PNAS
www.pnas.org

Supplementary Information for

Multi-omics of azacitidine-treated AML cells reveals variable and convergent targets that remodel the cell surface proteome

Kevin K Leung, Aaron Nguyen, Tao Shi, Lin Tang, Xiaochun Ni, Laure Escoubet, Kyle J MacBeth, Jorge DiMartino, James A Wells

Corresponding author:
James A Wells
Email: jim.wells@ucsf.edu

This PDF file includes:

Supplemental Text

Rationale for cell line

KG1a (FAB M1) was established as a subline of KG1 myeloblasts with minimal maturation, but unable to differentiate (1, 2); HL60 (FAB M2) was established from a promyelocyte that can differentiate into neutrophils, granulocyte-like cells, or monocyte/macrophage-like cells by various chemicals (3); HNT34 (FAB M4) was established from chronic myelomonocytic leukemia (CMMoL) (4); and AML193 (FAB M5) was established from acute monocytic leukemia (5). While FAB M3 represents a differentiation stage intermediate to the four cell lines studied, it is defined by a specific genetic mutation t(15;17)(q24;q21) resulting in the fusion protein Promyelocytic leukemia/retinoic acid receptor alpha (PML-RARA). It is highly treatable and not commonly treated by AZA, thus not included in the current study.

SMITE analysis

The SMITE algorithm (Significance-based Modules Integrating the Transcriptome and Epigenome) increases analysis power in identifying subnetworks of dysregulated genes by combining gene regulation information from both methylation and transcription levels. Here, the joint scores are mapped onto a gene interaction network. SMITE analysis revealed that each AML cell line had 16-27 functional modules identified and these functional modules encompass 481 ~ 612 genes (Dataset S3). In particular, analysis of the KG1a cell line benefited most from SMITE, where 528 genes were found to be dysregulated compared to 160 genes when analyzed using transcription data alone. This indicated wider functional impact of AZA treatment on the KG1a cell line when examining both sets of information in the context of gene interaction networks compared to transcriptomics alone. Also, when looking across all four cell lines, instead of the 5 common genes found to be transcriptionally dysregulated, SMITE identified 19 functional modules encompassing 310 genes. We observed recurrent themes of dysregulation of metabolism from multiple top functional modules identified across all four cell lines - HDC:HNMT, CTSA:GNS:IDS:NEU1, and PTGS2:LTC4S, representing genes involved in histamine metabolism, protein processing, and leukotrienes metabolism. Other top common functional modules affected were ‘natural killer cell mediated cytotoxicity’ and ‘osteoclast differentiation’. These findings were consistent with the impact of AZA on immune pathways and AML cell differentiation, and also identified novel functional modules. For example, HNMT encodes a methyltransferase that metabolizes histamine via N(tau)-methylation. Core genes from these functional modules would also serve as interesting therapeutic targets for follow up validation. AZA impacted different functional modules in the individual cell lines – top modules involved in various signaling pathways (mTOR, insulin, p53 etc) were affected in AML193 cells, while modules related to the MAPK signaling pathway was regulated in HL60 cells.

Supplemental Materials and Methods

Cell culture and azacitidine treatment

AML193, HL60, KG1a, and HNT34 cells were cultured in RPMI SILAC media (Thermo Fisher Scientific; Waltham, MA) containing L-[12C6,14N2]lysine and L-[12C6,14N4]arginine (light label) and treated with vehicle (DMSO) or 0.5 µM AZA daily for 3 days. Cells were cultured for

another 4 days in drug-free media before frozen cell pellets were harvested for DNA methylation (Infinium MethylationEPIC BeadChip, Illumina) and gene expression (GeneChip™ Human Genome U133 Plus 2.0 Array, Thermo Fisher Scientific) analyses.

DNA methylation and gene expression data analysis

For DNA methylation data, the standard beta values from Illumina's BeadStudio were used as our initial input and further corrected for different probe designs using Beta Mixture Quantile dilation (BMIQ) method (6). Probes with known SNPs and too many missing values were also removed from subsequent analyses. The final input data matrix contains 833622 probes. For differential methylation analyses, beta values were transformed to M value (a logit transformation of beta values) and limma package (7) was applied to assess the significance.

For gene expression data, gene expression values were calculated based on the Robust Multi-array Average (RMA) algorithm (8). For each gene, a single probe set was selected based on jetset algorithm (9) resulting the final input data matrix of 20517 genes. Functional enrichment of baseline gene expression profiles was assessed using GSVA package from Bioconductor (10). The package limma was used to assess differential gene expression (7) and gene set enrichment analysis was carried out using fast pre-ranked gene set enrichment analysis (fgsea) package from Bioconductor (11).

Hierarchical clustering based on Euclidean distance and Ward's minimum variance method was used to evaluate the similarity/difference between samples based on their global DNA methylation or gene expression profiles. The integrated analyses combining DNA methylation and gene expression data were carried out using SMITE package in Bioconductor (12). All data analyses were carried out using R: A language and environment for statistical computing (13). Methylome and transcriptome datasets have been deposited to NCBI's Gene Expression Omnibus database (https://www.ncbi.nlm.nih.gov/geo/)(14) and are available under accession number GSE123211.

SILAC surface proteomics sample preparation

For stable isotope labeling with amino acids in cell culture (SILAC) experiments, each cell line was cultured in RPMI SILAC media (Thermo Fisher Scientific) containing L-[13C6,15N2]lysine and L-[13C6,15N4] arginine (heavy label) (Thermo Fisher Scientific) or L-[12C6,14N2]lysine and L-[12C6,14N4]arginine (light label) for 5 passages to ensure full incorporation of the isotope labeling on cells. In the forward SILAC experiment, heavy labeled cells were treated with AZA while the light labeled cells were treated with DMSO control. In a parallel reverse SILAC experiment, light labeled cells were treated with AZA while the heavy labeled cells were treated with DMSO control. On day 7 of the aforementioned drug treatment, 40-60 x 106 cells from AZA or DMSO treated cells were mixed at a 1:1 cell count ratio for cell surface capture enrichment with several optimization (15). Briefly, live cells were treated with a sodium periodate buffer (2 mM $NaPO_4$, PBS pH 6.5) at 4°C for 20 mins to oxidize terminal sialic acids of glycoproteins. Aldehydes generated by periodate oxidation were then reacted with biocytin hydrazide in a labeling buffer (1 mM biocytin hydrazide (biotium), 10 mM analine (Sigma), PBS

pH 6.5) at 4°C for 90 mins. Cells were then washed four times in PBS pH 6.5 to remove excess biocytin-hydrazide and flash frozen.

