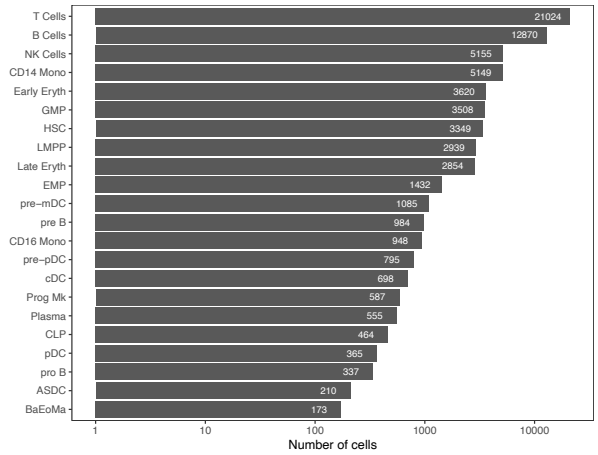
A Overview of the pipeline



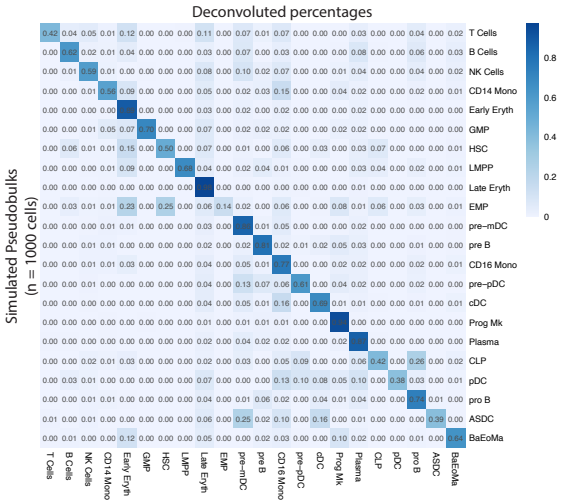
B Integrated single-cell healthy BM reference



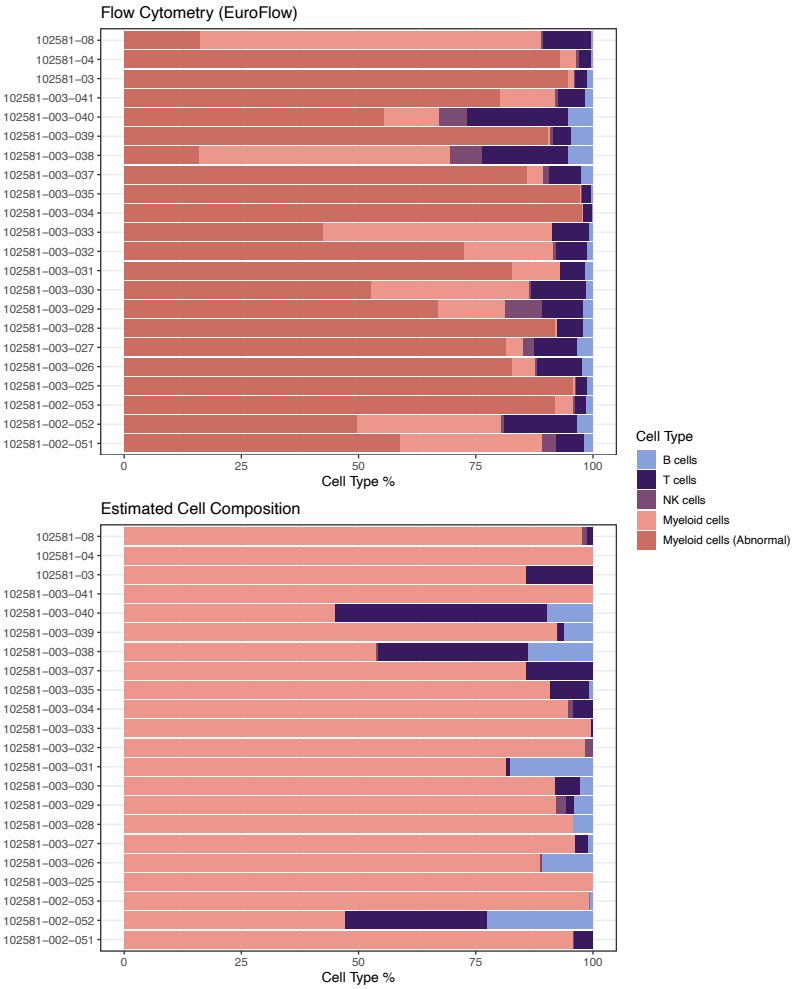
C Number of cells in each cell type (Healthy BM reference)



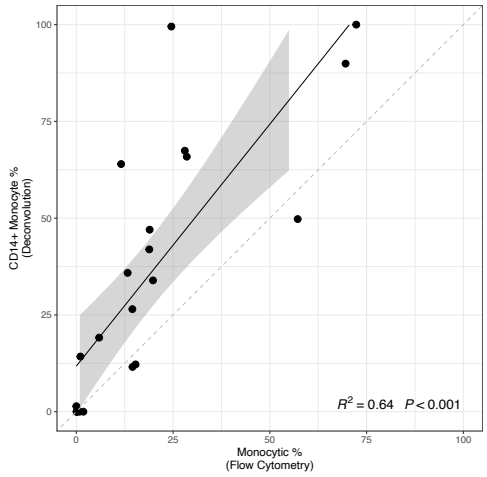
D Deconvolution of in-silico pseudobulks from Human Cell Atlas (HCA)



