Stratified computational meta-analysis of 2213 acute myeloid leukemia patients reveals age- and sex-dependent gene expression signatures

Raeuf Roushangar [1, 2], George I. Mias [1, 2*]

[1] Department of Biochemistry and Molecular Biology,

[2] Institute for Quantitative Health Science and Engineering,

Michigan State University, East Lansing MI 48824, USA

*Corresponding author

E-mail: gmias@msu.edu (GM)

In 2018 alone, an estimated 20,000 new acute myeloid leukemia (AML) patients were diagnosed, in the United States, and over 10,000 of them 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 in patients older than 60 years old as much as 70% of patients will die within a year of diagnosis. In this study, we conducted stratified computational meta-analysis of 2,213 acute myeloid leukemia patients compared to 548 healthy individuals, using curated publicly available data. We carried out analysis of variance of normalized batch corrected data, including considerations for disease, age, tissue and sex. We identified 974 differentially expressed probe sets and 4 significant pathways associated with AML. Additionally, we identified 70 sex- and 375 age-related probe set expression signatures relevant to AML. Finally, we used a machine learning model (KNN model) to classify AML patients compared to healthy individuals with 90+% achieved accuracy. Overall our findings provide a new reanalysis of public datasets, that enabled the identification of potential new gene sets relevant to AML that can potentially be used in future experiments and possible stratified disease diagnostics.

INTRODUCTION

Acute myeloid leukemia (AML) is a heterogeneous malignant disease of the hematopoietic system myeloid cell lineage. AML is best characterized by the 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[1,2]. 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[2,3]. The number of new AML cases is increasing each year – in 2018 alone, there have been an estimated about 20,000 new diagnosed AML patients, over 10,000 of them will die from the disease[4].

According to the 2016 World Health Organization (WHO) newly revised myeloid neoplasms and acute leukemia classification system[5], AML prognosis criteria for classification is highly dependent on the presence of chromosomal abnormalities, including chromosomal deletions, duplications, translocations, inversions, and gene fusion. Mostly, AML is diagnosed through microscopic, cytogenetics, and molecular genetic analyses of patients' blood and/or bone marrow samples. Microscopic examination is 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 CBFB-MYH11), and mutations in genes frequently mutated in AML (e.g., NPM1, CEBPA, RUNX1, FLT3)[6-8]. These cytogenetic and molecular genetic analyses are used to identify prognosis markers that can be used to classify AML patients into three risk categories: favorable, intermediate, and unfavorable. The largest group of AML patients (almost 50%) however, present normal karyotype and lack genetic abnormalities[7-10]. These patients are classified as intermediate risk, and often have heterogeneous clinical outcome with standard therapy with risk of AML relapse[11].

Additionally, AML prognosis worsens as age increases, and older patients respond less to current treatments with poorer clinical outcomes than their younger counterparts[12,13]. AML can occur in people of all ages but is primarily diagnosed among the elderly (>60 years), with a median age of 68 year at diagnosis[4]. Recent advances in AML biology expanded our understanding of its complex genetic landscape and led to significant improvement in prognoses and therapeutic strategy for younger patients[13,14]. However, in patients older than 60 years old, prognoses remain grim and therapeutic strategy has been nearly the same for more than 30 years[2,6,13-15]. Approximately 70% of AML patients 65 years of age or older die within a year following diagnosis[16]. 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's age[14,17,18]. Taking into consideration age considerations in the identification of changes in AML global gene expression can lead to improved early diagnosis and improvement in treatment approaches for elderly patients. Further complicating, AML has multiple driver mutations and competing clones that evolve over time, making it a very dynamic disease[19,20]

Multiple gene expression analyses of AML have been carried out, 25 of which these have been systematically compared by Miller and Stamatoyannopoulos[21], who analyzed information on 4918 genes, and identified 25 genes reported across multiple, with potential prognostic features. In this study, we performed comprehensive gene expression meta-analysis of 2213 acute myeloid leukemia patients and 548 healthy subjects using 34 publicly available gene expression microarray datasets (following strict inclusion criteria) to identify disease, sex- and age-related gene expression changes associated with AML. We identified sex- and age-related gene expression signatures that show similar alteration in gene expression levels and associated signaling pathways in AML and have used our results (gene sets) to predict 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 sex- and age-dependent biomarkers for AML prognosis compared to healthy, as well as the identification of new treatment targets with mechanisms of action different from those used in conventional chemotherapy

RESULTS

Data curation and gene expression preprocessing.

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, 643 selected (577 were excluded as non-Affymetrix, various platform arrays). From the 66 remaining, 34 studies were excluded due to lack of metadata, non-peripheral blood and non-bone marrow tissues, cell line or cell-type specific, 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. The sets were re-analyzed, starting from raw data, to perform a gene expression analysis of variance and functional pathway enrichment analysis (see online Methods). Table 1 provides a description on each dataset with a sub-table summary of all curated data used in our current study. After pre-processing each individual data set separately, Fig. 1b, we performed the statistical analysis on 44,754 probe sets which were common across all samples (Affymetrix expression array 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 information regarding the sample source (i.e. tissue, either bone marrow [BM] or peripheral blood [PB] annotation). To predict the missing sex and sample source meta-data, we trained and validated various machine learning supervised models, including logistic regression (LR) classification models. The prediction of missing annotations for these arrays was essential in our study, to increase the sample size, and statistical power[22]. The models were trained and verified using our annotated preprocessed expression data. Model training, parameters used in training, validation for this analysis are discussed in the Methods. Results from model training and predictions, including confusion matrix, model accuracy, and error can be viewed in Supplementary Table S1 online and results from classification for missing annotation are presented in Supplementary files 1 and 2 for sample source and sex annotations respectively.

Batch correction

Our pre-processed data, AML and healthy, were processed using a “dataset-wise” batch effect correction approach. The datasets used in 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. Assuming the batch effects due to each data set is a function of the number of samples in the data set (weight), normalizing sets of unevenly sized datasets may lead to 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 2), and pre-processed exactly as our curated data sets. These were used together with each of the remaining datasets to batch correct each dataset with respect the covariate reference using ComBat[23]. 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. The batch effect correction results were then compared to clustering results prior to batch effect correction (Supplementary Fig. 1)

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

Gene expression meta-analysis of AML disease state.

Following batch correction, we performed an analysis of differential expression (DE) on 34 data sets including 2213 AML patients and 548 healthy controls. Analysis of Variance (ANOVA)[24-26] was performed according to a linear model (see method section **Meta-analysis**), including factors for age, sample source (adjust for differences in tissue between AML and healthy), and sex, as well as binary interactions thereof. In the analysis we used probe sets to avoid assumptions on averaging over multiple probe sets corresponding the same gene symbol. 974 Statistically significant differentially expressed probe sets (DEPS) (with genes corresponding to 964 unique gene symbols) for AML versus healthy were selected based on a Bonferroni[27] adjusted p-value < 0.01 (accounting for multiple hypothesis testing), in conjunction with a two-tailed 5% quantile selection[28] based on the mean difference distribution between AML-healthy group comparisons (post-hoc analyses using Tukey's Honestly Significant Difference (HSD). The heatmap (Fig. 2a) shows the hierarchical clustering of genee 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 < $4.11\text{x}10^{-11}$ was the most DE up-regulated gene while CRISP3 (cysteine-rich secretory protein 3) with mean difference of -0.52 and adjusted p-value < $4.11\text{x}10^{-11}$ 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 Bonferroni 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 online.

(ii) Gene enrichment analysis AML disease state DEPS.

To identify signaling pathways associated DEPS in AML, gene enrichment analysis was performed on all 974 DEPS combined. Pathway over-representation analysis in Kyoto Encyclopedia of Genes and Genomes (KEGG)[29-31] signaling pathways, and Gene Ontology (GO) term[32,33] were carried out using the Database for Annotation, Visualization and Integrated Discovery (DAVID)[34,35]. Four KEGG signaling pathways were identified as enriched (Benjamini and Hochberg[36] 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. 2a-d), including unadjusted p-values and Benjamini and Hochberg[36] adjusted p-values. 56 DEPS including 27 up- and 29 down-regulated (Fig. 2c) were associated these signaling pathways, and the heatmap of their mean differences is shown in Fig. 2d. From our gene enrichment analysis for overrepresented biological GO terms, 21 GO terms were statistically significant 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 entire list of our enrichment analysis results (statistically significant over-representation in KEGG and GO terms) can be found as Supplementary Table S3 online.

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

Further analysis of gene expression and pathways enrichment were conducted in order to characterize sex- and age-specific gene expression changes in AML patients compared to healthy individuals, (i) **Analysis 2a:** “**Sex-relevance differential gene expression meta-analysis and associated signaling pathways in AML”,** and (ii) **Analysis 2b:** “**Age-dependent differential gene expression meta-analysis and associated signaling pathways in AML”**. We used the same filtering criteria in both analyses as those used in analysis 1 for significant 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 Bonferroni adjusted p-value from Tukey’s HSD < $2.2x10^{-7}$ (=0.01/44,754 probe sets tested).

Analysis 2a. Sex-relevance differential gene expression meta-analysis and associated signaling pathways in AML.

Gene expression meta-analysis was also used to identify DEPS that show sex differences between male AML patients as compared to female AML patients. 266 DEPS were regarded statistically significant (p-value < $2.2x10^{-7}$). A list of all 266 DEPS (including whether higher in either males or females, gene title and symbol, male-female mean difference, and Bonferroni corrected p-value) can be found as Supplementary Table S3 online. 70 DEPS were found to overlap between analysis 1 (AML disease state) and analysis 2 (Sex-relevance in AML). Figure 3a shows these 70 DEPS with gene symbol annotations, and their mean difference values in the heatmap, which displays differences in significance for a common DEPS in both analyses 1 and 2. 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 Figure 3c.

