


OPEN

Multi-study reanalysis of 2,213 acute myeloid leukemia patients reveals age- and sex-dependent gene expression signatures

Raeuf Roushangar^{1,2} & George I. Mias ^{1,2}

In 2019 it is estimated that more than 21,000 new acute myeloid leukemia (AML) patients will be diagnosed in the United States, and nearly 11,000 are expected to die from the disease. AML is primarily diagnosed among the elderly (median 68 years old at diagnosis). Prognoses have significantly improved for younger patients, but as much as 70% of patients over 60 years old will die within a year of diagnosis. In this study, we conducted a reanalysis of 2,213 acute myeloid leukemia patients compared to 548 healthy individuals, using curated publicly available microarray gene expression data. We carried out an analysis of normalized batch corrected data, using a linear model that included considerations for disease, age, sex, and tissue. We identified 974 differentially expressed probe sets and 4 significant pathways associated with AML. Additionally, we identified 375 age- and 70 sex-related probe set expression signatures relevant to AML. Finally, we trained a k nearest neighbors model to classify AML and healthy subjects with 90.9% accuracy. Our findings provide a new reanalysis of public datasets, that enabled the identification of new gene sets relevant to AML that can potentially be used in future experiments and possible stratified disease diagnostics.

Acute myeloid leukemia (AML) is a heterogeneous malignant disease of the hematopoietic system myeloid cell lineage^{1–5}. AML is best characterized by terminal differentiation in normal blood cells and excessive production and release of cells at various stages of incomplete maturation (leukemia cells). As a result of this faster than normal, and uncontrolled growth of leukemia cells, healthy myeloid precursors involved in hematopoiesis are suppressed, and ultimately can soar to death within months from diagnosis if untreated^{1,6}. AML accounts for 70% of myeloid leukemia and nearly 80% of acute leukemia cases, making it the most common form of both myeloid and acute leukemia^{1,7}. The number of new AML cases is increasing each year – in 2019 alone, an estimated 21,450 new AML patients will be diagnosed, and nearly 10,920 are expected to die from the disease⁸.

According to the 2016 World Health Organization (WHO) newly revised myeloid neoplasms and acute leukemia classification system⁹, AML prognosis criteria for classification are highly dependent on the presence of chromosomal abnormalities, including chromosomal deletions, duplications, translocations, inversions, and gene fusions. AML is diagnosed predominantly through microscopic, cytogenetic, and molecular genetic analyses of patients' blood, and/or bone marrow samples. Microscopic examination may be used to detect distinctive features (e.g. Auer rods) in cell morphology, cytogenetic analysis to identify chromosomal structural aberrations (e.g., t(8;21), inv(16), t(16;16), or t(9;11)), and molecular genetic analysis to identify gene fusion (e.g., RUNX1-RUNX1T1 and CBFβ-MYH11), and mutations in genes frequently mutated in AML (e.g., NPM1, CEBPA, RUNX1, FLT3)^{1,3,5,10–12}. Such cytogenetic and molecular genetic analyses are used to identify prognosis markers for classifying AML patients into three risk categories: favorable, intermediate, and unfavorable, currently based primarily on the European LeukemiaNet (ELN) 2017 classification^{3,10} (see Estey³ for a recent review, including ELN assessments). A large group of AML patients present normal karyotypes and lack chromosomal abnormalities^{3,5,10,11,13}. These patients are classified as intermediate risk, and often have heterogeneous clinical outcome with standard therapy with risk of AML relapse^{3,5,14}.

¹Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI 48824, USA.

²Institute for Quantitative Health Science and Engineering, Michigan State University, East Lansing, MI 48824, USA. Correspondence and requests for materials should be addressed to G.I.M. (email: gias@msu.edu)

Additionally, AML prognosis worsens with age, and older patients respond less to current treatments, with poorer clinical outcomes compared to younger patients^{15,16}. AML can occur in people of all ages but is primarily diagnosed among the elderly (>60 years old), with a median age of 68 years at diagnosis⁸. Recent advances in AML biology have expanded our understanding of its complex genetic landscape, and led to significant improvement in prognoses and therapeutic strategy for younger patients^{2,16}. For elderly patients, prognoses remain grim and the main therapeutic strategy, remission induction therapy followed by an intensive consolidation phase (post-remission), had remained nearly unchanged over the past three decades^{1,2,4,5,10,16,17}. More recently, however, new therapeutic agents have been approved for older AML patients^{4,5}, and these include venetoclax (combined with decitabine or azacitidine)¹⁸, midostaurin (combined with standard chemotherapy)¹⁹, and gilteritinib²⁰. It is expected that the new therapeutic agents will improve prognosis for older AML patients (where in the past up to 70% of AML patients aged 65 or older were reported to die within a year following diagnosis²¹). While it is apparent that the nature of AML changes with age, still little is known about the extent of these associations and how they vary with patient age^{2,22,23}, and current indications from ELN and the National Comprehensive Cancer Network (NCCN) essentially consider age as a surrogate variable that is used only in conjunction with other treatment-related mortality factors^{3,5,10}. Taking into consideration age in the identification of changes in AML global gene expression may lead to improved early diagnosis and improvement in treatment approaches for elderly patients. To further complicate matters, AML has multiple driver mutations and competing clones that evolve over time, making it a very dynamic disease^{13,24}.

Multiple gene expression analyses of AML have been carried out, 25 of which have been systematically compared by Miller and Stamatoyannopoulos²⁵, who analyzed information on 4,918 genes, and identified 25 genes reported across multiple studies, with potential prognostic features. In this study, we performed a comprehensive gene expression analysis of 2,213 AML patients and 548 healthy subjects, by re-analyzing publicly available gene expression microarray data from 37 curated studies (a reanalysis following strict inclusion criteria) and identified disease-, age- and sex-related gene expression changes associated with AML. The differentially expressed gene sets were associated to signaling pathways relevant in AML, and also used to train and test a predictive model of AML or healthy status. We believe that our results may lead to improved AML early detection, and diagnostic testing with target genes, which collectively can potentially serve as age- and sex-dependent biomarkers for AML prognosis, as well as new treatment targets with mechanisms of action different from those used in conventional chemotherapy.

Results

Data curation and gene expression pre-processing. We searched the Gene Expression Omnibus (GEO) public repository, based on our systematic workflow and inclusion criteria, Fig. 1a,b. Overall, 2,132 datasets were screened, and 643 selected (577 were excluded as non-Affymetrix, various platform arrays). From the 66 remaining corresponding studies, 34 were excluded due to: lack of metadata, using non-peripheral blood or non-bone marrow tissues, or being cell line or cell-type specific, or analyzing treated subjects. After this curation we obtained 34 age-annotated gene expression datasets from 32 different studies covering 2,213 AML patients and 548 healthy individuals. These 34 datasets were reanalyzed, starting from raw microarray data, to perform a gene expression analysis of variance and functional pathway enrichment analysis (see online Methods). Table 1 provides a description of each dataset with a sub-table summary of all curated data used in this study. After pre-processing each individual dataset separately, Fig. 1b, we performed the statistical analysis on 44,754 probe sets which were common across all samples (Affymetrix expression microarray data).

Classification of missing metadata annotation. Following the data curation step, 805 arrays (802 AML and 3 healthy) of 2,761 curated data were found to be missing sex annotation, and 737 arrays (all AML patients) were missing sample source annotation (i.e. whether the tissue from which RNA had been extracted was either bone marrow [BM] or peripheral blood [PB]). To predict the missing sex and sample source annotations, we trained and validated a logistic regression (LR) classification model. The prediction of missing annotations for these arrays was essential in our study, to increase the sample size, and statistical power²⁶. The trained models were cross-validated using our annotated preprocessed expression data, and were $96 \pm 8\%$ and $96.7 \pm 4\%$ accurate for sex and sample source predictions respectively (see Supplementary Table S1 and Fig. S1 for additional LR model performance metrics). Model training, parameters used in training, and validation for this analysis are discussed in the Online Methods. The results from classification for missing annotation were used for the downstream analysis of gene expression variability, and are presented in Supplementary Files S1 and S2 for sample source and sex annotations respectively.

Batch correction. The different datasets we curated for this study did not include within-study healthy controls, which would limit analysis of variance, and particularly the ability to separate biological from batch effects. To address this, we implemented an iterative batch effect correction approach, essentially employing a weight-based method for correcting batch effects – here we use the term “dataset-wise” batch effect correction for this approach. Assuming the batch effects due to each dataset are a function of the number of samples in the dataset (weight), normalizing sets of unevenly sized datasets may lead to an unbalanced batch correction. We used 5 additional datasets as a reference set, which we refer to as “covariate” hereafter. Each of the covariate reference datasets included within-study healthy controls. All 5 datasets together consisted of a total 613 arrays (455 AML and 158 healthy) (Table 1), and were pre-processed exactly as our curated datasets. Each of the remaining datasets was batch corrected with respect to the combined covariate datasets reference using ComBat²⁷. After this dataset-wise correction, the 5 covariate reference datasets were removed, and our expression data were clustered using principal component analysis (PCA), to visually examine the effect of covariate reference datasets on distributing the batch weight during batch correction (Supplementary Fig. S2).

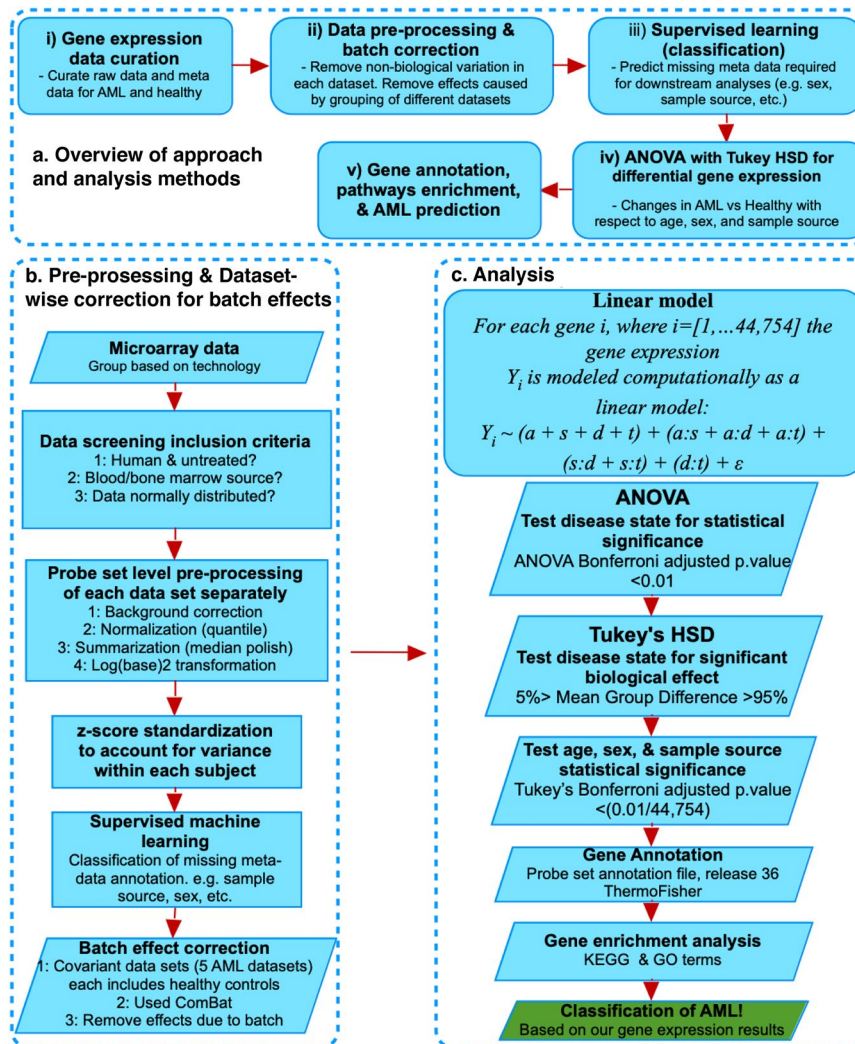


Figure 1. General approach, data curation, and analysis workflow summary. The flowchart shows in (a) the five main steps that summarize our method of approach for our study, and in (b) the curation and screening criteria for raw gene expression and annotation data files curation, data pre-processing, supervised machine learning for missing metadata prediction, and batch effects correction. (c) The analysis included a linear model analysis of variance (ANOVA) coupled with Tukey's Honestly Significant Difference (HSD) post-hoc tests, and KEGG pathway and GO enrichment. Finally, we performed a machine learning classification of AML based on our findings.