E Comparison of flow cytometry and estimated cell composition



F Monocytic levels with flow cytometry and estimated cell composition



G ECCs of TARGET-AML and -ALL

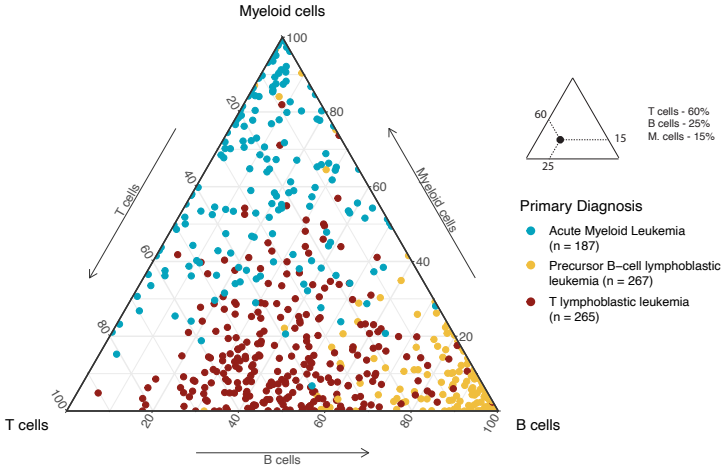


Figure 2

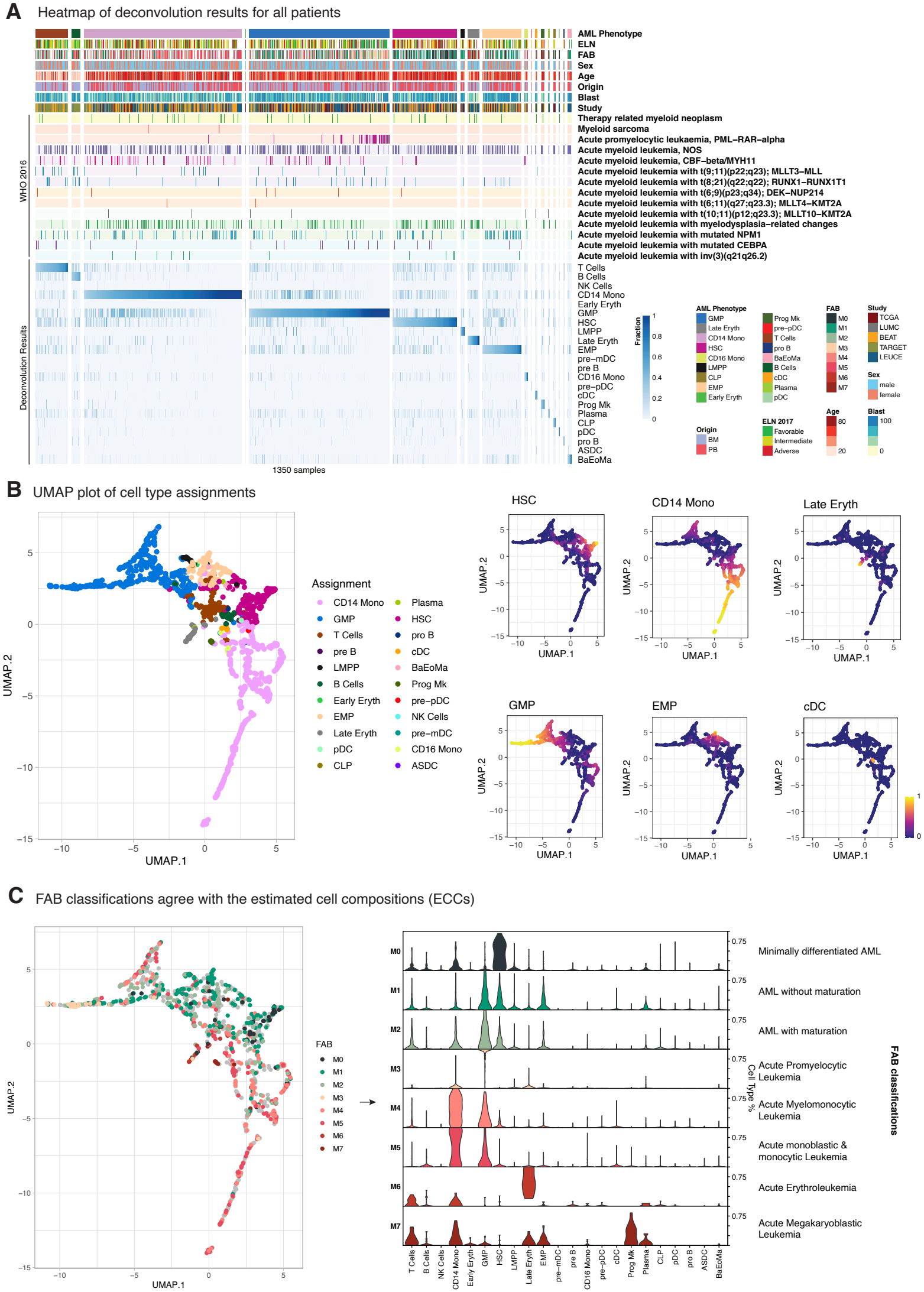

A
Heatmap of deconvolution results for all patients
AML Phenotype
ELN
FAB
Sex
Age
Origin
Blast
Study
WHO 2016
Therapy related myeloid neoplasm
Myeloid sarcoma
Acute promyelocytic leukaemia, PML–RAR–alpha
Acute myeloid leukemia, NOS
Acute myeloid leukemia, CBF–beta/MYH11
Acute myeloid leukemia with t(9;11)(p22;q23); MLLT3–MLL
Acute myeloid leukemia with t(8;21)(q22;q22); RUNX1–RUNX1T1
Acute myeloid leukemia with t(6;9)(p23;q34); DEK–NUP214
Acute myeloid leukemia with t(6;11)(q27;q23.3); MLLT4–KMT2A
Acute myeloid leukemia with t(10;11)(p12;q23.3); MLLT10–KMT2A
Acute myeloid leukemia with myelodysplasia–related changes
Acute myeloid leukemia with mutated NPM1
Acute myeloid leukemia with mutated CEBPA
Acute myeloid leukemia with inv(3)(q21q26.2)
Deconvolution Results
T Cells
B Cells
NK Cells
CD14 Mono
Early Eryth
GMP
HSC
LMPP
Late Eryth
EMP
pre–mDC
pre B
CD16 Mono
pre–pDC
cDC
Prog Mk
Plasma
CLP
pDC
pro B
ASDC
BaEoMa
1350 samples
Fraction
AML Phenotype
GMP
Late Eryth
CD14 Mono
HSC
CD16 Mono
LMPP
CLP
EMP
Early Eryth
Prog Mk
pre–pDC
T Cells
pro B
BaEoMa
B Cells
cDC
Plasma
pDC
FAB
M0
M1
M2
M3
M4
M5
M6
M7
Study
TCGA
LUMC
BEAT
TARGET
LEUCE
Sex
male
female
Origin
BM
PB
ELN 2017
Favorable
Intermediate
Adverse
Age
Blast
B
UMAP plot of cell type assignments
Assignment
CD14 Mono
GMP
T Cells
pre B
LMPP
B Cells
Early Eryth
EMP
Late Eryth
pDC
CLP
Plasma
HSC
pro B
cDC
BaEoMa
Prog Mk
pre–pDC
NK Cells
pre–mDC
CD16 Mono
ASDC
UMAP.1
UMAP.2
HSC
CD14 Mono
Late Eryth
GMP
EMP
cDC
C
FAB classifications agree with the estimated cell compositions (ECCs)
FAB
M0
M1
M2
M3
M4
M5
M6
M7
Minimally differentiated AML
AML without maturation
AML with maturation
Acute Promyelocytic Leukemia
Acute Myelomonocytic Leukemia
Acute monoblastic & monocytic Leukemia
Acute Erythroleukemia
Acute Megakaryoblastic Leukemia
Cell Type %
FAB classifications
T Cells
B Cells
NK Cells
CD14 Mono
Early Eryth
GMP
HSC
LMPP
Late Eryth
EMP
pre–mDC
pre B
CD16 Mono
pre–pDC
cDC
Prog Mk
Plasma
CLP
pDC
pro B
ASDC
BaEoMa