For enrichment analysis, we searched for common intersections in KEGG pathways and GO terms between the sex meta-analysis and the 974 DE probe sets from disease state in AML meta-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, and MS4A1 was higher in females and found in 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. The entire list of our enrichment analysis (statistically significant KEGG and GO terms) can be found as Supplementary Table S4.

Analysis 2b. Age-dependent differential gene expression meta-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 meta-analysis, 1395 unique probe sets across all age-groups were identified as statistically significant (Bonferroni adjusted 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 meta-analysis, accounting for an overall 1400 binary comparisons between the multiple age groups deemed statistically significant, based on Tukey HSD tests between age-group pairs. The entire list of 1400 identified pairwise differences between age groups and associated probe set/gene information can be found as Supplementary Table S6 online. The top 10 up- and down- regulated DEPS (labeled with gene symbols) from this analysis are shown in Fig. 4a. Additionally, Fig. 4b shows 75 DEPS with gene symbols identified to have appeared specifically in one age-group comparison. Utilizing results for KEGG analysis for signaling pathways from analysis 1, Fig. 4c shows 17 DE genes identified in all 4 KEGG pathways according to age groups (see also 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, Figure 4d shows the mean difference of 25 DEPS with respect to the 0-19 baseline across all other groups. The mean difference values between AML and healthy are shown in the right-most column of Fig. 4a, b and d for reference.

AML Classification Machine Learning Model

We used the 974 DEPS to train a k-nearest neighbor (KNN) algorithm in ClassificaIO[37]. All 34 datasets (16 AML and 18 healthy) were used for training, and testing was done on all 5 covariate reference datasets, include AML and healthy subjects. The KNN algorithm trained was 98% accurate, and >90% accurate in testing results (see online Methods for parameters and also details in Supplementary File 3).

DISCUSSION

In the present study, we aimed to establish, disease sex-linked and age-dependent biomarkers from genes with similar changes in gene expression levels and associated signaling pathways relevant to AML. Utilizing microarray gene expression data and combined with various machine learning models, respectively, our biomarkers were indicative of prognostic signature for AML prediction compared to healthy with 90+% achieved accuracy. We re-analyzed data aggregated from our curation of 34 publicly available microarray gene expression datasets covering 2213 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. We note here that 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: primarily 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.

i) Analysis 1: Gene expression meta-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 baseline indicator for AML disease state. 974 DEPS (487 up- and 487 down-regulated) were identified as significantly differentially expressed between AML patients and healthy individuals (Bonferroni adjusted p-value < 0.01). 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 with their respected Tukey's HSD mean difference and Bonferroni p-adjusted values. As shown in Figure 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 to cellular development and cell survival, and it has been known to be highly expressed with an oncogenic role in AML[38,39], 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 gland and bone marrow[40]. Recently, 80 genes were reported as "extracellular matrix specific genes" in leukemia, and CRISP3 was among the downregulated DE genes reported[41]. CRISP3 associations with AML merit further investigation.

The enrichment analysis for GO terms of the 974 DE probe sets (Fig. 2c) results showed 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[42,43].

Biological process (BP) category, GO terms included inflammatory and immune 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, and likely affects many biological processes associated to 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). The enriched pathways Fig. 2d shows the mean difference values of the 56 DE pathway-associated genes, including 27 genes up- and 29 down-regulated. These KEGG pathways are known to be involved in tumorigenesis. Additionally, the majority of the associated DE genes from AML meta-analysis with the identified signaling pathways are known to be abnormally expressed in AML. These findings are consistent with findings 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 studies[21], 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.

ii) Analysis 2a: Sex-dependent gene expression meta-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 top10 up- and down-regulated DE genes with respect to females include (Fig. 3c) – DDX3Y (DEAD-Box Helicase 3 Y-Linked), EIF1AY (Eukaryotic Translation Initiation Factor 1A Y-Linked), KDM5D (Lysine Demethylase 5D), RPS4Y1 (Ribosomal Protein S4 Y-Linked 1) with higher expression in males compared to females, and XIST (X Inactive Specific Transcript), TSIX (TSIX Transcript, XIST Antisense RNA), and PRKX (Protein Kinase X-Linked) were as higher in females. These genes are known to be sex-specific and show such 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.

iii) *Analysis 2b: Age-dependent gene expression meta-analysis and associated signaling pathways in AML compared to healthy individuals*, was carried out to identify common set of age-dependent genes and associated signaling pathways and to explore age-dependent trends in gene expression in AML. The age-dependent meta-analysis in AML using ANOVA, identified 1,395 DEPS (Bonferroni adjusted p-value <0.01). 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. 4a), and identified an overlap of 375 DEPS (Bonferroni adjusted p.value <0.01). As shown in Figure 4b, the top 10 most and least DE age-associate genes in AML according to the mean difference values in seven age-groups, including their corresponding values from AML disease state in column "AML - healthy" for comparisons. Interestingly, CRISP3 was among the down regulated genes specifically and involved in this analysis as well, specifically associated with differences in younger age groups, 20 to 49 years of age as compared to 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 to embryonic development and hematopoiesis, and associated with chromosomal abnormalities translocation and over-expression in AML[44,45]. Also identified with age-specific DE, was ORM1, which in Analysis 1 was among the top-10 most 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 inflammation[46]. 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.

In summary, our study successfully integrated multiple datasets to perform a study 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 974 DEPS, a KNN model allowed AML with 91.7% 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 annotation data using supervised machine learning, iv) Differential gene expression analysis, and v) Gene enrichment for pathway analysis that includes gene annotation, and finally gene expression-based prediction of AML (Fig. 1a).

Gene expression data curation and screening criteria.

Datasets used in this study were selected from the GEO public repository, maintained by the National Center for Biotechnology Information (NCBI)[47] (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 Biopython[48]. We used the script to navigate the GEO records, and download microarray gene expression datasets up to 10/18. 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, screen for availability of both raw and meta-data annotation files provided, human samples used from untreated subjects, and that the sample source was from either bone marrow (BM) and/or peripheral blood (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. 1 a-b.

Gene expression data sets used in our analysis.

The curation method is summarized in the Supplementary File 4 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 data set, 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, GSE12417[49] and GSE37642[50-53], 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 meta-analysis, data set GSE12417: (i) subgroup 1 included 73 BM and 5 PB samples, and (ii) subgroup 2 included 160 BM and 2PB. For dataset GSE37642 (i) subgroup 1 included 140 BM and (ii) subgroup 2 422 BM samples (Table 1).

Dataset annotation and preprocessing.

Figure 1b outlines the workflow of our preliminary data analysis including preprocessing. 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 and each array, which corresponded to either an AML patient or healthy individual, was verified and correctly annotated for sample source (BM or PB), platform technology used, age, sex, and disease state (AML or healthy). Raw CEL files from individual datasets were individually pre-processed using the RMA (Robust Multi-Array Average) algorithm[54-56]. Datasets with mixed sample source, i.e both BM and PB, were pre-processed together irrespective of sample source. Preprocessing 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, our expression data were first reduced to 44,754 probe sets that are common to and appeared in all data. Data sets were z-score standardized across all probe sets and arrays.

Prediction of missing sex- and sample source annotations from curated data sets.

805 arrays (802 from AML patients and 3 were healthy subjects) of curated data were not annotated for sex, while 737 arrays (all AML patients) were missing sample source information. 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 datasets[22]. To address this, we used supervised machine learning classifiers to predict metadata. For all prediction, we used ClassificalO[37], a machine learning for classification user interface, which we recently developed, to carry out the machine learning classification analyses utilizing the sklearn package in Python[57]

To predict sex pre-processed data sets, 1956 arrays (including both healthy and AML), that include 44,754 probe sets and their annotated sex information were used to train logistic regression (LR) classification models, and to predict 805 sex annotations. Additionally, 2024 arrays were used to train for sample source, and the prediction was performed on 737 arrays.

The supervised machine learning LR classifier we used with the following parameters:

random_state = None, shuffle = True, penalty = l2, multi_class = ovr, solver = liblinear, max_iter= 100, tol = 0.0001, intercept_scaling = 1.0, verbose = 0, n_jobs = 1, C = 1.0, fit_intercept = True, dual = False, warm_start = False, class_weight = None

The trained models for classification of missing sex and sample source annotation from curated data achieved > 95% classification accuracy with ~ 3-5% classification errors. Confusion matrix details, model accuracy and error for training and testing are presented in Supplementary Table S1 online, and results in Supplementary files 1 and 2. To account for training overfitting, we used 10-fold cross-validation on all 1,956 gene expression data arrays for training and validation.

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. We used this approach to account for the lack within-study healthy controls in the curated gene expression datasets. To address this issue, we used 5 additional datasets the included within-study controls, GEO accessions: GSE107968, GSE68172[58], GSE17054[59], GSE33223[60], and GSE15061[61] (Table 1B). We refer to the latter datasets hereafter as "covariate" reference datasets, as they were 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, totaling 455 AML and 158 healthy controls. We used ComBat[23] to correct for study batch effects, as its empirical Bayes-based algorithm uses both scale and mean center based methods, providing an appropriate algorithm[23]. 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. Covariate reference datasets were removed after the batch correction step and were not part of our downstream meta-analysis. (Supplementary Fig. S1).

Gene expression meta-analysis.