Frozen cell pellets were lysed using RIPA buffer (VWR) with protease inhibitor cocktail (Sigma-Aldrich; St. Louis, MO) at 4°C for 30 mins. Cell lysate was then sonicated, clarified, and incubated with 500µL of neutravidin agarose slurry (Thermo Fisher Scientific) at 4°C for 30 mins. The neutravidin beads were then extensively washed with RIPA buffer, high salt buffer (1M NaCl, PBS pH 7.5), and urea buffer (2M urea, 50mM ammonium bicarbonate) to remove non-specific proteins. Samples were then reduced on-bead with 5mM TCEP at 55°C for 30 mins and alkylated with 10mM iodoacetiamide at room temperature for 30 mins. To release bound proteins, we first performed an on-bead digestion using 20µg trypsin (Promega; Madison, WI) at room temperature overnight. The "tryptic" fraction was then eluted using spin column and the neutravidin beads were extensively washed again with RIPA buffer, high salt buffer (1M NaCl, PBS pH 7.5), and urea buffer (2M urea, 50mM ammonium bicarbonate). To release the remaining trypsin digested N-glycosylated peptides bound to the neutravidin beads, we performed a second on-bead digestion using 2500U PNGase F (New England Biolabs; Ipswich, MA) at 37°C for 3 hrs. Similarly, the "PNGase F" fraction was eluted using a spin column. Both tryptic and PNGase F fractions were then desalted using SOLA HRP SPE column (Thermo Fisher Scientific) using standard protocol, dried, and dissolved in 0.1% formic acid, 2% acetonitrile prior to LC-MS/MS analysis.

Mass spectrometry analysis

Approximately 1µg of peptides was injected to a pre-packed 0.75mm x 150mm Acclaimed Pepmap C18 reversed phase column (2µm pore size, Thermo Fisher Scientific) attached to a Q Exactive Plus (Thermo Fisher Scientific) mass spectrometer. For the "tryptic" fraction, peptides were separated using a linear gradient of 3-35% solvent B (Solvent A: 0.1% formic acid, solvent B: 80% acetonitrile, 0.1% formic acid) over 180 mins at 300µL/min. Similarly, the "PNGase F" fraction was separated using the same gradient over 120 mins. Data were collected in data-dependent mode using a top 20 method with dynamic exclusion of 35 secs and a charge exclusion setting that only samples peptides with a charge of 2, 3, or 4. Full (ms1) scans spectrums were collected as profile data with a resolution of 140,000 (at 200 m/z), AGC target of 3E6, maximum injection time of 120 ms, and scan range of 400 - 1800 m/z. MS-MS scans were collected as centroid data with a resolution of 17,500 (at 200 m/z), AGC target of 5E4, maximum injection time of 60 ms with normalized collision energy at 27, and an isolation window of 1.5 m/z with an isolation offset of 0.5 m/z.

Proteomics data processing

Peptide search for each individual dataset was performed using ProteinProspector (v5.13.2) against 20203 human proteins (Swiss-prot database, obtained March 5, 2015). Enzyme specificity was set to trypsin with up to two missed cleavage; cysteine carbamidomethyl was set as a fixed modification; methionine oxidation, lysine and arginine SILAC labels were set as variable modifications; asparagine deamidation was also set as variable modification for the PNGase F fraction; peptide mass tolerance was 6 ppm; fragment ion mass tolerance was 0.4 Da; peptide identification was filtered by peptide score of 0.0005 in Protein Prospector, resulting in

a false discovery rate (FDR) of <1% calculated by number of decoy peptides included in the database. To estimate the efficiency of the surface proteome enrichment method, a list of surface proteins was generated by searching for “membrane” but not “mitochondrial” or “nuclear” using Uniprot subcellular localization annotations. We found up to 60% of peptides identified in the tryptic fraction and up to 90% of peptides identified in the PNGase F fraction belonged to the surface proteome reflecting a high and expected surface protein enrichment. For proteins identified in both fractions, Pearson correlation between median log2 enrichment ratios found in tryptic fraction and PNGase F fractions were between 0.85 and 0.95 (Figure S7B).

Quantitative data analysis was performed using Skyline (UWashington) software with a ms1 filtering function. Specifically, spectral libraries from forward and reverse SILAC experiments were analyzed together such that ms1 peaks without an explicit peptide ID would be quantified based on aligned peptide retention time. An isotope dot product of at least 0.8 was used to filter out low quality peptide quantification, and a custom report from Skyline was then exported for further processing and analysis using R. In the tryptic fraction, only peptides with five or more well quantified peptides were included. In the PNGase F fraction, only peptides with N to D deamidation modification were included. Forward and reverse SILAC datasets for both tryptic and PNGase F fractions were then combined and reported as median log2 enrichment values normalized to a mean of zero for the AZA treated cells. The raw proteomics data, peaklist, Protein Prospector results, and Skyline quantification results have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the MassIVE partner repository with the data set identifier PXD011298.