Figure 3

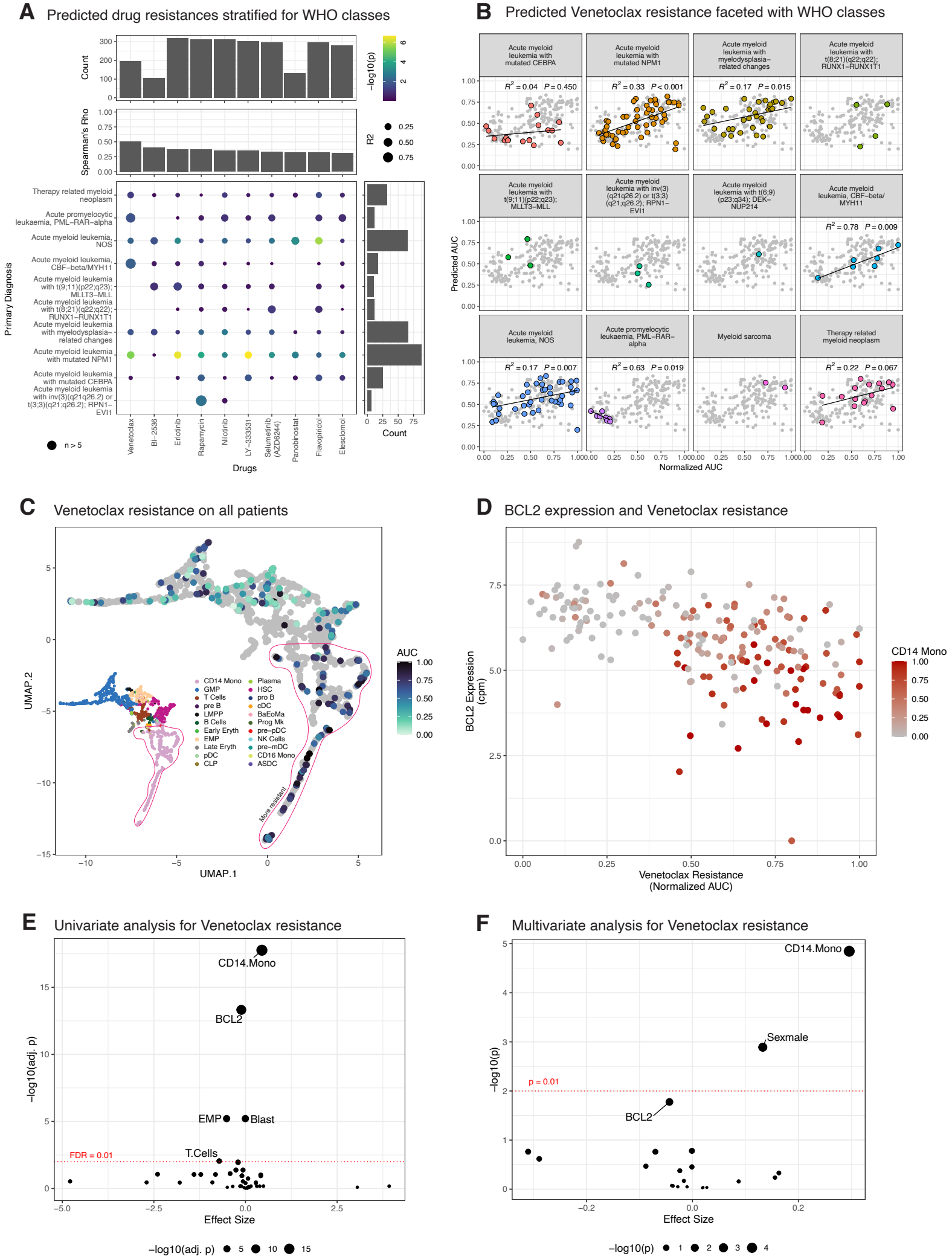

A
Predicted drug resistances stratified for WHO classes
Count
Spearman's Rho
−log10(p)
R2
Primary Diagnosis
Therapy related myeloid neoplasm
Acute promyelocytic leukaemia, PML−RAR−alpha
Acute myeloid leukemia, NOS
Acute myeloid leukemia, CBF−beta/MYH11
Acute myeloid leukemia with t(9;11)(p22;q23); MLLT3−MLL
Acute myeloid leukemia with t(8;21)(q22;q22); RUNX1−RUNX1T1
Acute myeloid leukemia with myelodysplasia−related changes
Acute myeloid leukemia with mutated NPM1
Acute myeloid leukemia with mutated CEBPA
Acute myeloid leukemia with inv(3)(q21q26.2) or t(3;3)(q21;q26.2); RPN1−EVI1
n > 5
Venetoclax
BI−2536
Erlotinib
Rapamycin
Nilotinib
LY−333531
Selumetinib (AZD6244)
Panobinostat
Flavopiridol
Elesclomol
Drugs
B
Predicted Venetoclax resistance faceted with WHO classes
Acute myeloid leukemia with mutated CEBPA
Acute myeloid leukemia with mutated NPM1
Acute myeloid leukemia with myelodysplasia−related changes
Acute myeloid leukemia with t(8;21)(q22;q22); RUNX1−RUNX1T1
Acute myeloid leukemia with t(9;11)(p22;q23); MLLT3−MLL
Acute myeloid leukemia with inv(3)(q21q26.2) or t(3;3)(q21;q26.2); RPN1−EVI1
Acute myeloid leukemia with t(6;9)(p23;q34); DEK−NUP214
Acute myeloid leukemia, CBF−beta/MYH11
Acute myeloid leukemia, NOS
Acute promyelocytic leukaemia, PML−RAR−alpha
Myeloid sarcoma
Therapy related myeloid neoplasm
$R^2 = 0.04$ $P = 0.450$
$R^2 = 0.33$ $P < 0.001$
$R^2 = 0.17$ $P = 0.015$
$R^2 = 0.78$ $P = 0.009$
$R^2 = 0.17$ $P = 0.007$
$R^2 = 0.63$ $P = 0.019$
$R^2 = 0.22$ $P = 0.067$
Predicted AUC
Normalized AUC
C
Venetoclax resistance on all patients
UMAP.1
UMAP.2
AUC
CD14 Mono
GMP
T Cells
pre B
LMPP
B Cells
Early Eryth
EMP
Late Eryth
pDC
CLP
Plasma
HSC
pro B
cDC
BaEoMa
Prog Mk
pre−pDC
NK Cells
pre−mDC
CD16 Mono
ASDC
More resistant
D
BCL2 expression and Venetoclax resistance
BCL2 Expression (cpm)
Venetoclax Resistance (Normalized AUC)
CD14 Mono
E
Univariate analysis for Venetoclax resistance
CD14.Mono
BCL2
EMP
Blast
T.Cells
FDR = 0.01
−log10(adj. p)
Effect Size
F
Multivariate analysis for Venetoclax resistance
CD14.Mono
Sexmale
BCL2
p = 0.01
−log10(p)
Effect Size

