ARTICLES
https://doi.org/10.1038/s41591-022-01819-x

nature medicine

Check for updates

A cellular hierarchy framework for understanding heterogeneity and predicting drug response in acute myeloid leukemia

Andy G. X. Zeng[1,2], Suraj Bansal[1,18], Liqing Jin[1,18], Amanda Mitchell[1], Weihsu Claire Chen[1,17], Hussein A. Abbas[3], Michelle Chan-Seng-Yue[1], Veronique Voisin[4], Peter van Galen[5,6,7,8], Anne Tierens[9], Meyling Cheok[10], Claude Preudhomme[10], Hervé Dombret[11], Naval Daver[3], P. Andrew Futreal[12], Mark D. Minden[1,13,14,15], James A. Kennedy[1,16], Jean C. Y. Wang[1,14,15] and John E. Dick[1,2] ✉

The treatment landscape of acute myeloid leukemia (AML) is evolving, with promising therapies entering clinical translation, yet patient responses remain heterogeneous, and biomarkers for tailoring treatment are lacking. To understand how disease heterogeneity links with therapy response, we determined the leukemia cell hierarchy makeup from bulk transcriptomes of more than 1,000 patients through deconvolution using single-cell reference profiles of leukemia stem, progenitor and mature cell types. Leukemia hierarchy composition was associated with functional, genomic and clinical properties and converged into four overall classes, spanning Primitive, Mature, GMP and Intermediate. Critically, variation in hierarchy composition along the Primitive versus GMP or Primitive versus Mature axes were associated with response to chemotherapy or drug sensitivity profiles of targeted therapies, respectively. A seven-gene biomarker derived from the Primitive versus Mature axis was associated with response to 105 investigational drugs. Cellular hierarchy composition constitutes a novel framework for understanding disease biology and advancing precision medicine in AML.

AML is a devastating disease characterized by extensive inter-patient and intra-patient heterogeneity. Poor outcomes are attributed to primary chemotherapy resistance and high relapse rates among patients achieving remission, highlighting the inadequacy of standard chemotherapy for curing most patients with AML. Recently, many promising new therapies targeting diverse cellular mechanisms have either been approved or are progressing through clinical trials, offering alternatives to chemotherapy. However, patient responses to these new therapies are also heterogeneous, and reliable ways to select the best therapy for each patient are lacking.

Historically, two distinct approaches have evolved for understanding AML heterogeneity and informing therapy selection: a genomic model and a stem cell model. The discovery of the Philadelphia chromosome in 1960[1] sparked a series of cytogenetic studies that identified distinct cytogenetic drivers of AML. More recently, advances in genome sequencing uncovered mutational drivers of AML and culminated in a prognostically informative genomic classification[2]. Although this genomic model accounts for a major source of inter-patient heterogeneity, cells sharing the same driver mutation can exhibit functional differences[3]. Moreover, although a subset of inhibitors target driver mutations, genomic profiling has limited predictive value for therapies targeted to specific biological processes or signaling pathways.

The discovery of hematopoietic stem cells (HSCs) in 1961 and the development of quantitative assays to interrogate stem cell function[4] enabled pioneering experiments demonstrating that blasts within individual patients exhibited functional differences in their cycling kinetics[5,6], differentiation state[7,8] and self-renewal capacity[9,10]. Collectively, these studies culminated in formal evidence for AML being sustained by rare leukemia stem cells (LSCs)[11–13]. LSCs have since been shown to mediate relapse[14], and LSC-based gene expression stemness scores have emerged as predictors of outcomes after chemotherapy[15–18]. Although LSCs are an important therapeutic target, this model offers limited guidance around therapy selection. Overall, these genomic and stem cell models provide complementary insight into AML heterogeneity, yet neither model alone is sufficient to guide therapy selection, particularly for newer agents. A new approach for personalized therapy selection that integrates the genomic and stem cell models is needed.

[1]Princess Margaret Cancer Centre, University Health Network, Toronto, ON, Canada. [2]Department of Molecular Genetics, University of Toronto, Toronto, ON, Canada. [3]Department of Leukemia, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [4]The Donnelly Centre, University of Toronto, Toronto, ON, Canada. [5]Division of Hematology, Brigham and Women's Hospital, Boston, MA, USA. [6]Department of Medicine, Harvard Medical School, Boston, MA, USA. [7]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [8]Ludwig Center at Harvard, Harvard Medical School, Boston, MA, USA. [9]Laboratory Medicine Program, Hematopathology, University Health Network, Toronto, ON, Canada. [10]University of Lille, CNRS, Inserm, CHU Lille, UMR9020-U1277 - CANTHER – Cancer Heterogeneity Plasticity and Resistance to Therapies, Lille, France. [11]Department of Hematology, Hôpital Saint-Louis, Assistance Publique-Hôpitaux de Paris, Université Paris Cité, Paris, France. [12]Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [13]Department of Medical Biophysics, University of Toronto, Toronto, ON, Canada. [14]Department of Medicine, University of Toronto, Toronto, ON, Canada. [15]Division of Medical Oncology and Hematology, University Health Network, Toronto, ON, Canada. [16]Division of Medical Oncology and Hematology, Sunnybrook Health Sciences Centre, Toronto, ON, Canada. [17]Present address: Biologics Discovery, Amgen British Columbia, Burnaby, BC, Canada. [18]These authors contributed equally: Suraj Bansal, Liqing Jin. ✉e-mail: john.dick@uhnresearch.ca

Cancer has long been recognized as a caricature of normal tissue development[19], and AML is the best-studied cancer system wherein leukemic cells are organized into a hierarchy resembling normal blood development. Cellular hierarchies in AML can be distorted in different ways, depending on their genetic alterations and cell of origin. For example, a strong differentiation block arising in a stem cell may result in a shallow, stem-cell-dominant hierarchy. In other cases, considerable, albeit aberrant, differentiation may occur, resulting in a steep hierarchy wherein rare LSCs generate a bulk blast population with mature myeloid features. In this way, the cellular composition of each patient's leukemic hierarchy likely reflects the functional consequences of specific mutations on disease-sustaining LSCs. Thus, interrogation of leukemic hierarchies may provide an opportunity to potentially integrate features of the genetic and stem cell models of AML[20]. Single-cell RNA sequencing (scRNA-seq) has emerged as a powerful tool for dissecting cellular hierarchies[21,22]; however, prohibitive costs restrict these studies to a limited number of patients. Without measuring these cellular hierarchies at scale in large clinical datasets, the relationship of hierarchy composition to therapy response remains unknown.

In this study, the cellular hierarchies of more than 1,000 patients with AML were characterized through gene expression deconvolution on bulk AML transcriptomes using single-cell reference profiles of distinct AML stem, progenitor and mature cell types. This approach to characterizing AML heterogeneity enabled integration of both the genomic and stem cell models of AML, resulting in a novel framework for understanding disease biology and predicting drug response.

Results

Heterogeneity among leukemia stem and progenitor cells. As a first step to uncover the organization of cellular hierarchies in AML, we re-analyzed the scRNA-seq profiles of 13,653 cells from 12 patients with AML at diagnosis[21] with a focus on primitive stem and progenitor blast populations (henceforth, leukemia stem and progenitor cells (LSPCs)). Using self-assembling manifolds (SAM), an unsupervised approach to prioritize biologically relevant features among comparatively homogenous cells[23], we previously identified two transcriptomic populations of normal human HSCs: a deeply quiescent population with low transcriptome diversity (Non-Primed) and another residing in a shallower state of quiescence with higher *CDK6* expression (Cycle-Primed)[24]. We applied SAM to analyze LSPCs and identified three distinct populations shared across the 12 patients (Fig. 1a and Extended Data Fig. 1a–c). One population had low transcriptome diversity and was enriched for core LSC programs but appeared otherwise inactive (Extended Data Fig. 1d). We named this population Quiescent LSPC. The second population was enriched for *CDK6* expression and E2F3 targets suggestive of cell cycle priming (Extended Data Fig. 1e) as well as inflammatory signatures suggestive of priming for myeloid differentiation[25] (Extended Data Fig. 1h). We named this population Primed LSPC. The third population exhibited enrichment for CTCF targets suggestive of stem cell activation[26] and broad enrichment of E2F targets indicating cell cycle progression, with 40% of cells classified as cycling (Fig. 1b and Extended Data Fig. 1f,g,i). We named this third population Cycling LSPC. The existence of distinct cellular states provides a molecular basis for the known functional heterogeneity that is found within the LSC compartment[27]. These new classes of Quiescent, Primed and Cycling LSPC led to higher classification performance compared to the prior 'HSC-like' and 'Progenitor-like' classification from van Galen et al.[21] (weighted accuracy: 0.93 versus 0.73; Extended Data Fig. 1j). We combined these new LSPC classes with the existing classification of more committed blasts by van Galen et al.[21] (that is, GMP-like blasts, ProMono-like blasts, Mono-like blasts and cDC-like blasts) to constitute a map of common leukemic blast states shared across these 12 patients with AML, with each leukemic state having distinct molecular properties.

Deconvolution of constituent cell populations in AML. We next sought to understand how these defined AML cell populations and the hierarchies into which they are organized relate to functional, biological and clinical properties of AML. To study this at scale, we employed gene expression deconvolution to infer the leukemic hierarchy composition from bulk AML transcriptomes (Fig. 1c). We performed benchmarking analysis of multiple scRNA-seq-based deconvolution methods and identified CIBERSORTx[28] as the highest-performing approach in the context of AML (Supplementary Note 1 and Extended Data Fig. 2). Additionally, we confirmed that deconvolution with the new LSPC classification of primitive AML cells improves discrimination of clinical and biological phenotypes compared to the prior HSC-like and Prog-like classification (Supplementary Note 2 and Extended Data Fig. 3). Thus, subsequent deconvolution was performed through CIBERSORTx using single-cell transcriptomes from seven leukemic cell types (Quiescent LSPC, Primed LSPC, Cycling LSPC, GMP-like, ProMono-like, Mono-like and cDC-like) and seven non-leukemic immune cell types (Natural Killer, Naive T, $CD8^+$ T, B, Plasma, Monocytes and cDCs) as a reference.

Functional LSCs associate with Quiescent LSPC abundance. We first sought to determine whether any of our newly defined LSPC cellular states were associated with LSC activity. The LSC state is functionally defined by whether a leukemic cell can initiate leukemia in vivo[29]. We, thus, performed RNA-seq on 111 AML fractions previously evaluated by microarray and where LSC activity was determined through xenotransplantation[17] and applied deconvolution to determine the cell type composition of each fraction (Fig. 1d). LSC^+ fractions were highly enriched for Quiescent LSPC ($P=8\times10^{-6}$) and Primed LSPC ($P=6\times10^{-4}$) but not Cycling LSPC ($P=0.74$) (Fig. 1e). Conversely, LSC^- fractions were highly enriched for Mono-like blasts ($P=2\times10^{-3}$) (Fig. 1e). Given that immunophenotype does not consistently predict LSC activity[16,17,30], we compared deconvolution against immunophenotype by training classifiers to predict LSC activity in AML fractions based on cell type composition versus CD34/CD38 status. Classifiers trained on immunophenotype were consistently outperformed by those trained on leukemia cell composition (median areas under the curve (AUCs) = 0.71 versus 0.86, $P<2\times10^{-16}$) and were even outperformed by models trained from Quiescent LSPC abundance as a single variable (median AUCs = 0.71 versus 0.77, $P<2\times10^{-16}$) (Fig. 1f and Supplementary Table 3). Finally, we found Quiescent LSPC to be associated with high LSC frequency in an independent dataset of bulk AML samples assessed through limiting dilution analysis[31] (Fig. 1g). Collectively, these findings establish a new link between transcriptomic LSPC states and functionally defined LSCs at the apex of the leukemia cell hierarchy, suggesting that LSC activity can be inferred through deconvolution of patient hierarchies.

Hierarchy composition is associated with AML genomics. The differentiation properties of the LSCs sustaining each patient's AML are reflected in the cellular composition of the hierarchies that they generate. To examine how these hierarchies vary across patient samples and how they relate to molecular and clinical features of AML, we applied our deconvolution approach to infer the abundance of seven leukemic cell types as well as seven non-leukemic immune populations (described above) for 864 patient samples from The Cancer Genome Atlas (TCGA)[32], BEAT-AML[33] and Leucegene cohorts[34]. Clustering patients based on the composition of their leukemia cell hierarchies revealed four distinct subtypes: Primitive (shallow hierarchy, LSPC-enriched), Mature (steep hierarchy, enriched for mature Mono-like and cDC-like blasts), GMP

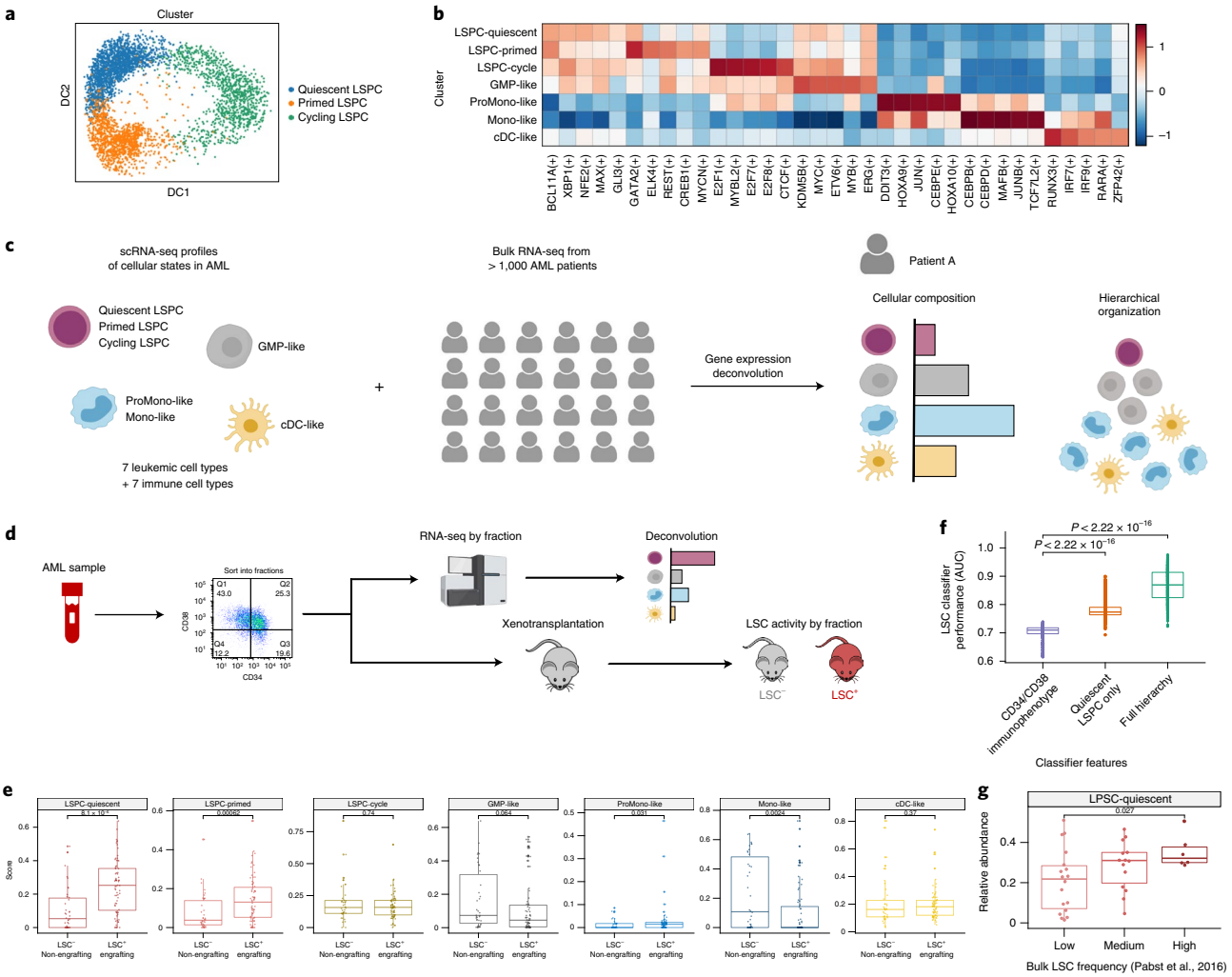


Fig. 1 | Functional significance of LSPC populations from scRNA-seq. a, Diffusion map of 4,163 AML LSPCs using SAM-derived feature weights. **b**, TF regulon activity in each leukemic cell type, inferred through PySCENIC. TF regulon enrichment scores were scaled, and the top five regulons for each cell type are depicted. **c**, Schematic depicting AML deconvolution approach using reference signatures from scRNA-seq populations. **d**, Experimental design for evaluating the relationship between AML cell states from scRNA-seq and functional LSC activity. In total, 111 sorted AML fractions previously evaluated for functional LSC activity through xenotransplantation in Ng et al.[17] were subject to RNA-seq and gene expression deconvolution. The relative abundance of each leukemic population was subsequently compared across LSC$^+$ (engrafting) and LSC$^-$ (non-engrafting) fractions. **e**, Enrichment of leukemic cell types across 72 LSC$^+$ (engrafting) and 38 LSC$^-$ (non-engrafting) AML fractions. Relative abundances of each cell type were compared through a two-sided Wilcoxon rank-sum test. **f**, Model performance (AUC) of RF classifiers predicting functional LSC activity in 110 sorted AML fractions. Classifiers were trained and evaluated through five-fold nested cross-validation with 1,000 repeats, as outlined in Extended Data Fig. 3a. Three types of classifiers were trained, each using different features to predict LSC activity: (1) using the CD34/CD38 immunophenotype of each fraction, (2) using the relative abundance of Quiescent LSPC alone and (3) using the relative abundance of all leukemic populations spanning the full AML hierarchy. AUC values were paired by repeat iteration and comparisons were performed using a two-sided Wilcoxon signed-rank test. Box plots indicate the range of the central 50% of the data, with the central line marking the median. Whiskers extend from each box to 1.5× the interquartile range. **g**, Relative abundance of Quiescent LSPC in patient samples with low ($n=18$), medium ($n=14$) and high ($n=6$) bulk LSC frequencies, as defined by Pabst et al.[31]. Comparisons were performed using a two-sided Wilcoxon rank-sum test. Box plots indicate the range of the central 50% of the data, with the central line marking the median. Whiskers extend from each box to 1.5× the interquartile range.

(dominated by GMP-like blasts) and Intermediate (balanced distribution) (Fig. 2a–c). Hierarchy composition was associated with multiple biological and clinical parameters, including age at diagnosis, white blood cell (WBC) differential counts and FAB class (Extended Data Fig. 4a,b). We focused on cytogenetic and mutational correlates to understand the cellular states and hierarchies generated by common genetic drivers of AML.

Patient hierarchies were separated along two principal components (PCs): PC1, spanning a continuum from Primitive to GMP (35% of variance), and PC2, spanning Primitive to Mature (28% of variance) (Fig. 2a). Hierarchies generated by cytogenetic alterations primarily separated along the Primitive versus GMP axis (PC1) (Fig. 2d and Extended Data Fig. 4e), wherein adverse cytogenetic alterations generated Primitive hierarchies and favorable cytogenetic alterations generated GMP-dominant hierarchies (Fig. 2d). Cellular hierarchies generated by genetic driver mutations and their combinations primarily separated along the Primitive versus Mature axis (PC2), reflecting their impact on the extent of AML differentiation (Fig. 2e and Extended Data Fig. 4c,d). Notably, different driver mutations in the same gene could also have different consequences on the resulting hierarchies. For example, *DNMT3A* R882 mutations were associated with more mature disease than

Fig. 2 | AML hierarchy composition correlates with genomics and survival. a, PCA of 864 patients with AML from the TCGA, BEAT-AML and Leucegene cohorts based on the composition of their cellular hierarchy. **b**, Relative abundance of each leukemic cell type in each patient. Each bar represents an individual patient, and the distribution of colors throughout each bar represents the distribution of leukemic cell populations within their leukemic hierarchy. **c**, Depictions of the cellular organization of Primitive, GMP and Mature hierarchies. **d**, Density plots depicting cytogenetic groups along the Primitive versus GMP axis (PC1). Cytogenetic alterations are colored by prognostic significance, wherein red indicates adverse prognosis and green indicates favorable prognosis. **e**, Density plots depicting common driver mutation combinations along the Primitive versus Mature axis (PC2). **f**, OS outcomes of AML hierarchy subtypes in the TCGA and BEAT-AML cohorts. Differences in survival across all subtypes were evaluated through a log-rank test. **g**, Univariate and pairwise HRs, for each AML hierarchy subtype across three patient cohorts: TCGA ($n=173$), BEAT-AML ($n=281$) and GSE6891 ($n=495$). Univariate HRs for each subtype and pairwise HRs between subtypes are depicted alongside their 95% confidence intervals. Pairwise comparisons are colored based on the reference subtype, which is always positioned lower than the query cluster. Combined HRs, obtained by pooling individual patient outcomes and performing Cox proportional hazards regression stratified by cohort, are also depicted by black squares alongside the HRs derived from each individual cohort.

other *DNMT3A* mutations (Extended Data Fig. 4f,g), suggesting that *DNMT3A* R882 may be more permissive of AML differentiation than other *DNMT3A* mutations. Collectively, these data demonstrate that depicting AML inter-patient heterogeneity through hierarchy composition can capture and potentially integrate both genomic and stem cell models of AML heterogeneity.