Supplemental Figures S1-S13 and Tables S1-S2

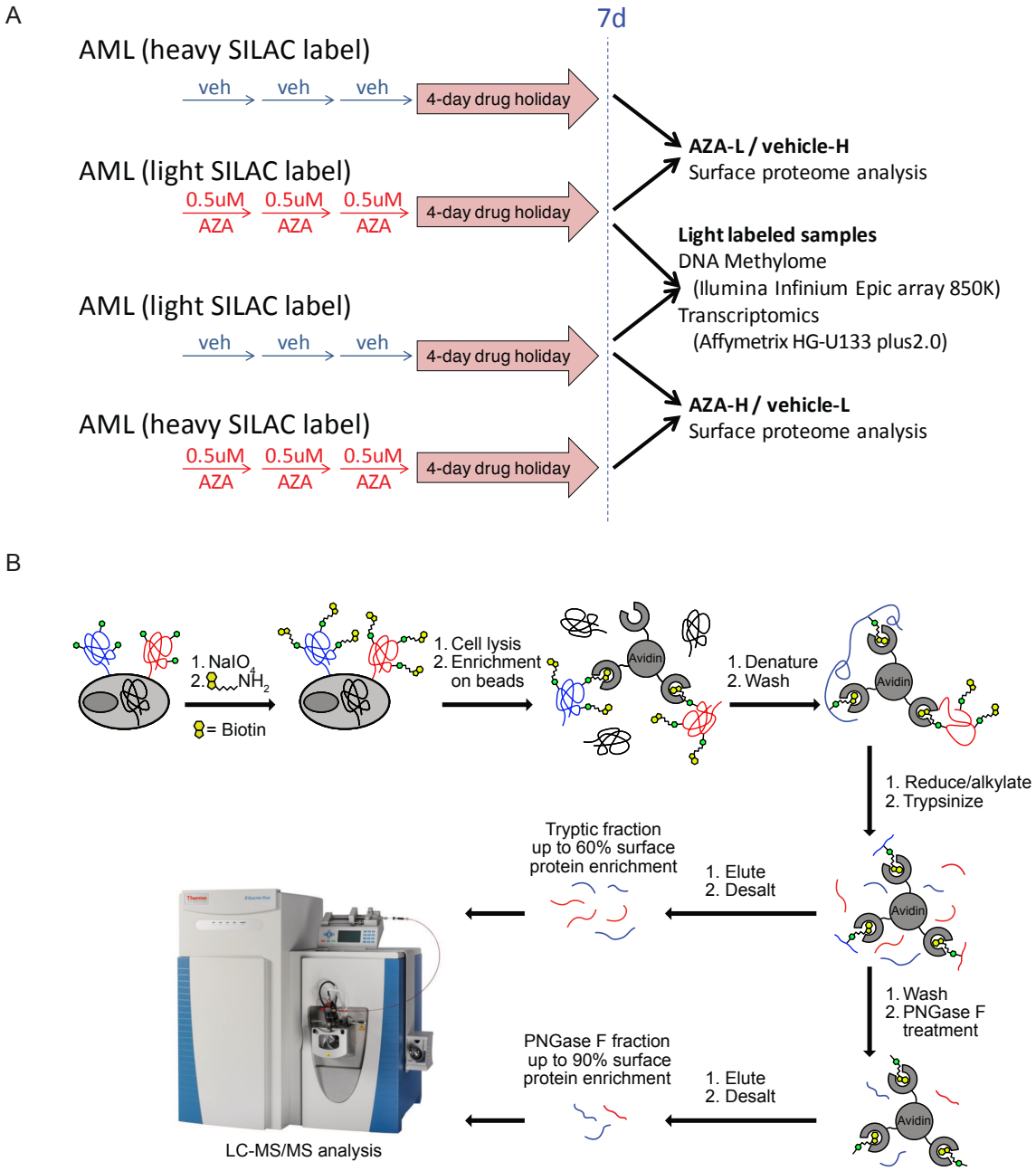


Figure S1. Azacitidine treatment schematic and surface proteome capture methods. (A) The AML cell lines were treated with 0.5μM fresh AZA or vehicle for three days and allowed to recover for four days. The AML AZA-vehicle pairs were grown in either heavy SILAC or light SILAC media. AML cell lines tested in the current study include KG1a, HL60, HNT34, and AML193. (B) Cell surface capture method by biocytin-hydrazide labeling of glycoproteins (see methods for detailed descriptions).

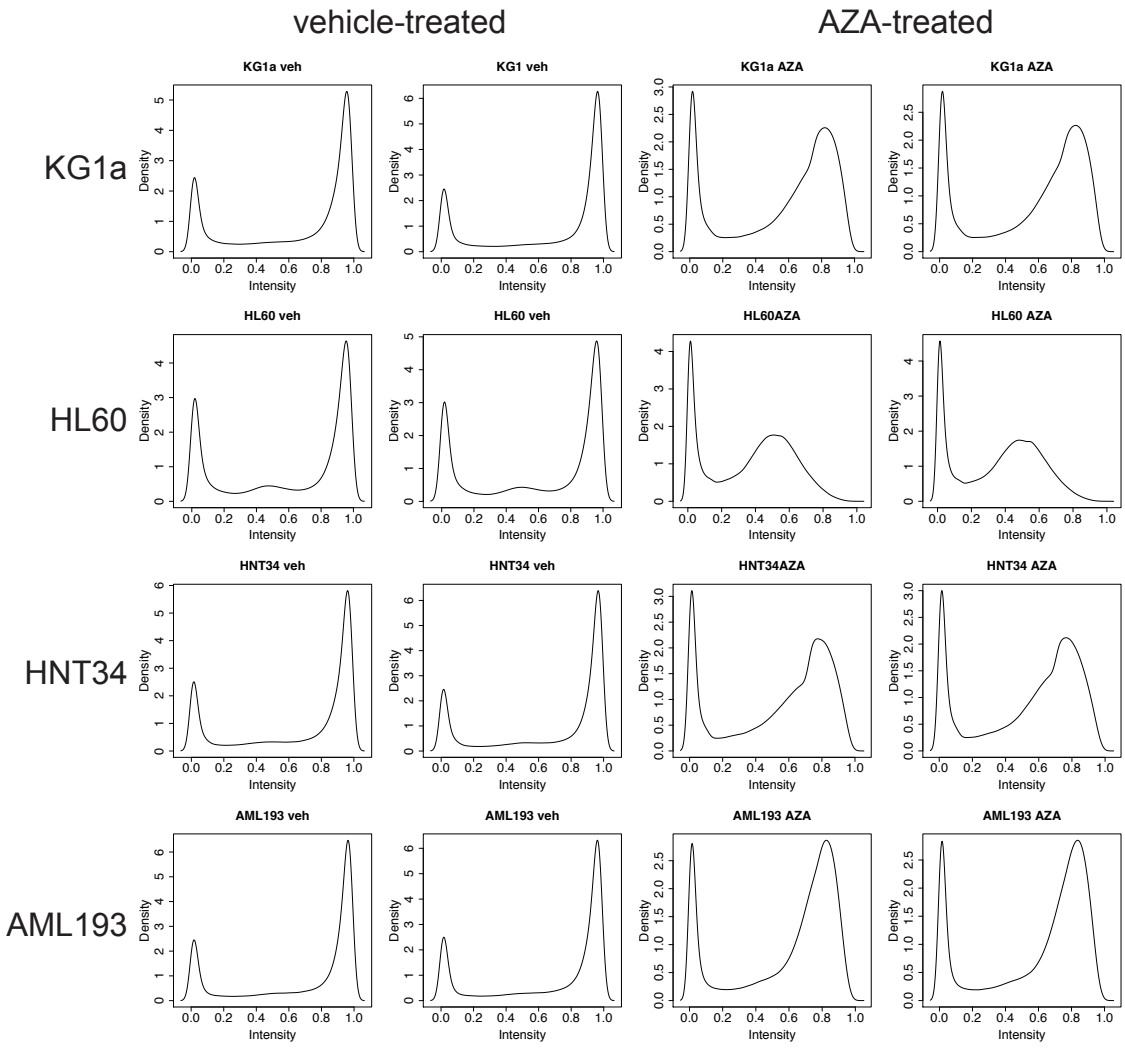


Figure S2. Global de-methylation of AML cell lines when treated by AZA. Density plot of beta values for each cell line sample before and after treatment indicate a general shift of the hypermethylation peak.

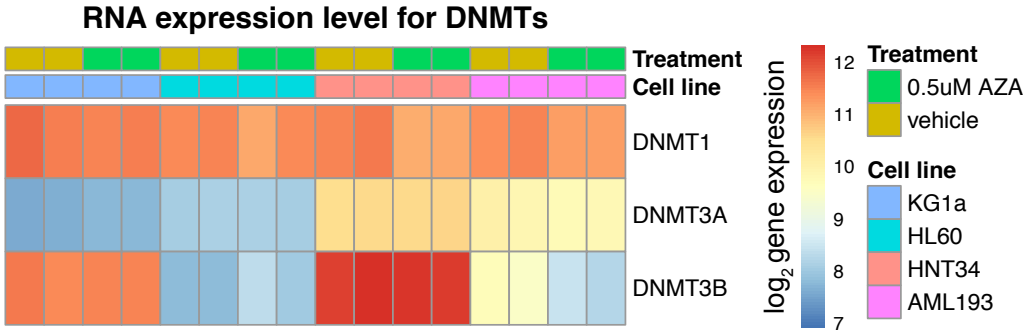


Figure S3. Expression of DNMT transcript in each cell line. Expression of DNMT1 transcript is high in all four cell lines. Expression of both *de novo* methyltransferases DMNT3A and DMNT3B were low in HL60.