Figure 4

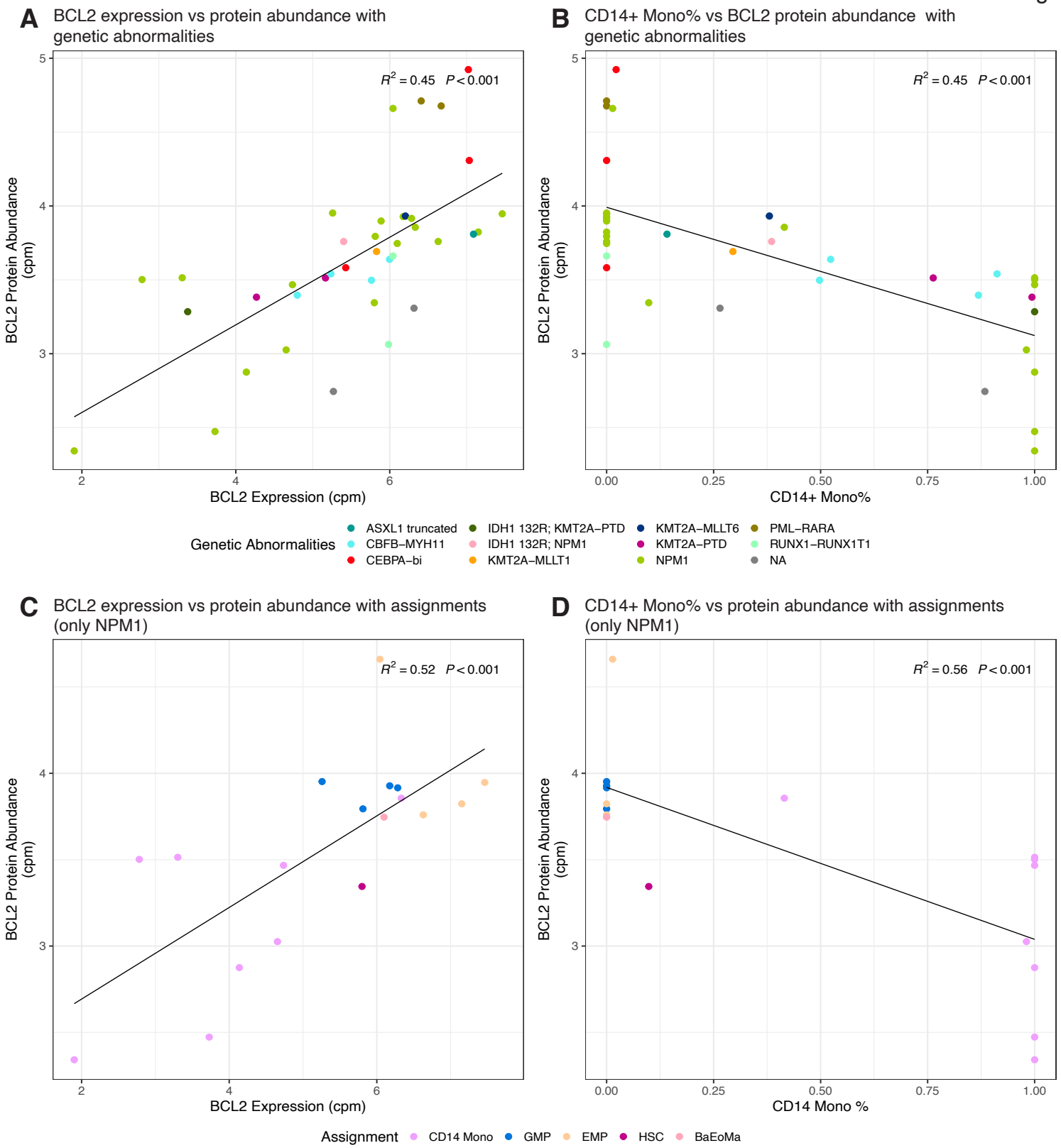

A
BCL2 expression vs protein abundance with genetic abnormalities
$R^2 = 0.45$ $P < 0.001$
BCL2 Protein Abundance (cpm)
BCL2 Expression (cpm)
B
CD14+ Mono% vs BCL2 protein abundance with genetic abnormalities
$R^2 = 0.45$ $P < 0.001$
BCL2 Protein Abundance (cpm)
CD14+ Mono%
Genetic Abnormalities
ASXL1 truncated
CBFB–MYH11
CEBPA–bi
IDH1 132R; KMT2A–PTD
IDH1 132R; NPM1
KMT2A–MLLT1
KMT2A–MLLT6
KMT2A–PTD
NPM1
PML–RARA
RUNX1–RUNX1T1
NA
C
BCL2 expression vs protein abundance with assignments (only NPM1)
$R^2 = 0.52$ $P < 0.001$
BCL2 Protein Abundance (cpm)
BCL2 Expression (cpm)
D
CD14+ Mono% vs protein abundance with assignments (only NPM1)
$R^2 = 0.56$ $P < 0.001$
BCL2 Protein Abundance (cpm)
CD14 Mono %
Assignment
CD14 Mono
GMP
EMP
HSC
BaEoMa

Figure S1

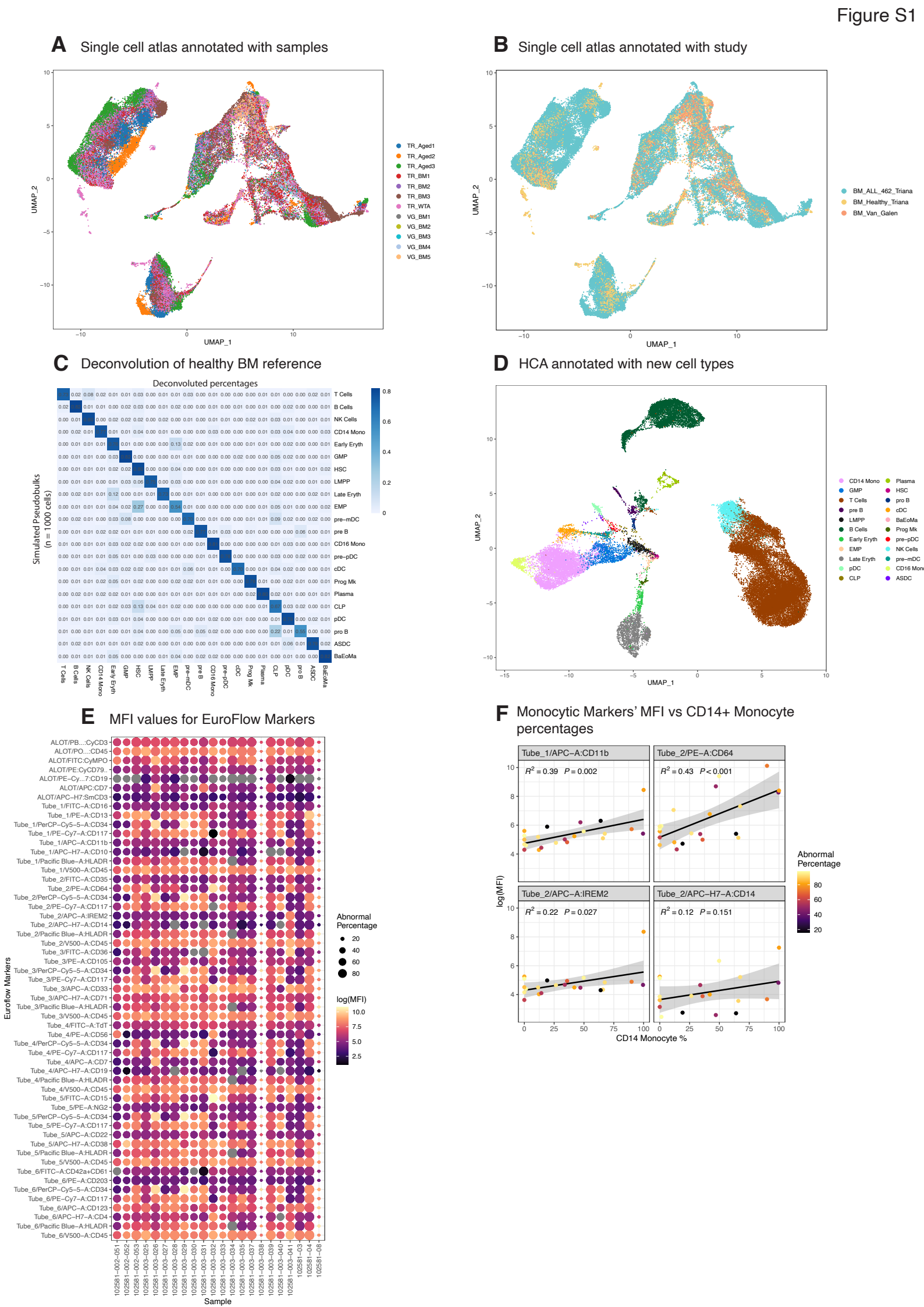

A
Single cell atlas annotated with samples
UMAP_1
UMAP_2
TR_Aged1
TR_Aged2
TR_Aged3
TR_BM1
TR_BM2
TR_BM3
TR_WTA
VG_BM1
VG_BM2
VG_BM3
VG_BM4
VG_BM5
B
Single cell atlas annotated with study
BM_ALL_462_Triana
BM_Healthy_Triana
BM_Van_Galen
C
Deconvolution of healthy BM reference
Deconvoluted percentages
Simulated Pseudobulks (n = 1000 cells)
T Cells
B Cells
NK Cells
CD14 Mono
Early Eryth
GMP
HSC
LMPP
Late Eryth
EMP
pre−mDC
pre B
CD16 Mono
pre−pDC
cDC
Prog Mk
Plasma
CLP
pDC
pro B
ASDC
BaEoMa
D
HCA annotated with new cell types
CD14 Mono
GMP
T Cells
pre B
LMPP
B Cells
Early Eryth
EMP
Late Eryth
pDC
CLP
Plasma
HSC
pro B
cDC
BaEoMa
Prog Mk
pre−pDC
NK Cells
pre−mDC
CD16 Mono
ASDC
E
MFI values for EuroFlow Markers
Euroflow Markers
Sample
Abnormal Percentage
log(MFI)
F
Monocytic Markers' MFI vs CD14+ Monocyte percentages
Tube_1/APC−A:CD11b
$R^2 = 0.39$ $P = 0.002$
Tube_2/PE−A:CD64
$R^2 = 0.43$ $P < 0.001$
Tube_2/APC−A:IREM2
$R^2 = 0.22$ $P = 0.027$
Tube_2/APC−H7−A:CD14
$R^2 = 0.12$ $P = 0.151$
log(MFI)
CD14 Monocyte %
Abnormal Percentage