Primitive versus GMP axis captures patient prognosis. In line with the observed associations with favorable and adverse cytogenetics, patients with different hierarchy subtypes also differed in their survival outcomes, wherein Primitive hierarchies were associated with worse outcomes and GMP-dominant hierarchies were associated with better outcomes (Fig. 2f,g and Supplementary Table 6). We validated these findings using microarray data from a cohort of genetically diverse adult patients with AML (GSE6891 (ref. [35]); Extended Data Fig. 5a,b) as well as an RNA-seq cohort of pediatric patients with AML (TARGET-AML[36]; Extended Data Fig. 5c,d). To identify leukemic cell types linked to patient survival, we performed regularized Cox regression on the TCGA and BEAT-AML datasets using leukemia cell type abundances. Quiescent LSPC abundance and Cycling LSPC abundance were predictive of adverse outcomes (coefficients: 0.34 and 0.72), and GMP-like abundance was predictive of favorable outcomes (coefficient: −1.54). Strikingly, the composite survival score that included all three of these populations was highly anti-correlated with PC1 ($r=-0.99$; Extended Data Fig. 5e). Indeed, PC1 was highly associated with survival outcomes in the TCGA, BEAT-AML and GSE6891 cohorts (Extended Data Fig. 5f) and retained significance in a multivariate meta-analysis incorporating all three datasets ($P=0.007$; Supplementary Table 7). Moreover, both pediatric and adult patients with AML who did not achieve complete remission after induction chemotherapy had higher Quiescent LSPC abundance and lower GMP-like abundance than patients who achieved remission (Extended Data Fig. 5g). In contrast, PC2 was not associated with patient survival ($P=0.412$; Supplementary Table 7).

We reasoned that the biology underlying the Primitive versus GMP axis may also underlie part of the variation captured by existing

prognostic scores in AML. Indeed, when we considered four recent prognostic gene expression scores for AML (LSC17 (ref. [17]), APS[37], 3-Gene[38] and CODEG22 (ref. [39])), we found convergence across all four scores, wherein patients with high scores, indicative of adverse prognosis, had high Quiescent LSPC abundance and low GMP-like abundance (Extended Data Fig. 5h,i). Finally, unbiased analysis of associations between the expression of individual genes and patient survival outcomes revealed that genes associated with shorter survival were enriched for HSC-specific programs, whereas genes associated with longer survival were enriched for GMP-specific programs (Extended Data Fig. 5j).

Although previous studies implicated primitiveness as reflecting poor outcomes, our data reveal more complexity, suggesting that the biological distinction between stem cells and GMP progenitors underlie prognosis in AML rather than stemness properties alone. Thus, our data argue that existing prognostic AML scores are linked to this specific axis, pointing to the clinical importance of the biological properties that determine hierarchy composition.

Hierarchy composition changes between diagnosis and relapse. Given the associations observed between hierarchy composition and clinical outcomes in AML, we asked whether the composition of these hierarchies evolve over the course of disease. To understand how AML hierarchies change throughout disease progression, we deconvoluted 44 pairs of AML samples collected at diagnosis and relapse after induction chemotherapy from four independent cohorts[14,40–42] (Fig. 3a and Extended Data Fig. 6a). At diagnosis, patients presented with diverse hierarchy compositions, yet, by relapse, most were classified as Primitive (Fig. 3b) with significant expansion of total LSPC populations ($P=1\times10^{-8}$) and, in particular, Quiescent LSPC ($P=9\times10^{-6}$) (Fig. 3c). To validate this finding at the single-cell level, we analyzed scRNA-seq data from eight patients with relapsed AML[43] and observed uniformly higher LSPC abundance as compared to scRNA-seq data from 12 diagnostic AML samples[21] (Fig. 3d and Extended Data Fig. 6b).

Although LSPC expansion at relapse is in line with prior functional xenotransplantation studies, both the consistency and magnitude of this phenotype were unexpected, with LSPC expansion observed in 89% of patients at relapse (39 of 44 pairs). Furthermore, all five patients for whom LSPC abundance decreased at relapse already had high LSPC abundance at diagnosis (median 78.9%, compared to 27.3% for other patients). To benchmark this finding, we also evaluated 12,441 biological signatures from the Molecular Signatures Database (MSigDB) spanning biological pathways, immune processes and cancer/AML-specific gene sets and found the enrichment in total LSPC at relapse to be two orders of magnitude more significant than the top-ranked signature from the MSigDB (Fig. 3e). Indeed, classifiers trained on the abundance of LSPC populations were able to achieve near-perfect performance in classifying paired diagnosis and relapse samples (median AUC = 0.96; Supplementary Table 8).

These changes in cellular composition from diagnosis to relapse can also help to contextualize patterns of clonal evolution in AML. For example, in the context of *NPM1*-mutant AML, *FLT3*-ITD alterations are recurrently gained at relapse, whereas *NRAS* and *FLT3*-TKD alterations are recurrently lost at relapse[44]. Indeed, *FLT3*-ITD with *NPM1c* generated Primitive hierarchies (Fig. 3f), whereas mutant *NRAS* or *FLT3*-TKD with *NPM1c* generated Mature hierarchies (Fig. 3g). For a subset of the patients we analyzed, changes in hierarchy composition from diagnosis to relapse were concordant with patterns of clonal evolution (Fig. 3h and Extended Data Fig. 6c–f). In other cases, shifts in hierarchy composition occurred in the absence of clear genetic changes, potentially due to non-genetic modes of evolution (Fig. 3i). Together, our findings establish LSPC population expansion as a common hallmark across diverse evolutionary paths to AML relapse after chemotherapy.

Primitive versus Mature axis captures ex vivo drug sensitivity. Having shown that survival outcomes after chemotherapy are tied to hierarchy composition (that is, the Primitive versus GMP axis), we asked whether AML samples with different cellular compositions varied in their vulnerability to newer investigational therapies. Ex vivo drug sensitivity data from two public datasets[33,45] were integrated with cell type abundance to generate drug sensitivity profiles for each leukemic cell type (Fig. 4a). This revealed large differences in drug responses between primitive blasts and mature blasts, with separation of drug responses occurring primarily along the Primitive versus Mature axis, in which PC2 significantly correlated (false discovery rate (FDR) < 0.05) with response to 37 drugs in the BEAT-AML screen and 64 drugs in a separate screen from Lee et al.[45] (total = 101; Fig. 4b and Extended Data Fig. 7a). By contrast, PC1 was not associated with sensitivity to any drug from either screen (Extended Data Fig. 7a).

The large number of drugs for which sensitivity was associated with the Primitive versus Mature axis suggested that this axis comprised the primary source of variation underlying ex vivo sensitivity to investigational drugs in AML. To test this hypothesis, we performed unsupervised clustering of patient samples from Lee et al.[45] based on their ex vivo sensitivity to 159 drugs and identified two patient clusters with global differences in their drug sensitivity profiles (Extended Data Fig. 7b). Differential expression (DE) analysis revealed that one cluster was highly enriched for primitive HSC programs, whereas the other was enriched for mature myeloid programs (Extended Data Fig. 7c), demonstrating that the Primitive versus Mature axis captures fundamental differences in drug sensitivity profiles among patients with AML.

Hierarchy-based gene expression scores predict drug response. As a first step to translate the association between the Primitive versus Mature axis and an individual's response to a specific drug into the clinic, we sought to capture this axis through simple gene expression scores. As a proof of concept, we turned to the LSC17 score, for which a CAP/CLIA-certified clinical assay has been developed on the NanoString platform[17,46]. Given that the LSC17 score was associated with leukemic hierarchy composition (Extended Data Fig. 5h,i), we reasoned that deriving a subscore from these 17 genes to estimate PC2 may provide a rapidly deployable tool to inform therapy selection using data from the existing LSC17 assay. We, thus, retrained the LSC17 genes on PC2 through LASSO regression to identify a seven-gene lineage classification subscore (hereafter, LinClass-7; Supplementary Table 9) (Fig. 4c). LinClass-7 correlated well with PC2 ($|r| = 0.82$) in the validation cohort and was significantly associated with sensitivity to 33 drugs from BEAT-AML as well as 72 drugs from Lee et al.[45] (total = 105; Fig. 4d), each of which targeted either primitive blasts (for example, venetoclax, azacytidine and mubritinib) or mature blasts (for example, MEK/mTOR inhibitors) (Fig. 4e). Notably, neither LSC17 nor other prognostic AML scores[37–39] were significantly associated with drug sensitivity (Extended Data Fig. 7d), and none effectively stratified patients by sensitivity to clinically relevant agents, such as venetoclax and azacitidine (Extended Data Fig. 7e).

To examine the clinical relevance of these Primitive versus Mature scores, we turned to gene expression data from the pivotal ALFA-0701 adult AML clinical trial of low fractionated doses of gemtuzumab ozogamicin (GO) in combination with standard chemotherapy[47,48]. We asked whether the LinClass-7 score could predict clinical benefit from GO treatment. In the ALFA-0701 trial, the addition of GO to standard chemotherapy conferred significant benefits in event-free survival (EFS) and relapse-free survival (RFS) (EFS hazard ratio (HR) = 0.64 (0.47–0.89), $P=0.008$; RFS HR = 0.65 (0.43–0.99), $P=0.045$; Supplementary Table 10), but differences in overall survival (OS) were not significant at the final time of follow-up. LinClass-7 effectively separated responders

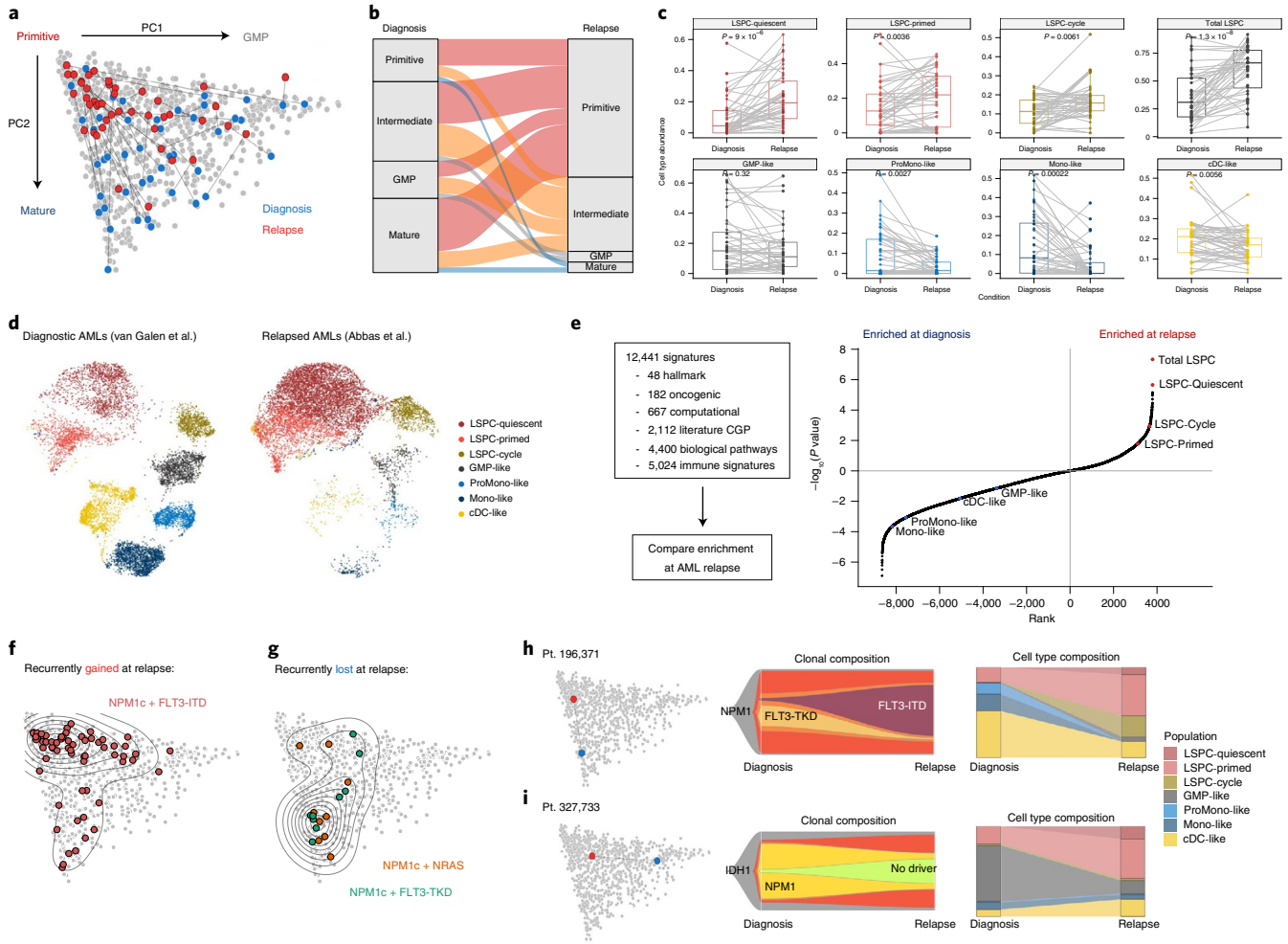


Fig. 3 | Transitions in hierarchy composition from diagnosis to relapse. a, Transitions in hierarchy composition from diagnosis to relapse in 44 paired AML samples from four independent cohorts. **b**, Alluvial diagram depicting distribution of hierarchy subtype from diagnosis to relapse. The width of each band reflects the number of patients transitioning from one subtype to another from diagnosis to relapse. **c**, Changes in leukemic cell type abundance, including total LSPC abundance, between 44 diagnosis and relapse pairs. Significance was evaluated using a two-sided Wilcoxon signed-rank test. Box plots indicate the range of the central 50% of the data, with the central line marking the median. Whiskers extend from each box to 1.5× the interquartile range. **d**, scRNA-seq of twelve diagnostic AMLs from van Galen et al.[21] compared to eight relapsed AMLs from Abbas et al.[43], classified, sub-sampled to 10,000 cells and projected onto a common embedding using scArches with scANVI. **e**, Benchmarking the significance and magnitude of changes in hierarchy composition from diagnosis to relapse against 12,441 signatures from the MSigDB. The rank and significance of enrichment for each leukemic cell population, as well as 12,441 signatures from diagnosis to relapse, are shown, with the *y* axis depicting the absolute value of the $\log_{10}$(*P* value) in the direction of the enrichment (positive for relapse, negative for diagnosis). *P* values were derived from paired *t*-tests; non-parametric Wilcoxon signed-rank tests were also performed to ensure that results were comparable. **f**, Patients with concurrent *NPM1c*+*FLT3*-ITD alterations ($n=64$ patients), which are recurrently expanded at relapse. **g**, Patients with concurrent *NPM1c*+*NRAS* ($n=9$ patients at VAF > 0.25) or *NPM1c*+*FLT3*-TKD ($n=9$ patients at VAF > 0.25) alterations, which are recurrently lost at relapse. **h**, **i**, Changes in clonal and cell type composition from diagnosis to relapse. These are depicted for a patient with concordant shifts in both clonal composition and cell type composition (**h**) as well as for a patient with substantial changes in cell type composition with minimal detected changes in clonal composition with respect to known driver mutations (**i**).

from non-responders: GO treatment led to significantly longer EFS and RFS (EFS HR = 0.57 (0.35–0.91), $P=0.018$; RFS HR = 0.53 (0.30–0.93), $P=0.028$; Supplementary Table 10) for patients with LinClass-7 low (Mature > Primitive) AML, although this association did not extend to OS. In contrast, patients with LinClass-7 high (Primitive > Mature) AML derived no significant survival benefit from GO (Fig. 4f and Supplementary Table 10). Notably, we observed no association between surface levels of GO target CD33 with either LinClass-7 (Fig. 4g) or PC2 (Extended Data Fig. 8b) and found most patients to be $CD33^+$ regardless of their hierarchy subtype (Fig. 4h). The LSC17 score has also been shown to predict clinical benefit from GO treatment[17]. Our analysis shows that the LSC17 and LinClass-7 scores captured different subsets of patients, and further subgroup analysis revealed that only patients who had low scores for both LinClass-7 and LSC17 derived clinical benefit from GO treatment (EFS HR = 0.44 (0.23–0.85), $P=0.014$; RFS HR = 0.35 (0.16–0.78), $P=0.009$; Extended Data Fig. 8c and Supplementary Table 10), demonstrating complementarity between LSC17 and LinClass-7 in the prediction of clinical benefit from GO.

In the context of adult AML, these analyses point to the utility of LinClass-7 as a companion score to LSC17, enabling prediction of response to an array of current and investigational drugs. Notably, the distribution of LSC17 and LinClass-7 scores across patient samples also loosely recapitulates the primary axes of variation in hierarchy composition, separating Primitive, GMP and Mature AMLs (Extended Data Fig. 9a). Thus, the LSC17 and LinClass-7

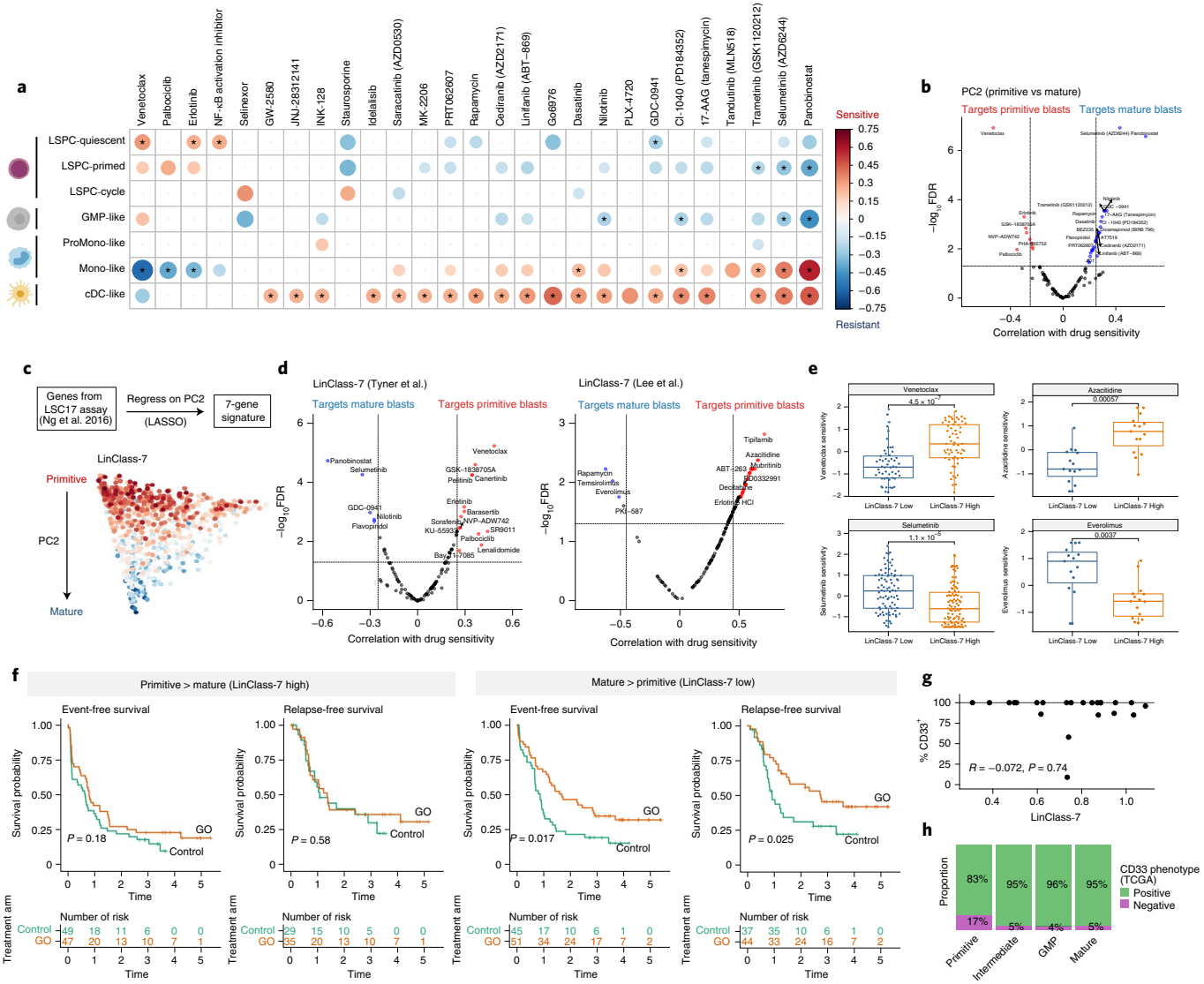


Fig. 4 | AML hierarchy composition as a determinant of targeted therapy response. a, Pearson correlation between cell type abundance and ex vivo drug sensitivity (−AUC) across 202 diagnostic patient samples in BEAT-AML, wherein color and size represent the direction and magnitude of the correlation. Only correlations with $P < 0.05$ are depicted; those with $q < 0.05$ are marked with an asterisk. **b**, Volcano plot of correlations between the Primitive versus Mature axis (PC2) and ex vivo drug sensitivities from the BEAT-AML screen, identifying drugs that preferentially target either primitive or mature AML blasts. **c**, LinClass-7 (trained on PC2) captures the Primitive versus Mature axis. **d**, Correlation with LinClass-7 identifies drugs targeting either primitive blasts or mature blasts from BEAT-AML (Tyner et al.[33]; $n = 202$) as well as a separate primary AML drug screen (Lee et al.[45]; $n = 30$). **e**, Scaled drug sensitivity (−AUC) of LinClass-7 high AMLs (Primitive > Mature) and LinClass-7 low AMLs (Mature > Primitive) for BCL2 inhibitor venetoclax ($n = 114$ from Tyner et al.[33]), hypomethylating agent azacitidine ($n = 30$ from Lee et al.[45]), MEK inhibitor selumetinib ($n = 178$ from Tyner et al.[33]) and MTOR inhibitor everolimus ($n = 30$ from Lee et al.[45]). Box plots indicate the range of the central 50% of the data, with the central line marking the median. Whiskers extend from each box to 1.5× the interquartile range. Significance was evaluated through a two-sided Wilcoxon rank-sum test. **f**, Subgroup analysis of the ALFA-0701 trial ($n = 192$), evaluating GO, a drug-conjugated antibody targeting CD33 in AML. Event-free survival and relapse-free survival of control patients (daunorubicin+cytarabine) compared to GO patients (daunorubicin+cytarabine+GO), stratified by LinClass-7 score into LinClass-7 High (Primitive > Mature) and LinClass-7 Low (Mature > Primitive). Significance was evaluated through a log-rank test. **g**, Lack of association between LinClass-7 and surface CD33 levels, evaluated by Pearson correlation across 23 patients with Toronto PMH AML for whom both RNA-seq and clinical flow information was available. **h**, CD33 positivity rates among 151 TCGA patients for whom CD33 surface marker phenotypes were available, split by AML hierarchy composition.

scores, measurable through the same CAP/CLIA-certified clinical assay[46] (Extended Data Fig. 9b), potentially enable both prognostic and predictive stratification of patient samples while also providing salient information on patient hierarchy composition.