Analysis 1: Gene expression analysis and enrichment analysis of AML disease state compared to healthy individuals.

Gene expression analysis of AML disease state. Following batch correction, we performed an analysis of differential expression (DE) on the 34 curated datasets including 2,213 AML patients and 548 healthy controls. Analysis of Variance (ANOVA)^{28–30} was performed according to a linear model (see Online Methods), including factors for age, sex, and sample source (to account for tissue differences between AML and healthy), as well as binary interactions thereof. To avoid assumptions on averaging over multiple probe sets corresponding to the same gene symbol, we analyzed probe sets with the linear model. We identified 974 statistically significant differentially expressed probe sets (DEPS) (corresponding to 964 unique gene symbols) for AML, based on post-hoc analyses (Tukey's Honestly Significant Difference (HSD) tests implemented in R, with adjusted p-value < 0.01), in conjunction with a two-tailed 5% quantile selection³¹ based on the mean difference distribution between AML-healthy group comparisons across probe sets (to identify largest biological effects). The heatmap (Fig. 2a) shows the hierarchical clustering of gene expression from the 974 DEPS, including 487 up- and 487 down-regulated with respect to AML as compared to healthy. From this analysis, WT1 (Wilms tumor 1) with mean difference of 0.26 and adjusted p-value < 4.11×10^{-11} was the most DE up-regulated gene while CRISP3 (cysteine-rich secretory protein 3) with mean difference of -0.52 and adjusted p-value < 4.11×10^{-11} was the least DE gene. Figure 2b shows the top 10 up- and down-regulated DEPS with corresponding gene symbols, that resulted from this analysis (also listed in Table 2, including mean difference and p-adjusted values from post-hoc analysis using Tukey's HSD tests). The entire list of all 974 DEPS can be found as Supplementary Table S2.

Author, Year		GEO accession		Disease Status*		Affymetrix platform id: Number of samples used & Sample source*		Refs*
(A) Curated datasets used in linear model analysis (34 datasets from 32 studies)								
Zatkova <i>et al.</i> , 2009		GSE10258		AML		GPL570: 8 BM		68
Tomasson <i>et al.</i> , 2008		GSE10358		AML		GPL570: 300 BM		69
Metzeler <i>et al.</i> , 2008		GSE12417		AML		GPL570: 73 BM & 5 PB GPL96/97: 160 BM & 2PB		55
Wouters <i>et al.</i> , 2009, Taskesen <i>et al.</i> , 2011		GSE14468		AML		GPL570: 482 BM & 43 PB		70,71
Figueroa <i>et al.</i> , 2009		GSE14479		AML		GPL570: 16 BM		72
Klein <i>et al.</i> , 2009		GSE15434		AML		GPL570: 231 BM & 20 PB		73
Lück <i>et al.</i> , 2011		GSE29883		AML		GPL570: 10 BM & 2 PB		74
Li <i>et al.</i> , 2013, Herold <i>et al.</i> , 2014, Janke <i>et al.</i> , 2014, Jiang <i>et al.</i> , 2016		GSE37642		AML		GPL570: 140 BM GPL96/97: 422 BM		56–59
Bullinger <i>et al.</i> , 2014		GSE39363		AML		GPL570: 11 BM & 2 PB		NYP
Opel <i>et al.</i> , 2015		GSE46819		AML		GPL570: 8 BM & 4 PB		75
TCGA <i>et al.</i> , 2015		GSE68833		AML		GPL570: 183 BM		NYP
Cao <i>et al.</i> , 2016		GSE69565		AML		GPL570: 12 PB		76
Bohl <i>et al.</i> , 2016		GSE84334		AML		GPL570: 25 BM & 20 PB		NYP
Li <i>et al.</i> , 2011		GSE23025		AML		GPL570: 21 BM & 13 PB		77
Warren <i>et al.</i> , 2009		GSE11375		Healthy		GPL570: 26 PB		78
Green <i>et al.</i> , 2009		GSE14845		Healthy		GPL570: 1 PB		NYP
Wu <i>et al.</i> , 2012		GSE15932		Healthy		GPL570: 8 PB		NYP
Karlovič <i>et al.</i> , 2009		GSE16028		Healthy		GPL570: 22 PB		79
Krug <i>et al.</i> , 2011		GSE17114		Healthy		GPL570: 14 PB		NYP
Kong <i>et al.</i> , 2012		GSE18123		Healthy		GPL570: 17 PB		80
Sharma <i>et al.</i> , 2009		GSE18781		Healthy		GPL570: 25 PB		81
Rosell <i>et al.</i> , 2011		GSE25414		Healthy		GPL570: 12 PB		82
Schmidt <i>et al.</i> , 2006		GSE2842		Healthy		GPL570: 2 PB		83
Meng <i>et al.</i> , 2015		GSE71226		Healthy		GPL570: 3 PB		NYP
Tasaki <i>et al.</i> , 2017		GSE84844		Healthy		GPL570: 30 PB		84
Leday <i>et al.</i> , 2018		GSE98793		Healthy		GPL570: 64 PB		85
Shamir <i>et al.</i> , 2017		GSE99039		Healthy		GPL570: 121 PB		86
Tasaki <i>et al.</i> , 2018		GSE93272		Healthy		GPL570: 35 PB		87
Clelland <i>et al.</i> , 2013		GSE46449		Healthy		GPL570: 24 PB		88
Lauwerys <i>et al.</i> , 2013 Ducreux <i>et al.</i> , 2016		GSE39088		Healthy		GPL570: 46 PB		89,90
Xiao <i>et al.</i> , 2011		GSE36809		Healthy		GPL570: 35 PB		91
Zhou <i>et al.</i> , 2010		GSE19743		Healthy		GPL570: 63 PB		92
(B) Covariate datasets (used for batch correction and for testing predictive models)								
Jiang <i>et al.</i> , 2018 [#]		GSE107968 [#]		2 AML; 1 Healthy		GPL570: 3 BM		NYP
Greiner <i>et al.</i> , 2015 [#]		GSE68172 [#]		20 AML; 5 Healthy		GPL570: 25 PB		64
Majeti <i>et al.</i> , 2009 [#]		GSE17054 [#]		9 AML; 4 Healthy		GPL570: 13 BM		65
Bacher <i>et al.</i> , 2012 [#]		GSE33223 [#]		20 AML; 10 Healthy		GPL570: 30 PB		66
Mills <i>et al.</i> , 2009 [#]		GSE15061 [#]		404 AML; 138 Healthy		GPL570: 542 BM		67
(C) Analysis datasets summary statistics								
Disease state		Sample source		Affymetrix platform id		Unique probe sets		
AML	Healthy	BM	PB	GPL570	GPL96/97	GPL570	GPL96/97	
2,213	548	2,090	671	2,177	584	54,675	44,760	

Table 1. Summary table gene expression datasets used in this study. Summary of datasets used in our analysis and disease classification. *GEO, Gene Expression Omnibus; AML, acute myeloid leukemia; Refs., references; NYP, not yet published; GPL570, Affymetrix Human Genome U133 Plus 2.0 Array; GPL96, Affymetrix Human Genome U133A Array; GPL97, Affymetrix Human Genome U133B Array; BM, Bone Marrow; PB, Peripheral Blood.

(ii) *Pathway and gene ontology enrichment analysis of DEPS.* We carried out overrepresentation analysis in Kyoto Encyclopedia of Genes and Genomes (KEGG)^{32–34} signaling pathways, and Gene Ontology (GO) terms^{35,36} on all 974 DEPS, using the Database for Annotation, Visualization and Integrated Discovery (DAVID)^{37,38}. Four KEGG signaling pathways were identified as enriched (Benjamini and Hochberg³⁹ adjusted p-value < 0.05), including Hematopoietic cell lineage, Cell cycle, p53 signaling pathway, and Transcriptional misregulation in cancer. The 4

KEGG signaling pathways are summarized in Table 3 (see also Supplementary Fig. S3a–d), including unadjusted p-values, and Benjamini-Hochberg³⁹ adjusted p-values. These signaling pathways were associated with 56 DEPS, including 27 up- and 29 down-regulated DEPS (Fig. 2c) - the heatmap of their mean differences (AML-healthy values) is shown in Fig. 2d. From our gene enrichment analysis for overrepresentation in GO terms, 21 GO terms were statistically significant (Benjamini and Hochberg³⁹ adjusted p-value < 0.05), with 727 DE unique identities (335 up- and 392 down-regulated). GO terms included protein and microtubule binding for the molecular function (MF) category, inflammatory and immune responses, mitotic nuclear division, and cell proliferation response for the biological process (BP) category, and finally, cytoplasm, extracellular exosome, cytosol, extracellular space, integral component of plasma membrane immune response, and others, for the cellular component (CC) category (Fig. 2e). The complete list of the enrichment analysis results is shown in Supplementary Table S3.

Analysis 2. Gene expression analysis and enrichment analysis of sex- and age-related DEPS in AML.

To characterize sex- and age-specific gene expression changes in AML patients compared to healthy individuals we conducted the following additional analyses detailed further below: (i) Analysis 2a: “Sex-relevance differential gene expression analysis and associated signaling pathways in AML”, and (ii) Analysis 2b: “Age-dependent differential gene expression analysis and associated signaling pathways in AML”. We used the same filtering criteria in both analyses as those used in Analysis 1 for identifying DEPS and signaling pathways between AML patients and healthy controls. In addition, DEPS were regarded as statistically significantly (up- or down-regulated) for each factor, sex and age, if they displayed p-value from Tukey’s HSD < 2.2×10^{-7} (Bonferroni⁴⁰ adjusted p-value of 0.01 divided by the number of probe sets tested, 44,754).

Analysis 2a. Sex-relevance differential gene expression analysis and associated signaling pathways in AML. We identified 266 DEPS that show sex differences between AML patients (p-value < 2.2×10^{-7}), as listed in Supplementary Table S4. 70 DEPS were found to overlap between Analysis 1 (AML disease state) and Analysis 2a (Sex-relevance in AML). Figure 3a shows these 70 DEPS with gene symbol annotations, and their mean difference values in the heatmap, which highlights differences in significance for common DEPS in both Analyses 1 and 2a. Figure 3b shows the hierarchical clustering of the 70 DEPS (rows) on sex and disease state of all 2,213 AML and 548 healthy subjects (columns) indicated by color bars above the heatmap. The top 10 DEPS higher in either males or females from this analysis are shown in Fig. 3c.