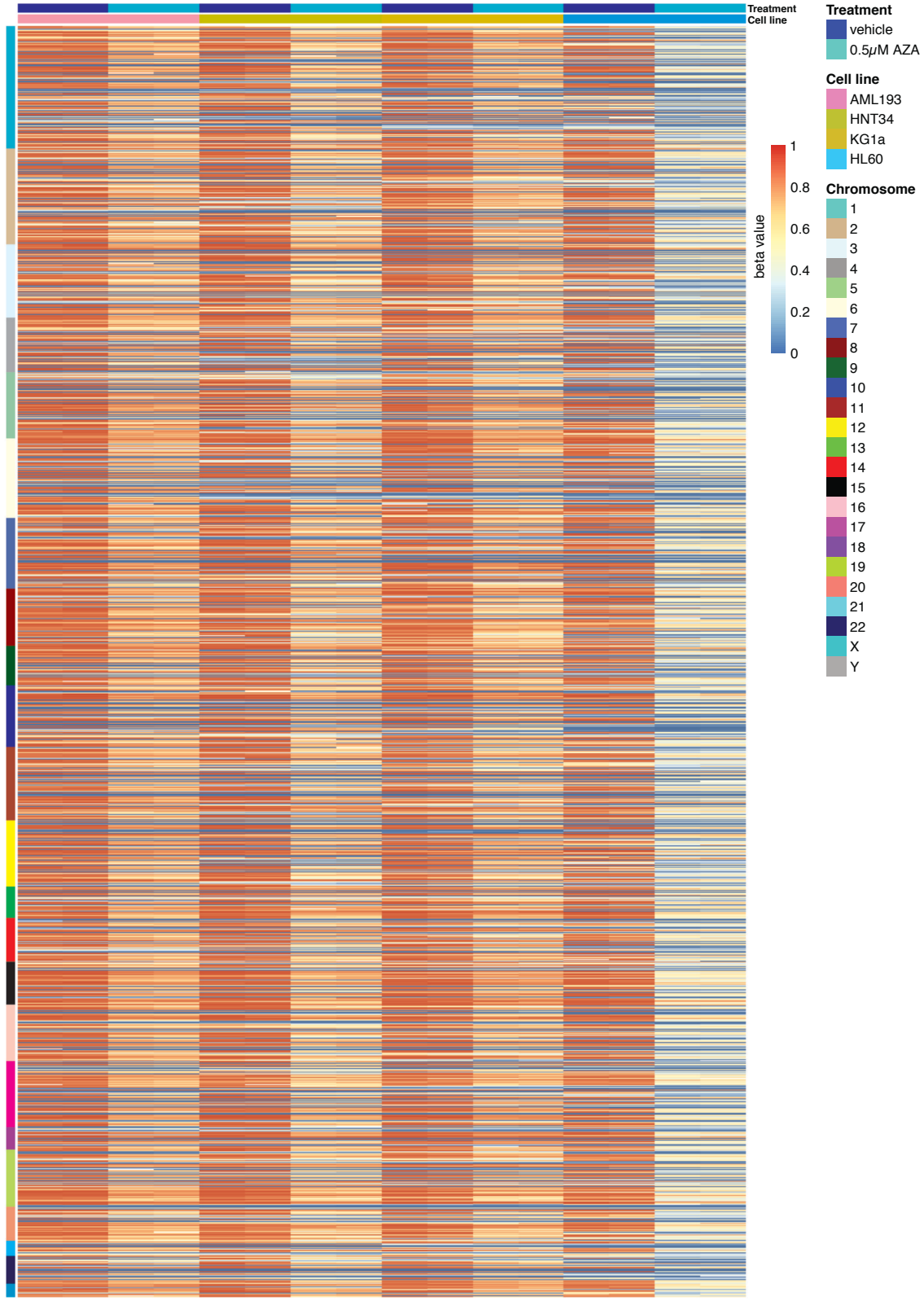


Figure S4. Heatmap of DNA methylation level (beta values) of all ~800,000 CpG arranged by their chromosomal locations in all samples.

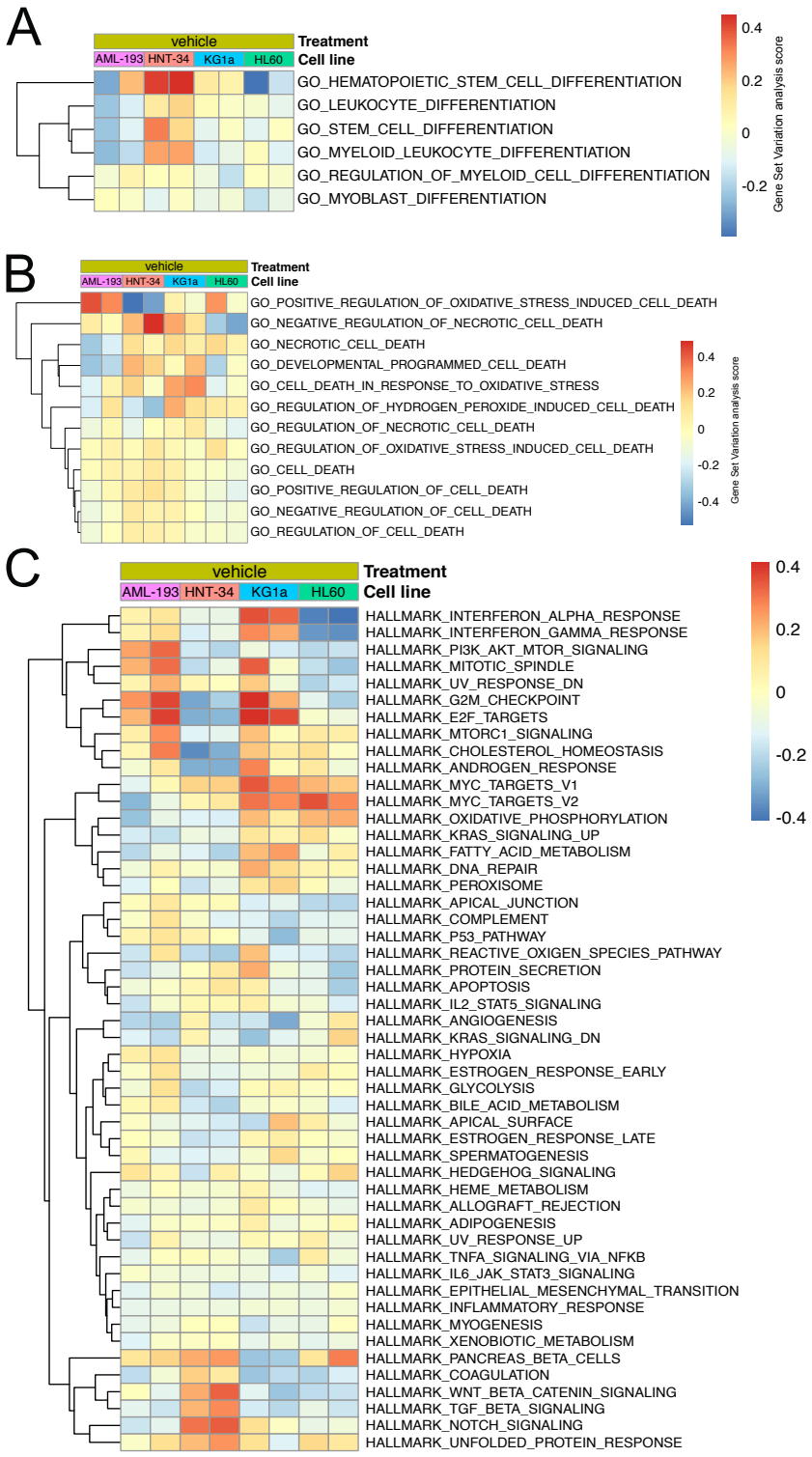


Figure S5. Baseline gene set variation analysis by gene expression data. (A-B) Gene Set Variation Analysis (GSVA) using Gene Ontology (GO) term analysis in differentiation and cell death pathways indicate differences between the cell lines at baseline. (C) GSVA of the top 50 hallmark gene set indicate different biological states between each cell line.