After batch correction step, we performed gene expression meta-analysis for differential expression on the merged datasets (34 data sets, 16 AML and 18 healthy), where the expression values for all 44,754 common probe sets were aggregated. The effects of patients' age, sex, and sample source, including their pairwise interactions were investigated using an analysis of variance (ANOVA)[8,62] . For each gene *i*, where *i*=[1,…44,754], the gene expression probe set Y_i was modeled computationally as 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. 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).

From the ANOVA analysis, genes were deemed to be disease state statistically significant (differentially expressed) if they displayed ANOVA Bonferroni-adjusted p-value < 0.01. Post-hoc analysis for significant genes was conducted for comparisons (between groups) using Tukey's Honestly Significant Difference (HSD) tests. Additionally, we performed a quantile-based effect filter, were genes were deemed to show biological effects in our analysis if they displayed mean difference values in the <5% and/or > 95% quantiles of the mean difference distributions of the binary group comparisons. Based on the post-hoc analysis, genes were deemed to be statistically significantly (up- or down-regulated) if they displayed Tukey HSD using a Bonferroni adjusted cutoff for p-value < 0.01/44,754.

Functional and pathway enrichment analysis

We carried our enrichment analysis for DEPS using the Database DAVID[34,35], the KEGG database[29-31] for signaling pathways, GO terms functional annotation for over representation of biological function [32,33] were utilized and signaling pathways were deemed significant based on Benjamini-Hochberg adjusted p-value < 0.05.

Using a k-nearest neighbor model to predict AML

Before gene expression data passed to the k-nearest neighbor (KNN) algorithm to train, gene expression signatures resulted from our meta-analysis were used to extract expression values. KNN in ClassificaIO[37] was used to carry out this analysis. All 34 data sets (16 AML and 18 healthy) were used for training, and testing was done on all 5 covariate data sets, include AML and healthy subjects.

Dependent, target , and testing data files were prepared in accordance with ClassificaIO[37] user guide. The KNN model used the following parameters (Supplementary File 3):

random_state = None, shuffle = True, metric = minkowski, weights = uniform, algorithm = auto, n_neighbors = 5, leaf_size = 30, n_jobs = 1, p = 2, metric_params = None

The trained model was 98% accurate, while testing was 91.7% accurate (details of training and testing are given in Supplementary File 3.

DATA AVAILABILITY STATEMENT

The datasets generated in the study, supplementary data, tables, figures and files are available online at http://doi.org/10.5281/zenodo.1492796

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

REFERENCES

1 Kumar, C. C. Genetic abnormalities and challenges in the treatment of acute myeloid leukemia. *Genes Cancer* **2**, 95-107, doi:10.1177/1947601911408076 (2011).

2 De Kouchkovsky, I. & Abdul-Hay, M. 'Acute myeloid leukemia: a comprehensive review and 2016 update'. *Blood Cancer J* **6**, e441, doi:10.1038/bcj.2016.50 (2016).

3 Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2018. *CA Cancer J Clin* **68**, 7-30, doi:10.3322/caac.21442 (2018).

4 Institute, N. C. SEER Cancer Stat Facts: Acute Myeloid Leukemia (Percent of New Cases by Age Group). [https://seer.cancer.gov/statfacts/html/amyl.html]. ((accessed 11.30.18), 2011-2015).

5 Arber, D. A. *et al.* The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia. *Blood* **127**, 2391-2405, doi:10.1182/blood-2016-03-643544 (2016).

6 Dohner, H. *et al.* Diagnosis and management of acute myeloid leukemia in adults: recommendations from an international expert panel, on behalf of the European LeukemiaNet. *Blood* **115**, 453-474, doi:10.1182/blood-2009-07-235358 (2010).

7 Grimwade, D. & Hills, R. K. Independent prognostic factors for AML outcome. *Hematology Am Soc Hematol Educ Program*, 385-395, doi:10.1182/asheducation-2009.1.385 (2009).

8 Dohner, H. Implication of the molecular characterization of acute myeloid leukemia. *Hematology Am Soc Hematol Educ Program*, 412-419, doi:10.1182/asheducation-2007.1.412 (2007).

9 Walter, M. J. *et al.* Acquired copy number alterations in adult acute myeloid leukemia genomes. *Proc Natl Acad Sci U S A* **106**, 12950-12955, doi:10.1073/pnas.0903091106 (2009).

10 Suela, J., Alvarez, S. & Cigudosa, J. C. DNA profiling by arrayCGH in acute myeloid leukemia and myelodysplastic syndromes. *Cytogenet Genome Res* **118**, 304-309, doi:10.1159/000108314 (2007).

11 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, doi:10.1016/j.blre.2012.11.001 (2013).

12 Klepin, H. D., Rao, A. V. & Pardee, T. S. Acute myeloid leukemia and myelodysplastic syndromes in older adults. *J Clin Oncol* **32**, 2541-2552, doi:10.1200/JCO.2014.55.1564 (2014).

13 Short, N. J., Rytting, M. E. & Cortes, J. E. Acute myeloid leukaemia. *Lancet* **392**, 593-606, doi:10.1016/S0140-6736(18)31041-9 (2018).

14 Dohner, H., Weisdorf, D. J. & Bloomfield, C. D. Acute Myeloid Leukemia. *N Engl J Med* **373**, 1136-1152, doi:10.1056/NEJMra1406184 (2015).

15 Reese, N. D. & Schiller, G. J. High-dose cytarabine (HD araC) in the treatment of leukemias: a review. *Curr Hematol Malig Rep* **8**, 141-148, doi:10.1007/s11899-013-0156-3 (2013).

16 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, doi:10.1007/s40258-013-0032-2 (2013).

17 Ferrara, F. & Schiffer, C. A. Acute myeloid leukaemia in adults. *Lancet* **381**, 484-495, doi:10.1016/S0140-6736(12)61727-9 (2013).

18 Appelbaum, F. R. *et al.* Age and acute myeloid leukemia. *Blood* **107**, 3481-3485, doi:10.1182/blood-2005-09-3724 (2006).

19 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, doi:10.1056/NEJMoa1301689 (2013).

20 Walter, M. J. *et al.* Clonal architecture of secondary acute myeloid leukemia. *N Engl J Med* **366**, 1090-1098, doi:10.1056/NEJMoa1106968 (2012).

21 Miller, B. G. & Stamatoyannopoulos, J. A. Integrative meta-analysis of differential gene expression in acute myeloid leukemia. *PLoS One* **5**, e9466, doi:10.1371/journal.pone.0009466 (2010).

22 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, doi:10.1371/journal.pmed.0050184 (2008).

23 Chen, C. *et al.* Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. *PLoS One* **6**, e17238, doi:10.1371/journal.pone.0017238 (2011).

24 Pavlidis, P. Using ANOVA for gene selection from microarray studies of the nervous system. *Methods* **31**, 282-289, doi:10.1016/S1046-2023(03)00157-9 (2003).

25 Pavlidis, P. & Noble, W. S. Matrix2png: a utility for visualizing matrix data. *Bioinformatics* **19**, 295-296, doi:DOI 10.1093/bioinformatics/19.2.295 (2003).

26 Mias, G. in *Mathematica for Bioinformatics: A Wolfram Language Approach to Omics* 193-226 (Springer International Publishing, 2018).

27 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, doi:DOI 10.1093/biomet/20A.3-4.263 (1928).

28 Waltman, L. & Schreiber, M. On the calculation of percentile-based bibliometric indicators. *J Am Soc Inf Sci Tec* **64**, 372-379, doi:10.1002/asi.22775 (2013).

29 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, doi:10.1093/nar/gkw1092 (2017).

30 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, doi:10.1093/nar/gkv1070 (2016).

31 Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30 (2000).

32 Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29, doi:10.1038/75556 (2000).

33 Carbon, S. *et al.* Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research* **45**, D331-D338, doi:10.1093/nar/gkw1108 (2017).

34 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, doi:10.1093/nar/gkn923 (2009).

35 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, doi:10.1038/nprot.2008.211 (2009).

36 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).

37 Roushangar, R. & Mias, G. I. ClassificaIO: machine learning for classification graphical user interface. *bioRxiv*, doi:10.1101/240184 (2017).

38 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, doi:10.1182/blood-2009-12-259390 (2010).

39 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, doi:10.1182/blood-2010-02-268953 (2010).

40 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).

41 Izzi, V. *et al.* An extracellular matrix signature in leukemia precursor cells and acute myeloid leukemia. *Haematologica* **102**, E245-E248, doi:10.3324/haematol.2017.167304 (2017).

42 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).

43 Rashidi, A. & Uy, G. L. Targeting the Microenvironment in Acute Myeloid Leukemia. *Curr Hematol Malig R* **10**, 126-131, doi:10.1007/s11899-015-0255-4 (2015).

44 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, doi:DOI 10.1038/ng0296-159 (1996).

45 Andreeff, M. *et al.* HOX expression patterns identify a common signature for favorable AML. *Leukemia* **22**, 2041-2047, doi:10.1038/leu.2008.198 (2008).

46 Fan, C., Stendahl, U., Stjernberg, N. & Beckman, L. Association between Orosomucoid Types and Cancer. *Oncology* **52**, 498-500 (1995).

47 Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res* **41**, D991-995, doi:10.1093/nar/gks1193 (2013).

48 Cock, P. J. A. *et al.* Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422-1423, doi:10.1093/bioinformatics/btp163 (2009).

49 Metzeler, K. H. *et al.* An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia. *Blood* **112**, 4193-4201, doi:10.1182/blood-2008-02-134411 (2008).

50 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, doi:10.1200/JCO.2012.44.3184 (2013).