For enrichment analysis, we searched for common intersections in KEGG pathways and GO terms between the sex analysis and the 974 DEPS from the disease state analysis. Sex-relevant DEPS were found in 3 different signaling pathways, including genes higher expressed in males: FLT3 and CD34 in Hematopoietic cell lineage, FLT3 in Transcriptional misregulation in cancer 1, and PMAIP1 in p53 signaling pathway 1. MS4A1 was higher in females and found in the Hematopoietic cell lineage pathway (Table 3). Figure 3d shows GO analysis results, where 15 overrepresented biological GO terms were overlapped, including terms for extracellular space, immune response, protein binding, spindle, and midbody.

Analysis 2b. Age-dependent differential gene expression analysis and associated signaling pathways in AML. The subjects were binned in 8 age-groups: 0–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, and 80–100 years old. From this analysis, 1395 unique probe sets across all age-groups had statistically significant differential expression (p-value < 2.2×10^{-7} , Supplementary Table S5). From these 375 unique DEPS (372 unique gene symbols) were found to overlap with the 974 DEPS probe sets from our AML disease state Analysis 1, accounting for an overall 1400 binary comparisons between the multiple age groups deemed statistically significant, based on Tukey HSD tests between age-group pairs. All 1400 identified pairwise differences between age groups and associated probe set/gene information can be found as Supplementary Table S6. The top 10 up- and down-regulated DEPS (labeled with gene symbols) from this analysis are shown in Fig. 4a. Additionally, 75 DEPS with gene symbols identified to have appeared specifically in one age-group comparison are shown in Fig. 4b. Through comparison with the results for KEGG analysis for signaling pathways from Analysis 1, 17 DE genes identified in all 4 KEGG pathways according to age groups (Fig. 4c, Table 4).

To investigate further the progression with age, pairwise correlations between age-groups were computed. The 0–19 age-group was used as a common comparison reference with respect to other groups. Using this 0–19 group as a baseline, the mean differences of 25 DEPS with respect to the 0–19 baseline across all other groups were calculated and visualized in Fig. 4d. The mean difference values between AML and healthy are shown in the right-most column of Fig. 4a,b,d for reference.

We also wanted to assess the interaction of age with disease state. We filtered the 375 unique DEPS which intersected between the age and disease statistically significant DEPS, to also have a statistically significant interaction based on the ANOVA results (p-value < 0.01). This resulted in 43 unique DEPS, statistically significant for age, and for disease, and for interaction between age and disease in the linear model (Supplementary Table S6).

AML classification machine learning model. We used the 974 DEPS from Analysis 1 to train a k-nearest neighbor (KNN) algorithm in ClassificalO⁴¹. All 34 datasets (16 AML and 18 healthy) were used for training, and testing was performed on the 5 covariate reference datasets, which included both AML and healthy subjects (Table 1). The trained KNN algorithm was $97.9 \pm 3\%$ accurate, and 92% accurate in testing results (see Online Methods for parameters, Supplementary Table S1 and Fig. S4).

We also identified a minimum DEPS set that can have good predictive power and sensitivity: We first sorted the 974 disease-related DEPS based on the absolute value of their effect size (mean difference between AML and healthy patients). We then iteratively trained and tested a KNN model on the top *n* DEPS post sorting (Supplementary Fig. S5), incrementing *n* by one in each iteration. Based on the results, we picked the top 10

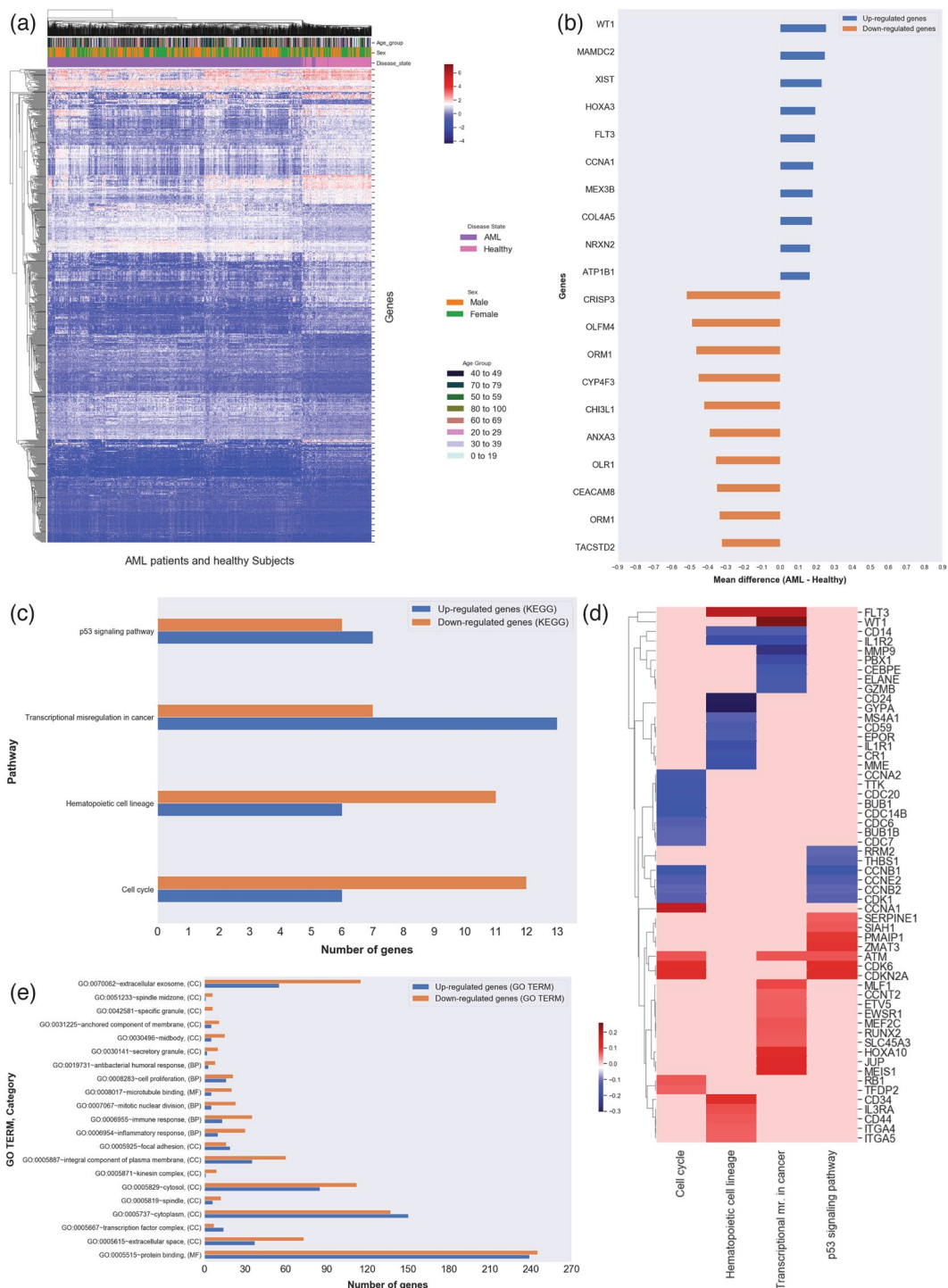


Figure 2. Functional classification of DEPS from AML analysis and associated KEGG and GO enrichment analysis. For all panels, normalized values are represented in blue for down-regulation and red for up-regulation, while light red/gray represents no reported specific direction. **(a)** Heatmap of 974 DEPS (rows) on 2,761 arrays (columns) including 2213 AML patients and 548 healthy individuals from AML analysis, using unsupervised hierarchical clustering and Euclidean distance for clustering. The age of each individual is illustrated in the color bar on the top (dark green for old and light blue for young). The disease state (AML vs healthy), sex of each subject and age-groups are also represented in color bars on the top. **(b)** Horizontal bar plot of the top 10 DEPS (gene symbols on vertical axis) from AML analysis with mean difference values between AML and healthy (horizontal axis). Enrichment analysis identified 4 KEGG signaling pathways **(c)** for our AML DEPS, also visualized as a heatmap **(d)** of DEPS mean difference values between AML and healthy DEPS (rows) identified in these 4 KEGG signaling pathways (columns). The GO enrichment analysis results are summarized in **(e)**.

Up-regulated*			
DEG name	DEPS Gene Symbol	Tukey's HSD Mean difference	p-adjusted value (HSD test in R)
Wilms tumor 1	WT1	0.255353	<4.11E-11
MAM domain containing 2	MAMDC2	0.248983	5.47E-09
X inactive specific transcript (non-protein coding)	XIST	0.230331	<4.11E-11
homeobox A3	HOXA3	0.195790	1.1E-06
fms-related tyrosine kinase 3	FLT3	0.193420	<4.11E-11
cyclin A1	CCNA1	0.185050	1.35E-07
mex-3 RNA binding family member B	MEX3B	0.181068	<4.11E-11
collagen, type IV, alpha 5	COL4A5	0.177721	1.7E-05
neurexin 2	NRXN2	0.166598	<4.11E-11
ATPase, Na ⁺ /K ⁺ transporting, beta 1 polypeptide	ATP1B1	0.165197	5.47E-09
Down-regulated			
cysteine-rich secretory protein 3	CRISP3	-0.51965625	<4.11E-11
olfactomedin 4	OLFM4	-0.489845396	<4.11E-11
orosomucoid 1	ORM1	-0.465232864	<4.11E-11
cytochrome P450, family 4, subfamily F, polypeptide 3	CYP4F3	-0.453467442	<4.11E-11
chitinase 3-like 1 (cartilage glycoprotein-39)	CHI3L1	-0.421520435	<4.11E-11
annexin A3	ANXA3	-0.390688999	<4.11E-11
oxidized low density lipoprotein (lectin-like) receptor 1	OLR1	-0.35525472	<4.11E-11
carcinoembryonic antigen-related cell adhesion molecule 8	CEACAM8	-0.351181264	<4.11E-11
orosomucoid 1	ORM1	-0.336303304	<4.11E-11
tumor-associated calcium signal transducer 2	TACSTD2	-0.323939961	<4.11E-11

Table 2. Top 10 up- and down-regulated of DEPS in AML from disease state. From the Post-hoc Tukey's test, gene expression means difference value < 5% or >95% between AML and healthy (AML - healthy) were selected for biological effect from the statistically significant differentially expressed genes for disease state - based on the analysis of variance of all 2,761 cases (2,213 AML patients and 548 healthy controls). *Significant DEPS (gene symbols) are listed in descending order of the mean difference value comparisons for disease state.

effect-sorted DEPS as a minimum set, as the graphs showed stabilization/saturation, with no substantial increase in performance after $n = 10$. The KNN model using these 10 effect-sorted DEPS had $96.1 \pm 4\%$ accuracy in training, and 90.9% accuracy in testing. (Supplementary Table S1, Fig. S5). The top 10 effect-sorted DEPS corresponded to the 10 top downregulated DEPS listed in Table 2.

Discussion

In the present study, we reanalyzed data aggregated from our curation of 34 publicly available microarray gene expression datasets covering 2,213 AML patients and 548 healthy individuals to identify changes in AML gene expression associated with disease state (AML compared to healthy), sex-linked (male compared to female), and age-dependent (across age-groups compared to baseline). We performed 3 differential probe set (gene) expression and gene enrichment analyses, as discussed below.