Although adult and pediatric AML are molecularly distinct diseases[36], the Primitive versus Mature axis captured through PC2 was also able to predict clinical benefit from GO among pediatric patients with AML in the TARGET-AML retrospective cohort[36]. Pediatric patients with high PC2 (Mature > Primitive) experienced longer OS and EFS outcomes with GO treatment than those who did not receive GO treatment (OS HR = 0.51 (0.30–0.89), $P = 0.017$; EFS HR = 0.58 (0.37–0.90), $P = 0.017$; Supplementary Table 12). In contrast, GO treatment did not influence survival outcomes of pediatric patients with low PC2 (Primitive > Mature) (OS HR = 1.20 (0.68–2.00),

$P=0.553$; EFS HR = 0.97 (0.62–1.50), $P=0.878$; Extended Data Fig. 8e,f and Supplementary Table 12). However, in the pediatric AML context, we observed lower correlation between LinClass-7 and PC2 ($r=0.51$) and found that prediction of clinical benefit from GO treatment in pediatric AML did not extend to either LinClass-7 or LSC17 (Extended Data Fig. 8e,f and Supplementary Table 12).

Given this, we asked whether a Primitive versus Mature score not constrained by the LSC17 genes could accurately recapitulate PC2 in both adult and pediatric AML. Starting with the top 50 PC2-correlated and top 50 PC2-anticorrelated genes, we trained a 34-gene score (termed PC2-34; Supplementary Table 9) that was highly correlated with PC2 ($|r| = 0.95$ in the validation cohort). PC2-34 performed similarly to LinClass-7 in capturing drug sensitivity in the BEAT-AML (48 drugs at FDR < 0.05) and Lee et al.[45] (82 drugs at FDR < 0.05) screens (Extended Data Fig. 7d,e). PC2-34 also predicted clinical benefit from GO treatment (Extended Data Fig. 8a and Supplementary Table 11) and demonstrated similar complementarity with LSC17 in further stratifying GO response (Extended Data Fig. 8d and Supplementary Table 11). Notably, the PC2-34 score remained well-correlated with the Primitive versus Mature axis in pediatric AML ($r=0.88$) and captured the same survival benefits from GO treatment among these pediatric patients (Extended Data Fig. 8e,f and Supplementary Table 12). Together, our data provide a proof of concept that gene expression scores can be readily generated to capture axes of variation in leukemic hierarchy composition, and these may potentially represent powerful biomarkers of response to non-chemotherapy agents. They are also one of only a few biomarkers that are broadly applicable in both adult and pediatric AML.

AML hierarchy framework to guide preclinical drug studies. We next sought to determine how our leukemic hierarchy framework can be deployed in the context of drug development. Drug candidates are often identified based on reduction in viability of bulk leukemia cells or cell lines, yet this measure lacks critical information pertaining to the subpopulations of cells that are targeted or that persist after treatment. To understand how drug treatment affects cellular composition, we deconvoluted RNA-seq data from 43 datasets in the Gene Expression Omnibus (GEO) and ArrayExpress with human AML cells sequenced before and after drug treatment (Fig. 5a,b and Supplementary Table 13). The changes in cellular composition after drug treatment were visualized in low-dimensional uniform manifold approximation and projection (UMAP) space, and treatments were clustered based on the changes they induced (Fig. 5c,d). Across 153 treatment conditions, 125 resulted in significant changes in cell type composition. Seventy-seven treatment conditions led to a significant increase in PC2, potentially reflecting differentiation, yet most of these treatments resulted in depletion of GMP-like blasts (69%), with fewer treatments depleting the more primitive Quiescent LSPC (30%) or Primed LSPC (14%) populations (Fig. 5d, Extended Data Fig. 10a,b and Supplementary Table 14). For example, ATRA induced differentiation predominantly from GMP-like blasts (Fig. 5e). In contrast, differentiation induced by the DHODH inhibitor Brequinar was accompanied by a reduction in Quiescent LSPC abundance, suggesting that this drug may better deplete the stem cell compartment (Fig. 5e).

In some cases, the cell population depleted by a drug corresponded to the expression of the drug target. Selinexor, a drug targeting the nuclear export protein XPO1, is an example (Fig. 5f). *XPO1* expression and nuclear export processes were enriched in the Cycling LSPC population at the single-cell level (Fig. 5g), and abundance of this cell population was correlated with ex vivo selinexor sensitivity in the BEAT-AML screen (Fig. 5h). Notably, treatment of primary AML samples with selinexor resulted in depletion of the Cycling LSPC population both in vitro[49] and in vivo[50] (Fig. 5i,j) across diverse genetic backgrounds. Together, these data shed light on the changes in cellular composition that follow drug treatment and offer a functionally relevant read-out for prioritizing candidate drugs in preclinical settings.

We next asked how hierarchy information can be used in preclinical studies that are more proximal to clinical translation, such as in the context of in vivo drug response. Our prior patient-derived xenograft (PDX) response data for two drugs were used: fedratinib[51], an approved JAK2 inhibitor for myeloproliferative neoplasms, and CC-90009 (ref. [52]), an immunomodulatory agent that induces cereblon-mediated degradation of GSPT1. AML samples ($n=32$ for fedratinib and $n=30$ for CC90009) were treated in 658 drug-treated or vehicle-treated PDX recipients. Deconvoluted RNA-seq profiles from the primary patient samples before xenotransplantation were clustered based on hierarchy composition and categorized as Primitive, Intermediate/GMP or Mature (Fig. 6a,b).

The primary target of fedratinib, JAK2, was predominantly expressed in Mono-like and cDC-like blasts (Fig. 6c). These mature blasts were, in turn, enriched in patient samples that responded well to fedratinib in vivo (Fig. 6d). Subgroup analysis of fedratinib response showed high efficacy in AMLs with Mature hierarchies (88% response rate), whereas response rates among other hierarchy subtypes were poor (46% for Primitive and 20% for Intermediate/GMP) (Fig. 6e). The target of CC-90009, GSPT1, was highly expressed in Cycling LSPC and GMP-like blasts (Fig. 6f). GMP-like blasts were enriched among responders, whereas Quiescent LSPCs were enriched among partial and non-responders (Fig. 6g). Subgroup analysis showed high CC-90009 efficacy in AMLs with Mature and Intermediate/GMP hierarchies, with 88% and 83% response rates, respectively. In contrast, those with Primitive hierarchies had heterogeneous responses at a rate of 40% (Fig. 6h).

To better understand the heterogeneous responses to fedratinib and CC-90009 among patient samples, we compared the genomic features of responding and non-responding AML samples for both fedratinib and CC-90009 treatment conditions. Among Primitive AML hierarchies, *NPM1c* mutations were associated with favorable response to fedratinib and poor response to CC-90009, whereas Primitive AMLs lacking *NPM1c* mutation demonstrated favorable response to CC-90009 and poor response to fedratinib (Fig. 6e,h). Notably, the association of *NPM1c* signatures with fedratinib and CC-90009 response among Primitive AMLs did not extend to other hierarchy subtypes. Given the *NPM1c*-based response dichotomy to fedratinib and CC-90009 among Primitive hierarchies, as well as the sensitivity of Intermediate/GMP hierarchies to CC-90009 and Mature hierarchies to both drugs, we reasoned that a combination of the two drugs may show efficacy against a broader range of samples than either drug alone. Therefore, PDX xenografts from eight patients with AML with diverse hierarchy compositions were treated with both drugs alone and in combination (median five mice per treatment per patient). Despite variable responses to single agents, PDXs from seven of the eight patients tested responded fully to combination treatment with virtual elimination of their leukemic grafts (Fig. 6i).

Overall, responses to fedratinib, CC-90009 and combination treatment in PDX models were all significantly associated with hierarchy composition (Fig. 6j). These data establish that stratifying AML by hierarchy composition can help to identify patient samples likely to benefit from specific therapies while also providing proof of concept for the design of combination regimens through pairing of drugs that exhibit complementarity in their hierarchy-based targeting profiles.

Discussion

In this study, we developed a new approach for understanding heterogeneity in AML by characterizing the cellular composition of each patient's leukemic hierarchy. Analysis of patient-specific

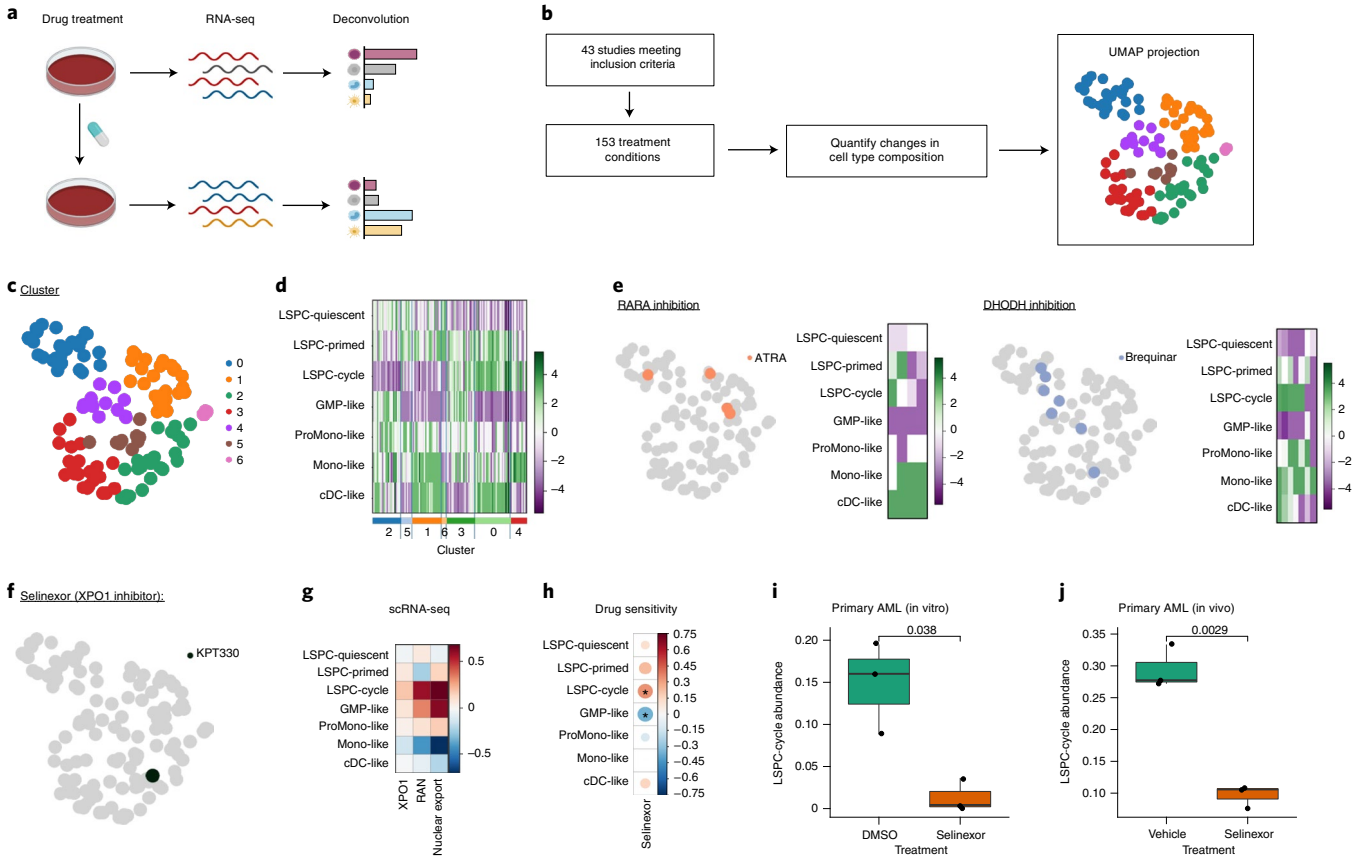
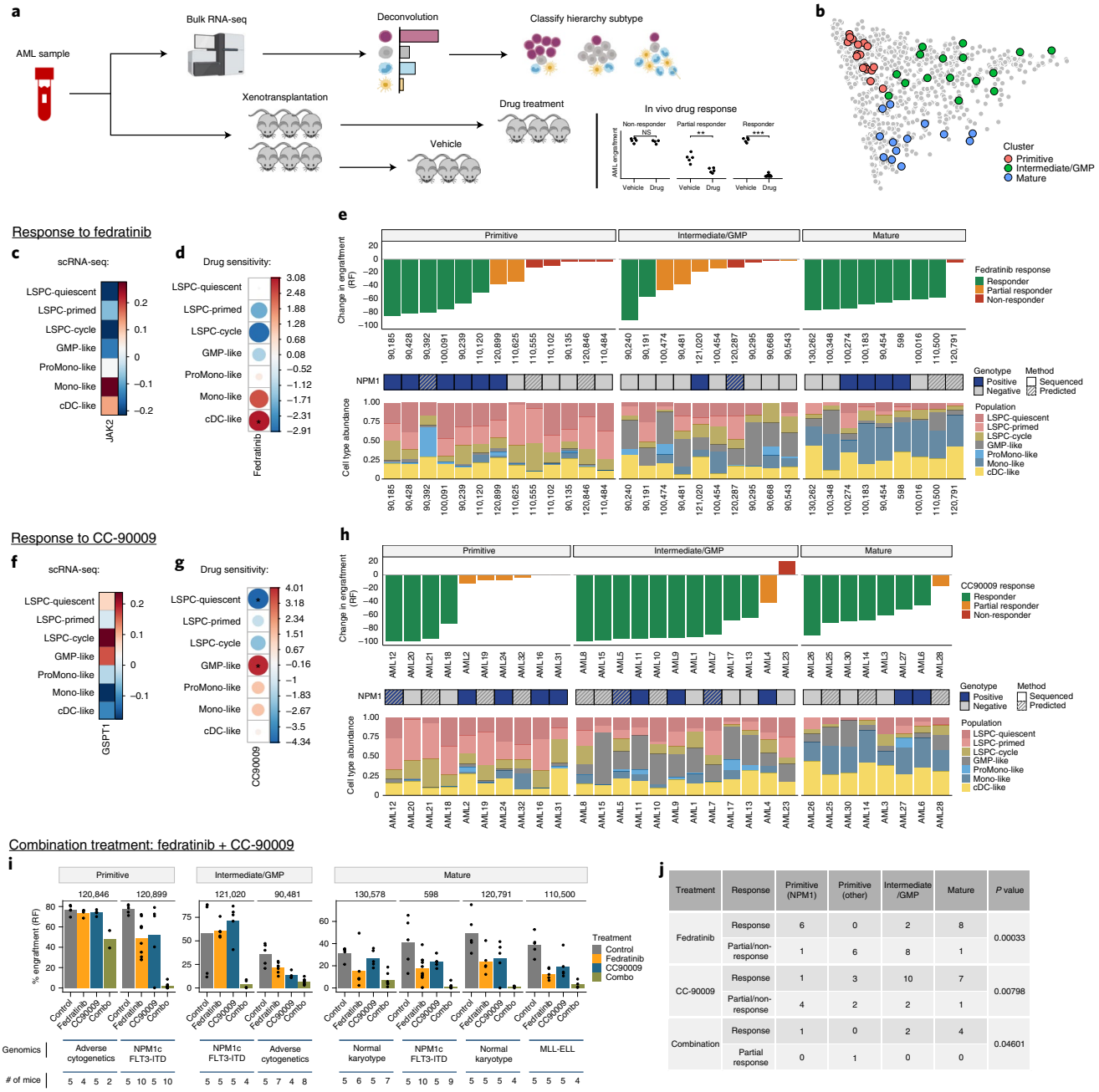


Fig. 5 | Changes in cellular composition after drug treatment. a, Experimental design of re-analyzed preclinical studies from the literature. Only studies of human AML with RNA-seq available before and after drug treatment were included to quantify changes in cell type composition. **b**, Schematic of re-analysis approach. Changes in the abundance of each cell type were quantified in each treatment condition, and treatments with significant changes in at least one cell type were used as input for dimensionality reduction with UMAP and subsequent clustering. **c**, Clustering of drug treatments on the basis of changes in cell type composition. **d**, Heat map depicting cell type composition changes of drug treatments within each cluster. Purple denotes decreased abundance after treatment, and green denotes increased abundance after treatment. **e**, Examples of the drug treatments targeting specific processes and the changes induced in the abundance of each cell type after treatment. **f**, Cellular composition changes after in vitro selinexor treatment in NPM1 mutant AMLs from Brunetti et al.[49]. **g**, Mean expression of XPO1 (the target of selinexor) and associated genes and pathways in AML blast populations from scRNA-seq. Nuclear export pathway gene set was obtained from Gene Ontology Biological Pathways. **h**, Pearson correlation between cell type abundance and ex vivo drug sensitivity in 40 of the 202 diagnostic BEAT-AML samples for which selinexor sensitivity was reported. Correlations with $P < 0.05$ are marked with an asterisk. **i**, LSPC-Cycle abundance in three primary AML samples treated with DMSO control or selinexor in vitro (treatment and RNA-seq from Brunetti et al.[49]). Significance was derived from a two-sided Student's *t*-test. Box plots indicate the range of the central 50% of the data, with the central line marking the median. Whiskers extend from each box to 1.5× the interquartile range. **j**, LSPC-Cycle abundance in three primary AML samples treated with DMSO control or selinexor in vivo (treatment from Etchin et al.[50]; RNA-seq from this study). Significance was derived from a two-sided Student's *t*-test. Box plots indicate the range of the central 50% of the data, with the central line marking the median. Whiskers extend from each box to 1.5× the interquartile range.

variation in hierarchy composition across large cohorts captured and integrated information pertaining to genomic profiles, functional stem cell properties and clinical outcomes within a single classification framework—something that could not be achieved by applying either of the genomic or stem cell models alone. Despite the wide diversity of genetic drivers, hierarchy composition could be distilled into four main subtypes, implying convergence in how mutations perturb LSC function and impair hematopoietic differentiation to generate a leukemia cell hierarchy. This new framework provides a means of understanding how different genetic subgroups relate to one another and, more broadly, how genetic alterations relate to LSC properties, enabling a more comprehensive view of biological heterogeneity in AML.