51 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, doi:10.1182/blood-2013-12-540716 (2014).

52 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**, doi:ARTN e89560
10.1371/journal.pone.0089560 (2014).

53 Jiang, X. *et al.* Eradication of Acute Myeloid Leukemia with FLT3 Ligand-Targeted miR-150 Nanoparticles. *Cancer Res* **76**, 4470-4480, doi:10.1158/0008-5472.CAN-15-2949 (2016).

54 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).

55 Irizarry, R. A. *et al.* Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res* **31**, e15 (2003).

56 Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4**, 249-264, doi:10.1093/biostatistics/4.2.249 (2003).

57 Pedregosa, F. *et al.* Scikit-learn: Machine Learning in Python. *J Mach Learn Res* **12**, 2825-2830 (2011).

58 Schneider, V. Z., L.; Markus, R.; Fekete, N.; Schrezenmeier, H.; Erle, A. ; Lars, B.; Hofmann, S.; Götz, M.; Döhner, K.; Ihme, S.; Döhner, H.; Buske, C.; Feuring-Buske, M.; Greiner, J. Leukemic progenitor cells are susceptible to targeting by stimulated cytotoxic T cells against immunogenic leukemia-associated antigens. (2015).

59 Majeti, R. *et al.* Dysregulated gene expression networks in human acute myelogenous leukemia stem cells. *Proc Natl Acad Sci U S A* **106**, 3396-3401, doi:10.1073/pnas.0900089106 (2009).

60 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, doi:10.1182/blood-2011-12-395574 (2012).

61 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, doi:10.1182/blood-2008-10-187203 (2009).

62 Tasaki, S. *et al.* Multi-omics monitoring of drug response in rheumatoid arthritis in pursuit of molecular remission. *Nat Commun* **9**, 2755, doi:10.1038/s41467-018-05044-4 (2018).

63 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, doi:10.1002/gcc.20658 (2009).

64 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, doi:10.1182/blood-2007-09-113027 (2008).

65 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, doi:10.1182/blood-2010-09-307280 (2011).

66 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, doi:10.1182/blood-2008-09-179895 (2009).

67 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, doi:10.1182/blood-2008-08-172387 (2009).

68 Klein, H. U. *et al.* Quantitative comparison of microarray experiments with published leukemia related gene expression signatures. *BMC Bioinformatics* **10**, 422, doi:10.1186/1471-2105-10-422 (2009).

69 Luck, S. C. *et al.* Deregulated apoptosis signaling in core-binding factor leukemia differentiates clinically relevant, molecular marker-independent subgroups. *Leukemia* **25**, 1728-1738, doi:10.1038/leu.2011.154 (2011).

70 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, doi:10.1002/ijc.29650 (2015).

71 Cao, Q. *et al.* BCOR regulates myeloid cell proliferation and differentiation. *Leukemia* **30**, 1155-1165, doi:10.1038/leu.2016.2 (2016).

72 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, doi:10.1016/j.ccr.2011.09.011 (2011).

73 Warren, H. S. *et al.* A genomic score prognostic of outcome in trauma patients. *Mol Med* **15**, 220-227, doi:10.2119/molmed.2009.00027 (2009).

74 Karlovich, C. *et al.* A longitudinal study of gene expression in healthy individuals. *BMC Med Genomics* **2**, 33, doi:10.1186/1755-8794-2-33 (2009).

75 Kong, S. W. *et al.* Characteristics and predictive value of blood transcriptome signature in males with autism spectrum disorders. *PLoS One* **7**, e49475, doi:10.1371/journal.pone.0049475 (2012).

76 Sharma, S. M. *et al.* Insights in to the pathogenesis of axial spondyloarthropathy based on gene expression profiles. *Arthritis Res Ther* **11**, R168, doi:10.1186/ar2855 (2009).

77 Rosell, A. *et al.* Brain perihematoma genomic profile following spontaneous human intracerebral hemorrhage. *PLoS One* **6**, e16750, doi:10.1371/journal.pone.0016750 (2011).

78 Schmidt, S. *et al.* Identification of glucocorticoid-response genes in children with acute lymphoblastic leukemia. *Blood* **107**, 2061-2069, doi:10.1182/blood-2005-07-2853 (2006).

79 Tasaki, S. *et al.* Multiomic disease signatures converge to cytotoxic CD8 T cells in primary Sjogren's syndrome. *Ann Rheum Dis* **76**, 1458-1466, doi:10.1136/annrheumdis-2016-210788 (2017).

80 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, doi:10.1016/j.biopsych.2017.01.021 (2018).

81 Shamir, R. *et al.* Analysis of blood-based gene expression in idiopathic Parkinson disease. *Neurology* **89**, 1676-1683, doi:10.1212/WNL.0000000000004516 (2017).

82 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, doi:10.1371/journal.pone.0069082 (2013).

83 Ducreux, 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, doi:10.1093/rheumatology/kew262 (2016).

84 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, doi:10.1002/art.37785 (2013).

85 Xiao, W. *et al.* A genomic storm in critically injured humans. *J Exp Med* **208**, 2581-2590, doi:10.1084/jem.20111354 (2011).

86 Zhou, B. *et al.* Analysis of factorial time-course microarrays with application to a clinical study of burn injury. *Proc Natl Acad Sci U S A* **107**, 9923-9928, doi: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 by 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 STATEMENT

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

ADDITIONAL INFORMATION

Competing interests. G.I.M. has consulted for Colgate-Palmolive North America and received compensation. R.R. declares no potential conflict of interest.

FIGURE LEGENDS

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 meta-analysis included a linear model analysis of variance

(ANOVA) coupled 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.

Figure 2: Functional classification of DEPS from AML meta-analysis and associated KEGG and GO enrichment analysis. For all panels, normalized values are represented in with 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 meta-analysis, using unsupervised hierarchical clustering and Euclidean distance for clustering. The age of each individual is displayed at the bottom and illustrated in the color bar on the top (dark green for young and yellow for old). The disease state (AML vs healthy), sex of each subject and age-groups are represented in color bars on the top. (**b)** Horizontal barplot of the top 10 DEPS (gene symbols on vertical axis) from AML meta-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)**.

Figure 3: Sex-related gene expression meta-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 barplot 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.

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

Table 1: Summary table of all 34 gene expression datasets used in this study.

Author, Year	GEO accession	Disease Status*	Affymetrix platform id: Number of samples used & Sample source*	Refs.
Zatkova et al, 2009	GSE10258	AML	GPL570: 8 BM	63
Tomasson et al, 2008	GSE10358	AML	GPL570: 300 BM	64
Metzeler et al, 2008	GSE12417	AML	GPL570: 73 BM & 5 PB GPL96/97: 160 BM & 2PB	49
Wouters et al, 2009, Taskesen et al, 2011	GSE14468	AML	GPL570: 482 BM & 43 PB	65,66
Figueroa et al, 2009	GSE14479	AML	GPL570: 16 BM	67
Klein et al, 2009	GSE15434	AML	GPL570: 231 BM & 20 PB	68
Lück et al, 2011	GSE29883	AML	GPL570: 10 BM & 2 PB	69
Li et al, 2013, Herold et al, 2014, Janke et al, 2014, Jiang et al, 2016	GSE37642	AML	GPL570: 140 BM GPL96/97: 422 BM	50-53
Bullinger et al, 2014	GSE39363	AML	GPL570: 11 BM & 2 PB	NYP
Opel et al, 2015	GSE46819	AML	GPL570: 8 BM & 4 PB	70
TCGA et al, 2015	GSE68833	AML	GPL570: 183 BM	NYP
Cao et al, 2016	GSE69565	AML	GPL570: 12 PB	71
Bohl et al, 2016	GSE84334	AML	GPL570: 25 BM & 20 PB	NYP
Li et al, 2011	GSE23025	AML	GPL570: 21 BM & 13 PB	72
Warren et al, 2009	GSE11375	Healthy	GPL570: 26 PB	73
Green et al, 2009	GSE14845	Healthy	GPL570: 1 PB	NYP
Wu et al, 2012	GSE15932	Healthy	GPL570: 8 PB	NYP
Karlovich et al, 2009	GSE16028	Healthy	GPL570: 22 PB	74
Krug et al, 2011	GSE17114	Healthy	GPL570: 14 PB	NYP
Kong et al, 2012	GSE18123	Healthy	GPL570: 17 PB	75
Sharma et al, 2009	GSE18781	Healthy	GPL570: 25 PB	76
Rosell et al, 2011	GSE25414	Healthy	GPL570: 12 PB	77
Schmidt et al, 2006	GSE2842	Healthy	GPL570: 2 PB	78
Meng et al, 2015	GSE71226	Healthy	GPL570: 3 PB	NYP
Tasaki et al, 2017	GSE84844	Healthy	GPL570: 30 PB	79
Leday et al, 2018	GSE98793	Healthy	GPL570: 64 PB	80
Shamir et al, 2017	GSE99039	Healthy	GPL570: 121 PB	81
Tasaki et al, 2018	GSE93272	Healthy	GPL570: 35 PB	62
Clelland et al, 2013	GSE46449	Healthy	GPL570: 24 PB	82
Lauwerys et al, 2013 Ducreux et al, 2016	GSE39088	Healthy	GPL570: 46 PB	83,84
Xiao et al, 2011	GSE36809	Healthy	GPL570: 35 PB	85
Zhou et al, 2010	GSE19743	Healthy	GPL570: 63 PB	86
Jiang et al, 2018[#]	GSE107968[*]	2 AML, 1 Healthy	GPL570: 3 BM	NYP
Greiner et al, 2015[#]	GSE68172[*]	20 AML, 5 Healthy	GPL570: 25 PB	58
Majeti et al, 2009[#]	GSE17054[*]	9 AML, 4 Healthy	GPL570: 13 BM	59
Bacher et al, 2012[#]	GSE33223[*]	20 AML, 10 Healthy	GPL570: 30 PB	60
Mills et al, 2009[#]	GSE15061[*]	404 AML, 138 Healthy	GPL570: 542 BM	61

Meta-analysis data sets summary

Disease state		Sample source		Affymetrix platform id		Unique probe sets	
AML	Healthy	BM	PB	GPL570	GPL96/97	GPL570	GPL96/97
2213	548	2090	671	2177	584	54,675	44,760

Table 1. A summary table of all our data sets using in our meta-analysis and disease classification.