Analysis 1. *Gene expression analysis and associated signaling pathways of AML disease state compared to healthy individuals*, was carried out to identify DEPS in AML disease state. The results from this analysis were then used as a baseline indicator for AML disease state. 974 DEPS (487 up- and 487 down-regulated) were identified as statistically significant differentially expressed between AML patients and healthy individuals (p-value < 0.01) and showing high effect size (5% 2-tailed quantile selection). Among these, 6 genes are known to be involved in AML functional pathways, including 4 up-regulated, JUP, CCNA1, FLT3, PIK3R1, and 2 down-regulated, CD14, CEBPE. The top 10 up- and down-regulated genes from this analysis are listed in Table 2. As shown in Fig. 2b of the top 10 up- and down-regulated DEPS and corresponding gene annotations, WT1 (Wilms tumor 1) was found to be the most expressed and CRISP3 (cysteine-rich secretory protein 3) was the most under-expressed gene. WT1 is a transcriptional regulatory protein essential for cellular development and cell survival, and it has been shown to be highly expressed with an oncogenic role in AML^{42,43}, in agreement with our findings. However, CRISP3's direct role in AML is still under investigation. CRISP3 is a member of the cysteine-rich secretory protein CRISP family with major role in female and male reproductive tract, and is mainly expressed in salivary glands and bone marrow⁴⁴. Recently, 80 genes were reported as "extracellular matrix specific genes" in leukemia, and CRISP3 was among the downregulated DE genes reported⁴⁵. CRISP3 associations with AML merit further investigation.

The enrichment analysis for GO terms of the 974 DE probe sets (Fig. 2c) resulted in 727 identifiers (335 up- and 392 down-regulated) enriched for 21 GO terms. 592 of these (257 up- and 335 down-regulated) were enriched in the cellular component (CC) categories mainly associated with cytoplasm, extracellular exosome, cytosol, and extracellular space. These terms are rather generic, but may still reflect relevance to AML development and progression^{46,47}. GO terms in the Biological process (BP) category included inflammatory and immune

AML Vs Healthy DEPS and associated signaling pathways					
Pathway	No. of genes*	Down-regulated	Up-regulated	p-value	p-value Benjamini adjusted
Hematopoietic cell lineage	11, 6	IL1R2, CD59, GYPA, MS4A1, EPOR, CD24, CD14, EPOR, IL1R1, MME, CRI	ITGA4, FLT3, CD34, IL3RA, ITGA5, CD44	2.3E-5	5.8E-3
Cell cycle	12, 6	CDC7, CDC6, CCNB1, CDC20, CCNA2, CCNE2, TTK, CDC14B, CDK1, BUB1, CCNB2, BUB1B	RB1, CCNA1, CDK6, ATM, TFPD2, CDKN2A	1.4E-4	1.2E-2
p53 signaling pathway	6, 7	THBS1, CCNB1, CCNE2, CDK1, RRM2, CCNB2	SLAH1, CDK6, ATM, SERPINE1, CDKN2A, PMAIP1, ZMAT3	1.0E-4	1.3E-2
Transcriptional misregulation in cancer	7, 13	IL1R2, GZMB, CD14, ELANE, MMP9, CEBPE, PBX1	WT1, RUNX2, ETV5, MEIS1, JUP, EWSR1, ATM, HOXA10, MLF1, FLT3, CCNT2, MEF2C, SLC45A3	6.5E-4	4.1E-2
AML sex relevant (male - female) DEPS & associated signaling pathways					
Pathway	No. of genes*	High in Females		High in Males	
Hematopoietic cell lineage	1, 2	—		FLT3, CD34	
p53 signaling pathway	—, 1	—		PMAIP1	
Transcriptional misregulation in cancer	—, 1	MS4A1		FLT3	

Table 3. KEGG pathway analysis of DEPS from meta-analysis of 34 gene expression datasets. Enrichment analysis was done using 974 DEPS, including KEGG enrichment analysis identified 4 statistically significant pathways from AML Vs Healthy analysis, shown with overlaps with sex-specific analysis. *Up and down regulated genes displayed.

responses, and cell proliferation, which are expected as AML is characterized by terminal differentiation of normal blood cells, and excessive proliferation and release of abnormally differentiated myeloid cells, which affects many biological processes associated with the immune system. The four statistically significant KEGG pathways identified in the pathway enrichment analysis encompassed 56 DEPS (Table 3). Transcriptional misregulation in cancer was the most up-regulated pathway in AML (13 up-regulated DE genes), while Hematopoietic cell lineage, and Cell cycle pathways were mostly down-regulated, and the p53 signaling pathway was balanced in terms of up/downregulated DE genes (Fig. 2c). For the enriched pathways, Fig. 2d shows the mean difference values of the 56 DE pathway-associated genes, including 27 up- and 29 down-regulated genes. These KEGG pathways are known to be involved in tumorigenesis. Additionally, the majority of the DE genes from the AML analysis associated with the identified signaling pathways are known to be abnormally expressed in AML. These findings are consistent with results from other studies, and our current understanding of AML pathogenesis.

The DEPS overlap with the 25 genes reported by Miller and Stamatoyannopoulos that were reported in at least 8 studies²⁵, namely HOXA10, CD34, MEIS1, VCAN, RBPMS and MN1. In terms of the genes reported in the same study for poor progression we also consistently identified as upregulated HOXA10, RBPMS, CD34, GNAI1, CLIP2, DAPK1, GUCY1A3, ANGPT1 and FLT3, and as downregulated UGCG. While these are known markers, with consistent expression differences, our additional results need to be investigated further and experimentally validated, including mechanistic considerations.

Analysis 2a. *Sex-dependent gene expression analysis and associated signaling pathways in AML compared to healthy individuals*, was performed to explore the relevance of patients' sex on gene expression and to identify sex-linked genes and associated signaling pathways in AML. A total of 266 DEPS were found statistically significant in this analysis, with 70 found to overlap with the DEPS from Analysis 1 (Fig. 3a,b). The top 10 up- and down-regulated DE genes with respect to females include (Fig. 3c): (i) DDX3Y (DEAD-Box Helicase 3 Y-Linked), EIF1AY (Eukaryotic Translation Initiation Factor 1 A Y-Linked), KDM5D (Lysine Demethylase 5D), RPS4Y1 (Ribosomal Protein S4 Y-Linked 1) with higher expression in males compared to females, and (ii) XIST (X Inactive Specific Transcript), TSIX (TSIX Transcript, XIST Antisense RNA), and PRKX (Protein Kinase X-Linked) with higher expression in females. These genes are known to be sex-specific and show expression differences and sex separation within the AML and the healthy groups respectively (Fig. 3d). The role of these genes as positive controls in studies with AML needs to be investigated further. We also reported sex and AML known genes that were statistically significant in our analysis, including FLT3 and MAL.

Analysis 2b. *Age-dependent gene expression analysis and associated signaling pathways in AML compared to healthy individuals*, was carried out to identify common set of age-dependent gene expression and associated signaling pathways and to explore age-dependent trends in AML. The age-dependent analysis using ANOVA, identified 1,395 DEPS (p-value < 2.2×10^{-7}). To identify age-related DEPS in AML we overlapped the 1,395 DEPS to our findings of 974 DEPS in AML disease state (Analysis 1) (Fig. 4), and identified an overlap of 375 DEPS (p-value < 2.2×10^{-7}). The top 10 up and down DE age-associate genes in AML according to the mean difference values in seven age-groups are shown in Fig. 4a (including their corresponding values from AML disease state in column "AML - healthy" for comparisons). Interestingly, CRISP3 was among the down regulated genes in this analysis as well, specifically associated with differences in younger age groups, 20 to 49 years of age as compared to the 0 to 19 age group. Other genes showing age-specific differences included HOXA3, HOXA5 and HOXA10-HOXA9, which belong to the homeobox genes (HOX) family of transcription factors, essential for

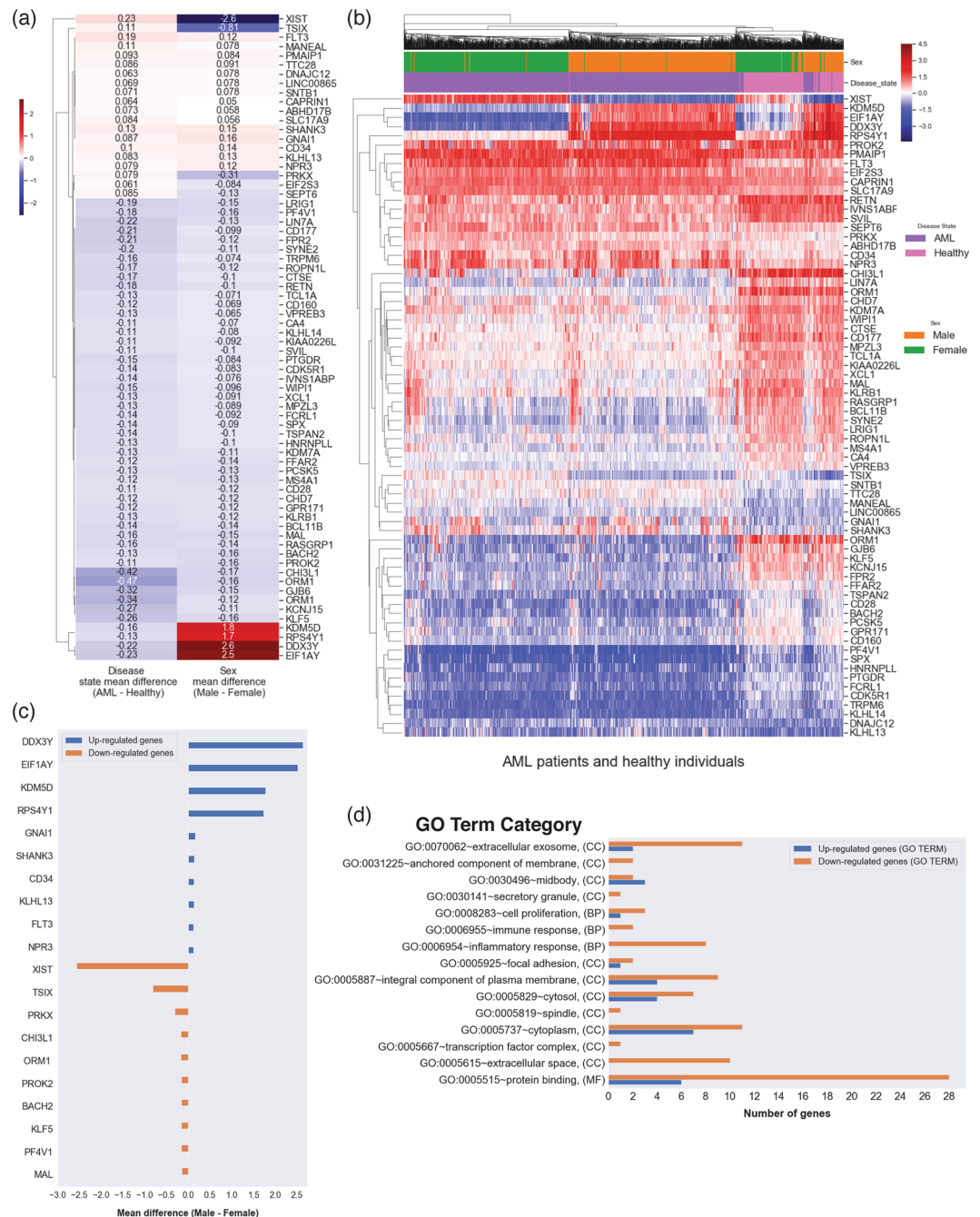


Figure 3. Sex-related gene expression analysis in AML. **(a)** The heatmap of mean difference values comparison between the 70 DE overlapping genes between Analysis 1 and Analysis 2a. **(b)** Heatmap the 70 DEPS expression (rows) on 2,761 arrays (columns) including 2213 AML patients and 548 healthy individuals from Analysis 2a of sex-relevance in AML (using unsupervised hierarchical clustering and Euclidean distance for clustering). The disease state (AML vs healthy) and sex of each subject are indicated in color bars at the top. **(c)** Horizontal bar plot of the top 10 DEPS (gene symbols on vertical axis), with the mean difference values between male-female (horizontal axis). **(d)** Enrichment analysis for statistically significant overrepresented biological GO terms on the 70 DE genes.

embryonic development and hematopoiesis, and associated with chromosomal abnormalities translocation and over-expression in AML^{48,49}. Also identified with age-specific DE, was ORM1, which in Analysis 1 was among the top-10 under-expressed genes, and was also among the 70 DE genes in analysis 2a. ORM1's direct role in AML also merits further investigation, given ORM1 involvement in immunosuppression and inflammation⁵⁰. Finally, we have identified 75 DEPS that show association with only one age-group, exclusively from all other age-groups, suggestive of potential age-specific differential gene expression signature.