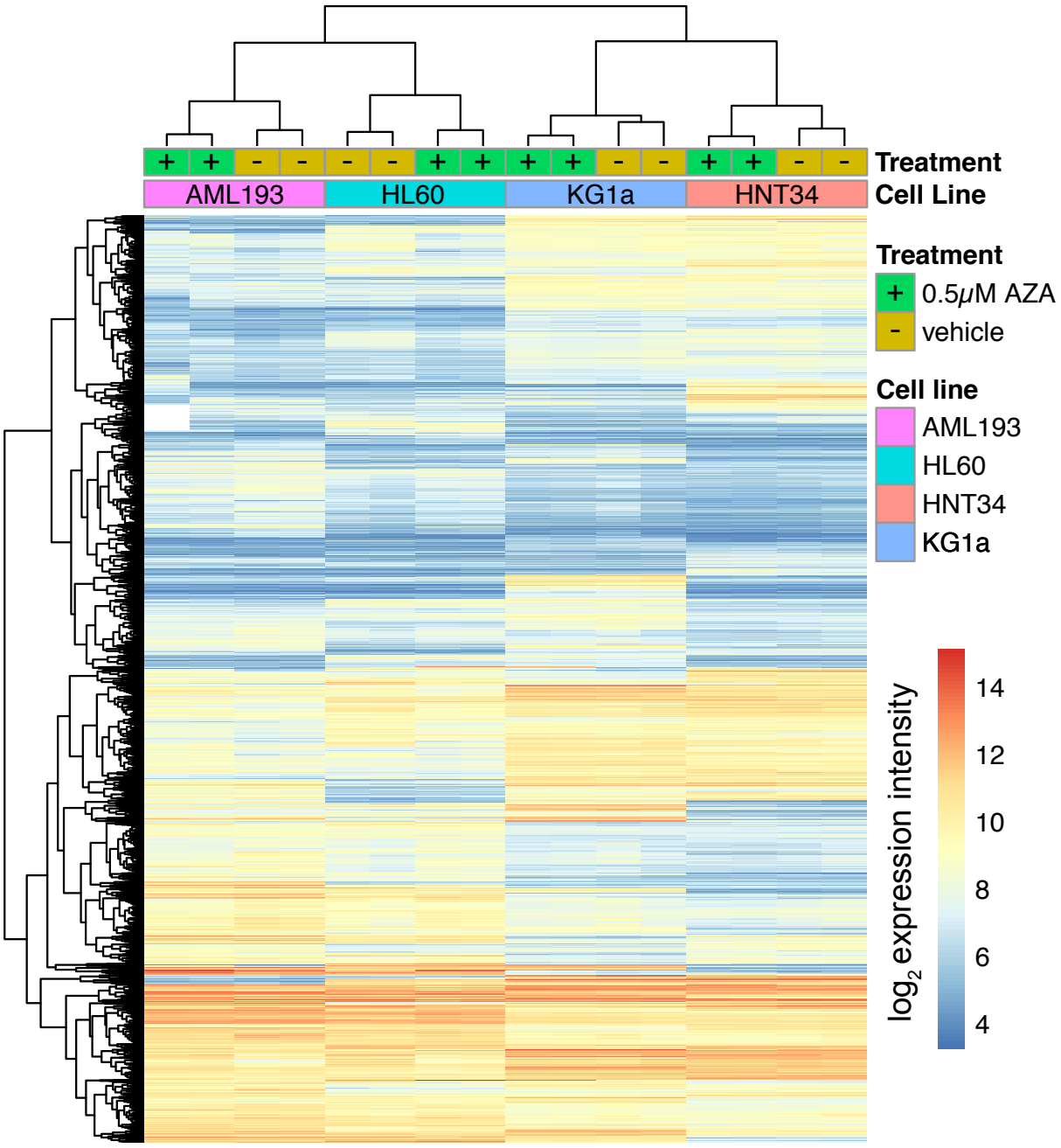


Figure S6. Hierarchal clustering of top variable gene expression profiles (Interquartile range of probesets between four cell lines > 1, n = 5747). The differences in gene expression varied more among cell lines than by AZA treatment.

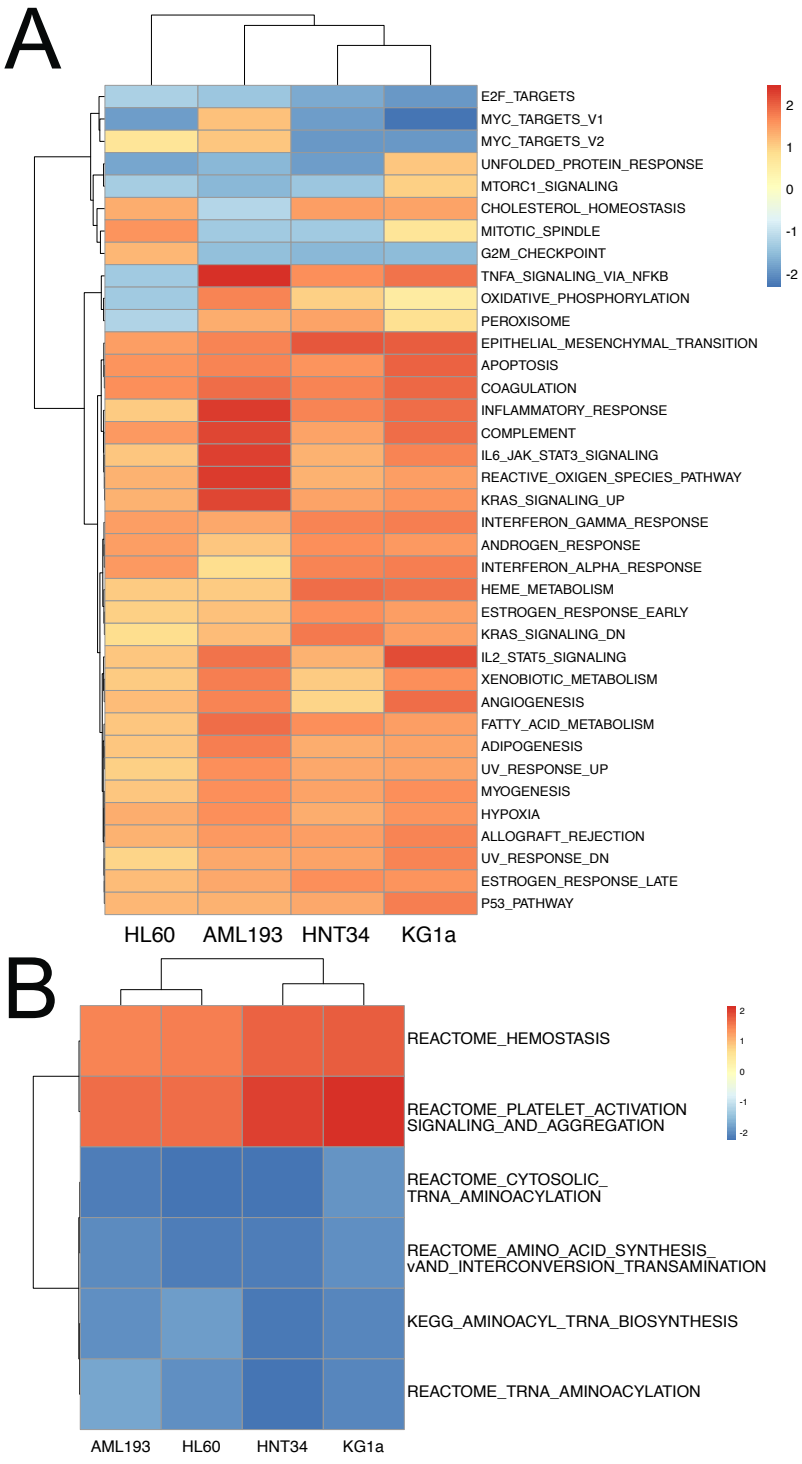


Figure S7. Gene Set Enrichment Analysis (GSEA) of gene expression data reveals common biologically processes induced by AZA. (A-B) GSEA of differentially expressed genes (adjusted p value < 0.05) using hallmark (A), reactome and KEGG gene sets (B).