[#] “Covariate reference data sets,” 5 data sets that were used during the batch correction step., datasets were used only during the batch effect correction steps.

*GEO, Gene Expression Omnibus; AML, acute myeloid leukemia; Ref. reference; 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.

Table 2. Top 10 up- and down-regulated of DEPS in AML from disease state

Up-regulated*			
DEG name	**DEPS Gene Symbol**	**Tukey's HSD Mean difference**	**Bonferroni (p-adjusted)**
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. From the Post-hoc Tukey's test, gene expression means difference value < 5% or > 95% between AML and healthy (AML - healthy) were deemed statistically significant for AML. Genes were considered disease state statistically significant from the analysis of all 2761 cases (2213 AML patients and 548 healthy controls) using. The p-values were adjusted based on Bonferroni correction for false discovery rate (FDR). Significant DEPS (gene symbols) are listed in descending order of the mean difference value comparisons for disease state.

Table 3. KEGG pathway analysis of DEPS from meta-analysis of 34 gene expression datasets.

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, CR1	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, TFDP2, CDKN2A	1.4E-4	1.2E-2
p53 signaling pathway	6, 7	THBS1, CCNB1, CCNE2, CDK1, RRM2, CCNB2	SIAH1, 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

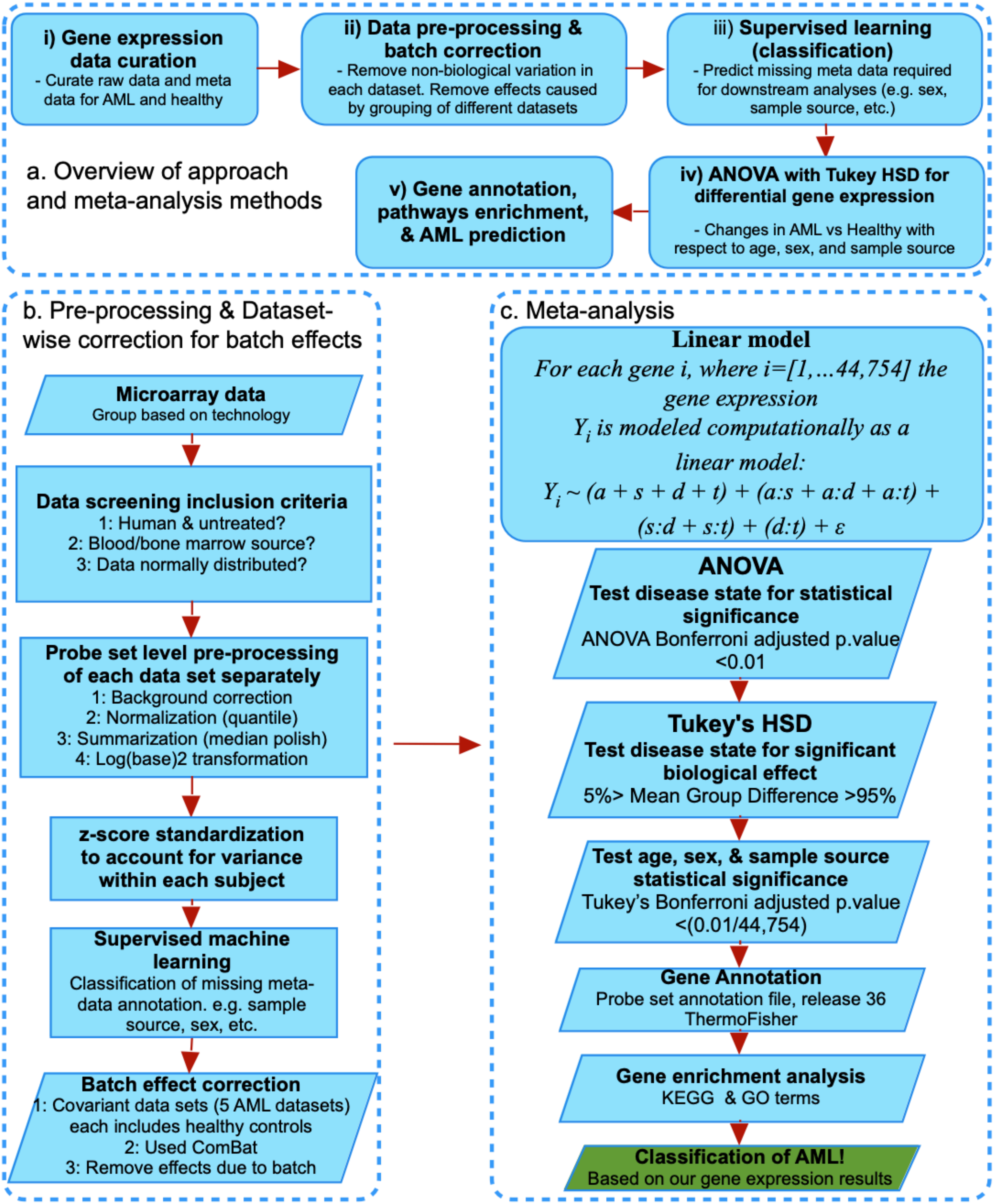
Pathway	No. of genes*	High in Females	High in Males
Hematopoietic cell lineage	1, 2	–	FLT3, CD34
p53 signaling pathway	–, 1	–	PMAIP1
Transcriptional misregulation in cancer	–, 1	MS4A1	FLT3

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

Table 4. KEGG pathway analysis of DEPS from meta-analysis of 34 gene expression datasets overlap with age-specific findings.

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 to19), (70 to 79) - (0 to 19), (70 to 79) - (20 to29), (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: Enrichment analysis was done using 974 DEPS overlapped with age-specific analysis
* up and down regulated genes displayed



a. Overview of approach and meta-analysis methods

i) Gene expression data curation
- Curate raw data and meta data for AML and healthy

ii) Data pre-processing & batch correction
- Remove non-biological variation in each dataset. Remove effects caused by grouping of different datasets

iii) Supervised learning (classification)
- Predict missing meta data required for downstream analyses (e.g. sex, sample source, etc.)

iv) ANOVA with Tukey HSD for differential gene expression
- Changes in AML vs Healthy with respect to age, sex, and sample source

v) Gene annotation, pathways enrichment, & AML prediction

b. Pre-processing & Dataset-wise correction for batch effects

Microarray data
Group based on technology

Data screening inclusion criteria
1: Human & untreated?
2: Blood/bone marrow source?
3: Data normally distributed?

Probe set level pre-processing of each data set separately
1: Background correction
2: Normalization (quantile)
3: Summarization (median polish)
4: Log(base)2 transformation

z-score standardization to account for variance within each subject

Supervised machine learning
Classification of missing meta-data annotation. e.g. sample source, sex, etc.

Batch effect correction
1: Covariant data sets (5 AML datasets) each includes healthy controls
2: Used ComBat
3: Remove effects due to batch

c. Meta-analysis

Linear model
For each gene i, where i=[1,...44,754] the gene expression Y_i is modeled computationally as a linear model:

$$Y_i \sim (a + s + d + t) + (a{:}s + a{:}d + a{:}t) + (s{:}d + s{:}t) + (d{:}t) + \varepsilon$$

ANOVA
Test disease state for statistical significance
ANOVA Bonferroni adjusted p.value <0.01