Figure S2

A Patient meta summary from all studies

Characteristic	Overall, N = 1,267[1]	BEAT, N = 412[1]	LEUCE, N = 452[1]	LUMC, N = 96[1]	TARGET, N = 156[1]	TCGA, N = 151[1]
Age	55 (0, 89)	62 (2, 87)	57 (17, 87)	NA (Inf, -Inf)	9 (0, 23)	56 (22, 89)
Unknown	97	1	0	96	0	0
Sex						
female	567/1,267 (45%)	183/412 (44%)	195/452 (43%)	44/96 (46%)	77/156 (49%)	68/151 (45%)
male	700/1,267 (55%)	229/412 (56%)	257/452 (57%)	52/96 (54%)	79/156 (51%)	83/151 (55%)
FAB						
M0	51/761 (6.7%)	5/90 (5.6%)	27/381 (7.1%)	0/0 (NA%)	4/140 (2.9%)	15/150 (10%)
M1	189/761 (25%)	7/90 (7.8%)	129/381 (34%)	0/0 (NA%)	17/140 (12%)	36/150 (24%)
M2	138/761 (18%)	10/90 (11%)	54/381 (14%)	0/0 (NA%)	37/140 (26%)	37/150 (25%)
M3	56/761 (7.4%)	11/90 (12%)	30/381 (7.9%)	0/0 (NA%)	0/140 (0%)	15/150 (10%)
M4	155/761 (20%)	24/90 (27%)	59/381 (15%)	0/0 (NA%)	43/140 (31%)	29/150 (19%)
M5	145/761 (19%)	31/90 (34%)	69/381 (18%)	0/0 (NA%)	30/140 (21%)	15/150 (10%)
M6	14/761 (1.8%)	0/90 (0%)	10/381 (2.6%)	0/0 (NA%)	2/140 (1.4%)	2/150 (1.3%)
M7	13/761 (1.7%)	2/90 (2.2%)	3/381 (0.8%)	0/0 (NA%)	7/140 (5.0%)	1/150 (0.7%)
Unknown	506	322	71	96	16	1
ELN						
Adverse	197/686 (29%)	148/390 (38%)	0/0 (NA%)	0/0 (NA%)	9/148 (6.1%)	40/148 (27%)
Favorable	203/686 (30%)	107/390 (27%)	0/0 (NA%)	0/0 (NA%)	64/148 (43%)	32/148 (22%)
Intermediate	286/686 (42%)	135/390 (35%)	0/0 (NA%)	0/0 (NA%)	75/148 (51%)	76/148 (51%)
Unknown	581	22	452	96	8	3
Blast	72 (0, 100)	60 (0, 98)	77 (1, 99)	75 (15, 99)	74 (14, 100)	71 (0, 100)
Unknown	144	109	31	0	4	0
Primary Diagnosis						
Acute myeloid leukemia with inv(3)(q21q26.2) or t(3;3)(q21;q26.2); RPN1-EVI1	17/1,267 (1.3%)	8/412 (1.9%)	7/452 (1.5%)	2/96 (2.1%)	0/156 (0%)	0/151 (0%)
Acute myeloid leukemia with mutated CEBPA	37/1,267 (2.9%)	28/412 (6.8%)	0/452 (0%)	0/96 (0%)	9/156 (5.8%)	0/151 (0%)
Acute myeloid leukemia with mutated NPM1	163/1,267 (13%)	96/412 (23%)	0/452 (0%)	27/96 (28%)	7/156 (4.5%)	33/151 (22%)
Acute myeloid leukemia with myelodysplasia-related changes	213/1,267 (17%)	88/412 (21%)	114/452 (25%)	11/96 (11%)	0/156 (0%)	0/151 (0%)
Acute myeloid leukemia with t(10;11)(p12;q23.3); MLLT10-KMT2A	9/1,267 (0.7%)	0/412 (0%)	0/452 (0%)	0/96 (0%)	9/156 (5.8%)	0/151 (0%)
Acute myeloid leukemia with t(6;11)(q27;q23.3); MLLT4-KMT2A	5/1,267 (0.4%)	0/412 (0%)	0/452 (0%)	0/96 (0%)	5/156 (3.2%)	0/151 (0%)
Acute myeloid leukemia with t(6;9)(p23;q34); DEK-NUP214	8/1,267 (0.6%)	3/412 (0.7%)	3/452 (0.7%)	1/96 (1.0%)	1/156 (0.6%)	0/151 (0%)
Acute myeloid leukemia with t(8;21)(q22;q22); RUNX1-RUNX1T1	60/1,267 (4.7%)	9/412 (2.2%)	20/452 (4.4%)	3/96 (3.1%)	21/156 (13%)	7/151 (4.6%)
Acute myeloid leukemia with t(9;11)(p22;q23); MLLT3-MLL	35/1,267 (2.8%)	9/412 (2.2%)	8/452 (1.8%)	4/96 (4.2%)	12/156 (7.7%)	2/151 (1.3%)
Acute myeloid leukemia, CBF-beta/MYH11	101/1,267 (8.0%)	26/412 (6.3%)	29/452 (6.4%)	10/96 (10%)	28/156 (18%)	8/151 (5.3%)
Acute myeloid leukemia, NOS	492/1,267 (39%)	91/412 (22%)	215/452 (48%)	35/96 (36%)	64/156 (41%)	87/151 (58%)
Acute promyelocytic leukaemia, PML-RAR-alpha	62/1,267 (4.9%)	17/412 (4.1%)	29/452 (6.4%)	2/96 (2.1%)	0/156 (0%)	14/151 (9.3%)
Myeloid sarcoma	3/1,267 (0.2%)	3/412 (0.7%)	0/452 (0%)	0/96 (0%)	0/156 (0%)	0/151 (0%)
Therapy related myeloid neoplasm	62/1,267 (4.9%)	34/412 (8.3%)	27/452 (6.0%)	1/96 (1.0%)	0/156 (0%)	0/151 (0%)

[1] Median (Range) or Frequency (%)

B AML studies on ternary plot

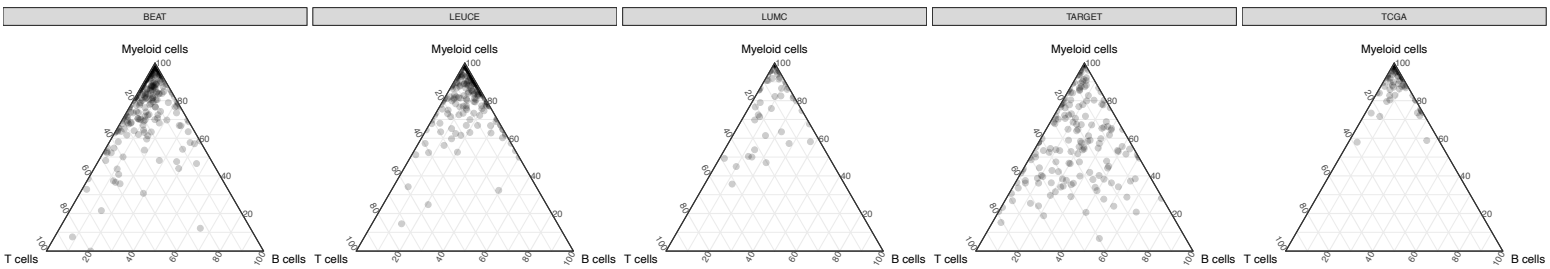
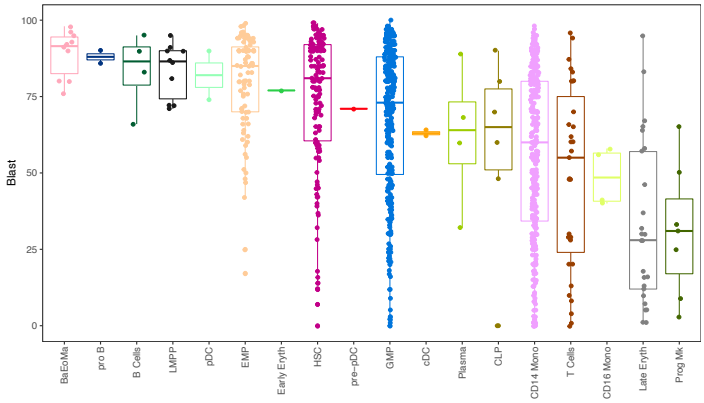
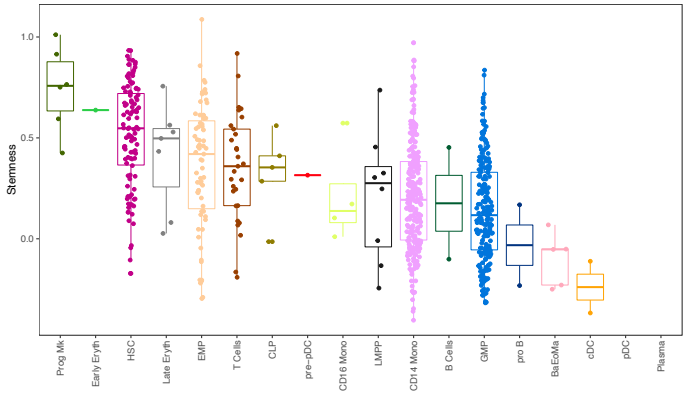