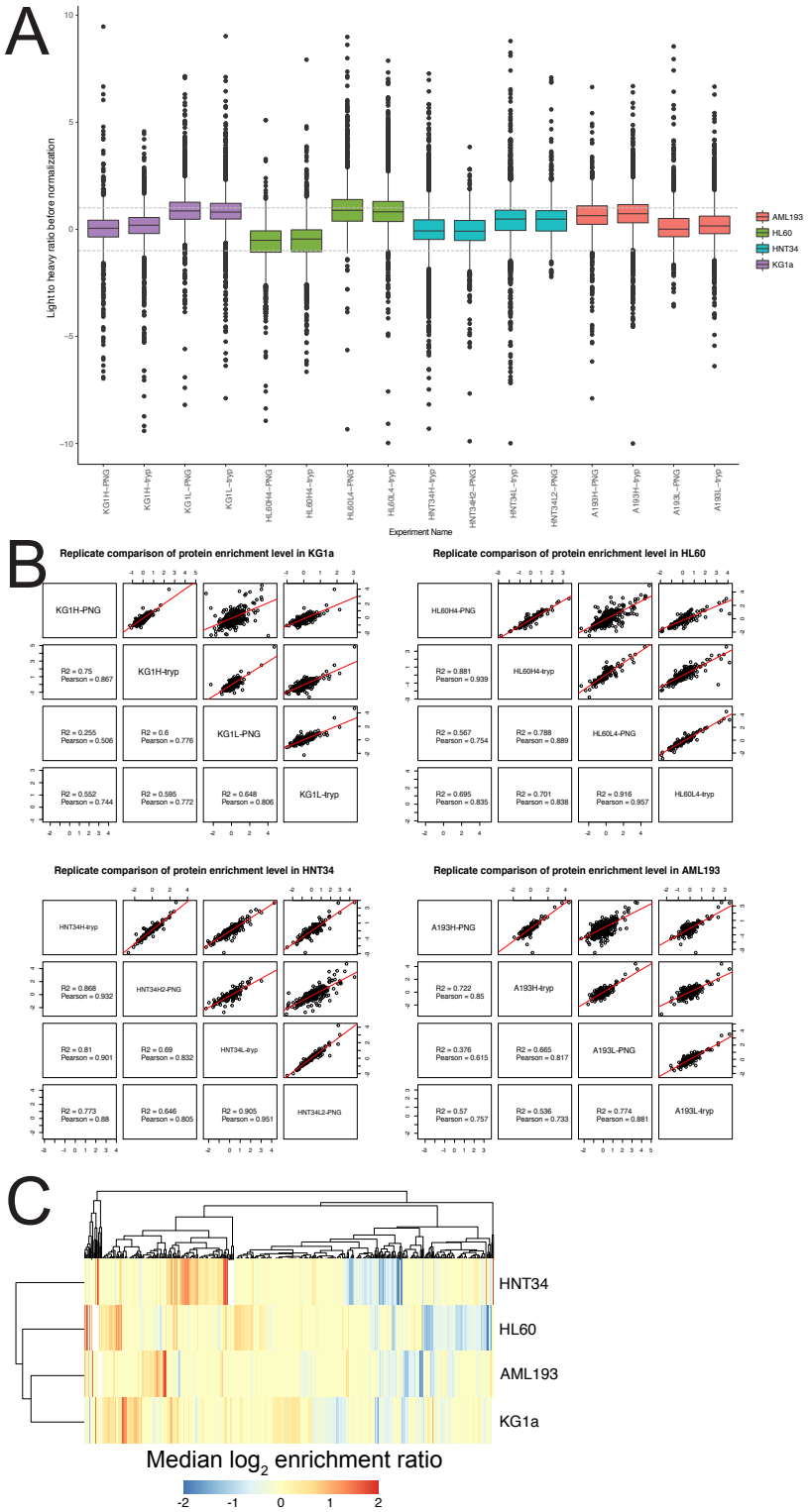
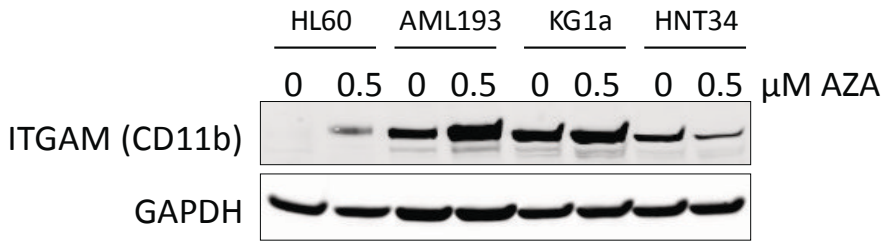
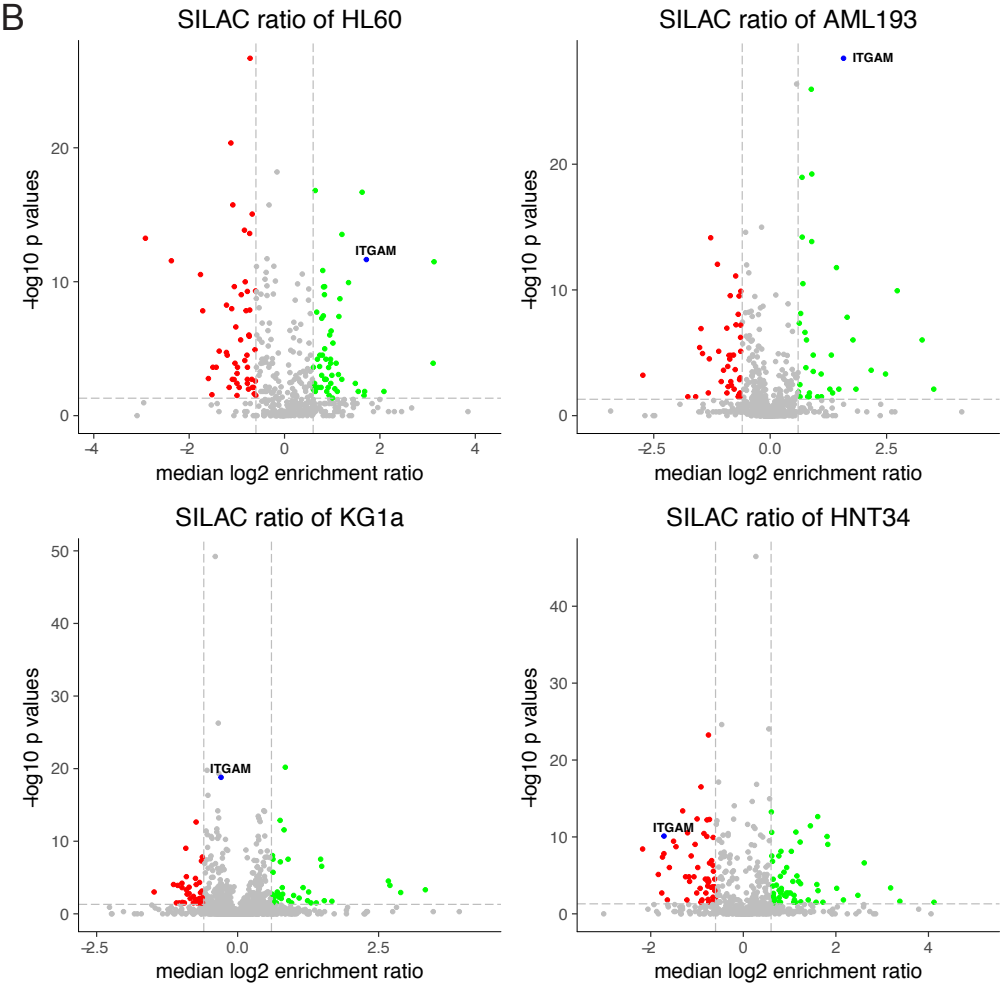