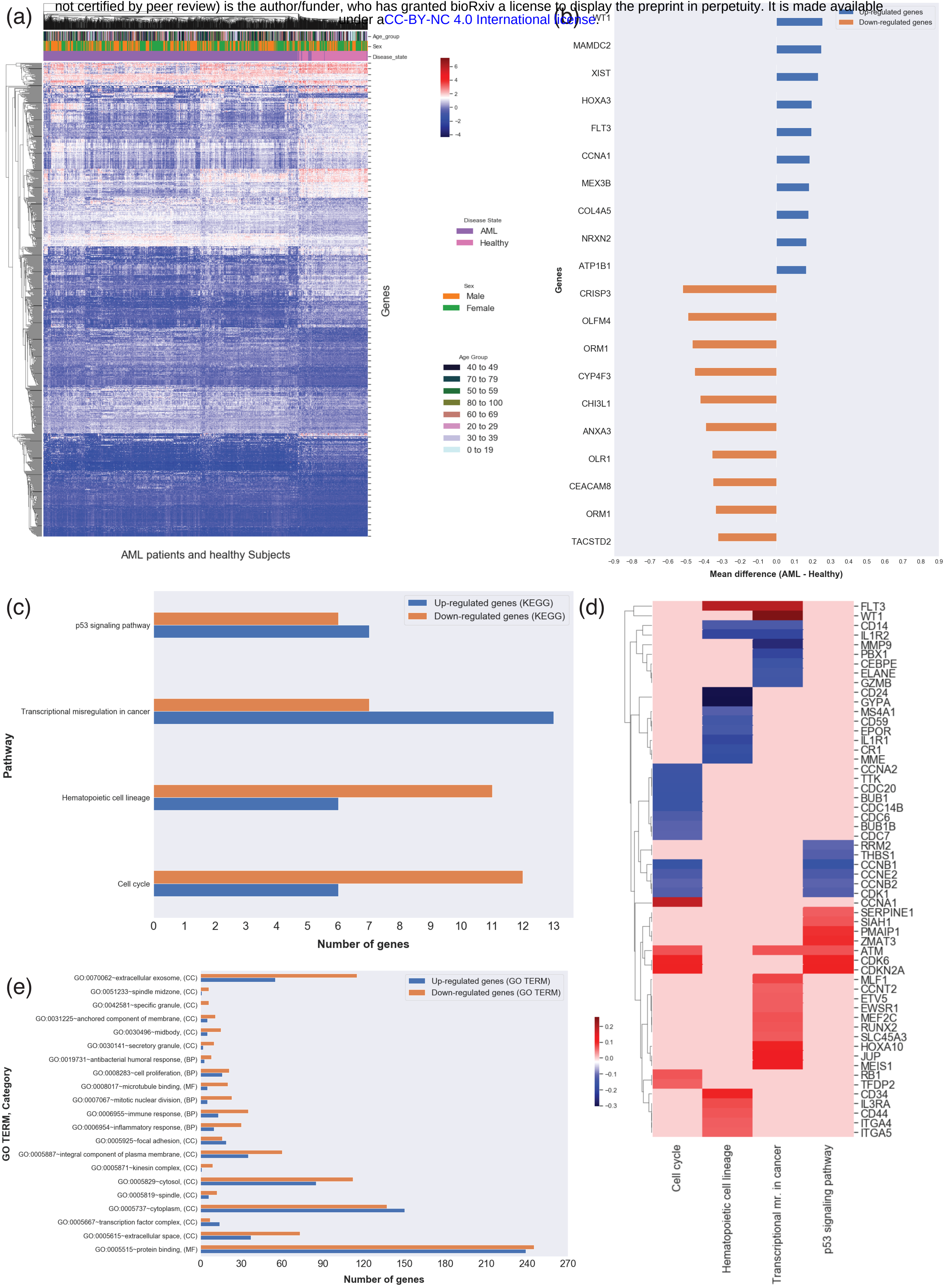
Tukey's HSD
Test disease state for significant biological effect
5%> Mean Group Difference >95%

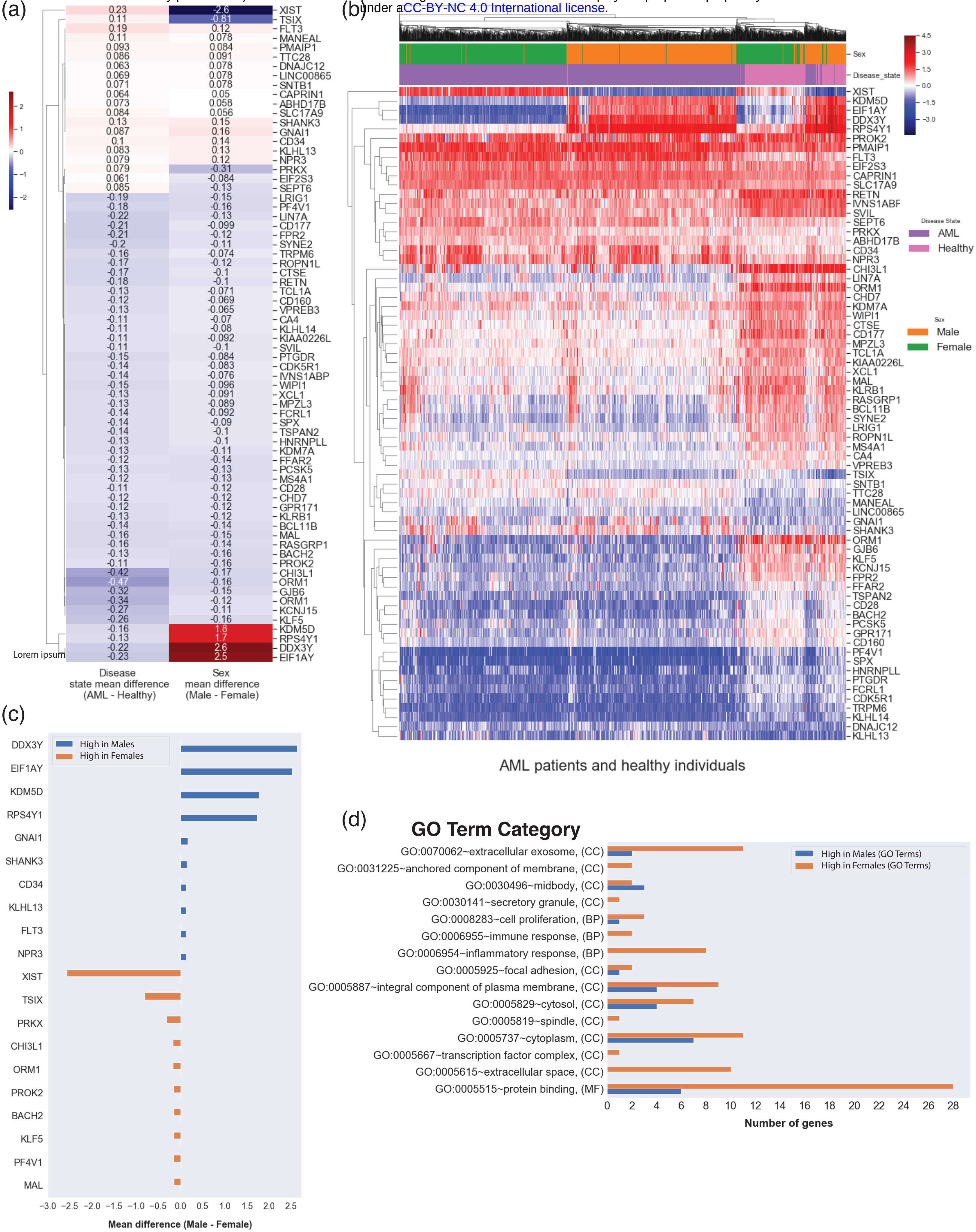
Test age, sex, & sample source statistical significance
Tukey's Bonferroni adjusted p.value <(0.01/44,754)

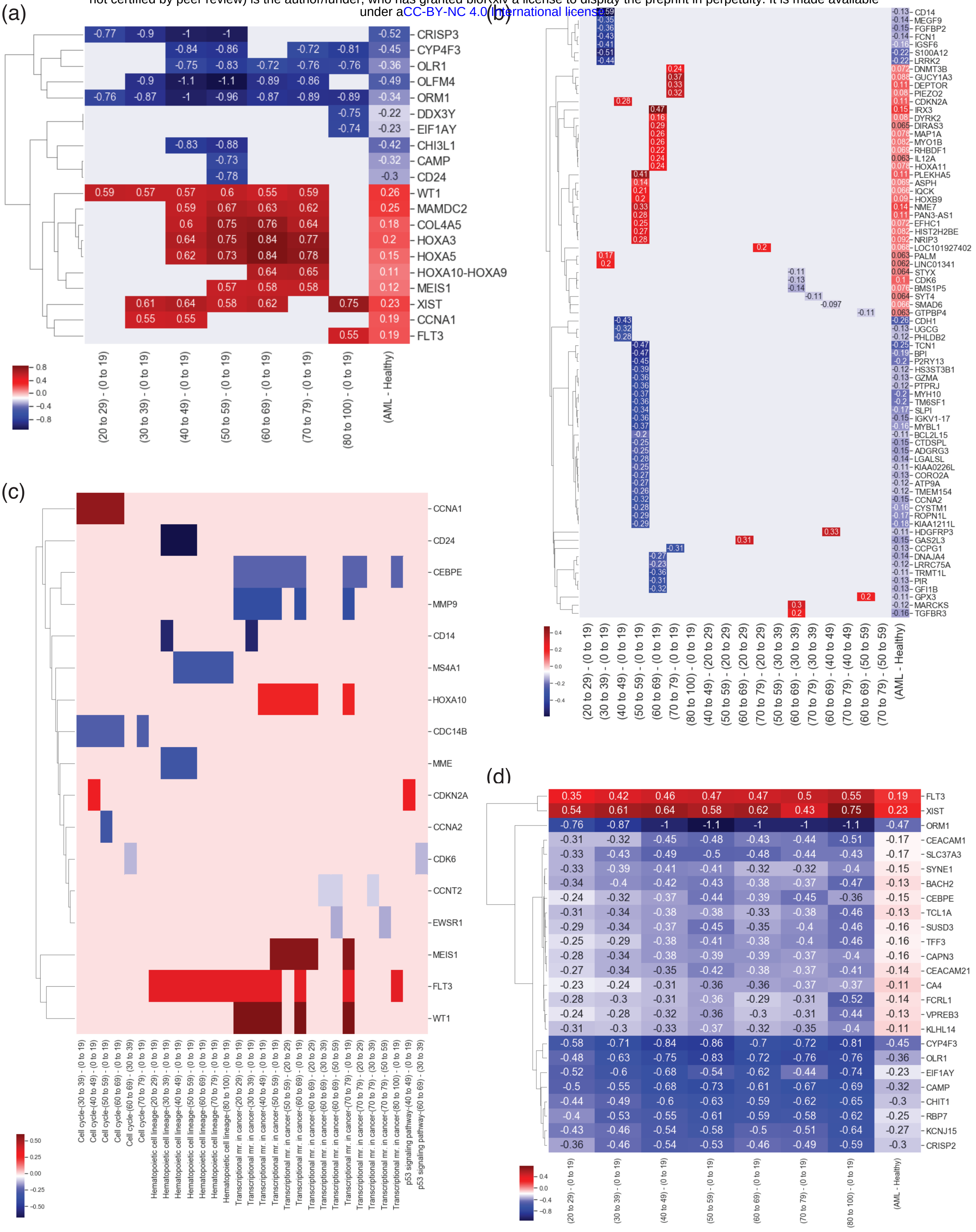
Gene Annotation
Probe set annotation file, release 36 ThermoFisher