Our analysis of diagnosis/relapse pairs demonstrates the value of longitudinal monitoring of AML hierarchy composition through disease progression. Although only a subset of relapsed AML cases are explained by clear patterns of genetic evolution, we show that LSPC expansion constitutes a hallmark of AML relapse after chemotherapy. This has important implications for trial design given that patients with relapsed AML, for whom Mature and GMP-dominant hierarchies are underrepresented, are often the first patients in which novel therapeutics are evaluated. This mismatch could lead to valuable drugs for patients with such hierarchies being discounted. Our findings also raise an interesting question on the cell types that bear stemness properties. Emerging data from relapse after venetoclax and azacitidine treatment show a loss of phenotypic LSC and the emergence of a promonocytic blast population that may also carry leukemic propagation potential[53]. Thus, an important and unresolved question pertains to the self-renewal capacity of distinct blast populations within these leukemia cell hierarchies. To address this, deeper functional studies of patient samples reflecting specific hierarchy subtypes will be



j

Treatment	Response	Primitive (NPM1)	Primitive (other)	Intermediate /GMP	Mature	*P* value
Fedratinib	Response	6	0	2	8	0.00033
	Partial/non-response	1	6	8	1	
CC-90009	Response	1	3	10	7	0.00798
	Partial/non-response	4	2	2	1	
Combination	Response	1	0	2	4	0.04601
	Partial response	0	1	0	0	

Fig. 6 | Hierarchy-based stratification predicts in vivo response to fedratinib and CC-90009. a, Experimental design for evaluating the relationship between AML hierarchy composition and drug response in PDX models. Response data of 658 drug-treated or vehicle-treated treated mice from prior studies (Surka et al.[52] and Chen et al.[51]) were integrated with hierarchy composition data from the primary patient samples, and heterogeneous in vivo drug response was re-analyzed in the context of patient hierarchy subtypes. **b**, Projected hierarchy composition of primary patient samples before in vivo drug treatment, categorized by subtype. **c**, Mean expression of JAK2, the target of fedratinib, in AML blast populations from scRNA-seq. **d**, Differences in cell type abundance between responders and partial/non-responders to fedratinib, represented as the absolute log(*P* value) from a two-sided Wilcoxon rank-sum test in the direction of the change. Red depicts enrichment in responders, and blue depicts enrichment in partial/non-responders. Significant differences ($P < 0.05$) are marked with an asterisk. **e**, Xenograft responses to fedratinib, stratified by leukemic hierarchy subtype. Bar plot depicts the mean difference in leukemic engraftment in fedratinib-treated mice compared to vehicle-treated mice. *NPM1* mutation status and cell type composition of each patient sample before treatment are depicted below each bar. **f**, Mean expression of GSPT1, the target of CC-90009, in AML blast populations from scRNA-seq. **g**, Differences in cell type abundance between responders and partial/non-responders to CC-90009, represented as the absolute log(*P* value) from a two-sided Wilcoxon rank-sum test in the direction of the change. Red depicts enrichment in responders, and blue depicts enrichment in partial/non-responders. Significant differences ($P < 0.05$) are marked with an asterisk. **h**, Xenograft responses to CC-90009, stratified by leukemic hierarchy subtype. Bar plot depicts the mean difference in leukemic engraftment in CC-90009-treated mice compared to vehicle-treated mice. *NPM1* mutation status and cell type composition of each patient before treatment are depicted below each bar. **i**, Response to fedratinib+CC-90009 combination treatment. Patients are stratified by hierarchy, and mean AML engraftment levels are depicted for each treatment condition. **j**, In vivo efficacy of fedratinib, CC-90009 and combination treatment of xenografted AML patient samples, stratified by patient hierarchy subtype. Significance was evaluated using a chi-squared test. NS, not significant.

necessary in order to pinpoint the specific populations that must be targeted to ensure long-term remission.

An unexpected finding of our study was the lack of association between features that were prognostic and features that were predictive of drug response, both of which were captured through hierarchy composition. The Primitive versus GMP axis was highly prognostic and indirectly captured by LSC17 (ref. [17]) and other modern prognostic biomarkers[37–39]. However, neither this axis nor existing prognostic biomarkers could adequately predict drug response to biologically targeted therapies, which was instead captured by the Primitive versus Mature axis. Future functional studies are needed to understand the biological mechanism of why this axis drives drug response prediction. Notably, this axis can be distilled into new gene expression scores, including LinClass-7, that have potential for rapid translation into the clinic. Hierarchy-based classification has implications for clinical trial design of investigational drugs: AMLs can potentially be stratified preclinically on the basis of hierarchy composition to identify patient subsets that may most likely benefit from a single drug or even a predicted combination. Furthermore, for drugs that are already in the clinic, such as venetoclax, azacitidine and GO, where response varies, this stratification could potentially be used to select patients most likely to benefit from these specific treatments.

A large number of targeted therapies are currently in development for AML and many investigational therapies are progressing to clinical trials. Collectively, our study suggests that biomarkers focused on the composition of each patient's leukemia cell hierarchy have strong potential to guide the development and selection of these therapies, thereby setting the foundation for a new precision medicine framework in AML.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-022-01819-x.

Received: 27 July 2021; Accepted: 7 April 2022;
Published online: 26 May 2022

References

1. Hungerford, D. A. & Nowell, P. C. A minute chromosome in human chronic granulocytic leukemia. *Science* **132**, 1497–1499 (1960).
2. Papaemmanuil, E. et al. Genomic classification and prognosis in acute myeloid leukemia. *N. Engl. J. Med.* **374**, 2209–2221 (2016).
3. Klco, J. M. et al. Functional heterogeneity of genetically defined subclones in acute myeloid leukemia. *Cancer Cell* **25**, 379–392 (2014).
4. Till, J. E. & McCULLOCH, E. A. A direct measurement of the radiation sensitivity of normal mouse bone marrow cells. *Radiat. Res.* **14**, 213–222 (1961).
5. Clarkson, B., Ohkita, T., Ota, K. & Fried, J. Studies of cellular proliferation in human leukemia. I. Estimation of growth rates of leukemic and normal hematopoietic cells in two adults with acute leukemia given single injections of tritiated thymidine. *J. Clin. Invest.* **46**, 506–529 (1967).
6. Minden, M. D., Till, J. E. & McCulloch, E. A. Proliferative state of blast cell progenitors in acute myeloblastic leukemia (AML). *Blood* **52**, 592–600 (1978).
7. Griffin, J. D., Larcom, P. & Schlossman, S. F. Use of surface markers to identify a subset of acute myelomonocytic leukemia cells with progenitor cell properties. *Blood* **62**, 1300–1303 (1983).
8. Wouters, R. & Löwenberg, B. On the maturation order of AML cells: a distinction on the basis of self-renewal properties and immunologic phenotypes. *Blood* **63**, 684–689 (1984).
9. Buick, R. N., Minden, M. D. & McCulloch, E. A. Self-renewal in culture of proliferative blast progenitor cells in acute myeloblastic leukemia. *Blood* **54**, 95–104 (1979).
10. Chang, L. J., Till, J. E. & McCulloch, E. A. The cellular basis of self renewal in culture by human acute myeloblastic leukemia blast cell progenitors. *J. Cell. Physiol.* **102**, 217–222 (1980).
11. McCulloch, E. A. Stem cells in normal and leukemic hemopoiesis (Henry Stratton Lecture, 1982). *Blood* **62**, 1–13 (1983).
12. Lapidot, T. et al. A cell initiating human acute myeloid leukaemia after transplantation into SCID mice. *Nature* **367**, 645–648 (1994).
13. Bonnet, D. & Dick, J. E. Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell. *Nat. Med.* **3**, 730–737 (1997).
14. Shlush, L. I. et al. Tracing the origins of relapse in acute myeloid leukaemia to stem cells. *Nature* **547**, 104–108 (2017).
15. Gentles, A. J., Plevritis, S. K., Majeti, R. & Alizadeh, A. A. Association of a leukemic stem cell gene expression signature with clinical outcomes in acute myeloid leukemia. *JAMA* **304**, 2706–2715 (2010).
16. Eppert, K. et al. Stem cell gene expression programs influence clinical outcome in human leukemia. *Nat. Med.* **17**, 1086–1093 (2011).
17. Ng, S. W. K. et al. A 17-gene stemness score for rapid determination of risk in acute leukaemia. *Nature* **540**, 433–437 (2016).
18. Elsayed, A. H. et al. A six-gene leukemic stem cell score identifies high risk pediatric acute myeloid leukemia. *Leukemia* **34**, 735–745 (2020).
19. Pierce, G. B. & Speers, W. C. Tumors as caricatures of the process of tissue renewal: prospects for therapy by directing differentiation. *Cancer Res.* **48**, 1996–2004 (1988).
20. Kreso, A. & Dick, J. E. Evolution of the cancer stem cell model. *Cell Stem Cell* **14**, 275–291 (2014).
21. van Galen, P. et al. Single-cell RNA-seq reveals AML hierarchies relevant to disease progression and immunity. *Cell* **176**, 1265–1281 (2019).
22. Wu, J. et al. A single-cell survey of cellular hierarchy in acute myeloid leukemia. *J. Hematol. Oncol.* **13**, 128 (2020).
23. Tarashansky, A. J., Xue, Y., Li, P., Quake, S. R. & Wang, B. Self-assembling manifolds in single-cell RNA sequencing data. *eLife* **8**, e48994 (2019).
24. Xie, S. Z. et al. Sphingolipid modulation activates proteostasis programs to govern human hematopoietic stem cell self-renewal. *Cell Stem Cell* **25**, 639–653 (2019).
25. Xie, S. Z. et al. Sphingosine-1-phosphate receptor 3 potentiates inflammatory programs in normal and leukemia stem cells to promote differentiation. *Blood Cancer Discov.* **2**, 32–53 (2021).
26. Takayama, N. et al. The transition from quiescent to activated states in human hematopoietic stem cells is governed by dynamic 3D genome reorganization. *Cell Stem Cell* **28**, 488–501 (2021).
27. Hope, K. J., Jin, L. & Dick, J. E. Acute myeloid leukemia originates from a hierarchy of leukemic stem cell classes that differ in self-renewal capacity. *Nat. Immunol.* **5**, 738–743 (2004).
28. Newman, A. M. et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nat. Biotechnol.* **37**, 773–782 (2019).
29. Dick, J. E. Stem cell concepts renew cancer research. *Blood* **112**, 4793–4807 (2008).
30. Quek, L. et al. Genetically distinct leukemic stem cells in human CD34$^-$ acute myeloid leukemia are arrested at a hemopoietic precursor-like stage. *J. Exp. Med.* **213**, 1513–1535 (2016).
31. Pabst, C. et al. GPR56 identifies primary human acute myeloid leukemia cells with high repopulating potential in vivo. *Blood* **127**, 2018–2027 (2016).
32. Cancer Genome Atlas Research Network et al. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* **368**, 2059–2074 (2013).
33. Tyner, J. W. et al. Functional genomic landscape of acute myeloid leukaemia. *Nature* **562**, 526–531 (2018).
34. Marquis, M. et al. High expression of *HMGA2* independently predicts poor clinical outcomes in acute myeloid leukemia. *Blood Cancer J.* **8**, 68 (2018).
35. Verhaak, R. G. W. et al. Prediction of molecular subtypes in acute myeloid leukemia based on gene expression profiling. *Haematologica* **94**, 131–134 (2009).
36. Bolouri, H. et al. The molecular landscape of pediatric acute myeloid leukemia reveals recurrent structural alterations and age-specific mutational interactions. *Nat. Med.* **24**, 103–112 (2017).
37. Docking, T. R. et al. A clinical transcriptome approach to patient stratification and therapy selection in acute myeloid leukemia. *Nat. Commun.* **12**, 2474 (2021).
38. Wagner, S. et al. A parsimonious 3-gene signature predicts clinical outcomes in an acute myeloid leukemia multicohort study. *Blood Adv.* **3**, 1330–1346 (2019).
39. Nehme, A. et al. Horizontal meta-analysis identifies common deregulated genes across AML subgroups providing a robust prognostic signature. *Blood Adv.* **4**, 5322–5335 (2020).
40. Li, S. et al. Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. *Nat. Med.* **22**, 792–799 (2016).
41. Christopher, M. J. et al. Immune escape of relapsed AML cells after allogeneic transplant. *N. Engl. J. Med.* **379**, 2330–2341 (2018).
42. Cocciardi, S. et al. Clonal evolution patterns in acute myeloid leukemia with *NPM1* mutation. *Nat. Commun.* **10**, 2031 (2019).
43. Abbas, H. A. et al. Single cell T cell landscape and T cell receptor repertoire profiling of AML in context of PD-1 blockade therapy. *Nat. Commun.* **12**, 6071 (2021).

44. Vosberg, S. & Greif, P. A. Clonal evolution of acute myeloid leukemia from diagnosis to relapse. *Genes Chromosomes Cancer* **58**, 839–849 (2019).
45. Lee, S.-I. et al. A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. *Nat. Commun.* **9**, 42 (2018).
46. Ng, S. W. K. et al. A clinical laboratory-developed LSC17 stemness score assay for rapid risk assessment of patients with acute myeloid leukemia. *Blood Adv.* **6**, 1064–1073 (2022).
47. Castaigne, S. et al. Effect of gemtuzumab ozogamicin on survival of adult patients with de-novo acute myeloid leukaemia (ALFA-0701): a randomised, open-label, phase 3 study. *Lancet* **379**, 1508–1516 (2012).
48. Lambert, J. et al. Gemtuzumab ozogamicin for de novo acute myeloid leukemia: final efficacy and safety updates from the open-label, phase III ALFA-0701 trial. *Haematologica* **104**, 113–119 (2019).
49. Brunetti, L. et al. Mutant NPM1 maintains the leukemic state through HOX expression. *Cancer Cell* **34**, 499–512 (2018).
50. Etchin, J. et al. Activity of a selective inhibitor of nuclear export, selinexor (KPT-330), against AML-initiating cells engrafted into immunosuppressed NSG mice. *Leukemia* **30**, 190–199 (2016).
51. Chen, W. C. et al. An integrated analysis of heterogeneous drug responses in acute myeloid leukemia that enables the discovery of predictive biomarkers. *Cancer Res.* **76**, 1214–1224 (2016).
52. Surka, C. et al. CC-90009, a novel cereblon E3 ligase modulator, targets acute myeloid leukemia blasts and leukemia stem cells. *Blood* **137**, 661–677 (2021).
53. Pei, S. et al. Monocytic subclones confer resistance to venetoclax-based therapy in patients with acute myeloid leukemia. *Cancer Discov.* **10**, 536–551 (2020).

Methods

Patient samples. All biological samples were collected with informed consent according to procedures approved by the Research Ethics Board of the University Health Network (UHN; REB no. 01-0573-C) and viably frozen in the Princess Margaret Hospital (PMH) Leukaemia Bank. No statistical methods were used to predetermine sample size. The investigators were not blinded to allocation during experiments and outcome assessment.

RNA-seq and pre-processing. RNA was extracted from bulk peripheral blood mononuclear cells using an RNeasy Micro Kit (Qiagen). Libraries were constructed using SMART-Seq (Clontech). A paired-end 50-base pair (bp) flow cell lane Illumina HiSeq 2000 yielded an average of 240 million aligned reads per sample. To align RNA-seq reads from samples used in selinexor and fedratinib treatments, Illumina paired-end sequence data were analyzed with BWA/version 0.6 alignment software with option (-s) to disable Smith–Waterman alignment. Reads were mapped onto GRCh37-lite reference genome and exon–exon junction reference whose coordinates were defined based on transcript annotations in Ensembl/version 59. Reads with mapping quality <10 were discarded, and duplicate reads were tagged using the Picard's MarkDuplicates program. JAGuaR 2.1 was used to incorporate reads spanning multiple exons into the alignment by introducing large alignment gaps. All transcripts of a given gene were collapsed into a single gene model such that exonic bases were the union of exonic bases that belonged to all known transcripts of the gene. Read counts and subsequently RPKM counts were obtained by counting the fraction of each read that overlapped with an exonic region for that gene. To align RNA-seq reads from functionally annotated LSC fractions, sequence data were aligned against GRCh38, and transcript sequences were downloaded from Ensembl build 90 using STAR 2.5.2a. Default parameters were used except for the following: '–chimSegmentMin 12 –chimJunctionOverhangMin 12 –alignSJDBoverhangMin 10 –alignMatesGapMax 100000 –alignIntronMax 100000 –chimSegmentReadGapMax parameter 3 –alignSJstitchMismatchNmax 5 -1 5 5'. Counts were obtained using HTSeq version 0.9.1. RNA-seq reads from four AML samples previously profiled by scRNA-seq from van Galen et al.[21] were aligned to GRCh38 with transcript annotations from GENCODE version 37 using STAR version 2.7.9a. Default parameters were used except for the following: '–chimSegmentMin 12 –chimJunctionOverhangMin 8 –alignSJDBoverhangMin 10 –alignSJstitchMismatchNmax 5 -1 5 5'. Raw counts were obtained using HTSeq version 0.7.2.

Re-clustering of LSPCs. scRNA-seq data from 12 patients with AML at diagnosis were obtained from van Galen et al.[21] (GSE116256). scRNA-seq count data were normalized using the R package 'scran'[54], log-transformed with an offset value of 1 and scaled. Leukemic cells labeled as 'HSC-like' and 'Prog-like' (hereafter, LSPCs) from the original study were subject to re-analysis using the SAM algorithm[23]. SAM was applied individually to the four patient samples with the highest number of LSPCs (based on a cutoff of >100 HSC-like cells) to assign weights to each gene based on how well they can demarcate emerging transcriptomic states. Feature weights for each gene were averaged across the four samples and subsequently applied to LSPCs from all 12 patients. No batch correction was applied. Using the 'scanpy' package[55], weighted expression data were subject to dimensionality reduction and neighborhood detection based on the cell–cell correlation. Leiden clustering[56] was performed with a resolution of 0.15 to identify three clusters of LSPCs shared across the patient samples. Re-annotated LSPC labels are included in Supplementary Table 1.

Evaluation of LSPC clustering. To evaluate the new cluster assignments, cell type classifiers were built and evaluated for the new and prior classifications using the R package 'SingleCellNet'[57]. For each classification, scran-normalized gene expression values were used as input, and 800 cells from each leukemic cell type were used as a training set. For each cell type, paired products of the top 25 genes for each cell type were calculated, and the 50 top gene pairs for each cell type were used to train the random-forest-based model with nTrees = 1,000. Models trained on the new and prior cell type classification were subsequently evaluated on a held-out dataset of at least 250 remaining cells for each cell type.

Regulon analysis and signature enrichment. To infer transcription factor (TF) regulon activity in scRNA-seq data, regulon analysis was performed using SCENIC[58]. The Docker image of pySCENIC was run as per the guidelines from Van de Sande et al.[59]: log-transformed counts from leukemic AML cells were used as the input, and candidate TFs were identified using a list of human TFs from Lambert et al.[60], with default parameters. To prune putative TF-target links within each regulon using annotations of TF motifs, CisTarget was applied using databases of known human TF motifs annotated at 500 bp, 5 kilobases (kb) and 10 kb of transcriptional start sites. Drop-out masking was also applied during this step. Enrichment of refined TF regulons was inferred using AUCell, and enrichment scores were scaled for visualization.

Characterization of scRNA-seq AML populations. For biological characterization of the re-annotated leukemic cell types, single-cell enrichment scores of hallmark gene sets, as well as custom gene sets from Ng et al. (LSC$^+$ AML fractions)[17] and Xie et al. (S1PR3 overexpression in LT-HSCs)[25], were calculated using AUCell[58]. Cell cycle status was determined using the original annotations from van Galen et al.[21], in which cell cycle scoring and classification were performed. Shannon diversity of single-cell transcriptomes was calculated from raw count data using the Python package 'skbio' after downsampling each cell to 1,000 unique molecular identifers (UMIs).

Gene expression deconvolution. Raw gene expression counts from 13,653 cells belonging to any of seven leukemic populations (LSPC-Quiescent, LSPC-Primed, LSPC-Cycle, GMP-like, ProMono-like, Mono-like and cDC-like) or seven non-leukemic immune populations (T, CTL, NK, B, Plasma and wild-type Monocyte and cDCs) were used as input for signature matrix generation with CIBERSORTx[28]. Default settings were used, with the exception of the minimum expression parameter, which was set to 0.25. Deconvolution was performed on TPM-normalized bulk RNA-seq data using S-mode batch correction and Absolute mode. Due to differences in S-mode batch correction performance between the CIBERSORTx web portal and the CIBERSORTx Docker image, we exclusively used the web portal for our analyses. For downstream analysis, the abundance of the seven leukemic populations were normalized to a sum of 1, wherein the score for each population represents the estimated proportion of all leukemic cells. For bulk RNA-seq samples composed entirely of leukemic blasts (cell lines or sorted primary samples), a second signature matrix with seven leukemic populations and no immune populations was used. Benchmarking analysis and deconvolution with other approaches is outlined in Supplementary Note 1.

Clinical AML datasets. Publicly available clinical RNA-seq datasets used for deconvolution analysis are outlined in Supplementary Table 4. All gene expression data were subject to TPM normalization before deconvolution with CIBERSORTx. Clinical and mutational data were extracted from the GDC Data Portal for TCGA (https://portal.gdc.cancer.gov/projects/TCGA-LAML) and from supplemental materials in Tyner et al.[33] for BEAT-AML. For the BEAT-AML cohort, we focused exclusively on pre-treatment samples collected at AML diagnosis ($n = 281$). For the Leucegene cohort, clinical and mutational annotations were extracted from supplemental materials of 13 publications (Supplementary Table 4) and linked based on sample ID.