Figure S8. Proteomics analysis statistics. (A) Boxplot of light/heavy SILAC ratio for all datasets. (B) Replicate comparison between each cell line and each fraction. (C) Hierarchal clustering of log_2 SILAC surface protein enrichment ratio showed a distinct response across the four cell lines when treated with AZA and did not correspond to AML lineage marker similarity.

A



B



C

ITGAM fold enrichment	HL60	AML193	KG1a	HNT34
By western blot	up	up	no change	down
By mass spectrometry (log2)	1.7	1.58	-0.29	-1.71

Figure S9. Validation of ITGAM regulation by western blot. (A) Western blot of ITGAM in each cell line with vehicle or AZA treatment indicate different response in each cell line. (B) Volcano plots of surface proteomics data with ITGAM labeled in blue in each cell line. (C) Summary of ITGAM regulation indicating up-regulation in HL60 and AML193 cells, no change in KG1a cells and is down-regulation in HNT34 cells.

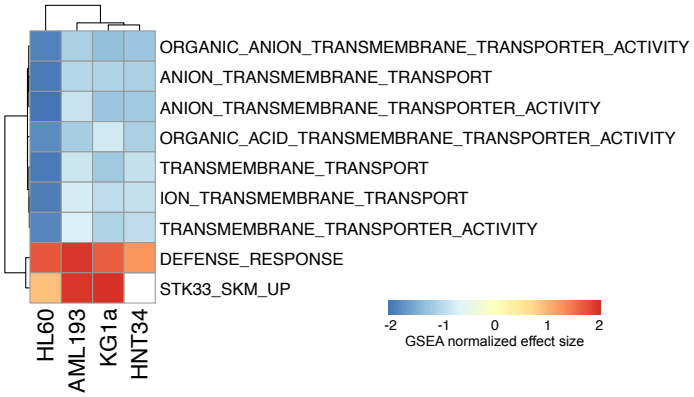


Figure S10. Top enriched gene sets affected by AZA treatments identified by GSEA of the proteomics dataset using GO term analysis shows common functional changes despite few specific proteins that overlapped.

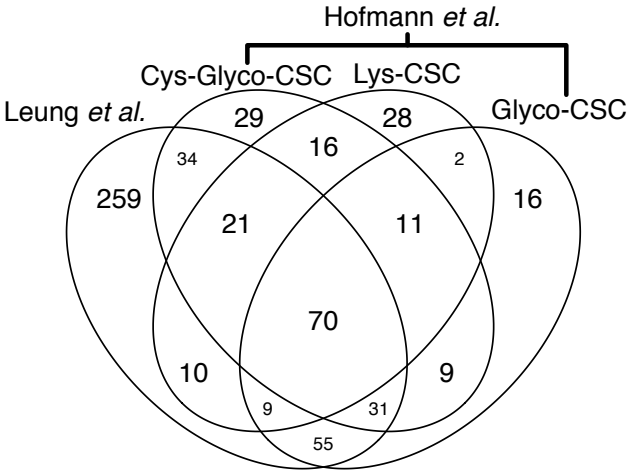


Figure S11. Comparison of surface proteomics data on HL60 cells. Overlap of proteins identified in current study compared to those identified by Hofmann *et al.* (16). A total of 230 proteins were commonly identified in both studies.