Gene enrichment analysis
KEGG & GO terms

Classification of AML!
Based on our gene expression results


(a)
Age_group
Sex
Disease_state
6
4
2
0
−2
−4
Genes
AML patients and healthy Subjects
Disease State
AML
Healthy
Sex
Male
Female
Age Group
40 to 49
70 to 79
50 to 59
80 to 100
60 to 69
20 to 29
30 to 39
0 to 19
(b)
Up-regulated genes
Down-regulated genes
WT1
MAMDC2
XIST
HOXA3
FLT3
CCNA1
MEX3B
COL4A5
NRXN2
ATP1B1
CRISP3
OLFM4
ORM1
CYP4F3
CHI3L1
ANXA3
OLR1
CEACAM8
ORM1
TACSTD2
Genes
-0.9 -0.8 -0.7 -0.6 -0.5 -0.4 -0.3 -0.2 -0.1 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9
Mean difference (AML - Healthy)
(c)
Up-regulated genes (KEGG)
Down-regulated genes (KEGG)
p53 signaling pathway
Transcriptional misregulation in cancer
Hematopoietic cell lineage
Cell cycle
Pathway
0 1 2 3 4 5 6 7 8 9 10 11 12 13
Number of genes
(d)
FLT3
WT1
CD14
IL1R2
MMP9
PBX1
CEBPE
ELANE
GZMB
CD24
GYPA
MS4A1
CD59
EPOR
IL1R1
CR1
MME
CCNA2
TTK
CDC20
BUB1
CDC14B
CDC6
BUB1B
CDC7
RRM2
THBS1
CCNB1
CCNE2
CCNB2
CDK1
CCNA1
SERPINE1
SIAH1
PMAIP1
ZMAT3
ATM
CDK6
CDKN2A
MLF1
CCNT2
ETV5
EWSR1
MEF2C
RUNX2
SLC45A3
HOXA10
JUP
MEIS1
RB1
TFDP2
CD34
IL3RA
CD44
ITGA4
ITGA5
Cell cycle
Hematopoietic cell lineage
Transcriptional mr. in cancer
p53 signaling pathway
0.2
0.1
0.0
−0.1
−0.2
−0.3
(e)
Up-regulated genes (GO TERM)
Down-regulated genes (GO TERM)
GO:0070062~extracellular exosome, (CC)
GO:0051233~spindle midzone, (CC)
GO:0042581~specific granule, (CC)
GO:0031225~anchored component of membrane, (CC)
GO:0030496~midbody, (CC)
GO:0030141~secretory granule, (CC)
GO:0019731~antibacterial humoral response, (BP)
GO:0008283~cell proliferation, (BP)
GO:0008017~microtubule binding, (MF)
GO:0007067~mitotic nuclear division, (BP)
GO:0006955~immune response, (BP)
GO:0006954~inflammatory response, (BP)
GO:0005925~focal adhesion, (CC)
GO:0005887~integral component of plasma membrane, (CC)
GO:0005871~kinesin complex, (CC)
GO:0005829~cytosol, (CC)
GO:0005819~spindle, (CC)
GO:0005737~cytoplasm, (CC)
GO:0005667~transcription factor complex, (CC)
GO:0005615~extracellular space, (CC)
GO:0005515~protein binding, (MF)
GO TERM, Category
0 30 60 90 120 150 180 210 240 270
Number of genes


(a)
0.23 -2.6 XIST
0.11 -0.81 TSIX
0.19 0.12 FLT3
0.11 0.078 MANEAL
0.093 0.084 PMAIP1
0.086 0.091 TTC28
0.063 0.078 DNAJC12
0.069 0.078 LINC00865
0.071 0.078 SNTB1
0.064 0.05 CAPRIN1
0.073 0.058 ABHD17B
0.084 0.056 SLC17A9
0.13 0.15 SHANK3
0.087 0.16 GNAI1
0.1 0.14 CD34
0.083 0.13 KLHL13
0.079 0.12 NPR3
0.079 -0.31 PRKX
0.061 -0.084 EIF2S3
0.085 -0.13 SEPT6
-0.19 -0.15 LRIG1
-0.18 -0.16 PF4V1
-0.22 -0.13 LIN7A
-0.21 -0.099 CD177
-0.21 -0.12 FPR2
-0.2 -0.11 SYNE2
-0.16 -0.074 TRPM6
-0.17 -0.12 ROPN1L
-0.17 -0.1 CTSE
-0.18 -0.1 RETN
-0.13 -0.071 TCL1A
-0.12 -0.069 CD160
-0.13 -0.065 VPREB3
-0.11 -0.07 CA4
-0.11 -0.08 KLHL14
-0.11 -0.092 KIAA0226L
-0.11 -0.1 SVIL
-0.15 -0.084 PTGDR
-0.14 -0.083 CDK5R1
-0.14 -0.076 IVNS1ABP
-0.15 -0.096 WIPI1
-0.13 -0.091 XCL1
-0.13 -0.089 MPZL3
-0.14 -0.092 FCRL1
-0.14 -0.09 SPX
-0.14 -0.1 TSPAN2
-0.13 -0.1 HNRNPLL
-0.13 -0.11 KDM7A
-0.12 -0.14 FFAR2
-0.13 -0.13 PCSK5
-0.12 -0.13 MS4A1
-0.11 -0.12 CD28
-0.12 -0.12 CHD7
-0.12 -0.12 GPR171
-0.13 -0.12 KLRB1
-0.14 -0.14 BCL11B
-0.16 -0.15 MAL
-0.16 -0.14 RASGRP1
-0.13 -0.16 BACH2
-0.11 -0.16 PROK2
-0.42 -0.17 CHI3L1
-0.47 -0.16 ORM1
-0.32 -0.15 GJB6
-0.34 -0.12 ORM1
-0.27 -0.11 KCNJ15
-0.26 -0.16 KLF5
-0.16 1.8 KDM5D
-0.13 1.7 RPS4Y1
-0.22 2.6 DDX3Y
-0.23 2.5 EIF1AY
2
1
0
-1
-2
Lorem ipsum
Disease state mean difference (AML - Healthy)
Sex mean difference (Male - Female)
(b)
Sex
Disease_state
4.5
3.0
1.5
0.0
−1.5
−3.0
XIST
KDM5D
EIF1AY
DDX3Y
RPS4Y1
PROK2
PMAIP1
FLT3
EIF2S3
CAPRIN1
SLC17A9
RETN
IVNS1ABP
SVIL
SEPT6
PRKX
ABHD17B
CD34
NPR3
CHI3L1
LIN7A
ORM1
CHD7
KDM7A
WIPI1
CTSE
CD177
MPZL3
TCL1A
KIAA0226L
XCL1
MAL
KLRB1
RASGRP1
BCL11B
SYNE2
LRIG1
ROPN1L
MS4A1
CA4
VPREB3
TSIX
SNTB1
TTC28
MANEAL
LINC00865
GNAI1
SHANK3
ORM1
GJB6
KLF5
KCNJ15
FPR2
FFAR2
TSPAN2
CD28
BACH2
PCSK5
GPR171
CD160
PF4V1
SPX
HNRNPLL
PTGDR
FCRL1
CDK5R1
TRPM6
KLHL14
DNAJC12
KLHL13
Disease State
AML
Healthy
Sex
Male
Female
AML patients and healthy individuals
(c)
High in Males
High in Females
DDX3Y
EIF1AY
KDM5D
RPS4Y1
GNAI1
SHANK3
CD34
KLHL13
FLT3
NPR3
XIST
TSIX
PRKX
CHI3L1
ORM1
PROK2
BACH2
KLF5
PF4V1
MAL
−3.0 −2.5 −2.0 −1.5 −1.0 −0.5 0.0 0.5 1.0 1.5 2.0 2.5
Mean difference (Male - Female)
(d)
GO Term Category
High in Males (GO Terms)
High in Females (GO Terms)
GO:0070062~extracellular exosome, (CC)
GO:0031225~anchored component of membrane, (CC)
GO:0030496~midbody, (CC)
GO:0030141~secretory granule, (CC)
GO:0008283~cell proliferation, (BP)
GO:0006955~immune response, (BP)
GO:0006954~inflammatory response, (BP)
GO:0005925~focal adhesion, (CC)
GO:0005887~integral component of plasma membrane, (CC)
GO:0005829~cytosol, (CC)
GO:0005819~spindle, (CC)
GO:0005737~cytoplasm, (CC)
GO:0005667~transcription factor complex, (CC)
GO:0005615~extracellular space, (CC)
GO:0005515~protein binding, (MF)
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28
Number of genes