Figure S3

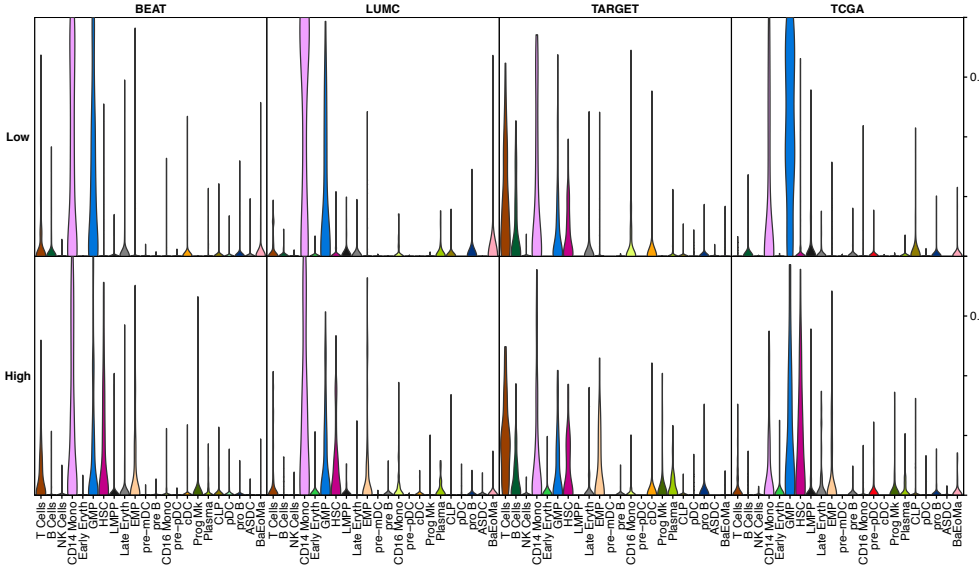
A Blast percentage vs AML phenotypes

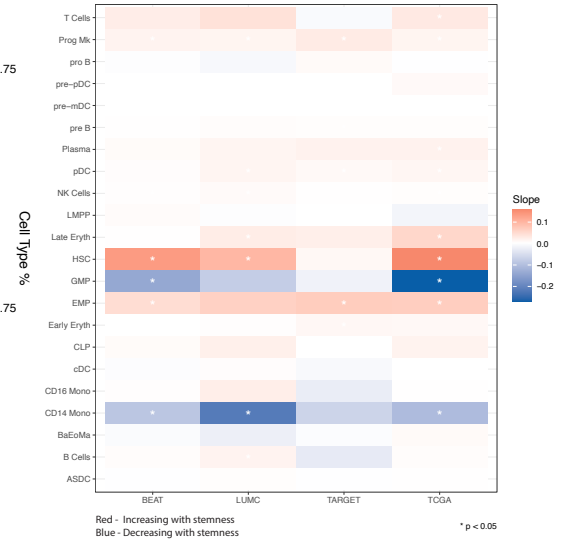


B Stemness score vs AML phenotypes



C Compositional changes for stemness score





D UMAP plot of compositions with different annotations

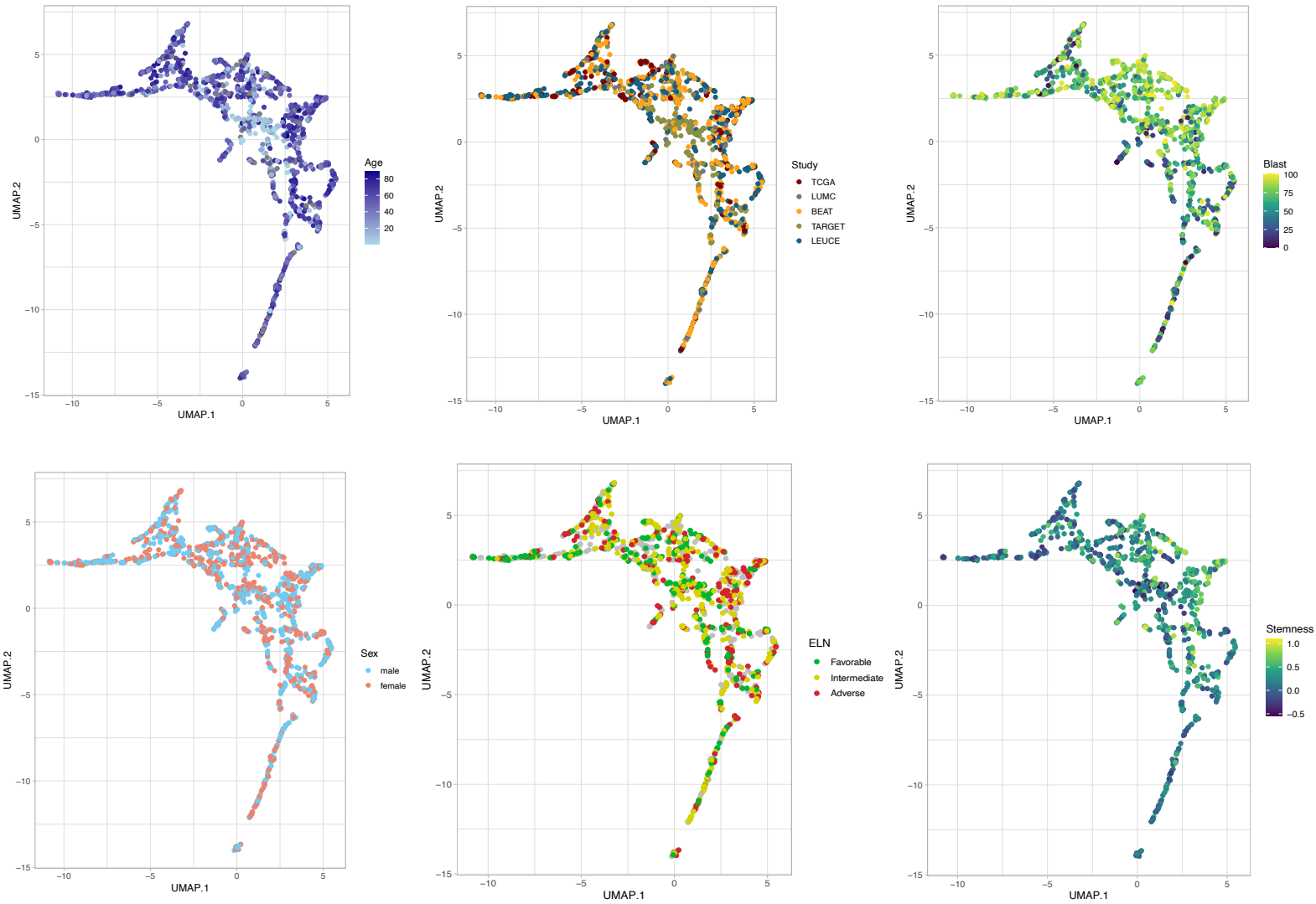
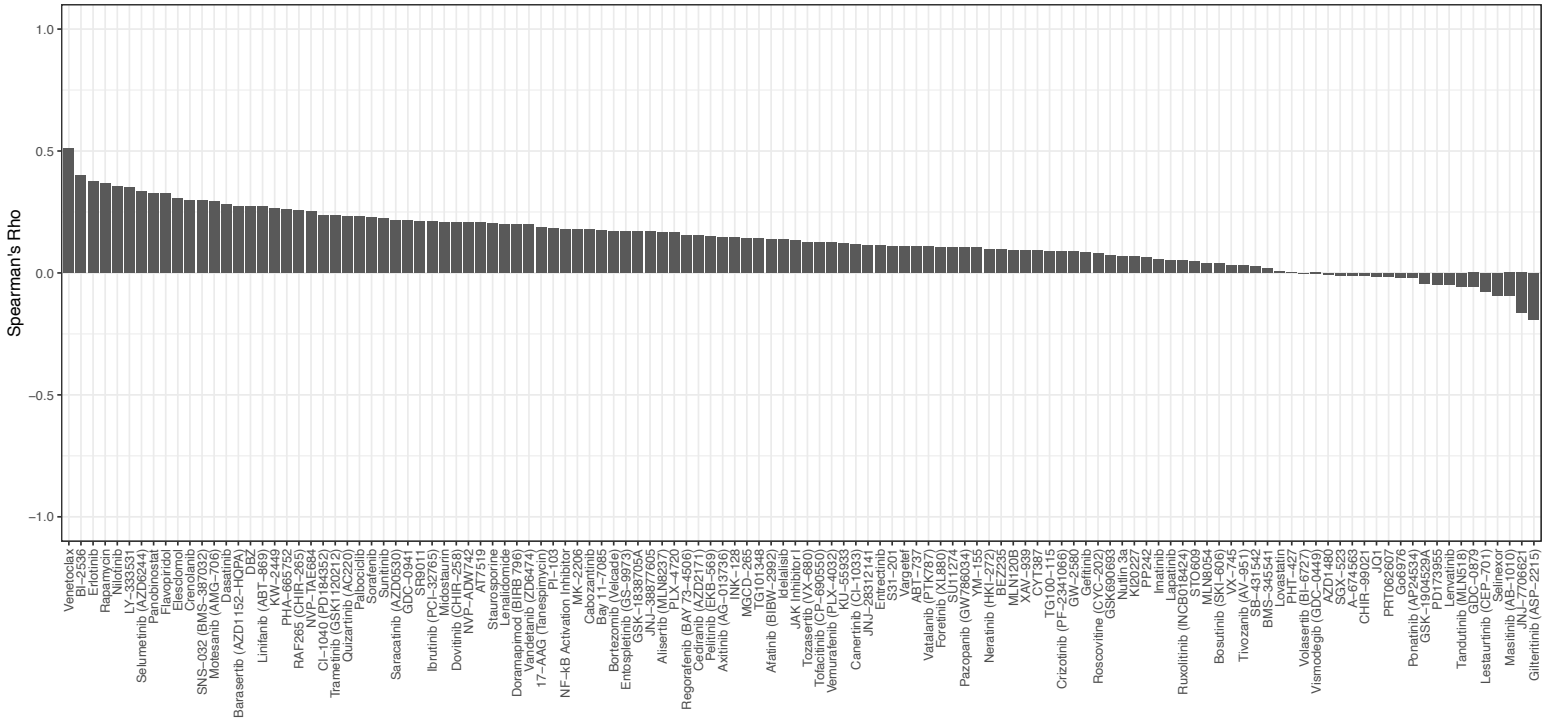