Mapping and clustering AML hierarchy composition. To map patients with AML based on the composition of their leukemic hierarchies, only deconvolution results pertaining to leukemic AML populations were used. In these cases, estimated abundances from leukemic populations were normalized to 1, such that the value associated with each cell type represents the proportion of total leukemic blasts that it constitutes. Patients from TCGA, BEAT-AML and Leucegene were used. Principal component analysis (PCA) was performed on the normalized leukemic cell type compositions of these patients. Neighbors were calculated using Euclidean distance with a local neighborhood size of 30. To determine the optimal number of clusters, the package 'NbClust' was used to calculate 30 clustering metrics for values of k from 2 to 10, and $k = 4$ was selected by majority rule. Leiden clustering was subsequently performed at a resolution of 0.4 to obtain four hierarchy clusters. Cluster assignments, hierarchy compositions and genomic annotations for TCGA, BEAT-AML and Leucegene are included in Supplementary Table 5.

To project hierarchies onto the reference map from the three AML cohorts (TCGA, BEAT-AML and Leucegene), normalized leukemic cell type abundances from the query dataset were combined with the reference dataset, and batch correction was applied using ComBat[61]. After this, the ingest function from scanpy was used to project the batch-corrected query dataset onto the PCs of the batch-corrected reference dataset and assign cluster labels.

Hierarchy classification of microarray cohorts. To enumerate patient hierarchy composition from microarray data, we performed CIBERSORTx deconvolution on SCAN-normalized microarray data[62] (see Supplemental Note 1 for benchmarking between microarray normalizations).

Given the comparatively low accuracy of deconvolution with microarray data, we employed three approaches to classify microarray samples from the query cohort as Primitive, Intermediate, Mature or GMP. The first approach involved direct projection of query data onto the hierarchy map from the reference cohorts as described in the previous section. A second approach involved projection of query data onto deconvoluted microarray data from TCGA, for which cluster assignments were available from bulk RNA-seq for 158 overlapping samples. Deconvolution results from the query cohort were batch corrected with the TCGA reference data using ComBat, and cluster labels were projected using the ingest function from scanpy. As a third approach, cluster classifiers were trained from the microarray expression data from the TCGA cohort, using the top ten marker genes for each cluster based on a Wilcoxon test. Microarray expression data from the query cohort were batch corrected with the TCGA reference data using ComBat, and L1-penalty logistic regression (L1-LR), L2-penalty logistic regression (L2-LR), support vector machine (SVM), k-nearest neighbor (KNN) and random forest (RF) classifiers were subsequently trained from these marker genes with hyperparameter tuning performed through a grid search with ten-fold cross-validation for each model. For cluster assignment based on gene expression data in the query

cohort, the majority vote of all five models was used. To obtain the final cluster assignments, the predictions from all three approaches were combined. Within the GSE6891 cohort[35], most samples (372/537, 69%) were assigned to a cluster unanimously by all three approaches, whereas the remaining samples (165/537, 31%) had conflicting assignments between approaches. These ambiguous samples were primarily positioned at the boundary between the Intermediate cluster and other clusters (Primitive, Mature and GMP). These ambiguous cases were reclassified through a KNN approach trained on the 372 high-confidence samples using the ingest function from scanpy to obtain the final classifications.

Classification benchmarking. To benchmark classification performance for biological phenotypes (for example, LSC activity, Relapse, Adverse Cytogenetics) as outlined in Extended Data Fig. 3a, a repeated nested cross-validation approach was employed to obtain high-confidence estimates of model performance. Samples were subject to a five-fold split (outer cross-validation), wherein each 20% split was used as a held-out set, with the remaining 80% used as a training set. Within each split, LR or RF classifiers were trained with hyperparameter optimization performed through a grid search with five-fold internal cross-validation, leading to a total of five separate AUC values. The mean of these five values was calculated as a summary AUC metric, and this nested cross-validation process was repeated for a total of 1,000 iterations, with samples being randomly shuffled between each iteration. Together, this produces a distribution of 1,000 summary AUC metrics, enabling statistical comparisons of model performance across different sets of features. Comparisons were performed through Wilcoxon signed-rank tests, with AUC metrics paired on each iteration, for which the cross-validations splits were the same.

Clinical and morphological correlates. For associations of leukemic and immune cell type abundance with clinical features in TCGA, Pearson correlations were calculated between the absolute abundance of each leukemic and immune population with each clinical feature. Only correlations with uncorrected $P < 0.05$ were retained. To characterize the hierarchy compositions of distinct FAB morphological classes, we visualized 378 AMLs from TCGA, BEAT-AML and Leucegene for which FAB annotations were available. Samples labeled as M5 (without specifying M5A or M5B) were excluded; samples labeled as M6 ($n = 4$) or M7 ($n = 4$) were also excluded due to sample size.

Survival analysis. OS was defined as the time from diagnosis until death or last follow-up. Differences in OS between hierarchy classes in each cohort were evaluated using Mantel–Cox log-rank tests using the R package 'survival', and survival curves for each cluster were visualized using Kaplan–Meier plots using the R package 'survminer'. Univariate and pairwise HRs for each cluster were derived from Cox proportional hazards regression. For combined HRs, individual patient data were pooled, and stratified Cox regression was performed with the patient cohort (TCGA, BEAT-AML and GSE6891) set as the stratifier. For multivariate survival meta-analysis, we included covariates that were available across all three cohorts (Cytogenetic Risk, Age, WBC, *NPM1* status and *FLT3*-ITD status) and performed multivariate stratified Cox regression, with patient cohort as the stratifier. We determined whether hierarchy information (for example, Cluster, PC1 or PC2) adds value in addition to baseline covariates through a likelihood ratio test to assess model improvement after incorporating hierarchy information.

To benchmark the prognostic value of the new LSPC annotation compared to the prior HSC-like/Prog-like annotation, repeated nested cross-validation was performed as described in the classification benchmarking section, using stratified Cox regression to predict OS within the TCGA and BEAT-AML cohorts from the abundances of primitive AML populations. L1 (LASSO) or L2 (Ridge) penalties were applied using partial likelihood deviance as the loss function, and five-fold internal cross-validation was performed to identify the optimal lambda value. Model performance was estimated through the mean likelihood ratio test statistic across the five outer cross-validation splits. This was repeated for 1,000 iterations.

To identify the leukemic cell types associated with survival, we performed stratified Cox regression on the TCGA and BEAT-AML cohorts using an L1 (LASSO) penalty with partial likelihood deviance loss. Because this process was not repeated over multiple iterations, this score was trained on the full dataset, and leave-one-out cross-validation was employed to determine the optimal lambda value. Coefficients for each leukemic population were subsequently used to determine feature importance.

For gene set enrichment analysis (GSEA) of genes ranked by their association with OS, only genes that were detected in TCGA and BEAT-AML were evaluated. Univariate Wald tests were performed to evaluate the association with log(TPM + 1)-normalized expression of each gene with OS in each of the TCGA and BEAT-AML cohorts. The Wald test statistics from each cohort were averaged for each gene and used as a rank statistic for GSEA analysis using Stem versus GMP signatures from normal hematopoiesis (top 200 upregulated and top 200 downregulated genes from limma DE analysis of healthy LT-HSC versus GMP sorted fractions from umbilical cord blood[25]) and malignant hematopoiesis (top 200 correlated and top 200 anti-correlated genes with PC1).

Calculation of prognostic AML scores. We calculated LSC17 and other prognostic AML scores using log(TPM + 1)-normalized expression values from TCGA, BEAT-AML and Leucegene and normalized microarray expression values from Lee et al.[45]. In cases where specific genes were missing from a dataset, we calculated the score with those genes removed. To ensure high concordance of these partial scores, we calculated the correlation between each partial score and the full score in other datasets to ensure high concordance: we observed a median correlation of $r = 0.99$ between partial and complete scores, with the lowest correlation being $r = 0.95$. Patients were classified into high and low groups for each score based on a median split within each cohort.

Mutation analysis. Cytogenetic and driver mutation annotations from TCGA, BEAT-AML and Leucegene were used to correlate hierarchy composition with genomic profiles. Mutation combinations between driver mutations were identified, and all combinations present in at least five patients were retained and visualized along hierarchy axes PC1 and PC2 using the R package 'ggridges'. Due to missing variant allele frequency (VAF) information in an appreciable subset of mutation calls from genomic annotations, samples were considered mutated as long as the mutation was called. This analysis was repeated exclusively using mutation calls where VAF > 0.25 to confirm that the observed trends remained the same.

scRNA-seq classification in relapsed AMLs. scRNA-seq profiles of blast cells from eight relapsed AML patient samples were obtained from Abbas et al.[43]. To project these cells onto our cell types defined from diagnostic AML samples from van Galen et al.[21], we used a transfer learning approach implemented through the scANVI[63] and scArches[64] packages. First, semi-supervised dimensionality reduction was performed with scANVI using unnormalized scRNA-seq data from diagnostic AML samples filtered for 3,000 variable genes with malignant cell type annotations and patient batch as a covariate. For scANVI, an initial unsupervised neural network was trained over 500 epochs with patience for early stopping set to 10 epochs, followed by a semi-supervised neural network incorporating cell type annotations that was trained over 200 epochs with a patience of 10 epochs. Transfer learning with scArches was subsequently applied to update the scANVI neural network using scRNA-seq data from the relapsed AML samples, and training was performed over 500 epochs with a patience of 10 epochs. The updated model was subsequently applied to both diagnostic and relapsed AML samples to generate a shared latent representation, and this latent representation was used for further dimensionality reduction with UMAP. For visualization purposes, the diagnostic and relapsed AML data were each downsampled to 10,000 cells.

Benchmarking relapse phenotypes. To benchmark the changes in cellular composition from diagnosis to relapse, we obtained 12,441 gene sets from the MSigDB corresponding to hallmark gene sets ($n = 48$), oncogenic signatures ($n = 182$), computationally derived signatures ($n = 667$), chemical and genetic perturbations (CGP) from prior literature ($n = 2{,}112$), Gene Ontology Biological Pathways ($n = 4{,}400$) and previously published immune signatures ($n = 5{,}024$). The relative expression of each signature was scored in each individual diagnosis and relapse sample ($n = 88$) through gene set variation analysis (GSVA) to generate a single-sample enrichment score for each signature. GSVA enrichment scores for each of the MSigDB signatures, alongside the inferred abundance of each leukemic cell type, were compared between diagnosis and relapse through paired *t*-tests based on the significance of their enrichment at relapse (absolute value of the $\log_{10}(P)$). Each signature was ranked, and relapse enrichment of each leukemic subpopulation was subsequently compared against relapse enrichment of each of the MSigDB signatures. Non-parametric Wilcoxon signed-rank tests were also performed for each signature to ensure comparable results.

To benchmark classification performance from using hierarchy information to discriminate between diagnosis/relapse samples, we performed repeated nested cross-validation as outlined in Extended Data Fig. 3a and described in the 'Classification benchmarking' section. This was performed first on individual samples without paired information ($n = 88$) or on paired patient samples, wherein the changes in cell composition were provided, and the classifier was required to identify whether that change in composition corresponded to a transition from diagnosis to relapse ($n = 44$) or from relapse to diagnosis ($n = 44$).

Clonal evolution analysis. Clonal analysis of paired diagnosis and relapse samples from four independent cohorts was performed using annotated single-nucleotide variant calls derived from targeted sequencing[40], whole-exome sequencing[42] or whole-genome sequencing[14,41] data. Genetic clones were identified using PhyloWGS[65], selecting the phylogenetic tree with the highest log likelihood (LLH) value. In cases of tied LLH values, the simplest tree with the most representative branching patterns among the top candidates was manually selected. Graphical representations of evolution of genetic clones were depicted using the R package 'Fishplot'[66], whereas representations of changes in cell type composition were depicted using the R package 'ggAlluvial'.

Association with drug sensitivity. Ex vivo drug response in AML patient samples from BEAT-AML[33] and Lee et al.[45] was measured through the AUC metric, wherein a low AUC corresponds to sensitivity and a high AUC corresponds to resistance. AUC values were scaled and multiplied by −1 to represent sensitivity in each

treatment condition. Pearson correlation was used to measure association between cell type abundance and drug sensitivities. Associations were depicted using the R package 'corrplot', and drug sensitivity volcano plots were generated using the R package 'EnhancedVolcano'.

Unsupervised clustering by drug sensitivity. Unsupervised clustering of 30 AML patient samples from Lee et al.[45] was performed on the basis of their ex vivo drug sensitivity values to 159 drugs. AUC values for each patient were scaled for each drug, and dimensionality reduction with PCA was applied and neighbors were calculated with a local neighborhood size of 5. Leiden clustering with a resolution of 0.3 was used to determine the final clusters. DE analysis between drug response clusters was performed using limma[67] from normalized microarray expression values obtained from GSE107465. The moderated *t*-statistic for each gene was subsequently used as the rank statistic for GSEA analysis using Stem versus Mature Myeloid signatures from normal hematopoiesis (top 200 upregulated and top 200 downregulated genes from limma DE analysis of healthy LT-HSC versus Granulocyte/Monocyte sorted fractions from umbilical cord blood[25]) and malignant hematopoiesis (top 200 correlated and top 200 anti-correlated genes with PC2).

Derivation of PC2-based gene expression scores. For derivation of PC2-based gene expression scores, we used log(CPM + 1)-normalized gene expression values, which we found to improve model performance during training. To derive the LinClass-7 score, logCPM-normalized expression of 16 genes from the LSC17 assay were used as input features for LASSO regression: DNMT3B, GPR56, NGFRAP1, CD34, DPYSL3, SOCS2, MMRN1, KIAA0125, EMP1, NYNRIN, LAPTM4B, CDK6, AKR1C3, ZBTB46, CPXM1 and ARHGAP22. The 17th gene, C19orf77, was excluded due to a lack of expression data in the Leucegene cohort. LASSO regression was performed on negative PC2 (high in Primitive and low in Mature) with leave-one-out cross-validation using the LassoCV function from scikit-learn with a path length of 0.1 to determine the optimal lambda value. Patients from TCGA and Leucegene were combined into a training set, and patients from BEAT-AML were used as a validation set to evaluate the strength of the association between LinClass-7 and PC2.

To train the PC2-34 score, we started with the top 50 correlated and top 50 anti-correlated genes with PC2, based on the average Pearson correlation between the TCGA and Leucegene cohorts. LASSO regression was performed on PC2 with leave-one-out cross-validation to determine the lambda value corresponding to the lowest mean square prediction error. To further reduce the number of features in the model, the largest lambda within one standard error of the lowest root mean square prediction error (RMSE) was selected instead of the lambda directly corresponding to the lowest RMSE. This resulted in a 34-gene score (PC2-34), which was then evaluated in the BEAT-AML validation set.

The LinClass-7 and PC2-34 scores were calculated by multiplying the expression of each constituent gene by its specific weight and calculating the total sum. Constituent genes and weights for both scores are included in Supplementary Table 9.

Literature screen for drug-treated RNA-seq datasets. To identify RNA-seq datasets collected from AML samples before and after drug treatment, applying the search terms 'Acute Myeloid Leukemia' and 'AML' with the 'Homo Sapien' and 'RNA-sequencing' flags on GEO and ArrayExpress, we identified 95 datasets posted before 17 June 2021. From these, 53 were drug studies that met the inclusion criteria of human AML samples with available RNA-seq data collected before and after drug treatment. Datasets with only DE results or bigWig files were excluded. Datasets with fewer than three samples in each treatment group were also excluded, resulting in a total of 47 datasets included in the final analysis. Detailed information on included datasets is available in Supplementary Table 13. Each dataset was processed and underwent TPM normalization and deconvolution with CIBERSORTx using a signature matrix of seven leukemic cell types (LSPC-Quiescent, LSPC-Primed, LSPC-Cycle, GMP-like, ProMono-like, Mono-like and cDC-like). For quality control among cell line samples, the deconvolution correlation values from each sample across every dataset were compared, and the Jenks natural breaks algorithm was employed to identify cutoffs demarcating low, medium and high correlation bins. Cell line samples classified as 'low-correlation', with a correlation value below 0.437, were excluded from further analysis, leaving 43 datasets spanning 153 treatment conditions.

Quantifying hierarchy composition changes after drug treatment. Relative changes in cell type abundance in each treatment condition were evaluated using Wilcoxon rank-sum tests for technical replicates or Wilcoxon signed-rank tests for biological replicates with paired treatment conditions. For dimensionality reduction with UMAP, we focused exclusively on changes in cell type abundance where $P < 0.05$ to emphasize the key changes in cell type composition induced by each drug, resulting in 125 treatment conditions spanning 38 studies. Absolute log *P* values were used to represent the magnitude of the shift in cell type abundance, and cell type changes where $P > 0.05$ were assigned a magnitude of 0. We then applied UMAP with the following parameters (n_neighbors = 13, min_dist = 0.05) to generate the final representation, and Leiden clustering was applied with a resolution of 1. We note that UMAP was selected for visualization rather than PCA, because, despite the low number of features, PCA did not adequately capture the variability between clusters. Cell type composition changes for treatment conditions were visualized with the R package 'ComplexHeatmap' and are also included in Supplementary Table 14.

Fedratinib and CC90009—hierarchy classification. Using normalized leukemic cell type composition data for 46 patient samples used for in vivo fedratinib or CC-90009 treatment, dimensionality reduction was performed and clustering was assigned using the Leiden algorithm with a resolution of 0.7, yielding three clusters: Primitive, Mature and Intermediate/GMP. Owing to an under-representation of engrafting samples with GMP hierarchies, we did not attempt to divide the Intermediate/GMP cluster into Intermediate and GMP groups. Samples were subsequently projected on the reference map for visualization and confirmation of cluster assignments.

Fedratinib and CC90009—response classification. Patient samples were classified into response categories by comparing the relative reduction (RR) of AML engraftment in drug-treated mice versus vehicle-treated mice, as per Galkin et al.[68]. RR was calculated as: ((mean % engraftment in control mice) − (mean % engraftment in drug treated mice)) / (mean % engraftment in control mice). Patient samples were classified as Responders if RR in the injected femur (right femur) was >50%; classified as Partial Responders if we observed 20–50% RR in the right femur or >20% in the non-injected femur (bone marrow) only; and classified as Non-Responders if there was no statistically significant difference in engraftment levels between control-treated and drug-treated mice or if RR was <20% in both right femur and bone marrow.

Fedratinib and CC90009—imputation of *NPM1* mutation status. Patient samples from PMH in Toronto were classified as *NPM1*-mutant (*NPM1*-mut) or *NPM1*-wild-type (*NPM1*-wt) based on clinical sequencing results. For patient samples where targeted sequencing data were unavailable, we predicted *NPM1* status using gene-expression-based classifiers. First, log(TPM + 1)-normalized RNA-seq data from PMH samples and the three reference cohorts (TCGA, BEAT-AML and Leucegene) were combined and batch corrected using ComBat[61]. Two groups of classifiers were trained: the first group comprised LR, SVM and RF classifiers trained on *NPM1* status from the reference cohorts; the second group comprised LR, SVM and RF classifiers trained on *NPM1* status from 46 PMH samples for which *NPM1* genotype was available (out of 88 total samples).

To select features for the first group of classifiers, DE analysis was performed using DESeq2 (ref. [69]) with AML cohort and patient hierarchy cluster as covariates, and DE genes were selected using an absolute $\log_2$ fold change threshold >1 and FDR < 0.01. The top 50 *NPM1*-mut and top 50 *NPM1*-wt genes were then used to train LR, SVM and RF classifiers. Hyperparameter tuning was performed through a grid search with ten-fold cross-validation for each model. The final group-1 classifiers were subsequently evaluated on the 46 PMH samples with *NPM1* genotype information. To further account for batch effects between reference cohorts and the PMH cohort, optimal classification thresholds were identified based on the receiver operating characteristic curves for the 46 genotyped PMH samples, yielding final classification accuracies of 0.87 (LR), 0.89 (SVM) and 0.93 (RF). These thresholds were subsequently used for prediction of the remaining 42 PMH samples for which *NPM1* status was missing.

The second group of classifiers was trained directly on the 46 PMH samples for which *NPM1* genotype was available. Given that a subset of the PMH samples did not have raw counts available for DESeq2, we identified *NPM1* mutation-specific marker genes through a Wilcoxon rank-sum test with the log(TPM + 1)-normalized expression. Genes with significant differences in expression between *NPM1*-mut and *NPM1*-wt PMH samples at FDR < 0.05 were subsequently filtered to keep genes that were also significantly differentially expressed in the reference cohorts at an absolute $\log_2$ fold change threshold >1 and FDR < 0.01, leaving 63 high-confidence genes. LR, SVM and RF classifiers were subsequently trained from these 63 genes, and hyperparameter tuning was performed through a grid search with ten-fold cross-validation for each model. Classification performance was evaluated by ten-fold nested cross-validation with 100 repeats, yielding median accuracies of 0.98 (LR), 1.00 (SVM) and 0.98 (RF). These models were subsequently used to predict NPM1 status in the remaining 42 PMH samples for whom *NPM1* status was missing.