(a)
-0.77 -0.9 -1 -1 -0.52 CRISP3
-0.84 -0.86 -0.72 -0.81 -0.45 CYP4F3
-0.75 -0.83 -0.72 -0.76 -0.76 -0.36 OLR1
-0.9 -1.1 -1.1 -0.89 -0.86 -0.49 OLFM4
-0.76 -0.87 -1 -0.96 -0.87 -0.89 -0.89 -0.34 ORM1
-0.75 -0.22 DDX3Y
-0.74 -0.23 EIF1AY
-0.83 -0.88 -0.42 CHI3L1
-0.73 -0.32 CAMP
-0.78 -0.3 CD24
0.59 0.57 0.57 0.6 0.55 0.59 0.26 WT1
0.59 0.67 0.63 0.62 0.25 MAMDC2
0.6 0.75 0.76 0.64 0.18 COL4A5
0.64 0.75 0.84 0.77 0.2 HOXA3
0.62 0.73 0.84 0.78 0.15 HOXA5
0.64 0.65 0.11 HOXA10-HOXA9
0.57 0.58 0.58 0.12 MEIS1
0.61 0.64 0.58 0.62 0.75 0.23 XIST
0.55 0.55 0.19 CCNA1
0.55 0.19 FLT3
0.8
0.4
0.0
-0.4
-0.8
(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)
(AML - Healthy)
(b)
-0.59 -0.35 -0.36 -0.43 -0.41 -0.51 -0.44
0.24 0.37 0.33 0.32
0.28
0.47 0.16 0.29 0.26 0.26 0.22 0.24 0.24
0.41 0.14 0.21 0.2 0.33 0.28 0.25 0.27 0.28
0.2
0.17 0.2
-0.11 -0.13 -0.14
-0.11
-0.097
-0.11
-0.43 -0.32 -0.28
-0.47 -0.47 -0.45 -0.39 -0.36 -0.36 -0.37 -0.36 -0.34 -0.36 -0.37 -0.2 -0.25 -0.25 -0.28 -0.25 -0.27 -0.27 -0.26 -0.32 -0.28 -0.29 -0.29
0.33
0.31
-0.31
-0.27 -0.23 -0.36 -0.31 -0.32
0.2
0.3 0.2
-0.13 CD14
-0.14 MEGF9
-0.15 FGFBP2
-0.14 FCN1
-0.16 IGSF6
-0.22 S100A12
-0.22 LRRK2
0.072 DNMT3B
0.088 GUCY1A3
0.11 DEPTOR
0.08 PIEZO2
0.11 CDKN2A
0.15 IRX3
0.08 DYRK2
0.065 DIRAS3
0.078 MAP1A
0.082 MYO1B
0.069 RHBDF1
0.063 IL12A
0.078 HOXA11
0.11 PLEKHA5
0.069 ASPH
0.066 IQCK
0.09 HOXB9
0.14 NME7
0.11 PAN3-AS1
0.072 EFHC1
0.082 HIST2H2BE
0.092 NRIP3
0.068 LOC101927402
0.063 PALM
0.062 LINC01341
0.064 STYX
0.1 CDK6
0.076 BMS1P5
0.064 SYT4
0.066 SMAD6
0.063 GTPBP4
-0.26 CDH1
-0.13 UGCG
-0.12 PHLDB2
-0.25 TCN1
-0.19 BPI
-0.2 P2RY13
-0.12 HS3ST3B1
-0.13 GZMA
-0.12 PTPRJ
-0.2 MYH10
-0.2 TM6SF1
-0.17 SLPI
-0.15 IGKV1-17
-0.16 MYBL1
-0.11 BCL2L15
-0.15 CTDSPL
-0.15 ADGRG3
-0.14 LGALSL
-0.11 KIAA0226L
-0.13 CORO2A
-0.12 ATP9A
-0.12 TMEM154
-0.15 CCNA2
-0.16 CYSTM1
-0.17 ROPN1L
-0.18 KIAA1211L
-0.11 HDGFRP3
-0.15 GAS2L3
-0.13 CCPG1
-0.14 DNAJA4
-0.12 LRRC75A
-0.11 TRMT1L
-0.13 PIR
-0.13 GFI1B
-0.11 GPX3
-0.12 MARCKS
-0.16 TGFBR3
0.4
0.2
0.0
-0.2
-0.4
(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)
(40 to 49) - (20 to 29)
(50 to 59) - (20 to 29)
(60 to 69) - (20 to 29)
(70 to 79) - (20 to 29)
(50 to 59) - (30 to 39)
(60 to 69) - (30 to 39)
(70 to 79) - (30 to 39)
(60 to 69) - (40 to 49)
(70 to 79) - (40 to 49)
(60 to 69) - (50 to 59)
(70 to 79) - (50 to 59)
(AML - Healthy)
(c)
CCNA1
CD24
CEBPE
MMP9
CD14
MS4A1
HOXA10
CDC14B
MME
CDKN2A
CCNA2
CDK6
CCNT2
EWSR1
MEIS1
FLT3
WT1
0.50
0.25
0.00
-0.25
-0.50
Cell cycle-(30 to 39) - (0 to 19)
Cell cycle-(40 to 49) - (0 to 19)
Cell cycle-(50 to 59) - (0 to 19)
Cell cycle-(60 to 69) - (0 to 19)
Cell cycle-(60 to 69) - (30 to 39)
Cell cycle-(70 to 79) - (0 to 19)
Hematopoietic cell lineage-(20 to 29) - (0 to 19)
Hematopoietic cell lineage-(30 to 39) - (0 to 19)
Hematopoietic cell lineage-(40 to 49) - (0 to 19)
Hematopoietic cell lineage-(50 to 59) - (0 to 19)
Hematopoietic cell lineage-(60 to 69) - (0 to 19)
Hematopoietic cell lineage-(70 to 79) - (0 to 19)
Hematopoietic cell lineage-(80 to 100) - (0 to 19)
Transcriptional mr. in cancer-(20 to 29) - (0 to 19)
Transcriptional mr. in cancer-(30 to 39) - (0 to 19)
Transcriptional mr. in cancer-(40 to 49) - (0 to 19)
Transcriptional mr. in cancer-(50 to 59) - (0 to 19)
Transcriptional mr. in cancer-(50 to 59) - (20 to 29)
Transcriptional mr. in cancer-(60 to 69) - (0 to 19)
Transcriptional mr. in cancer-(60 to 69) - (20 to 29)
Transcriptional mr. in cancer-(60 to 69) - (30 to 39)
Transcriptional mr. in cancer-(60 to 69) - (50 to 59)
Transcriptional mr. in cancer-(70 to 79) - (0 to 19)
Transcriptional mr. in cancer-(70 to 79) - (20 to 29)
Transcriptional mr. in cancer-(70 to 79) - (30 to 39)
Transcriptional mr. in cancer-(70 to 79) - (50 to 59)
Transcriptional mr. in cancer-(80 to 100) - (0 to 19)
p53 signaling pathway-(40 to 49) - (0 to 19)
p53 signaling pathway-(60 to 69) - (30 to 39)
(d)
0.35 0.42 0.46 0.47 0.47 0.5 0.55 0.19 FLT3
0.54 0.61 0.64 0.58 0.62 0.43 0.75 0.23 XIST
-0.76 -0.87 -1 -1.1 -1 -1 -1.1 -0.47 ORM1
-0.31 -0.32 -0.45 -0.48 -0.43 -0.44 -0.51 -0.17 CEACAM1
-0.33 -0.43 -0.49 -0.5 -0.48 -0.44 -0.43 -0.17 SLC37A3
-0.33 -0.39 -0.41 -0.41 -0.32 -0.32 -0.4 -0.15 SYNE1
-0.34 -0.4 -0.42 -0.43 -0.38 -0.37 -0.47 -0.13 BACH2
-0.24 -0.32 -0.37 -0.44 -0.39 -0.45 -0.36 -0.15 CEBPE
-0.31 -0.34 -0.38 -0.38 -0.33 -0.38 -0.46 -0.13 TCL1A
-0.29 -0.34 -0.37 -0.45 -0.35 -0.4 -0.46 -0.16 SUSD3
-0.25 -0.29 -0.38 -0.41 -0.38 -0.4 -0.46 -0.16 TFF3
-0.28 -0.34 -0.38 -0.39 -0.39 -0.37 -0.4 -0.16 CAPN3
-0.27 -0.34 -0.35 -0.42 -0.38 -0.37 -0.41 -0.14 CEACAM21
-0.23 -0.24 -0.31 -0.36 -0.36 -0.37 -0.37 -0.11 CA4
-0.28 -0.3 -0.31 -0.36 -0.29 -0.31 -0.52 -0.14 FCRL1
-0.24 -0.28 -0.32 -0.36 -0.3 -0.31 -0.44 -0.13 VPREB3
-0.31 -0.3 -0.33 -0.37 -0.32 -0.35 -0.4 -0.11 KLHL14
-0.58 -0.71 -0.84 -0.86 -0.7 -0.72 -0.81 -0.45 CYP4F3
-0.48 -0.63 -0.75 -0.83 -0.72 -0.76 -0.76 -0.36 OLR1
-0.52 -0.6 -0.68 -0.54 -0.62 -0.44 -0.74 -0.23 EIF1AY
-0.5 -0.55 -0.68 -0.73 -0.61 -0.67 -0.69 -0.32 CAMP
-0.44 -0.49 -0.6 -0.63 -0.59 -0.62 -0.65 -0.3 CHIT1
-0.4 -0.53 -0.55 -0.61 -0.59 -0.58 -0.62 -0.25 RBP7
-0.43 -0.46 -0.54 -0.58 -0.5 -0.51 -0.64 -0.27 KCNJ15
-0.36 -0.46 -0.54 -0.53 -0.46 -0.49 -0.59 -0.3 CRISP2
0.4
0.0
-0.4
-0.8
(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)
(AML - Healthy)