Figure S4

A Drug resistance predictions for all drugs

Leave one out cross validation within each drug



B Drug resistance predictions stratified for primary diagnosis

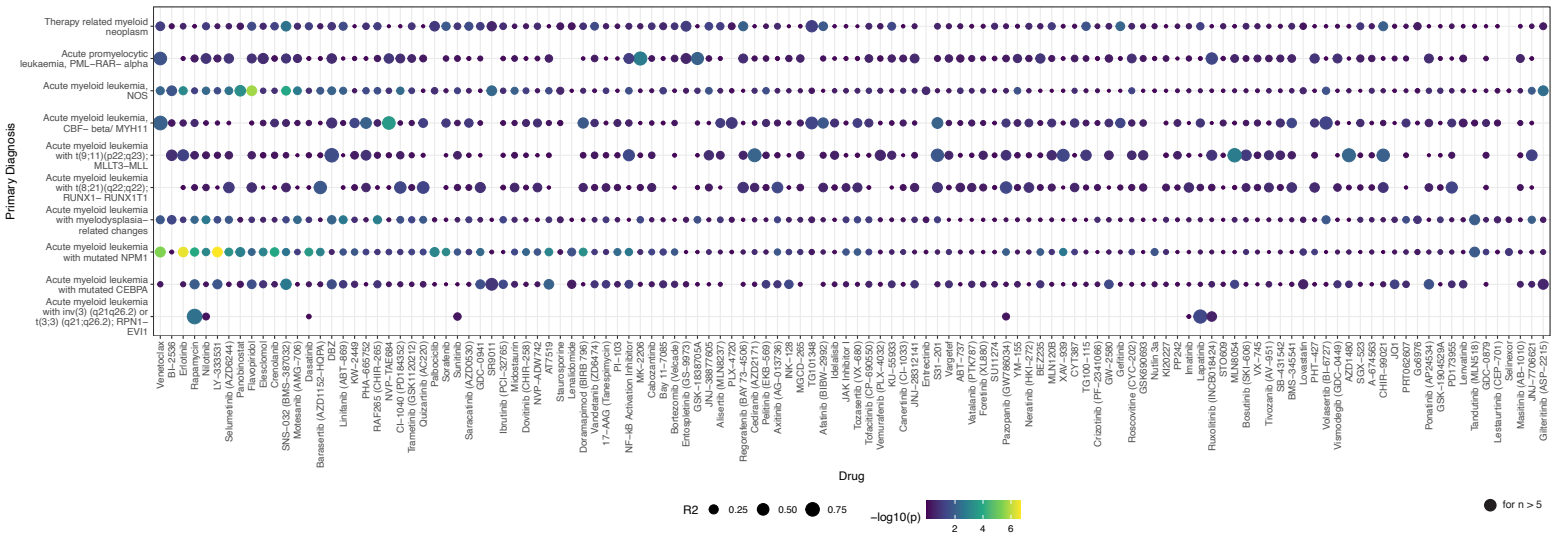


Figure S5

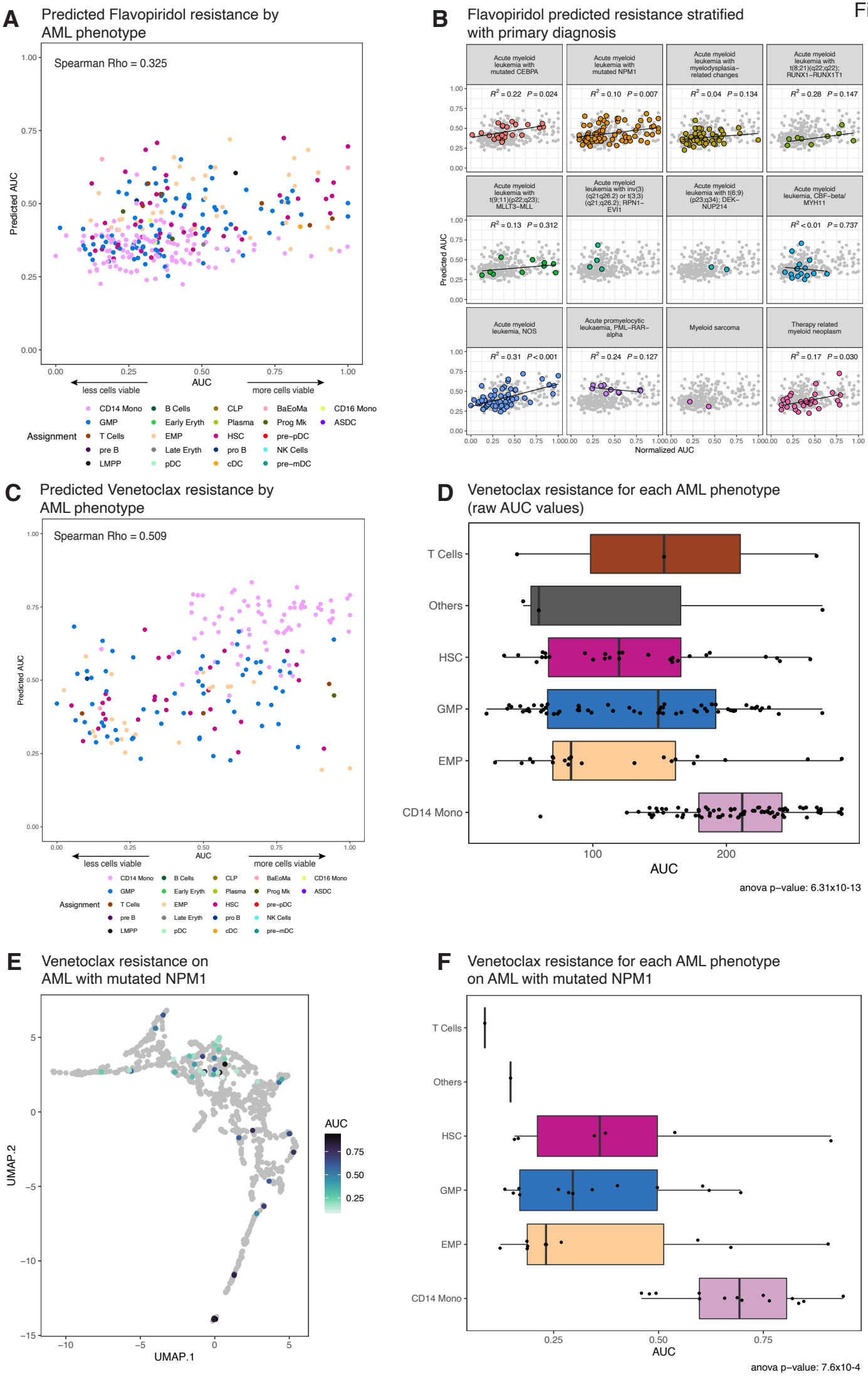
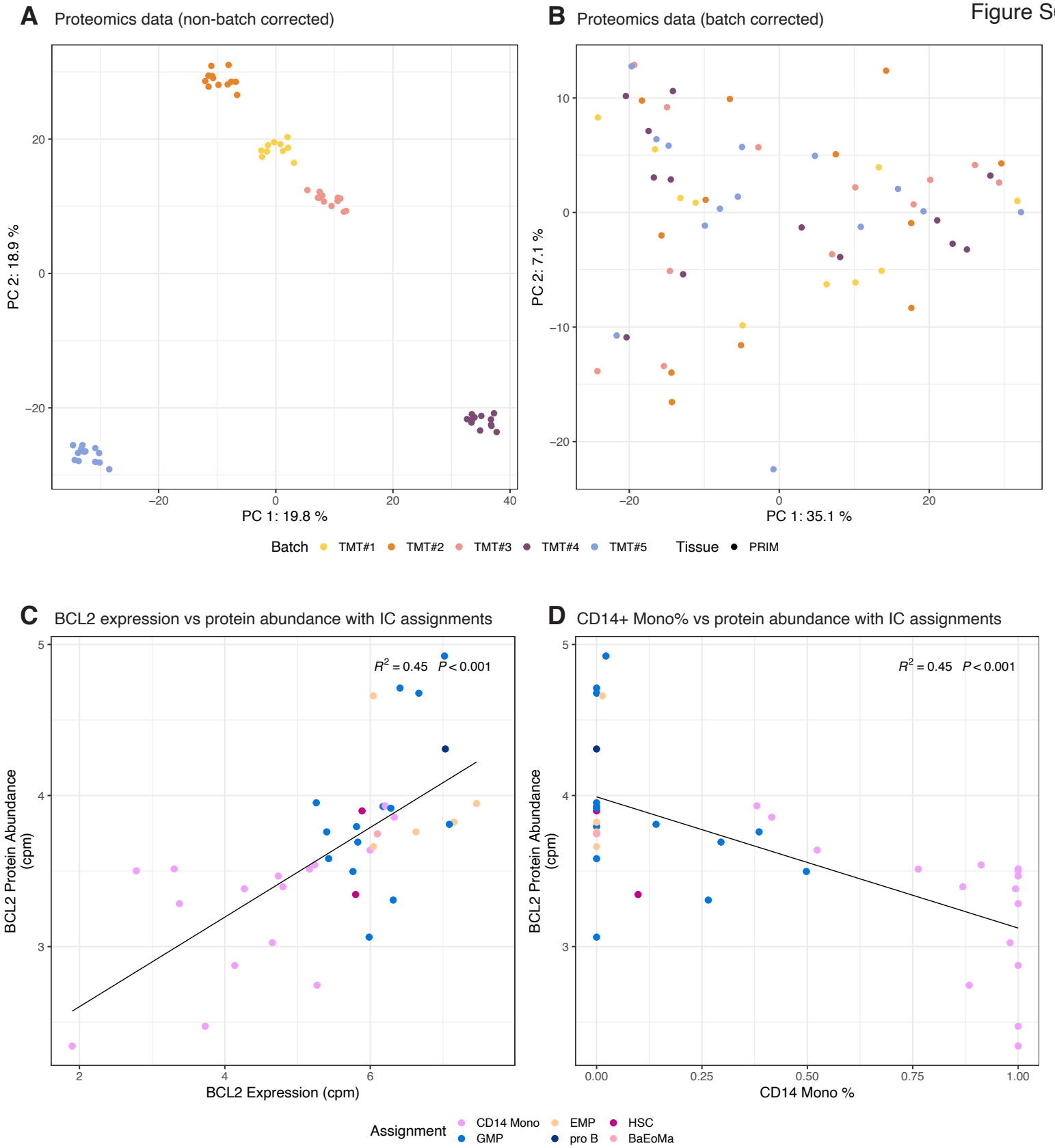

A
Predicted Flavopiridol resistance by AML phenotype
Spearman Rho = 0.325
Predicted AUC
AUC
less cells viable
more cells viable
Assignment
CD14 Mono
GMP
T Cells
pre B
LMPP
B Cells
Early Eryth
EMP
Late Eryth
pDC
CLP
Plasma
HSC
pro B
cDC
BaEoMa
Prog Mk
pre−pDC
NK Cells
pre−mDC
CD16 Mono
ASDC
B
Flavopiridol predicted resistance stratified with primary diagnosis
Acute myeloid leukemia with mutated CEBPA
Acute myeloid leukemia with mutated NPM1
Acute myeloid leukemia with myelodysplasia−related changes
Acute myeloid leukemia with t(8;21)(q22;q22); RUNX1−RUNX1T1
Acute myeloid leukemia with t(9;11)(p22;q23); MLLT3−MLL
Acute myeloid leukemia with inv(3)(q21q26.2) or t(3;3)(q21;q26.2); RPN1−EVI1
Acute myeloid leukemia with t(6;9)(p23;q34); DEK−NUP214
Acute myeloid leukemia, CBF−beta/MYH11
Acute myeloid leukemia, NOS
Acute promyelocytic leukaemia, PML−RAR−alpha
Myeloid sarcoma
Therapy related myeloid neoplasm
$R^2 = 0.22$ $P = 0.024$
$R^2 = 0.10$ $P = 0.007$
$R^2 = 0.04$ $P = 0.134$