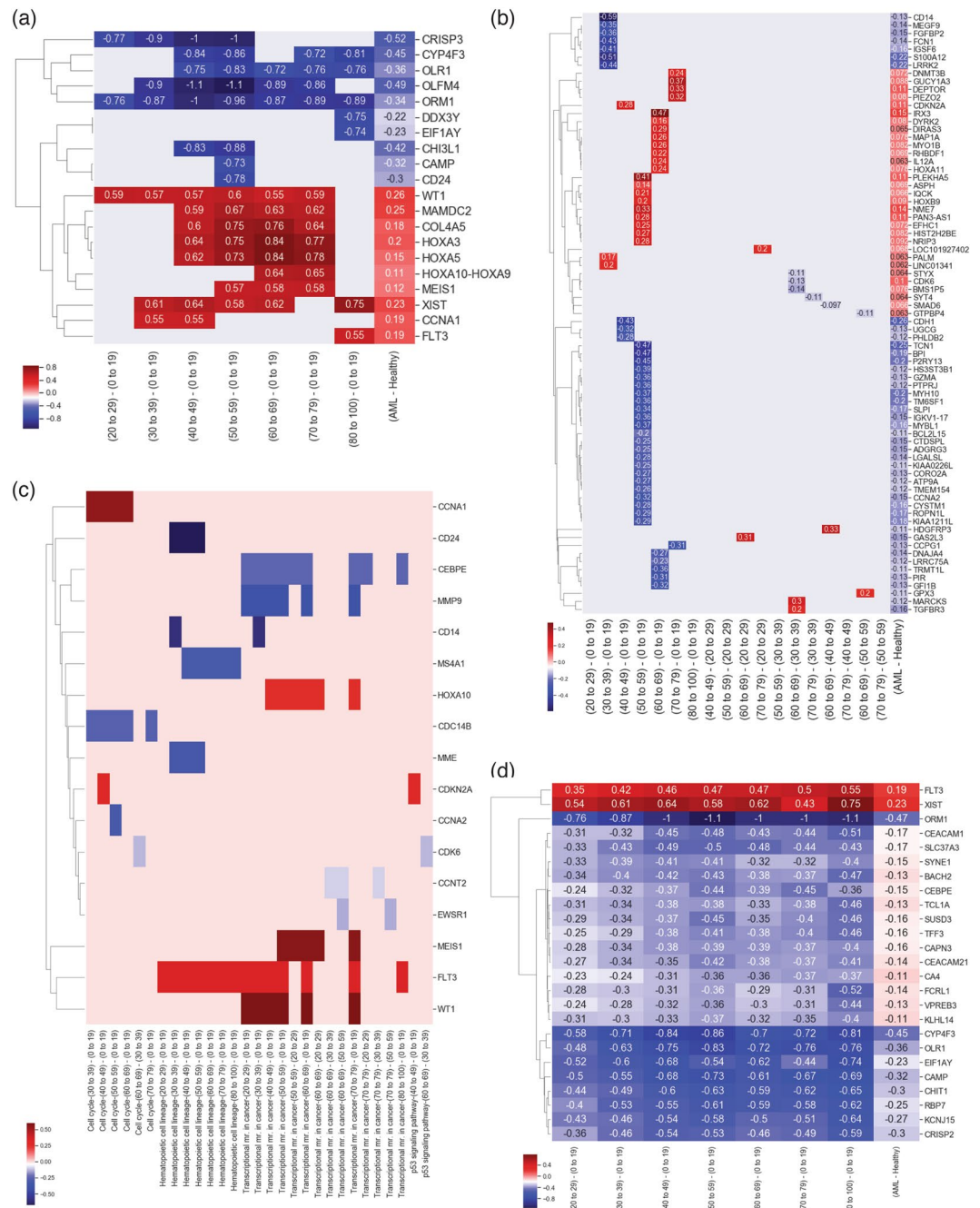


Figure 4. Age-related gene expression analysis in AML. (a) The top 10 up- and down- regulated DEPS overlapping AML and age-related analyses. (b) 75 DEPS specific to a single age-group comparison. (c) Overlaps over KEGG pathways of 17 DE genes identified in 4 KEGG pathways according to age groups. (d) The mean difference of 25 DEPS with respect to the 0–19 baseline across all other groups are plotted to illustrate changes with aging. The mean difference values between AML and healthy cohorts are shown in the right-most column of panes (a, b and d) for reference comparisons.

We further wanted to assess the interaction of age with disease state. From the 375 unique DEPS which intersected between the age and disease analysis, 43 unique DEPS showed statistically significant interaction between age and disease in the linear model ($p < 0.01$, Supplementary Table S6). Among the 43 DEPS are 13 immune disease related genes^{37,38} such as SOCS5 and SOCS6 (suppressors of cytokine signaling, whose role in cancer is still under investigation⁵¹), EBF1 (early B-cell Factor 1), CD160, TCL1A (T-cell leukemia/lymphoma 1A), VPREB3 (pre-B lymphocyte 3), KLF10 (Kruppel-like factor 10), NTM (neurotrimin), PLXNA4 (plexin A4), SLC25A21, SYT4 (synaptotagmin 4) and TCERG1 (transcription elongation regulator 1). While these genes/gene families have been associated with cancer^{13,52}, their potential role in AML is still under and merits further investigation. These 43 DEPS with statistically significant age-disease interactions may be important in AML development, particularly for detecting early markers of AML, potentially identifying preleukemic conditions, and using these markers as treatment targets.

AML age-dependent (AML - healthy) DEPS & associated signaling pathways			
Pathway	No. of genes*	Down-regulated Age-group	Up-regulated Age-group
Hematopoietic cell lineage	4, 1	CD14 (30 to 39)–(0 to 19)	FLT3 (20 to 29)–(0 to 19), (30 to 39)–(0 to 19), (40 to 49)–(0 to 19), (50 to 59)–(0 to 19), (60 to 69)–(0 to 19), (70 to 79)–(0 to 19), (80 to 100)–(0 to 19)
		MME (30 to 39)–(0 to 19), (40 to 49)–(0 to 19), (50 to 59)–(0 to 19)	
		CD24 (30 to 39)–(0 to 19), (40 to 49)–(0 to 19), (50 to 59)–(0 to 19)	
		MS4A1 (40 to 49)–(0 to 19), (50 to 59)–(0 to 19), (60 to 69)–(0 to 19), (70 to 79)–(0 to 19), (80 to 100)–(0 to 19)	
Cell cycle	3, 2	CCNA2 (50 to 59)–(0 to 19)	CCNA1 (30 to 39)–(0 to 19), (40 to 49)–(0 to 19), (50 to 59)–(0 to 19), (60 to 69)–(0 to 19)
		CDK6 (60 to 69)–(30 to 39)	
		CDC14B (30 to 39)–(0 to 19), (40 to 49)–(0 to 19), (50 to 59)–(0 to 19), (60 to 69)–(0 to 19), (70 to 79)–(0 to 19)	CDKN2A (40 to 49)–(0 to 19)
p53 signaling pathway	1, 1	CDK6 (60 to 69)–(30 to 39)	CDKN2A (40 to 49)–(0 to 19)
Transcriptional misregulation in cancer	5, 4	CD14 (30 to 39)–(0 to 19)	MEIS1 (50 to 59)–(0 to 19), (50 to 59)–(20 to 29), (60 to 69)–(0 to 19), (60 to 69)–(20 to 29), (70 to 79)–(0 to 19)
		MMP9 (20 to 29)–(0 to 19), (30 to 39)–(0 to 19), (40 to 49)–(0 to 19), (50 to 59)–(0 to 19), (60 to 69)–(0 to 19), (70 to 79)–(0 to 19)	
		EWSR1 (60 to 69)–(50 to 59), (70 to 79)–(50 to 59)	WT1 (20 to 29)–(0 to 19), (30 to 39)–(0 to 19), (40 to 49)–(0 to 19), (50 to 59)–(0 to 19), (60 to 69)–(0 to 19), (70 to 79)–(0 to 19)
		CEBPE (20 to 29)–(0 to 19), (30 to 39)–(0 to 19), (40 to 49)–(0 to 19), (50 to 59)–(0 to 19), (50 to 59)–(20 to 29), (60 to 69)–(0 to 19), (70 to 79)–(0 to 19), (70 to 79)–(20 to 29), (80 to 100)–(0 to 19)	FLT3 (20 to 29)–(0 to 19), (30 to 39)–(0 to 19), (40 to 49)–(0 to 19), (50 to 59)–(0 to 19), (60 to 69)–(0 to 19), (70 to 79)–(0 to 19), (80 to 100)–(0 to 19)
		CCNT2 (60 to 69)–(30 to 39), (70 to 79)–(30 to 39), (60 to 69)–(50 to 59)	HOXA10 (40 to 49)–(0 to 19), (50 to 59)–(0 to 19), (50 to 59)–(20 to 29), (60 to 69)–(0 to 19), (60 to 69)–(20 to 29), (70 to 79)–(0 to 19)

Table 4. KEGG pathway analysis of DEPS from analysis of 34 gene expression datasets overlap with age-specific findings. Enrichment analysis was done using 974 DEPS overlapped with age-specific analysis. *Up and down regulated genes displayed.

Disease status classification. Using the differential expression analysis results combined with various machine learning models, we were able to identify gene expression signatures for AML that we used for training a predictive KNN model of health status (AML/healthy) with 96.1 ± 4% training accuracy. The model uses a minimal set of 10 DEPS (determined through iteration using an increasing number of DEPS ranked by their absolute effect difference (mean differences between AML and healthy – see Online Methods and Results). The feature set coincides with the top 10 down-regulated DEPS for Analysis 1 (disease), Table 2. The trained KNN model was tested on the 5 covariate datasets, with high performance: 90% accuracy, 92.97% specificity, 84.81% sensitivity, 80.7% precision, and a receiver operating characteristic (ROC) area under the curve (AUC) of 88.89 (Supplementary Fig. S5). The set of genes can have a diagnostic impact, but will need to be validated experimentally, and additionally in same-tissue cohorts.