For the final prediction of *NPM1* status, the classifier with the lowest accuracy (group-1 LR) was excluded and the five remaining classifiers voted on the *NPM1* genotype, with the majority vote being assigned as the final prediction. Together, this resulted in high-confidence predictions of *NPM1* genotypes for patients for whom targeted sequencing data were not available. Imputed *NPM1* genotypes for each of these patients are presented in Fig. 6 alongside *NPM1* genotypes obtained from sequencing.

Fedratinib and CC90009 combination treatment. All animal experiments were performed in accordance with guidelines approved by the UHN animal care committee. NOD.SCID mice were bred and housed in a controlled environment designated exclusively for immunocompromised mice within the UHN animal care facility with a 12-hour light/dark cycle, including a 30-minute transition, a room temperature of 21–23 °C, humidity of 30–60% and constant access to

dry laboratory food and water. Ten-week-old NOD.SCID mice were irradiated (225 cGy) and pre-treated with anti-CD122 antibody (200 µg per mouse) 24 hours before transplantation. Viably frozen mononucleated cells from patients with AML were thawed, counted and intrafemorally injected at the dose of 5 million cells per mouse. At day 21 after transplantation, treatment of either CC-90009 or fedratinib alone with vehicle, or in combination, was initiated twice a day for 2 weeks. CC-90009 was given by intraperitoneal injections at the dose of 2.5 mg kg^{-1}, and fedratinib was dissolved in 0.5% methylcellulose and orally gavaged at 60 mg kg^{-1}. After treatment, levels of AML engraftment were assessed to determine the efficacy of drug treatment against the disease in the mice. Cells collected from the injected right femur and non-injected bone marrow of each individual mouse were stained with human-specific antibodies and evaluated by flow cytometry. Antibodies used for assessment of human AML engraftment include: APC-anti-CD45 (1:100), V450-anti-CD19 (1:100), APC-Cy7-anti-CD34 (1:300), FITC-anti-CD15 (1:100), PE-Cy5-anti-CD33 (1:100) and PE-anti-CD14 (1:100).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Processed and raw RNA-seq data generated in this study are available in the GEO under Superseries GSE199336. Citations and links of re-analyzed data from all clinical datasets analyzed in this study are provided in Supplementary Table 4. Citations and links of re-analyzed data from the literature screen are provided in Supplementary Table 13. Pathways and signatures used for relapse benchmarking were obtained from the MSigDB (https://www.gsea-msigdb.org/gsea/msigdb/). All deconvolution results are available with the analysis code on GitHub (https://github.com/andygxzeng/AMLHierarchies).

Code availability

Analysis notebooks for the main figures, as well as instructions for applying AML deconvolution, are available at: https://github.com/andygxzeng/AMLHierarchies.

References

54. Lun, A. T. L., Bach, K. & Marioni, J. C. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. *Genome Biol.* **17**, 75 (2016).
55. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
56. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).
57. Tan, Y. & Cahan, P. SingleCellNet: a computational tool to classify single cell RNA-seq data across platforms and across species. *Cell Syst.* **9**, 207–213 (2019).
58. Aibar, S. et al. SCENIC: single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
59. Van de Sande, B. et al. A scalable SCENIC workflow for single-cell gene regulatory network analysis. *Nat. Protoc.* **15**, 2247–2276 (2020).
60. Lambert, S. A. et al. The human transcription factors. *Cell* **172**, 650–665 (2018).
61. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2006).
62. Piccolo, S. R. et al. A single-sample microarray normalization method to facilitate personalized-medicine workflows. *Genomics* **100**, 337–344 (2012).
63. Xu, C. et al. Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol. Syst. Biol.* **17**, e9620 (2021).
64. Lotfollahi, M. et al. Mapping single-cell data to reference atlases by transfer learning. *Nat. Biotechnol.* **40**, 121–130 (2022).
65. Deshwar, A. G. et al. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. *Genome Biol.* **16**, 35 (2015).
66. Miller, C. A. et al. Visualizing tumor evolution with the fishplot package for R. *BMC Genomics* **17**, 880 (2016).
67. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
68. Galkin, O. et al. SIRPαFc treatment targets human acute myeloid leukemia stem cells. *Haematologica* **106**, 279–283 (2021).
69. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

Acknowledgements

We thank C. Jones, A. Tikhonova, N. Iscove, S. Chan, S. Abelson, B. Haibe-Kains, M. Anders and all members of the Dick Laboratory for valuable feedback on the manuscript. We thank P. Valk and the HOVON-SAKK trial group for providing the clinical annotations associated with GSE6891. A.G.X.Z. is supported by a CIHR Vanier scholarship. J.E.D. is supported by funds from the Ontario Institute for Cancer Research through funding provided by the Government of Ontario; the Canadian Institutes for Health Research (RN380110 - 409786); the International Development Research Centre Ottawa Canada; the Canadian Cancer Society (grant 703212 (end date 2019) and grant 706662 (end date 2025)); the Terry Fox New Frontiers Program Project Grant (Project 1106); University of Toronto's Medicine by Design initiative with funding from the Canada First Research Excellence Fund; a Canada Research Chair; Princess Margaret Cancer Centre; The Princess Margaret Cancer Foundation; and the Ontario Ministry of Health. The Leukemia Tissue Bank is supported by funds to M.D.M. from the Wendy Eisen Princess Margaret Hospital Foundation Account and the Philip Orsino Chair in Leukemia Research. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

A.G.X.Z. and J.E.D. conceived the project. A.G.X.Z. carried out the analysis. S.B. contributed to analyses of post-treatment changes in cell composition. L.J. performed fedratinib and CC-90009 combination treatments. A.M. and W.C.C. performed RNA-seq extraction and library preparation for LSC fractions and fedratinib samples, respectively. M.C.-S.-Y. and V.V. performed alignment and pre-processing for LSC fractions and fredratinib RNA-seq data. H.A.A., N.D. and A.F. provided scRNA-seq data and leukemic cell annotations for relapsed AMLs. P.v.G. provided bulk RNA-seq data for AML samples. A.T. analyzed clinical flow data from the PMH Leukemia cohort. M.C., C.P. and H.D. provided GE and clinical data for the ALFA-0701 trial cohort. M.D.M. provided PMH AML samples. J.A.K. and M.D.M. provided clinical annotations for the PMH AML cohort. A.G.X.Z., J.A.K., J.C.Y.W. and J.E.D. interpreted the data. A.G.X.Z. and J.E.D. wrote the paper. J.A.K. and J.C.Y.W. revised the paper.

Competing interests

J.E.D. reports research funding from Celgene/Bristol Myers Squibb and advisory board/royalities from Trillium Therapeutics. All other authors declare no competing interests.

Additional information

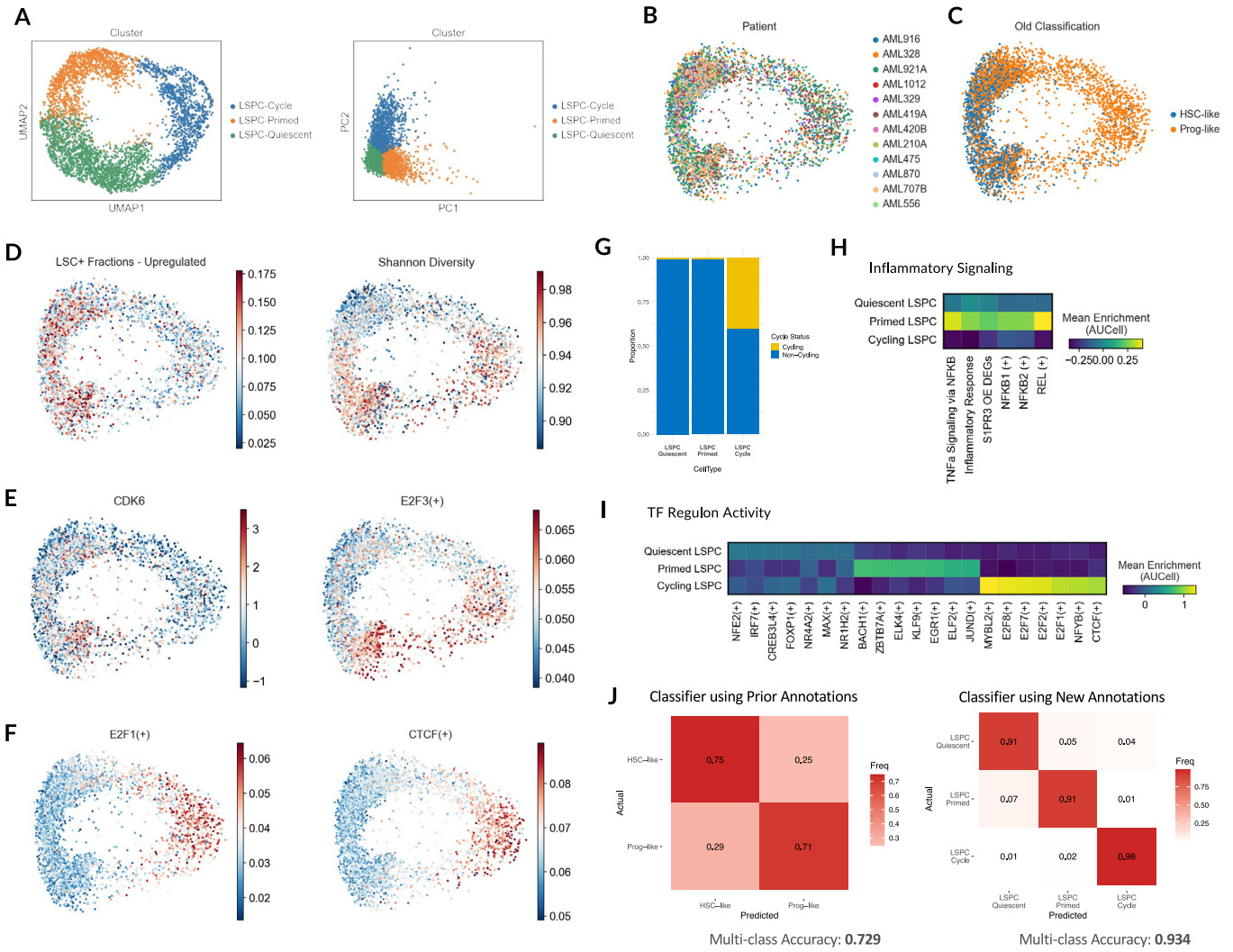
Extended data is available for this paper at https://doi.org/10.1038/s41591-022-01819-x.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-022-01819-x.

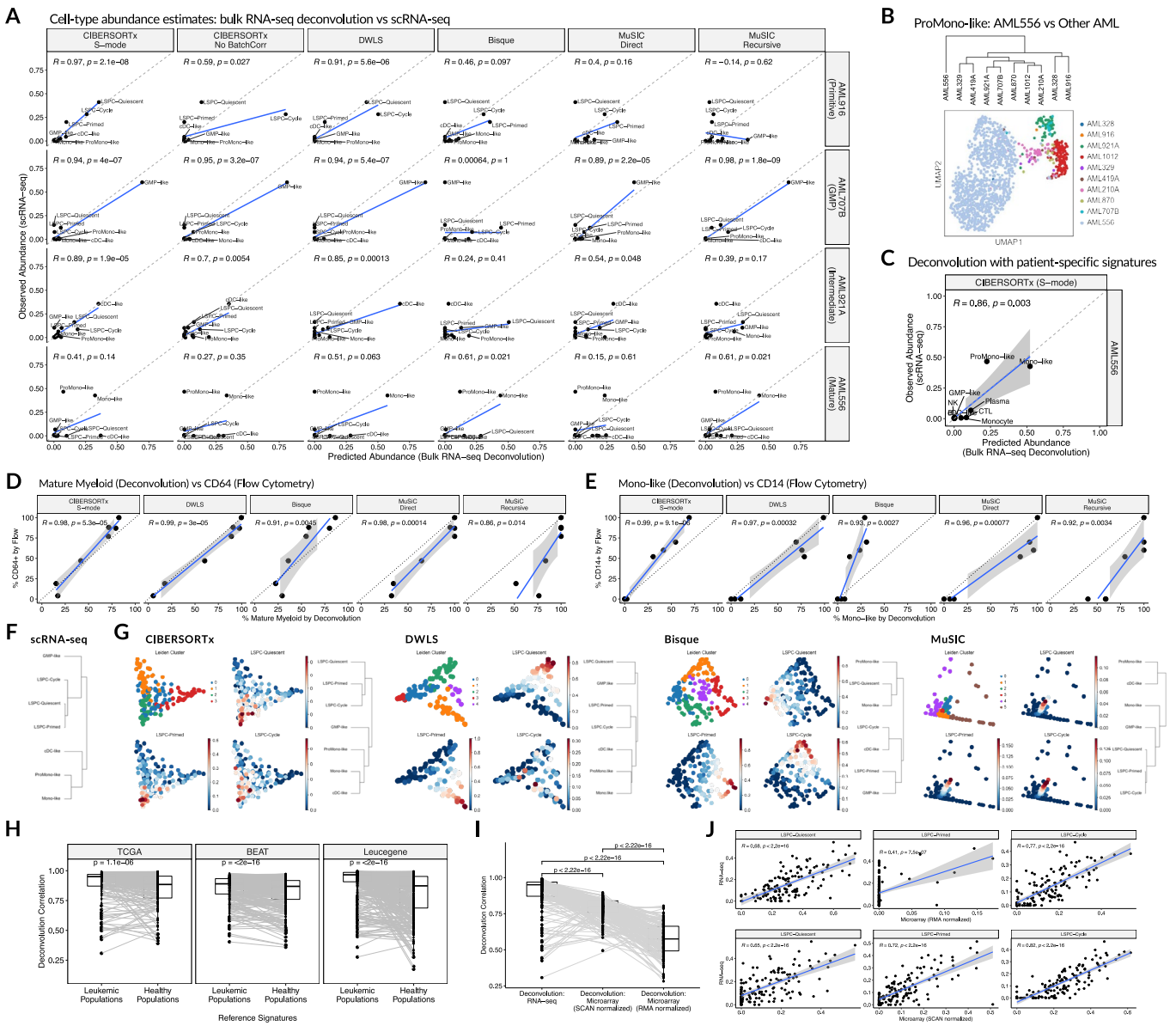
Correspondence and requests for materials should be addressed to John E. Dick.

Peer review information *Nature Medicine* thanks Craig Jordan, Simon Haas and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Editor recognition statement: primary handling editors were Javier Carmona and Michael Basson, in collaboration with the *Nature Medicine* team.

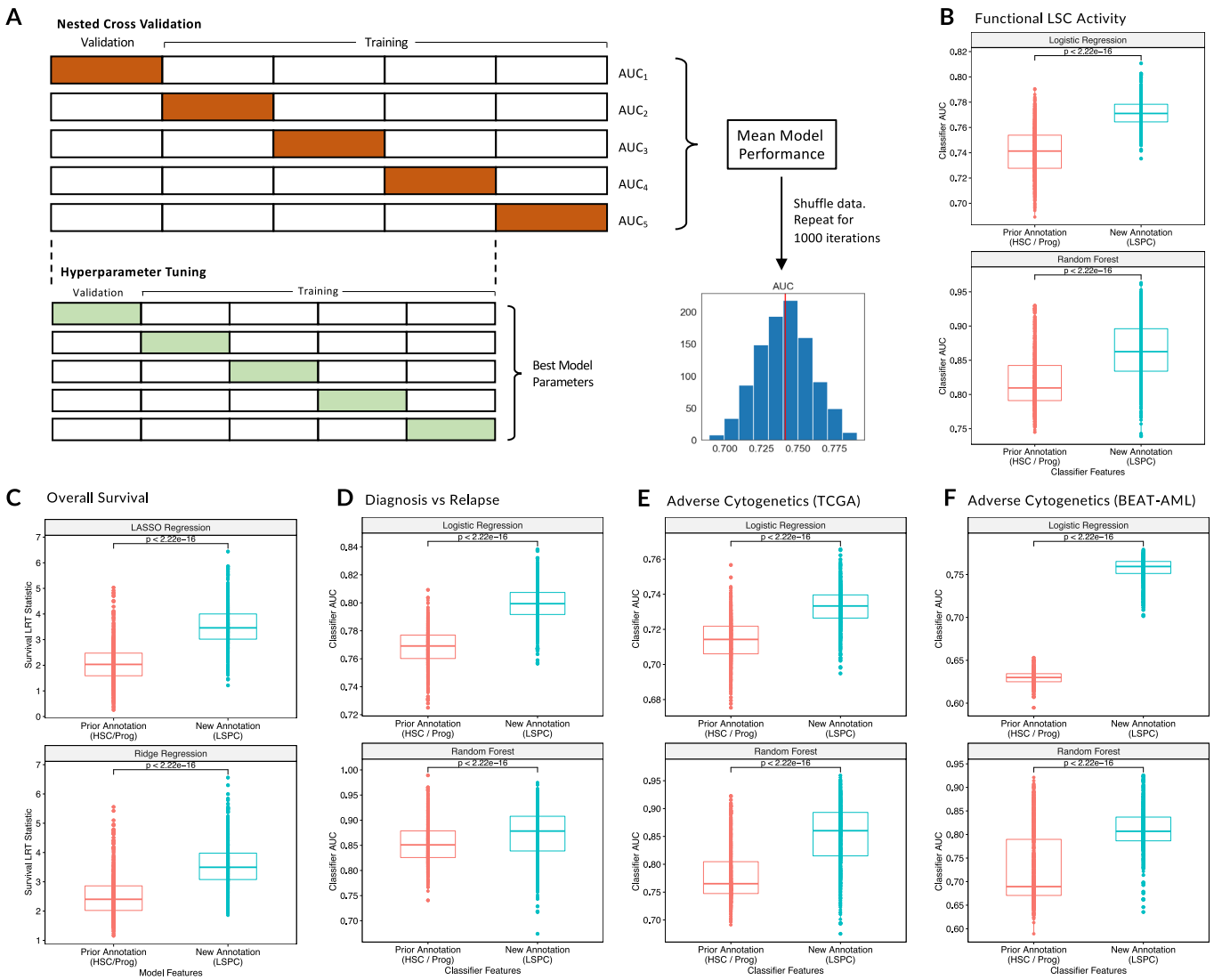
Reprints and permissions information is available at www.nature.com/reprints.



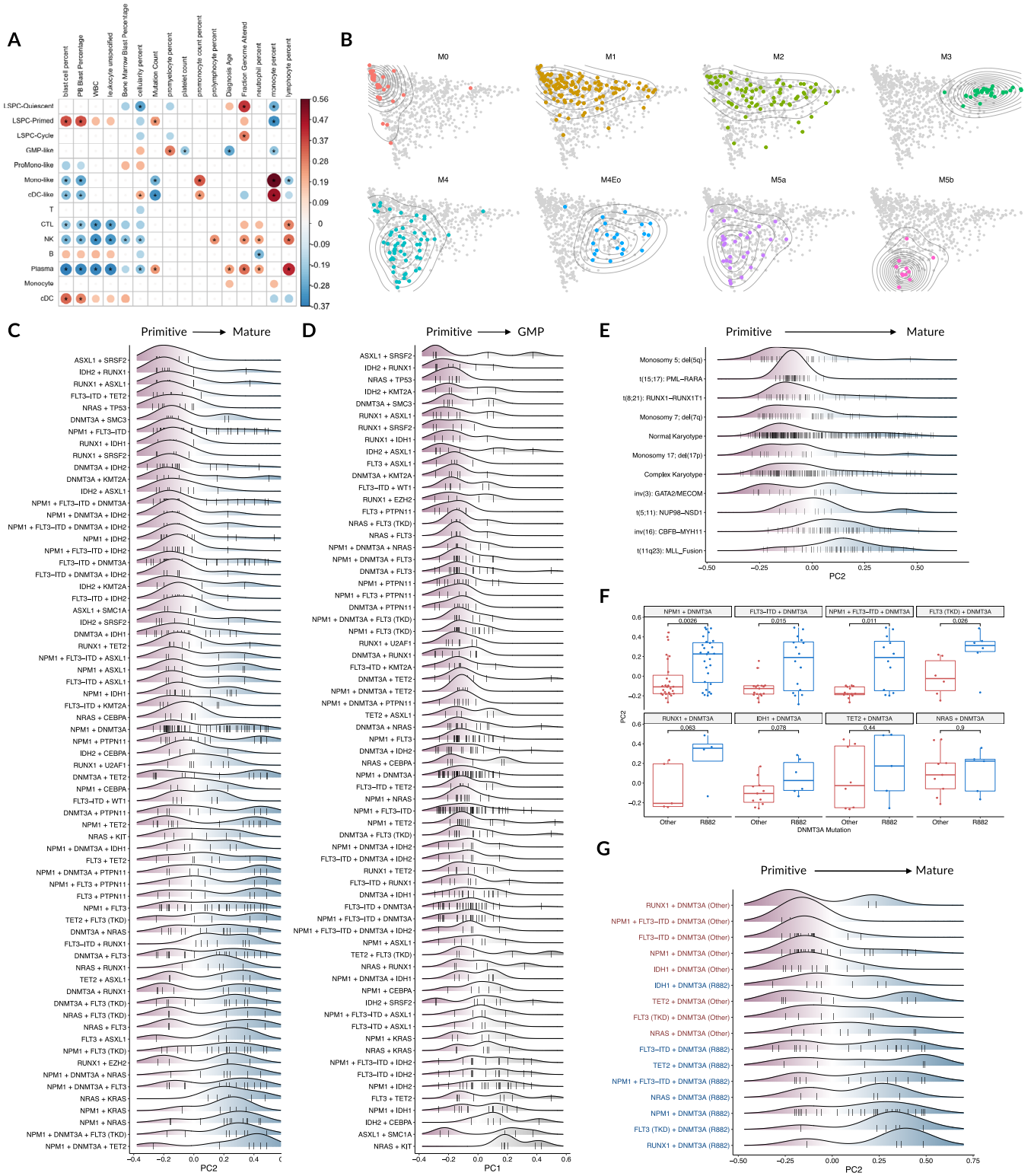
Extended Data Fig. 1 | Features of leukemia stem and progenitor cell populations from scRNA-seq. A) UMAP and PCA embeddings of 4163 AML LSPCs after feature weight derivation with the Self-Assembling Manifolds (SAM) algorithm. **B-F)** Diffusion map of re-annotated LSPC populations using SAM-derived feature weights, depicting: **B)** patient identity, **C)** prior cell type annotation, **D)** enrichment of LSC-specific genes from Ng *et al* (2016) and Shannon Diversity Index, **E)** scaled CDK6 expression and enrichment of the E2F3 regulon, and **F)** enrichment of E2F1 and CTCF regulons. **G)** Cell cycle status of Quiescent (n = 1855), Primed (n = 1240), and Cycling LSPCs (n = 1068). **H)** Enrichment of inflammatory signaling pathways and regulons in LSPCs. **I)** Transcription factor regulon activity, inferred through pySCENIC, specific to each LSPC. **J)** Normalized confusion matrix depicting classifier accuracy of prior and new cell type annotations for primitive AML cells. The classifier was built using SingleCellNet, an Ensemble classifier for scRNA-seq data trained from the top pairs of genes unique to each cell type. 800 cells from each cell type were used for training and 250 were used for validation.