$R^2 = 0.28$ $P = 0.147$
$R^2 = 0.13$ $P = 0.312$
$R^2 < 0.01$ $P = 0.737$
$R^2 = 0.31$ $P < 0.001$
$R^2 = 0.24$ $P = 0.127$
$R^2 = 0.17$ $P = 0.030$
Predicted AUC
Normalized AUC
C
Predicted Venetoclax resistance by AML phenotype
Spearman Rho = 0.509
Predicted AUC
AUC
less cells viable
more cells viable
Assignment
CD14 Mono
GMP
T Cells
pre B
LMPP
B Cells
Early Eryth
EMP
Late Eryth
pDC
CLP
Plasma
HSC
pro B
cDC
BaEoMa
Prog Mk
pre−pDC
NK Cells
pre−mDC
CD16 Mono
ASDC
D
Venetoclax resistance for each AML phenotype (raw AUC values)
T Cells
Others
HSC
GMP
EMP
CD14 Mono
AUC
anova p−value: 6.31x10-13
E
Venetoclax resistance on AML with mutated NPM1
UMAP.2
UMAP.1
AUC
F
Venetoclax resistance for each AML phenotype on AML with mutated NPM1
T Cells
Others
HSC
GMP
EMP
CD14 Mono
AUC
anova p−value: 7.6x10-4

Figure S6


A
Proteomics data (non-batch corrected)
PC 2: 18.9 %
PC 1: 19.8 %
B
Proteomics data (batch corrected)
PC 2: 7.1 %
PC 1: 35.1 %
Batch
TMT#1
TMT#2
TMT#3
TMT#4
TMT#5
Tissue
PRIM
C
BCL2 expression vs protein abundance with IC assignments
$R^2 = 0.45$ $P < 0.001$
BCL2 Protein Abundance (cpm)
BCL2 Expression (cpm)
D
CD14+ Mono% vs protein abundance with IC assignments
$R^2 = 0.45$ $P < 0.001$
BCL2 Protein Abundance (cpm)
CD14 Mono %
Assignment
CD14 Mono
GMP
EMP
pro B
HSC
BaEoMa