Our study identified multiple potentially significant DEPS, with age and sex related differences associated with AML. While our findings may generate further hypothesis-driven investigations, we need to also identify the study's limitations: The primary limitation is that the analysis of AML and healthy subjects involved bone-marrow and blood samples respectively in each disease group. We tried to account for this utilizing tissue as an effect in our linear model, and including multiple interactions. Other limitations include an unbalanced AML/healthy ratio, as well as the lack of in-study healthy controls. To address these we attempted to account for batch effects using a dataset-wise iterative batch correction transformation, as discussed in the methods. Finally, we also included binary interactions between the factors in the analysis to account for interaction-related confounding effects. Additionally, the study is limited by the available data, particularly for prognostic utility. The low numbers of PB data, as well as the lack of healthy BM data do not allow for an equal-footing comparison of differences in AML between PB and BM cell gene expression signatures. Furthermore, the prognostic utility of the study is limited by the lack of uniformly reported or sparse clinical data, including progression/relapse-free survival, (in-)effective therapeutic intervention, such as bone marrow transplantation or pharmaceutical treatment, or

mutational status. More uniform reporting of published metadata would greatly improve the utility of publicly available datasets. Additionally, more data would be necessary to fully evaluate mutational status and AML classifications. We anticipate that the use of sequencing data now being generated (e.g. RNA-sequencing data) as well as single cell level gene expression, will allow better concurrent determination of mutational status, global gene expression, and cell-type specific evaluation of patient samples.

In summary, our study successfully integrated multiple datasets to perform an analysis of gene expression in AML, across multiple factors that included disease, sex and age considerations, and identified interesting genes, both known and not previously reported as differentially expressed in each factor. We identified 974 DEPS and 4 associated significant pathways involved in AML, and 70 sex- and 375 age-related DE signatures. Using the 10 of the 974 DEPS, a KNN model allowed AML classification with 90.9% accuracy. We hope that these findings may provide additional relevant targets for further experimental mechanistic studies, and to help identify new markers and therapeutic targets for AML.

Methods

The generalized workflow consisted of five main steps: i) Curation of microarray gene expression data, ii) Preprocessing of raw data files followed by batch effect correction, iii) Predictions of missing annotations using supervised machine learning, iv) Differential gene expression analysis, and v) Pathway analysis, that included gene annotation, and finally gene expression-based prediction of AML (Fig. 1a).

Gene expression data curation and screening criteria. The datasets used in this study were selected from the GEO public repository, maintained by the National Center for Biotechnology Information (NCBI)⁵³ (<https://www.ncbi.nlm.nih.gov/geo/>). To facilitate speed of search and keep up-to-date with possible new and relevant datasets, as soon as they were released, a Python script was used that utilized functions from the Entrez Utilities from Biopython⁵⁴. We used the script to navigate the GEO records, and downloaded microarray gene expression datasets up to October 2018. We additionally utilized Python packages, including Pandas, NumPy, and Matplotlib for data structure, numerical computing for data processing, and data visualization respectively. We used strict inclusion criteria to maintain consistency in each dataset selection, screening for availability of both raw and data annotation files provided, human samples used from untreated subjects, and that the sample source was from either BM or PB. Array platform was restricted to Affymetrix, which was found to have the most available data, and to avoid cross-platform normalization issues. Inclusion criteria and the data curation workflow are illustrated in Fig. 1a,b.

Gene expression datasets used in our analysis. The curation method is summarized in the Supplementary File S3 flowchart and in the Results section. For our analysis we included 34 age-dependent datasets from 32 different studies, 16 included AML and 18 healthy subjects respectively. From the 34 datasets, 32 were produced from Affymetrix GeneChip Human Genome U133 Plus 2.0 (GPL570) and 2 conducted on Affymetrix GeneChip Human Genome U133 Array Set (GPL96 & GPL97) arrays. Table 1 provides detailed information about each dataset, including the number of samples used from each dataset, sample tissue source, as well as the total number of AML patients and healthy subjects. Two studies, GSE12417⁵⁵ and GSE37642^{56–59}, were originally conducted on two different Affymetrix array types (GPL570, and GPL96 & GPL97), so each was separated into two subgroups and each subgroup was considered as individual dataset in our analysis, dataset GSE12417: (i) subgroup 1 included 73 BM and 5 PB samples, and (ii) subgroup 2 included 160 BM and 2 PB. For dataset GSE37642: (i) subgroup 1 included 140 BM and (ii) subgroup 2 included 422 BM samples (Table 1).

Dataset annotation and pre-processing. Figure 1b outlines the workflow of our preliminary data analysis including pre-processing. For each dataset used in our analysis, raw microarray CEL files were downloaded from GEO, metadata was reviewed, and the data was manually curated to guarantee that each array corresponded to either an AML patient or healthy individual, was verified as correctly annotated for sample source (BM or PB), platform technology used, age, sex, and disease state (AML or healthy). For each individual dataset, raw CEL files were grouped and pre-processed together using the RMA (Robust Multi-Array Average) algorithm^{60–62}. Datasets with mixed sample source, i.e. both BM and PB, were pre-processed together irrespective of sample source. Pre-processing consisted of: correction for background noise using RMA background correction on perfect match (PM) raw intensities; quantile normalization to obtain the same empirical distribution of intensities for each array; median polish summarization of probes into probe sets to estimate gene-level expression value; and logarithm base-2 transformations of gene expression values to facilitate data interpretation (normal distributions) and comparisons between arrays. Additionally, the expression data were reduced to 44,754 probe sets that overlapped across all datasets. Finally, within each dataset we standardized across all probe sets, by subtracting the mean and dividing by the standard deviation to obtain a Z-score.

Prediction of missing sex and sample source annotations from curated datasets. From the curated datasets, 805 arrays (802 AML patients and 3 healthy subjects) and 737 arrays (all AML patients) were missing sex or sample source annotations respectively. Without these metadata, we would have to discard the data, which in turn would limit the statistical power for the study, and our ability to correct for biases stemming from individual datasets²⁶. To address this, we used supervised machine learning classifiers to predict both sex and sample annotations. For all predictions, we used ClassificalO⁴¹, a machine learning for classification graphical user interface, which we recently developed, that utilizes the scikit-learn machine learning package in Python⁶³.

To predict sex in pre-processed datasets, 1956 arrays (including both healthy and AML), that included 44,754 probe sets and their annotated sex information, were used to train a logistic regression (LR) classification model,

and to predict 805 sex annotations. Additionally, 2,024 arrays were used to train for sample source, with prediction performed on 737 arrays.

The supervised machine learning LR classifier we used had the following parameters (descriptions based on scikit-learn documentation⁶³):

- *random_state = None*: specifies whether a seed should be used for the pseudorandom generator in selecting training and testing subsets.
- *shuffle = True*: determines that data will be shuffled before splitting for training and testing.
- *penalty = l2*: determines that an L2 norm should be used for penalization.
- *multi_class = ovr*: specifies that a binary problem is fit for each label.
- *solver = liblinear*: specifies that liblinear is the algorithm used in the optimization.
- *max_iter = 100*: maximum number of iterations for the solvers to converge.
- *tol = 0.0001*: tolerance for stopping criteria.
- *fit_intercept = True*: a constant (bias/intercept) is added to the decision function.
- *intercept_scaling = 1.0*: a “synthetic” feature with constant value equal to *intercept_scaling* is appended to the instance vector.
- *Verbose = 0*: turns off verbosity in evaluation.
- *n_jobs = 1*: number of CPU cores used.
- *C = 1.0*: inverse of regularization strength.
- *dual = False*: dual or primal formulation.
- *warm_start = False*: do not reuse the solution of the previous call to fit as initialization.
- *class_weight = None*: all classes assumed to have weight one.

Confusion matrix details, model accuracy and error for training and testing are presented in Supplementary Table S1, Fig. S1 and results in Supplementary Files S1 and S2. To account for training overfitting, we used 10-fold cross-validation on all 1,956 gene expression data arrays for training and validation, implemented automatically in scikit-learn.

Dataset-wise correction approach for batch effects correction. Batch correction was done using a dataset-wise correction. Here we refer to the term “dataset-wise correction” to indicate performing batch correction iteratively on one dataset at a time, against a reference set of datasets chosen to account for variability. To account for the lack of within-study healthy controls in the curated gene expression datasets, we used 5 additional datasets that included within-study controls, with GEO accessions: GSE107968, GSE68172⁶⁴, GSE17054⁶⁵, GSE33223⁶⁶, and GSE15061⁶⁷ (Table 1). We refer to the latter datasets as “covariate” reference datasets, as they were used as the reference datasets in the batch correction. Our approach aimed to balance/distribute the weight of batch effects exerted by each dataset, as this is dependent on the number of observations within a given dataset. Combined, the covariate reference datasets included 613 total arrays (455 AML and 158 healthy controls). We used ComBat²⁷ to correct for study batch effects, as its empirical Bayes-based algorithm uses both scale and mean center based methods, providing an appropriate algorithm²⁷. Covariate reference datasets were treated as the covariate for batch during batch correction, to improve performance in correcting for batch effects rather than biological variation. After batch correction, we used principal component analysis (PCA), visualizing components in both 2 and 3 dimensions, to compare the clustering results for corrections (Supplementary Fig. S2). Covariate reference datasets were removed after the batch correction step and were not part of our downstream linear model analysis (as they lacked age annotations). The covariate datasets were used for testing of the AML prediction models discussed below.

Gene expression linear model analysis. After the batch correction step, we performed differential gene expression analysis on the merged datasets (34 datasets, 16 AML and 18 healthy), using the expression values for all 44,754 common probe sets. The effects of patients’ age, sex, and sample source, including their pairwise interactions were investigated using an analysis of variance (ANOVA)^{28,30}. For each probe set i , where $i = [1, 2, \dots, 44, 754]$, the expression Y_i was modeled with a linear model:

$$Y_i \sim a + s + d + t + a:s + a:d + a:t + s:d + s:t + d:t + \varepsilon,$$

where d is the disease state (AML or healthy), a is age (between 0 to 100 years), s is sex (female or male), t is sample source (BM or PB), and ε is a random error term, and colons represent interactions between factors. We note that the model includes sample source and its interactions to address comparisons involving different tissues in AML and healthy subjects (BM or PB respectively). The selection of using a linear model was based on having multiple factors to capture in the analysis, and also having a large number of samples (by integrating multiple datasets) – in that the Central Limit theorem allows for the assumptions for F-test to hold for ANOVA. We also evaluated fit residuals’ distribution for normality by plotting Quantile-Quantile (QQ) plots and density distributions.

Based on the ANOVA we first identified statistically significant differences for the disease state factor (p-value < 0.01). To identify statistically significant level differences (between AML and healthy) we then carried out post-hoc analyses for each statistically significant probe set using Tukey’s HSD tests implemented in R, (selecting probe sets with Tukey HSD p-value < 0.01). Finally, to focus on biological effects, we filtered the results to have mean difference values (i.e. differences between the means of AML and healthy groups) in the <5% and/or >95% quantiles of the overall mean difference distribution across probe sets. The final set of the results are referred to as differentially expressed probesets (DEPS) with respect to the disease.

Pathway enrichment analysis and functional annotation. We carried out enrichment analysis (overrepresentation) for DEPS using the database DAVID^{37,38} for KEGG signaling pathways^{32–34} and GO functional annotation terms^{35,36}. Pathways and terms identified were deemed statistically significant based on Benjamini-Hochberg adjusted p-value < 0.05.