Extended Data Fig. 2 | Benchmarking gene expression deconvolution approaches for AML. A) Pearson correlation between observed relative abundance of 14 cell types from scRNA-seq and predicted abundance of each cell type from deconvolution of matched bulk RNA-seq data, analyzed by patient. Gene expression deconvolution using CIBERSORTx (S-mode or No Batch Correction), DWLS, Bisque, or MuSIC (direct or recursive) were benchmarked across these samples. **B)** scRNA-seq of 1389 ProMono-like cells across 10 patients, demonstrating separation between AML556 and other patients. **C)** Pearson correlation depicting deconvolution performance of CIBERSORTx for AML556's bulk RNA-seq profile using patient-specific reference signatures derived from scRNA-seq data from AML556. **D-E)** Correlation between deconvolution and clinical flow cytometry for 7 AML patients from the Toronto PMH cohort. Deconvolution using scRNA-seq reference profiles was performed on RNA-seq data and matched with clinical flow cytometry data, both obtained from peripheral blood. **D)** Pearson correlation between total mature myeloid abundance (ProMono-like + Mono-like + cDC-like) from deconvolution with pan-myeloid surface marker CD64. **E)** Pearson correlation between mono-like abundance from deconvolution with monocyte-specific surface marker CD14. **F)** Dendrogram depicting associations between leukemic cell-types across scRNA-seq samples from 12 diagnostic AML patients. **G)** Observed associations between leukemic cell types from deconvolution analysis of 173 patients within the TCGA cohort, depicted for each deconvolution tool. MuSIC Direct was excluded due to multiple cell types having a detection rate of zero in bulk RNA-seq. **H-I)** Correlation between observed transcriptomic profiles and synthetic transcriptomic profiles reconstructed based on predicted cell-type abundance from CIBERSORTx. Higher correlation suggests greater deconvolution confidence. Box plots indicate the range of the central 50% of the data, with the central line marking the median. Whiskers extend from each box to 1.5*(interquartile range). Comparisons were performed through two-sided Wilcoxon signed-rank tests. These correlations are depicted for **H)** Deconvolution of 864 patients across three AML cohorts using reference signatures from leukemic populations compared to deconvolution with reference signatures from matched healthy populations, and **I)** RNA-seq compared to microarray from 158 matched TCGA patient samples. Prior to deconvolution, microarray data was normalized through either chip-based (RMA) or single-sample (SCAN) normalization approaches. **J)** Pearson correlation of estimated LSPC abundances between RNA-seq deconvolution and Microarray deconvolution, normalized with either RMA or SCAN, among 158 matched patient samples from the TCGA cohort.

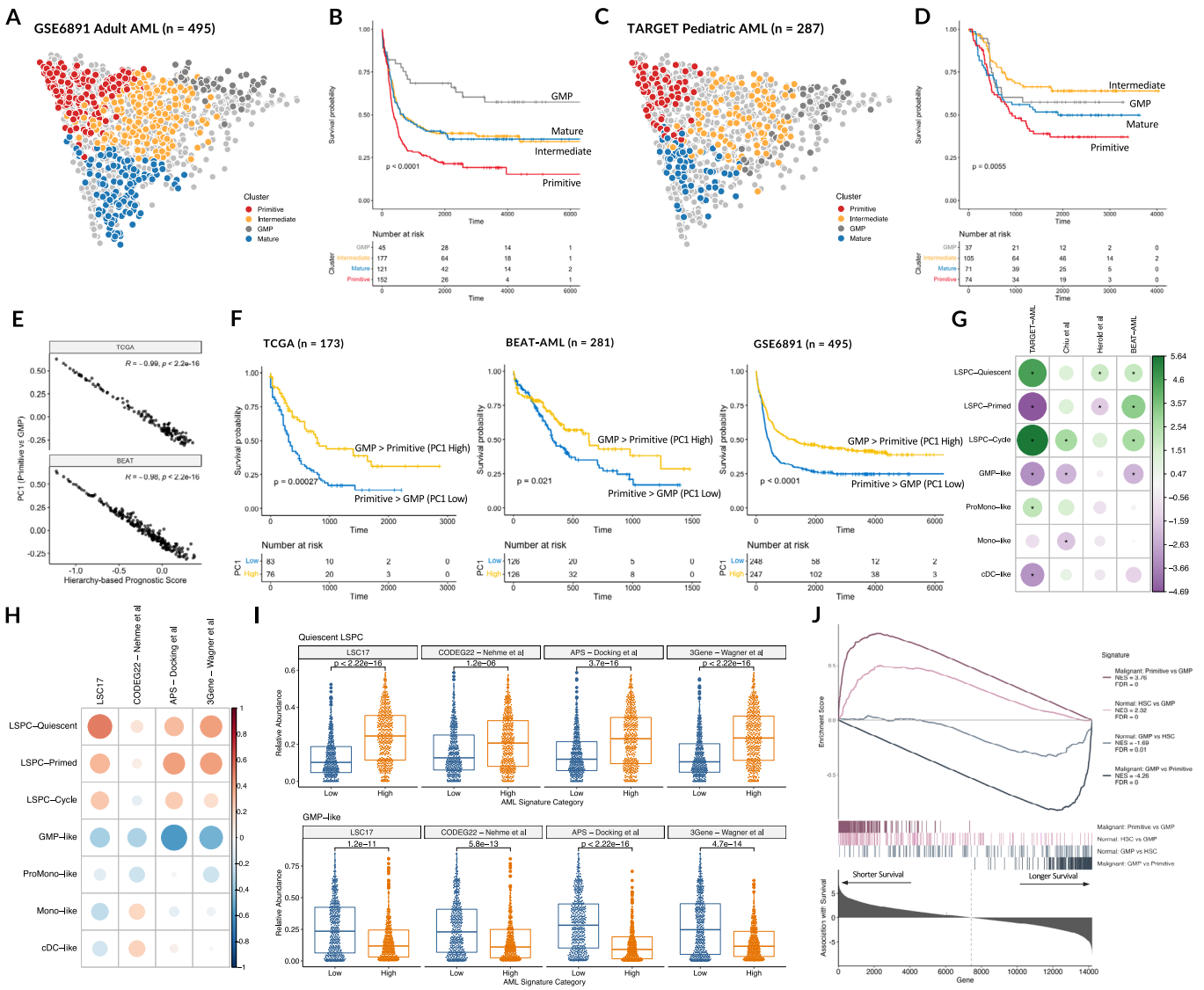


Extended Data Fig. 3 | Comparison of new and prior leukemia stem and progenitor cell annotations for discerning biological phenotypes. A) Workflow to compare prior (HSC-like and Prog-like) annotations and new (Quiescent LSPC, Primed LSPC, and Cycling LSPC) annotations with regard to their utility in predicting important biological phenotypes in AML. This was measured through the performance of logistic regression and random forest models trained on the relative abundance of these populations. Models were trained using nested cross-validation wherein samples were subject to a 5-fold split, in which 80% of samples (white boxes) were used to train a model and 20% of samples (orange boxes) were used to evaluate the model. Within each training set, hyperparameter optimization was performed by grid search with 5-fold internal cross validation. The model AUC from each outer cross-validation split was averaged to estimate overall classifier performance. This nested cross-validation process was repeated over 1000 iterations, with samples being shuffled between each iteration, to generate a distribution of AUC metrics. **B-F)** Model performance for predicting key biological and clinical phenotypes from either HSC-like and Prog-like abundance or Quiescent, Primed, and Cycling LSPC abundance. Performance metrics are paired by iteration, wherein sample order and cross validation splits were identical for each model. Box plots indicate the range of the central 50% of the data, with the central line marking the median. Whiskers extend from each box to 1.5*(interquartile range). Statistical significance was evaluated through a two-sided Wilcoxon signed-rank test. **B)** Prediction of functional LSC activity measured by leukemic engraftment from 72 LSC+ fractions and 38 LSC- fractions. **C)** Prediction of overall survival in the TCGA and BEAT-AML cohorts, evaluated through the likelihood ratio statistic from stratified cox regression (combined n = 454). In this case LASSO and Ridge regression was performed and these models were trained on splits of the TCGA and BEAT-AML cohorts, stratified by cohort. The repeated nested cross validation approach remained the same. **D)** Prediction of diagnosis vs relapse status from 44 relapsed and 44 diagnostic samples. **E)** Prediction of Adverse cytogenetic status in TCGA from 37 patients with Adverse cytogenetics and 131 patients with Intermediate or Favorable cytogenetics. **F)** Prediction of Adverse cytogenetic status in BEAT-AML from 53 patients with Adverse cytogenetics and 175 patients with Intermediate or Favorable cytogenetics.

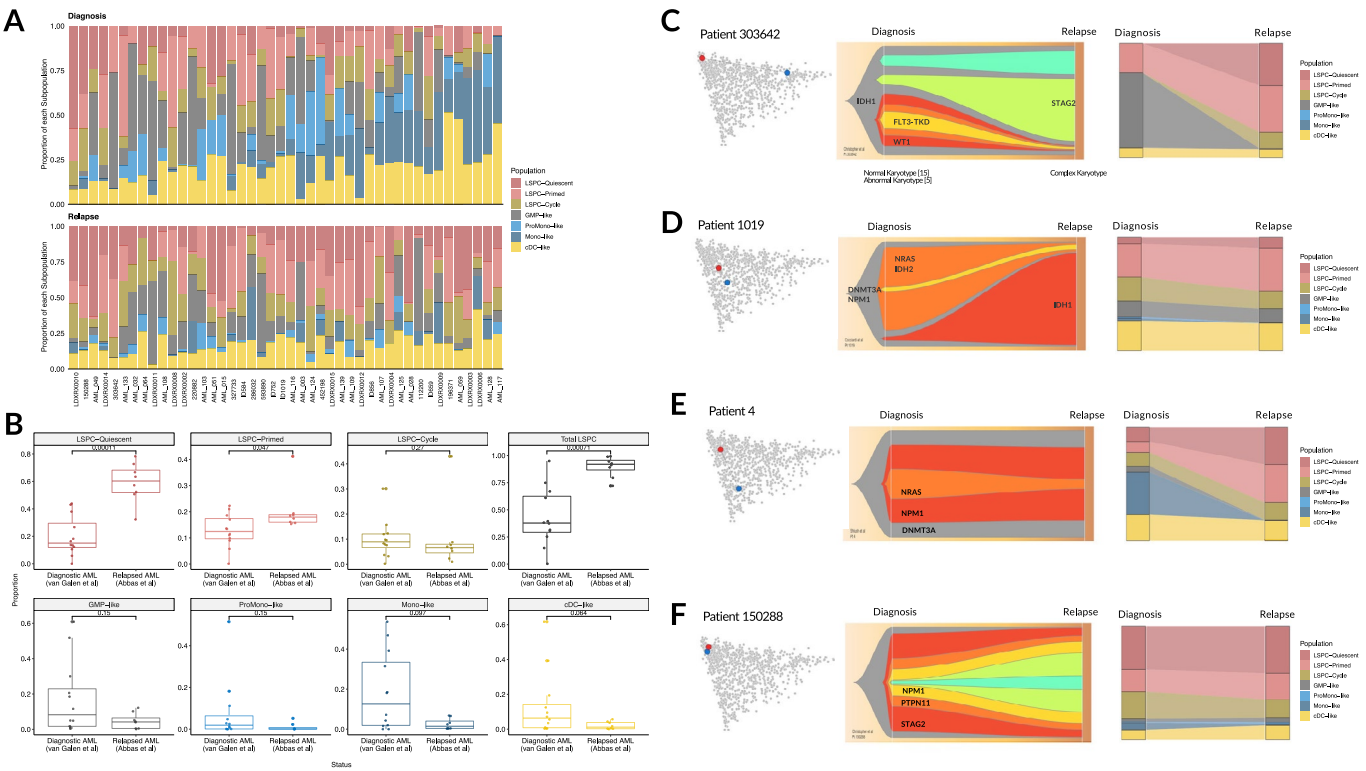


Extended Data Fig. 4 | See next page for caption.

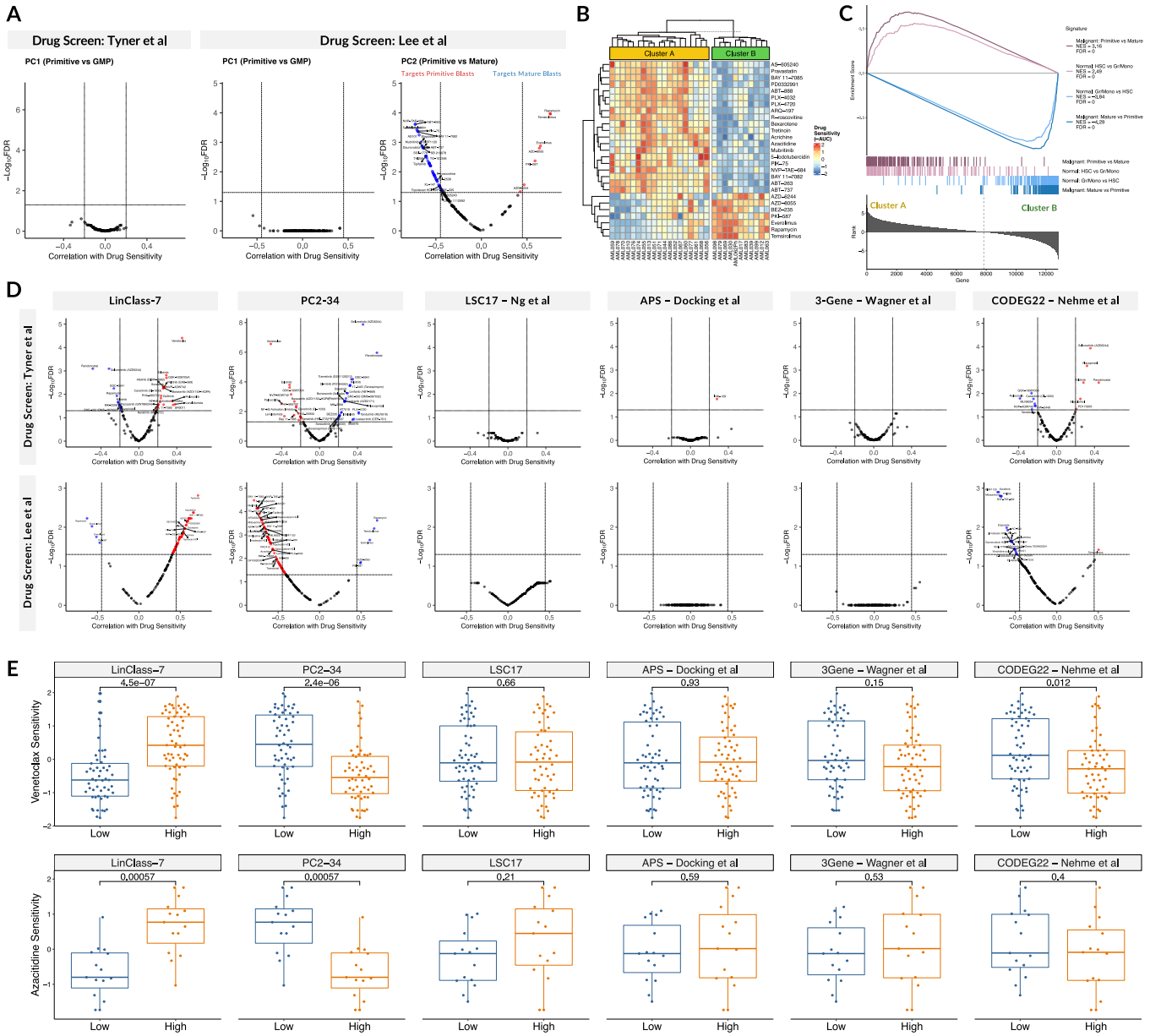
Extended Data Fig. 4 | Biological and genomic correlates of AML hierarchies. A) Correlation between deconvoluted abundance of leukemic and immune cell types with clinical features in TCGA (n = 173). Only correlations with $P < 0.05$ are depicted, and correlations with FDR < 0.05 are noted with an asterisk. **B)** Cellular hierarchy projections of patient samples classified as FAB M0 (n = 30), M1 (n = 122), M2 (n = 77), M3 (n = 30), M4 (n = 58), M4Eo (n = 22), M5A (n = 28), or M5B (n = 11) **C)** Density plots depicting all mutation combinations along the Primitive versus Mature axis (PC2). **D)** Density plots depicting all mutation combinations along the Primitive versus GMP axis (PC1). **E)** Density plots depicting all cytogenetic alterations along the Primitive versus Mature axis (PC2). **F-G)** Impact of *DNMT3A* R882 mutations compared to other *DNMT3A* mutations on leukemic hierarchy organization along the Primitive versus Mature axis (PC2). **F)** Boxplot comparing PC2 of AMLs with *DNMT3A* R882 mutations (n = 84) compared to other *DNMT3A* (n = 96) mutations, split by mutational partner. Box plots indicate the range of the central 50% of the data, with the central line marking the median. Whiskers extend from each box to 1.5*(interquartile range). Statistical significance was evaluated through a two-sided Wilcoxon rank sum test. **G)** Density plot depicting PC2 of mutational combinations with either *DNMT3A* R882 or other *DNMT3A* mutations.