Using a k nearest neighbor model to predict AML. To predict AML health status, normalized intensities from DEPS (with respect to disease) were used as features for training a k-nearest neighbor (KNN) model (implemented in ClassifiaIO⁴¹). All 34 datasets (16 AML and 18 healthy) were used as training data. Testing of the model was done independently of training on all 5 covariate datasets. The KNN model used the following parameters (please refer to scikit-learn documentation for further details⁶³):

- *random_state = None*: specifies whether a seed should be used for the pseudorandom generator in selecting training and testing subsets.
- *shuffle = True*: determines whether or not data will be shuffled before splitting for training and testing.
- *metric = minkowski and p = 2*: define which metric to use. The *minkowski* metric is using the Minkowski distance of order *p* between two *n*-dimensional vectors $X = \{x_1, x_2, \dots, x_n\}$, and $Y = \{y_1, y_2, \dots, y_n\}$, which is defined as $d(X, Y) = (\sum_i |x_i - y_i|^p)^{\frac{1}{p}}$.
- *weights = uniform*, defines that uniform weights will be used so that all points in each neighborhood are weighted equally.
- *metric_params = None*: additional metric parameters (none used in this case).
- *algorithm = auto*: automatically determines the algorithm to use for computing nearest neighbors, can internally use a *BallTree* or *KDTree* or brute force algorithm.
- *n_neighbors = 30*: number of nearest neighbors to be used.
- *leaf_size = 30*: leaf size passed to *BallTree* or *KDTree* algorithms.
- *n_jobs = 1*: number of parallel jobs to run for neighbors search.

Details of training and testing are given in Supplementary Table S1 and Fig. S4.

To identify a minimum set of DEPS with good predictive power and sensitivity, we first ranked the 974 disease-related DEPS based on the absolute value of their effect size (mean difference between AML and healthy patients). We then iteratively trained and tested a KNN model on the top *n* DEPS (Supplementary Fig. S5), incrementing *n* by one in each iteration. Based on the results, we picked the top 10 effect-ranked DEPS as a minimum set, as the graphs showed stabilization/saturation, with no substantial increase in performance after *n* = 10. We then trained a KNN model using these 10 effect-sorted DEPS, using the same parameters as listed above (Supplementary Table S1, Fig. S5).

Data Availability

The datasets generated in the study, supplementary data, tables, figures and files are available online at <https://doi.org/10.5281/zenodo.3257786>.

Datasets re-analyzed in the study are publicly available on the Gene Expression Omnibus repository, at <https://www.ncbi.nlm.nih.gov/geo/> under the accessions summarized in Table 1.

References

1. De Kouchkovsky, I. & Abdul-Hay, M. Acute myeloid leukemia: a comprehensive review and 2016 update. *Blood Cancer J* **6**, e441, <https://doi.org/10.1038/bcj.2016.50> (2016).
2. Dohner, H., Weisdorf, D. J. & Bloomfield, C. D. Acute Myeloid Leukemia. *N Engl J Med* **373**, 1136–1152, <https://doi.org/10.1056/NEJMra1406184> (2015).
3. Estey, E. H. Acute myeloid leukemia: 2019 update on risk-stratification and management. *Am J Hematol* **93**, 1267–1291, <https://doi.org/10.1002/ajh.25214> (2018).
4. Watts, J. & Nimer, S. Recent advances in the understanding and treatment of acute myeloid leukemia. *F1000Res* **7**, <https://doi.org/10.12688/f1000research.14116.1> (2018).
5. O'Donnell, M. R. *et al.* Acute Myeloid Leukemia, Version 3.2017, NCCN Clinical Practice Guidelines in Oncology. *J Natl Compr Canc Netw* **15**, 926–957, <https://doi.org/10.6004/jnccn.2017.0116> (2017).
6. Kumar, C. C. Genetic abnormalities and challenges in the treatment of acute myeloid leukemia. *Genes Cancer* **2**, 95–107, <https://doi.org/10.1177/1947601911408076> (2011).
7. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2018. *CA Cancer J Clin* **68**, 7–30, <https://doi.org/10.3322/caac.21442> (2018).
8. National Cancer Institute. SEER Cancer Stat Facts: Acute Myeloid Leukemia (Percent of New Cases by Age Group), <https://seer.cancer.gov/statfacts/html/amyl.html>. (Accessed 06.16.19).
9. Arber, D. A. *et al.* The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* **127**, 2391–2405, <https://doi.org/10.1182/blood-2016-03-643544> (2016).
10. Dohner, H. *et al.* Diagnosis and management of AML in adults: 2017 ELN recommendations from an international expert panel. *Blood* **129**, 424–447, <https://doi.org/10.1182/blood-2016-08-733196> (2017).
11. Grimwade, D. & Hills, R. K. Independent prognostic factors for AML outcome. *Hematology Am Soc Hematol Educ Program*, 385–395, <https://doi.org/10.1182/asheducation-2009.1.385> (2009).
12. Dohner, H. Implication of the molecular characterization of acute myeloid leukemia. *Hematology Am Soc Hematol Educ Program*, 412–419, <https://doi.org/10.1182/asheducation-2007.1.412> (2007).
13. Cancer Genome Atlas Research, N. *et al.* Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N Engl J Med* **368**, 2059–2074, <https://doi.org/10.1056/NEJMoa1301689> (2013).
14. Martelli, M. P., Sportoletti, P., Tiacci, E., Martelli, M. F. & Falini, B. Mutational landscape of AML with normal cytogenetics: biological and clinical implications. *Blood Rev* **27**, 13–22, <https://doi.org/10.1016/j.blre.2012.11.001> (2013).
15. Klepin, H. D., Rao, A. V. & Pardee, T. S. Acute myeloid leukemia and myelodysplastic syndromes in older adults. *J Clin Oncol* **32**, 2541–2552, <https://doi.org/10.1200/JCO.2014.55.1564> (2014).

16. Short, N. J., Rytting, M. E. & Cortes, J. E. Acute myeloid leukaemia. *Lancet* **392**, 593–606, [https://doi.org/10.1016/S0140-6736\(18\)31041-9](https://doi.org/10.1016/S0140-6736(18)31041-9) (2018).
17. Reese, N. D. & Schiller, G. J. High-dose cytarabine (HD araC) in the treatment of leukemias: a review. *Curr Hematol Malign Rep* **8**, 141–148, <https://doi.org/10.1007/s11899-013-0156-3> (2013).
18. DiNardo, C. D. *et al.* Venetoclax combined with decitabine or azacitidine in treatment-naïve, elderly patients with acute myeloid leukemia. *Blood* **133**, 7–17, <https://doi.org/10.1182/blood-2018-08-868752> (2019).
19. Stone, R. M. *et al.* Midostaurin plus Chemotherapy for Acute Myeloid Leukemia with a FLT3 Mutation. *N Engl J Med* **377**, 454–464, <https://doi.org/10.1056/NEJMoa1614359> (2017).
20. Perl, A. E. *et al.* Selective inhibition of FLT3 by gilteritinib in relapsed or refractory acute myeloid leukaemia: a multicentre, first-in-human, open-label, phase 1–2 study. *Lancet Oncol* **18**, 1061–1075, [https://doi.org/10.1016/S1470-2045\(17\)30416-3](https://doi.org/10.1016/S1470-2045(17)30416-3) (2017).
21. Meyers, J., Yu, Y., Kaye, J. A. & Davis, K. L. Medicare fee-for-service enrollees with primary acute myeloid leukemia: an analysis of treatment patterns, survival, and healthcare resource utilization and costs. *Appl Health Econ Health Policy* **11**, 275–286, <https://doi.org/10.1007/s40258-013-0032-2> (2013).
22. Ferrara, F. & Schiffer, C. A. Acute myeloid leukaemia in adults. *Lancet* **381**, 484–495, [https://doi.org/10.1016/S0140-6736\(12\)61727-9](https://doi.org/10.1016/S0140-6736(12)61727-9) (2013).
23. Appelbaum, F. R. *et al.* Age and acute myeloid leukemia. *Blood* **107**, 3481–3485, <https://doi.org/10.1182/blood-2005-09-3724> (2006).
24. Walter, M. J. *et al.* Clonal architecture of secondary acute myeloid leukemia. *N Engl J Med* **366**, 1090–1098, <https://doi.org/10.1056/NEJMoa1106968> (2012).
25. Miller, B. G. & Stamatoyannopoulos, J. A. Integrative meta-analysis of differential gene expression in acute myeloid leukemia. *PLoS One* **5**, e9466, <https://doi.org/10.1371/journal.pone.0009466> (2010).
26. Ramasamy, A., Mondry, A., Holmes, C. C. & Altman, D. G. Key issues in conducting a meta-analysis of gene expression microarray datasets. *PLoS Med* **5**, e184, <https://doi.org/10.1371/journal.pmed.0050184> (2008).
27. Chen, C. *et al.* Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS One* **6**, e17238, <https://doi.org/10.1371/journal.pone.0017238> (2011).
28. Pavlidis, P. Using ANOVA for gene selection from microarray studies of the nervous system. *Methods* **31**, 282–289, [https://doi.org/10.1016/S1046-2023\(03\)00157-9](https://doi.org/10.1016/S1046-2023(03)00157-9) (2003).
29. Pavlidis, P. & Noble, W. S. Matrix2png: a utility for visualizing matrix data. *Bioinformatics* **19**, 295–296, <https://doi.org/10.1093/bioinformatics/19.2.295> (2003).
30. Mias, G. In *Mathematica for Bioinformatics: A Wolfram Language Approach to Omics* 193–226 (Springer International Publishing, 2018).
31. Waltman, L. & Schreiber, M. On the calculation of percentile-based bibliometric indicators. *J Am Soc Inf Sci Tec* **64**, 372–379, <https://doi.org/10.1002/asi.22775> (2013).
32. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**, D353–D361, <https://doi.org/10.1093/nar/gkw1092> (2017).
33. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research* **44**, D457–D462, <https://doi.org/10.1093/nar/gkv1070> (2016).
34. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27–30 (2000).
35. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25–29, <https://doi.org/10.1038/75556> (2000).
36. Carbon, S. *et al.* Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research* **45**, D331–D338, <https://doi.org/10.1093/nar/gkw1108> (2017).
37. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research* **37**, 1–13, <https://doi.org/10.1093/nar/gkn923> (2009).
38. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44–57, <https://doi.org/10.1038/nprot.2008.211> (2009).
39. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *J Roy Stat Soc B Met* **57**, 289–300 (1995).
40. Neyman, J. & Pearson, E. S. On the use and interpretation of certain test criteria for purposes of statistical inference. Part II. *Biometrika* **20a**, 263–294, <https://doi.org/10.1093/biomet/20A.3-4.263> (1928).
41. Roushangar, R. & Mias, G. I. ClassificalO: machine learning for classification graphical user interface. *bioRxiv*, 240184, <https://doi.org/10.1101/240184> (2018).
42. Hou, H. A. *et al.* WT1 mutation in 470 adult patients with acute myeloid leukemia: stability during disease evolution and implication of its incorporation into a survival scoring system. *Blood* **115**, 5222–5231, <https://doi.org/10.1182/blood-2009-12-259390> (2010).
43. Ho, P. A. *et al.* Prevalence and prognostic implications of WT1 mutations in pediatric acute myeloid leukemia (AML): a report from the Children's Oncology Group. *Blood* **116**, 702–710, <https://doi.org/10.1182/blood-2010-02-268953> (2010).
44. Udby, L., Calafat, J., Sorensen, O. E., Borregaard, N. & Kjeldsen, L. Identification of human cysteine-rich secretory protein 3 (CRISP-3) as a matrix protein in a subset of peroxidase-negative granules of neutrophils and in the granules of eosinophils. *J Leukocyte Biol* **72**, 462–469 (2002).
45. Izzi, V. *et al.* An extracellular matrix signature in leukemia precursor cells and acute myeloid leukemia. *Haematologica* **102**, E245–E248, <https://doi.org/10.3324/haematol.2017.167304> (2017).
46. Buggins, A. G. *et al.* Microenvironment produced by acute myeloid leukemia cells prevents T cell activation and proliferation by inhibition of NF-kappaB, c-Myc, and pRb pathways. *J Immunol* **167**, 6021–6030 (2001).
47. Rashidi, A. & Uy, G. L. Targeting the Microenvironment in Acute Myeloid. *Leukemia. Curr Hematol Malign R* **10**, 126–131, <https://doi.org/10.1007/s11899-015-0255-4> (2015).
48. Borrow, J. *et al.* The t(7;11)(p15;p15) translocation in acute myeloid leukaemia fuses the genes for nucleoporin NUP98 and class I homeoprotein HOXA9. *Nature Genetics* **12**, 159–167, <https://doi.org/10.1038/ng0296-159> (1996).
49. Andreeff, M. *et al.* HOX expression patterns identify a common signature for favorable AML. *Leukemia* **22**, 2041–2047, <https://doi.org/10.1038/leu.2008.198> (2008).
50. Fan, C., Stendahl, U., Stjernberg, N. & Beckman, L. Association between Orosomucoid Types and Cancer. *Oncology* **52**, 498–500 (1995).
51. Sasi, W., Sharma, A. K. & Mokbel, K. The role of suppressors of cytokine signalling in human neoplasms. *Mol Biol Int* **2014**, 630797, <https://doi.org/10.1155/2014/630797> (2014).
52. Cancer Genome Atlas Research, N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet* **45**, 1113–1120, <https://doi.org/10.1038/ng.2764> (2013).
53. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* **41**, D991–995, <https://doi.org/10.1093/nar/gks1193> (2013).
54. Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423, <https://doi.org/10.1093/bioinformatics/btp163> (2009).
55. Metzeler, K. H. *et al.* An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia. *Blood* **112**, 4193–4201, <https://doi.org/10.1182/blood-2008-02-134411> (2008).

56. Li, Z. *et al.* Identification of a 24-gene prognostic signature that improves the European LeukemiaNet risk classification of acute myeloid leukemia: an international collaborative study. *J Clin Oncol* **31**, 1172–1181, <https://doi.org/10.1200/JCO.2012.44.3184> (2013).
57. Herold, T. *et al.* Isolated trisomy 13 defines a homogeneous AML subgroup with high frequency of mutations in spliceosome genes and poor prognosis. *Blood* **124**, 1304–1311, <https://doi.org/10.1182/blood-2013-12-540716> (2014).
58. Janke, H. *et al.* Activating FLT3 Mutants Show Distinct Gain-of-Function Phenotypes *In Vitro* and a Characteristic Signaling Pathway Profile Associated with Prognosis in Acute Myeloid Leukemia. *Plos One* **9**, <https://doi.org/10.1371/journal.pone.0089560> (2014).
59. Jiang, X. *et al.* Eradication of Acute Myeloid Leukemia with FLT3 Ligand-Targeted miR-150 Nanoparticles. *Cancer Res* **76**, 4470–4480, <https://doi.org/10.1158/0008-5472.CAN-15-2949> (2016).
60. Bolstad, B. M., Irizarry, R. A., Astrand, M. & Speed, T. P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).
61. Irizarry, R. A. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* **31**, e15 (2003).
62. Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249–264, <https://doi.org/10.1093/biostatistics/4.2.249> (2003).
63. Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12**, 2825–2830 (2011).
64. Schneider, V. Z. *et al.* Leukemic progenitor cells are susceptible to targeting by stimulated cytotoxic T cells against immunogenic leukemia-associated antigens (2015).
65. Majeti, R. *et al.* Dysregulated gene expression networks in human acute myelogenous leukemia stem cells. *Proc Natl Acad Sci USA* **106**, 3396–3401, <https://doi.org/10.1073/pnas.0900089106> (2009).
66. Bacher, U. *et al.* Multilineage dysplasia does not influence prognosis in CEBPA-mutated AML, supporting the WHO proposal to classify these patients as a unique entity. *Blood* **119**, 4719–4722, <https://doi.org/10.1182/blood-2011-12-395574> (2012).
67. Mills, K. I. *et al.* Microarray-based classifiers and prognosis models identify subgroups with distinct clinical outcomes and high risk of AML transformation of myelodysplastic syndrome. *Blood* **114**, 1063–1072, <https://doi.org/10.1182/blood-2008-10-187203> (2009).
68. Zatkova, A. *et al.* AML/MDS with 11q/MLL amplification show characteristic gene expression signature and interplay of DNA copy number changes. *Genes Chromosomes Cancer* **48**, 510–520, <https://doi.org/10.1002/gcc.20658> (2009).
69. Tomasson, M. H. *et al.* Somatic mutations and germline sequence variants in the expressed tyrosine kinase genes of patients with de novo acute myeloid leukemia. *Blood* **111**, 4797–4808, <https://doi.org/10.1182/blood-2007-09-113027> (2008).
70. Taskesen, E. *et al.* Prognostic impact, concurrent genetic mutations, and gene expression features of AML with CEBPA mutations in a cohort of 1182 cytogenetically normal AML patients: further evidence for CEBPA double mutant AML as a distinctive disease entity. *Blood* **117**, 2469–2475, <https://doi.org/10.1182/blood-2010-09-307280> (2011).
71. Wouters, B. J. *et al.* Double CEBPA mutations, but not single CEBPA mutations, define a subgroup of acute myeloid leukemia with a distinctive gene expression profile that is uniquely associated with a favorable outcome. *Blood* **113**, 3088–3091, <https://doi.org/10.1182/blood-2008-09-179895> (2009).
72. Figueroa, M. E. *et al.* Genome-wide epigenetic analysis delineates a biologically distinct immature acute leukemia with myeloid/T-lymphoid features. *Blood* **113**, 2795–2804, <https://doi.org/10.1182/blood-2008-08-172387> (2009).
73. Klein, H. U. *et al.* Quantitative comparison of microarray experiments with published leukemia related gene expression signatures. *BMC Bioinformatics* **10**, 422, <https://doi.org/10.1186/1471-2105-10-422> (2009).
74. Luck, S. C. *et al.* Deregulated apoptosis signaling in core-binding factor leukemia differentiates clinically relevant, molecular marker-independent subgroups. *Leukemia* **25**, 1728–1738, <https://doi.org/10.1038/leu.2011.154> (2011).
75. Opel, D. *et al.* Targeting inhibitor of apoptosis proteins by Smac mimetic elicits cell death in poor prognostic subgroups of chronic lymphocytic leukemia. *Int J Cancer* **137**, 2959–2970, <https://doi.org/10.1002/ijc.29650> (2015).
76. Cao, Q. *et al.* BCOR regulates myeloid cell proliferation and differentiation. *Leukemia* **30**, 1155–1165, <https://doi.org/10.1038/leu.2016.2> (2016).
77. Li, L. *et al.* Altered hematopoietic cell gene expression precedes development of therapy-related myelodysplasia/acute myeloid leukemia and identifies patients at risk. *Cancer Cell* **20**, 591–605, <https://doi.org/10.1016/j.ccr.2011.09.011> (2011).
78. Warren, H. S. *et al.* A genomic score prognostic of outcome in trauma patients. *Mol Med* **15**, 220–227, <https://doi.org/10.2119/molmed.2009.00027> (2009).
79. Karlovich, C. *et al.* A longitudinal study of gene expression in healthy individuals. *BMC Med Genomics* **2**, 33, <https://doi.org/10.1186/1755-8794-2-33> (2009).
80. Kong, S. W. *et al.* Characteristics and predictive value of blood transcriptome signature in males with autism spectrum disorders. *PLoS One* **7**, e49475, <https://doi.org/10.1371/journal.pone.0049475> (2012).
81. Sharma, S. M. *et al.* Insights in to the pathogenesis of axial spondyloarthritis based on gene expression profiles. *Arthritis Res Ther* **11**, R168, <https://doi.org/10.1186/ar2855> (2009).
82. Rosell, A. *et al.* Brain perihematoma genomic profile following spontaneous human intracerebral hemorrhage. *PLoS One* **6**, e16750, <https://doi.org/10.1371/journal.pone.0016750> (2011).
83. Schmidt, S. *et al.* Identification of glucocorticoid-response genes in children with acute lymphoblastic leukemia. *Blood* **107**, 2061–2069, <https://doi.org/10.1182/blood-2005-07-2853> (2006).
84. Tasaki, S. *et al.* Multiomic disease signatures converge to cytotoxic CD8 T cells in primary Sjogren's syndrome. *Ann Rheum Dis* **76**, 1458–1466, <https://doi.org/10.1136/annrheumdis-2016-210788> (2017).
85. Leday, G. G. R. *et al.* Replicable and Coupled Changes in Innate and Adaptive Immune Gene Expression in Two Case-Control Studies of Blood Microarrays in Major Depressive Disorder. *Biol Psychiatry* **83**, 70–80, <https://doi.org/10.1016/j.biopsych.2017.01.021> (2018).
86. Shami, R. *et al.* Analysis of blood-based gene expression in idiopathic Parkinson disease. *Neurology* **89**, 1676–1683, <https://doi.org/10.1212/WNL.0000000000004516> (2017).
87. Tasaki, S. *et al.* Multi-omics monitoring of drug response in rheumatoid arthritis in pursuit of molecular remission. *Nat Commun* **9**, 2755, <https://doi.org/10.1038/s41467-018-05044-4> (2018).
88. Clelland, C. L. *et al.* Utilization of never-medicated bipolar disorder patients towards development and validation of a peripheral biomarker profile. *PLoS One* **8**, e69082, <https://doi.org/10.1371/journal.pone.0069082> (2013).
89. Ducieux, J. *et al.* Interferon alpha kinoid induces neutralizing anti-interferon alpha antibodies that decrease the expression of interferon-induced and B cell activation associated transcripts: analysis of extended follow-up data from the interferon alpha kinoid phase I/II study. *Rheumatology (Oxford)* **55**, 1901–1905, <https://doi.org/10.1093/rheumatology/kew262> (2016).
90. Lauwerys, B. R. *et al.* Down-regulation of interferon signature in systemic lupus erythematosus patients by active immunization with interferon alpha-kinoid. *Arthritis Rheum* **65**, 447–456, <https://doi.org/10.1002/art.37785> (2013).
91. Xiao, W. *et al.* A genomic storm in critically injured humans. *J Exp Med* **208**, 2581–2590, <https://doi.org/10.1084/jem.20111354> (2011).
92. Zhou, B. *et al.* Analysis of factorial time-course microarrays with application to a clinical study of burn injury. *Proc Natl Acad Sci USA* **107**, 9923–9928, <https://doi.org/10.1073/pnas.1002757107> (2010).

Acknowledgements

R.R. has been supported by The Paul and Daisy Soros Fellowship for New Americans. G.I.M. and research reported in this publication have been supported by a Jean P. Schultz Endowed Biomedical Research Fund award, and previously NIH/NHGRI Grant HG0006785. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health and other funders.

Author Contributions

R.R. and G.I.M. wrote the main manuscript text and prepared the figures. All authors reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-019-48872-0>.

Competing Interests: G.I.M. has consulted for Colgate-Palmolive North America and received compensation. R.R. is founder and chief architect at MetaGentex.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2019