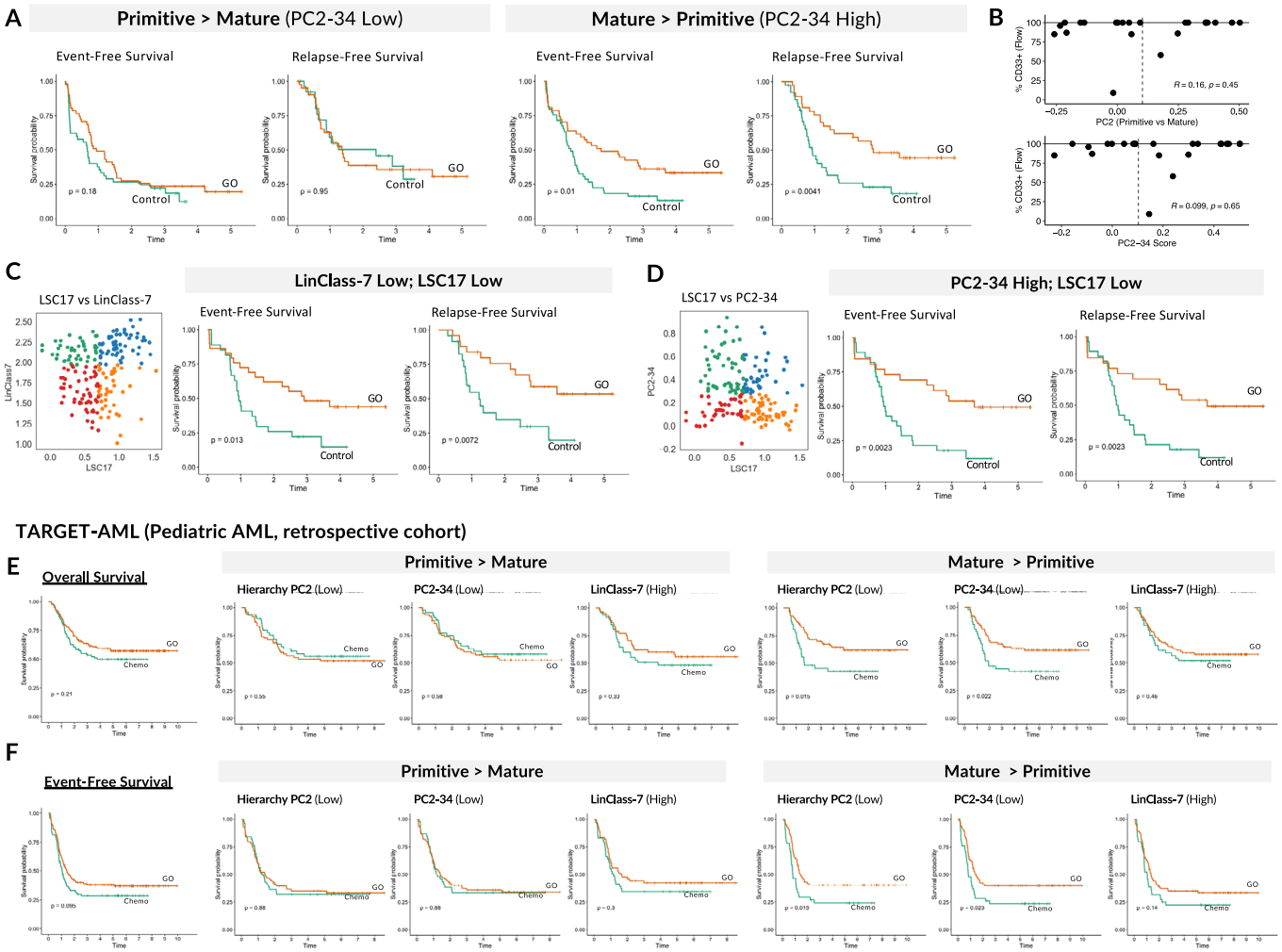
Extended Data Fig. 5 | The Primitive versus GMP axis governs AML prognosis. A) 495 adult AML patients from GSE6891, projected by hierarchy composition and classified based on the reference cohorts. **B)** Overall survival outcomes of 495 AML patients from GSE6891, stratified by hierarchy subtype. **C)** 287 pediatric AML patients from the TARGET-AML cohort, projected by hierarchy composition and classified based on the reference cohorts. **D)** Overall survival outcomes 287 pediatric AML patients from TARGET-AML, stratified by hierarchy subtype. **E)** Correlation between a prognostic score trained by regularized cox regression using leukemic cell type abundances with the Primitive versus GMP axis (PC1) within the TCGA and BEAT-AML cohorts (combined n = 454). **F)** Overall survival outcomes of patients stratified by PC1 within the TCGA (n = 173), BEAT-AML (n=281), and GSE6891 (n = 495) cohorts. **G)** Association between cell-type abundance and induction failure in four independent studies: Bolouri *et al* 2018 (pediatric AML; n = 257 remission, n = 30 induction failure), Chiu *et al* 2019 (adult AML; n = 18 remission, n = 18 induction failure), Herold *et al* 2018 (adult AML; n = 164 remission, n = 86 induction failure), and Tyner *et al* 2018 (adult AML; n = 140 remission, n = 63 induction failure), represented through the test statistic from a two-sided Wilcoxon rank sum test. Green denotes higher relative abundance in induction failure patients compare to patients who achieved complete remission, while purple denotes lower relative abundance in induction failure patients. Differences with an uncorrected $P < 0.10$ are noted with an asterisk. **H)** Correlation between four prognostic AML scores with the relative abundance of each leukemic cell type across the TCGA, BEAT-AML, and Leucegene cohorts (combined n = 864). **I)** Relative abundance of Quiescent LSPC and GMP-like blasts among 864 AML patients split into high and low risk categories by median split for four prognostic scores. Significance was evaluated through two-sided Wilcoxon rank-sum tests. Box plots indicate the range of the central 50% of the data, with the central line marking the median. Whiskers extend from each box to 1.5*(interquartile range). **J)** GSEA analysis with gene signatures derived from the Primitive versus GMP axis in normal and malignant hematopoiesis, performed on genes ranked by univariate associations with overall survival within the TCGA and BEAT-AML cohorts.



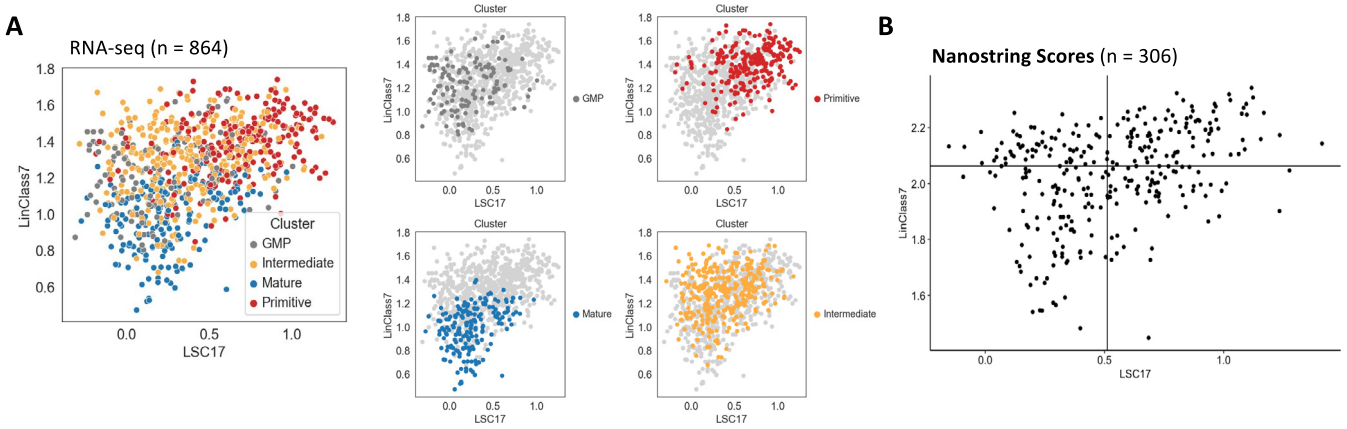
Extended Data Fig. 6 | Changes in hierarchy composition at AML relapse. A) Hierarchy composition of 44 matched diagnosis and relapse pairs. Top row depicts hierarchy composition at diagnosis while the bottom row depicts hierarchy composition at relapse. Samples from the same patient are aligned vertically. **B)** Relative abundance of each leukemic cell population from scRNA-seq of 12 diagnostic AMLs (van Galen *et al*, 2019) compared with 8 relapsed AMLs (Abbas *et al*, 2021). Statistical significance was evaluated through two-sided Wilcoxon rank-sum tests. Box plots indicate the range of the central 50% of the data, with the central line marking the median. Whiskers extend from each box to 1.5*(interquartile range). **C-F)** Evolution of paired diagnosis and relapse AML samples depicted through shifts in cellular hierarchies, evolution of genetic subclones, and changes in cell-type composition. **C)** Patient 303642, in which significant genetic evolution is accompanied by a dramatic shift in cellular hierarchy from GMP to primitive. **D)** Patient 1019, in which replacement of an NRAS and IDH2 positive clone with an IDH1 positive clone is associated with a modest shift in cellular hierarchy. **E)** Patient 4, in which a loss of monocytic blasts is accompanied by a modest decrease in the the size of an NRAS bearing clone. **F)** Patient 150288, in which extensive linear genetic evolution is not associated with any appreciable change in cell type composition.



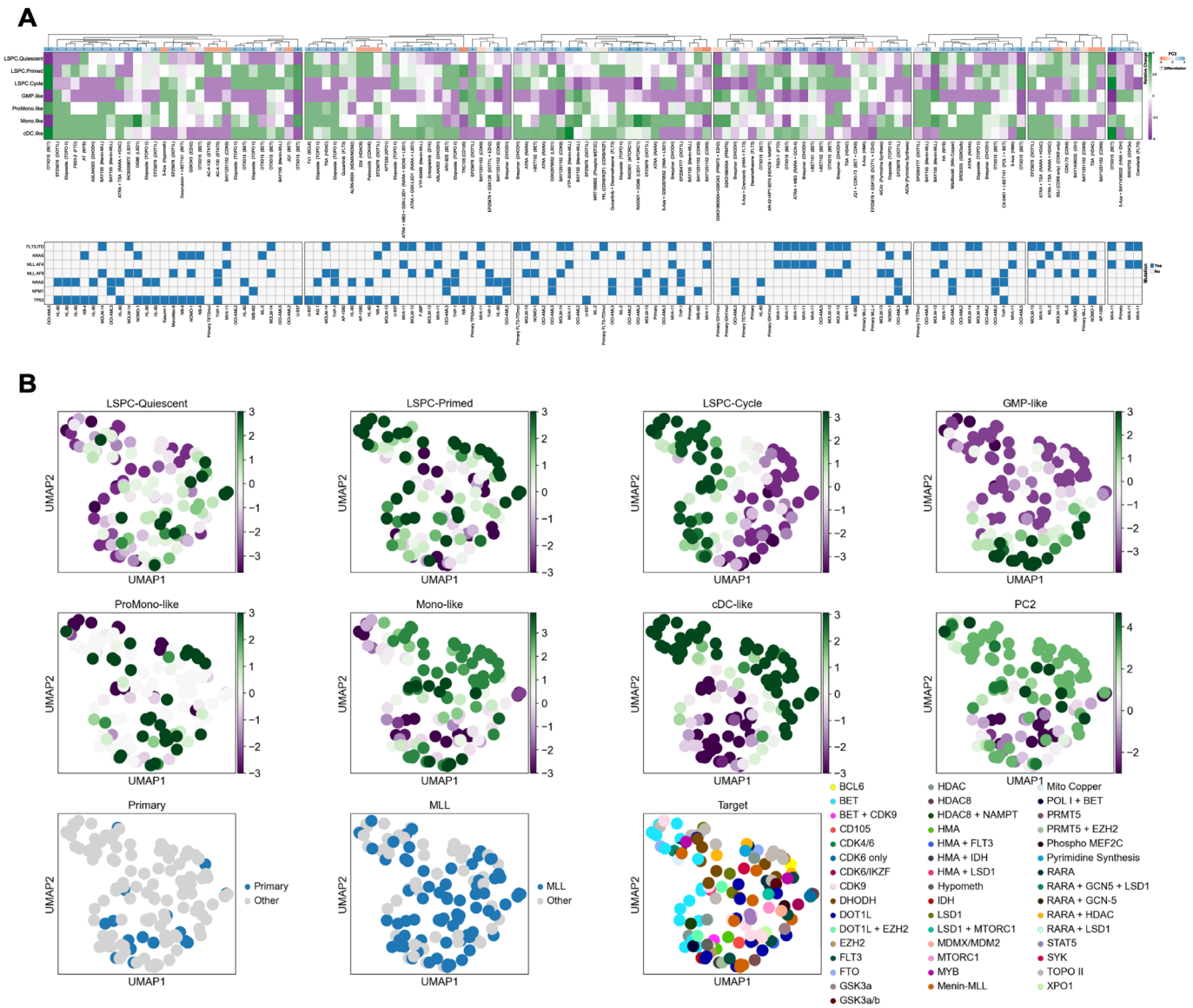
Extended Data Fig. 7 | The Primitive versus Mature axis governs ex vivo drug sensitivity. A) Volcano plot depicting associations between Primitive versus GMP axis (PC1) and ex vivo drug sensitivity from the BEAT-AML (Tyner *et al* 2018) drug screen (n = 202 patients), and between PC1 and PC2 and ex vivo drug sensitivity from Lee *et al* 2018 (n = 30 patients). **B)** Unsupervised clustering of 30 primary AML patients from Lee *et al* on the basis of ex vivo sensitivity to 159 drugs. Drug sensitivity values (scaled negative AUC) are depicted for the top drugs corresponding to each patient cluster. Red denotes sensitivity while blue denotes lower sensitivity. **C)** GSEA analysis with gene signatures derived from the Primitive vs Mature axis in normal and malignant hematopoiesis, performed on genes ranked by differential expression between the two drug response clusters from (B). **D)** Correlations of AML gene expression scores with ex vivo drug sensitivities from two drug screens (Tyner *et al*, n = 202; Lee *et al*, n = 30) performed on primary AML samples. **E)** Ex vivo drug sensitivity to Venetoclax (n = 114 from Tyner *et al*) and Azacytidine (n = 30 from Lee *et al*) of primary patient samples grouped into 'High' or 'Low' based on median splits of patient scores for each AML gene expression score. Significance was evaluated through two-sided Wilcoxon rank-sum tests. Box plots indicate the range of the central 50% of the data, with the central line marking the median. Whiskers extend from each box to 1.5*(interquartile range).



Extended Data Fig. 8 | The Primitive versus Mature axis predicts clinical benefit from Gemtuzumab-Ozogamicin in both adult and pediatric AML. A) Subgroup analysis of randomized clinical trial ALFA-0701 (n = 192) after stratification by PC2-34. Event-free survival (EFS) and Relapse-free survival (RFS) curves comparing chemotherapy only (Control arm) against chemotherapy + Gemtuzumab-Ozogamicin (GO arm). **B)** Lack of correlation between CD33 levels by flow cytometry and Primitive versus Mature axis (PC2 and PC2-34 score), evaluated across 23 Toronto PMH AML patients for which both RNA-seq and clinical flow information was available. **C)** Stratification of ALFA-0701 patients on the basis of both LSC17 and LinClass-7. EFS and RFS for the LinClass-7 Low (Mature > Primitive) and LSC17 Low subgroup (n = 56) is depicted as this was the only group to derive significant benefit from GO treatment. **D)** Stratification of ALFA-0701 patients on the basis of both LSC17 and PC2-34. EFS and RFS for the PC2-34 High (Mature > Primitive) and LSC17 Low subgroup (n = 54) is depicted as this was the only group to derive significant benefit from GO treatment. **E-F)** Subgroup analysis of a retrospective cohort of pediatric AML patients treated with either GO (n = 154) or Chemo (n = 91), stratified by the PC2 Primitive versus Mature axis and related gene expression scores. Outcomes are depicted for both overall survival **(E)** and event-free survival **(F)**.



Extended Data Fig. 9 | LinClass-7 as a companion score for LSC17. A) LSC17 and LinClass-7 scores of 864 AML patients by RNA-seq. Patients belonging to each hierarchy subtype (Primitive, Intermediate, GMP, Mature) are also depicted. (**B**) LSC17 and LinClass-7 scores measured through a 17-gene NanoString assay. Normalized NanoString-derived LSC17 and LinClass-7 scores from 306 Toronto PMH patients from Ng *et al* (2016) are depicted.



Extended Data Fig. 10 | Literature screen to identify treatment-induced changes in cellular composition. A) ComplexHeatmap depicting changes in cell type composition following drug treatment from 43 preclinical studies in human AML, represented as the absolute log(*P* value) from a two-sided Wilcoxon rank sum test in the direction of the change. Green depicts an increase in cell type abundance and purple depicts a decrease in cell type abundance. Each treatment is labeled with its target(s) in parentheses. Changes in PC2 (Primitive versus Mature) are depicted above the heatmap, and candidate differentiation drugs (increase in PC2 with uncorrected $P < 0.05$) are denoted with an asterisk. AML sample type and key genomic characteristics are also depicted for each treatment condition. **B)** UMAP coordinates for each drug treatment condition depicting changes in each cell type after drug treatment compared to control, represented as the absolute log(*P* value) from a two-sided Wilcoxon rank sum test in the direction of the change. Tissue source (Primary vs Cell Line), MLL translocation status, and drug target are also depicted for each treatment condition.

nature research

Corresponding author(s): John E. Dick

Last updated by author(s): 03/26/2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Datasets used in this study are outlined in Supplemental Tables S2 (n = 25 clinical datasets) and S4 (n = 43 preclinical datasets). Deconvolution: CIBERSORTx (no version info), MuSIC (v0.2.0), BisqueRNA (v1.0.5), DWLS (no version info). Fedratinib and Selinexor RNA-seq processing: BWA (v0.6), JAGuaR (v2.1) LSC Fraction RNA-seq processing: STAR (v2.5.2a), HTSeq (v0.9.1) van Galen bulk RNA-seq processing: STAR (v2.7.9a), HTSeq (v0.7.2)
Data analysis	Analysis code for the study is available at: https://github.com/andygxzeng/AMLHierarchies Software and versions for data analysis: clustering analysis: scran (v1.20.1), SAM (v0.7.1), scanpy (v1.5.1), scArches (v0.3.5), scikit-bio (v0.5.6), scikit-learn (v0.24.2), NbClust (v.3.0) survival analysis: survival (v3.2.11), survminer (v0.4.9), glmnet (v4.1.2), clonal evolution: phyloWGS (no version info), fishPlot (v0.5.1) data visualization: ggplot2 (v3.3.5), ggpubr (v0.4.0), ggridges (v0.5.3), ggalluvial (v0.12.3), corrplot (v0.90), EnhancedVolcano (v1.10.0), ComplexHeatmap (v2.8.0)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Processed and raw RNA-seq data generated in this study is available in the Gene Expression Omnibus (GEO) under Superseries GSE199336.
Citations and links of re-analyzed data from all clinical datasets analyzed in this study are provided in Table S4 (25 studies).
Citations and links of re-analyzed data from the literature screen are provided in Table S13 (43 studies).
Pathways and signatures used for relapse benchmarking were obtained from MSigDB (https://www.gsea-msigdb.org/gsea/msigdb/).
All deconvolution results are available on github (https://github.com/andygxzeng/AMLHierarchies).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	All publicly available datasets included in our study had sample sizes > 3 within any condition; datasets with smaller sample sizes were not included in the analysis as they could not be used for statistical inference. For data generated in this study including in vivo drug treatment data, no statistical approaches were used to pre-determine sample size but a minimum of n = 3 samples was required for any of the results reported in the manuscript in order to perform statistical inference. For reporting mutation combinations, the minimum number of patients was set to n = 5.
Data exclusions	Data were not excluded from analysis.
Replication	All replication attempts were successful and most of our findings were validated across multiple independent cohorts, particularly findings around survival outcomes (1236 patients across four independent cohorts), chemotherapy response (776 patients across four independent cohorts), relapse trends (44 patients across four independent cohorts), and drug sensitivity (232 patients across 2 independent cohorts). For in vivo drug treatment data generated in this study, each patient sample was transplanted into 10-20 immunodeficient mice which were independently subjected with either drug or control treatment.
Randomization	All available data was analyzed without sub-sampling or randomization as we utilized all available data to maximize the sample size for each analysis. For drug treatment experiments samples were chosen based on material availability.
Blinding	Researchers undertaking any data collection did not have access to the hierarchy subtype of each patient until after the data was generated. Data preprocessing was performed without patient subtype information. However by necessity the individuals performing the final stages of data analysis did have information pertaining to patient subtypes in order to identify associations between patient subtypes and clinical/ biological correlates.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	APC-anti-CD45: BD Cat#340943 (1:100) V450-anti-CD19: BD Cat#560354 (1:100) APC-Cy7-anti-CD34: BD Cat#NA (custom made by BD) (1:300) FITC-anti-CD15: BD Cat#555401 (1:100) PE-Cy5-anti-CD33: Beckman Coulter Cat#IM2647U (1:100) PE-anti-CD14: Beckman Coulter Cat#IM0650U (1:100)
Validation	All antibodies were purchased from commercial vendors as specified above with the respective specificity and validation documentation provided. All antibodies used for flow cytometry have been previously been used in published studies by us and others - in example, all antibodies listed above were previously used in Surka, Jin, et al Blood 2021. No further validation was performed.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	Ten-week old female NOD-SCID mice were used for xenotransplantation experiments.
Wild animals	This study did not involve wild animals.
Field-collected samples	This study did not involve field-collected samples.
Ethics oversight	Animal experiments were performed in accordance with institutional guidelines approved by the University Health Network Animal Care Committee.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Primary human samples were collected from peripheral blood from patients diagnosed with AML treated at the University Health Network, without any pre-selection for specific genomic or morphological groups. For each patient, written informed consent was obtained in accordance with the Declaration of Helsinki.
Recruitment	Patients were admitted to the Princess Margaret Cancer Centre for examination and treatment for AML. Only consenting patients were included in the study. AML patients were not pre-selected by any pre-requisite cytogenetic or genomic features, however only samples that were capable of leukemic engraftment in immunodeficient mice were used for xenotransplantation studies.
Ethics oversight	This study was approved by the research ethics board of the University Health Network.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*

Gating strategy

Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.