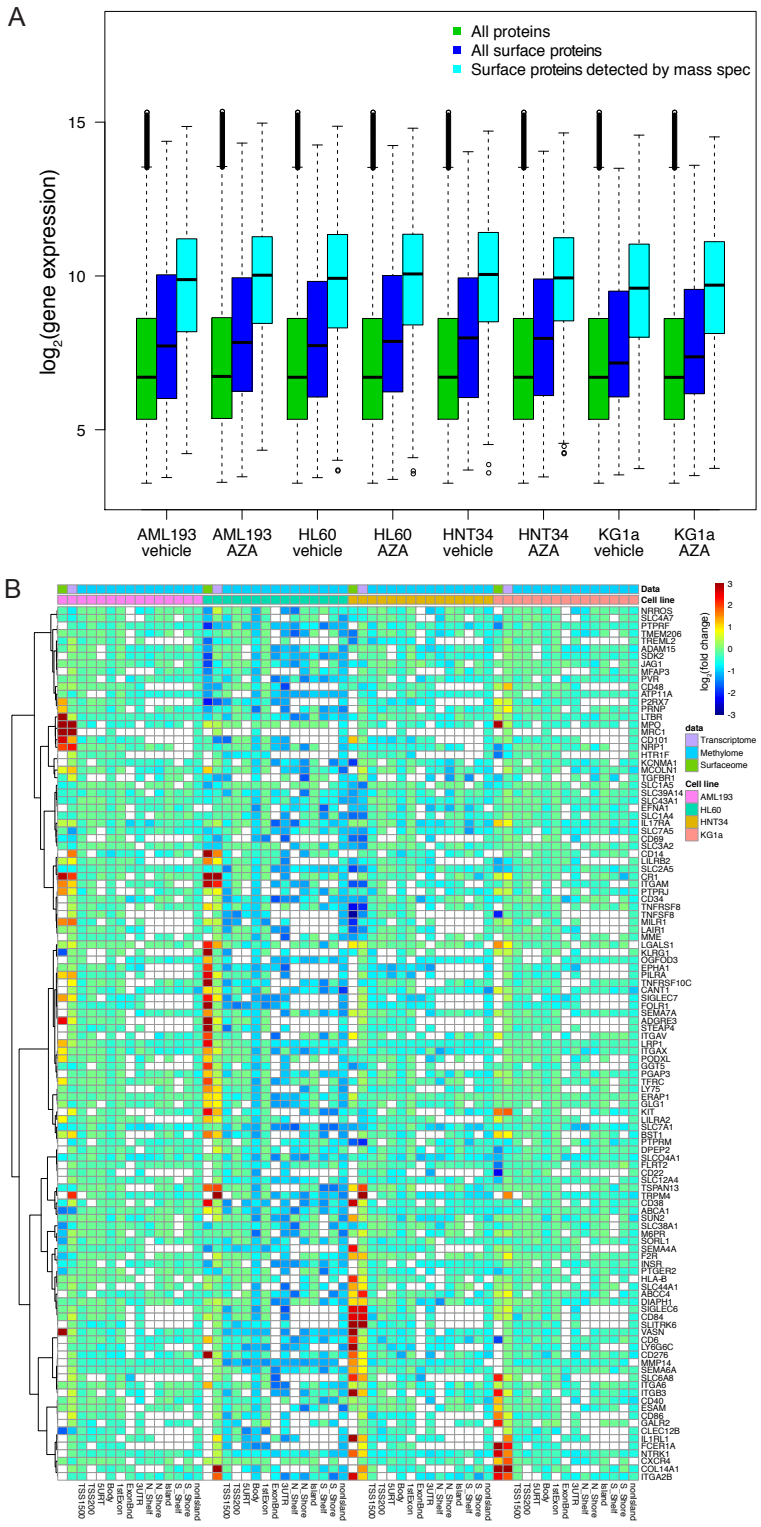


Figure S12. (A) Comparison of $\log_2$ RNA expression profile of all genes (green), of all genes annotated to be surface proteins (blue), and of genes identified by mass spectrometry experiment (cyan) across all cell lines. Whiskers extends up to 1.5 interquartile range above or below the box. (B) Extension to Figure 4D, showing changes in protein, gene expression, and DNA methylation for all surface protein

observed are shown in a heatmap. In addition to TSS1500, methylation changes in other regions such as TSS200, 5'UTR, gene body, 1st exon, exon boundary, 3'UTR, north shelf, north shore, CpG island, south shore, south shelf, and non CpG island are also included.

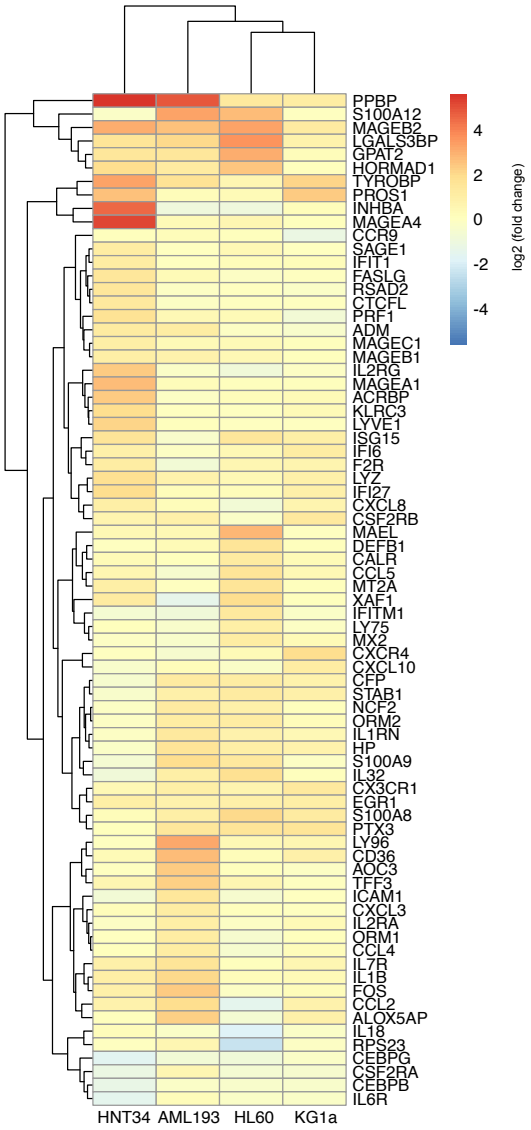


Figure S13. Expression of specific transcripts in the AIM gene set shows an overall enrichment of the AIM gene set to various degree.

Table S1. Viable cell count at the end of AZA treatment.

	Condition	Viable Cells/mL (x10^6) (Day 7)	% of Control
KG1a	vehicle	1	100.0
	0.5uM AZA	0.3	30.0
HL-60	vehicle	2.6	100.0
	0.5uM AZA	1.9	73.1
HNT-34	vehicle	1.3	100.0
	0.5uM AZA	0.35	26.9
AML-193	vehicle	1.39	100.0
	0.5uM AZA	0.22	15.8

Table S2. Comparison of AML cell lines according to known CD markers, gene expression, and peptide spectrum match.

		CD3	CD4	CD13	CD14	CD19	CD33	CD34
CD Marker[a]	KG1a	-	-	+	-	-	+	+
	HL60	-	+	+	-	-	+	-
	HNT34	-	+	+	-	-	+	+
	AML193	-	+	-	-	-	+	-
Gene expression[b]	KG1a	8.6	7.8	9.6	5.8	6.1	9.5	12.5
	HL60	8.1	11.3	11.4	8.4	6	11.3	7.6
	HNT34	8.2	8.6	11.4	5.8	6	9.4	11.7
	AML193	7.5	9.2	5.2	5.8	5.8	10.7	7.7
Peptide spectrum match[c]	KG1a	0	0	52	0	0	20	21
	HL60	0	10	62	0	0	15	0
	HNT34	0	2	72	0	0	6	4
	AML193	0	10	0	0	0	20	0

[a] CD markers detected by flow cytometry taken from DMSZ cell line repository (www.dsmz.de)

[b] Normalized gene expression profile in $\log_2$ scale

[c] Number of well-quantified peptides spectrum matches identified by mass spectrometry in cells with vehicle treatment

References

1. Furley AJ, et al. (1986) Divergent molecular phenotypes of KG1 and KG1a myeloid cell lines. *Blood* 68(5):1101–7.
2. Koeffler HP, Billing R, Lusis AJ, Sparkes R, Golde DW (1980) An undifferentiated variant derived from the human acute myelogenous leukemia cell line (KG-1). *Blood* 56(2):265–73.
3. Gallagher R, et al. (1979) Characterization of the continuous, differentiating myeloid cell line (HL-60) from a patient with acute promyelocytic leukemia. *Blood* 54(3):713–733.
4. Hamaguchi H, et al. (1997) Establishment of a novel human myeloid leukaemia cell line (HNT-34) with t(3;3)(q21;q26), t(9;22)(q34;q11) and the expression of EVI1 gene, P210 and P190 BCR/ABL chimaeric transcripts from a patient with AML after MDS with 3q21q26 syndrome. *Br J Haematol* 98(2):399–407.
5. Lange B, et al. (1987) Growth factor requirements of childhood acute leukemia: establishment of GM-CSF-dependent cell lines. *Blood* 70(1):192–9.
6. Teschendorff AE, et al. (2013) A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* 29(2):189–196.
7. Ritchie ME, et al. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* 43(7):e47.
8. Irizarry RA, et al. (2012) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Sel Work Terry Speed* (February):601–616.
9. Li Q, Birkbak NJ, Gyorffy B, Szallasi Z, Eklund AC (2011) Jetset: Selecting the optimal microarray probe set to represent a gene. *BMC Bioinformatics* 12. doi:10.1186/1471-2105-12-474.
10. Hänzelmann S, Castelo R, Guinney J (2013) GSVA: Gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics* 14. doi:10.1186/1471-2105-14-7.
11. Sergushichev A (2016) An algorithm for fast preranked gene set enrichment analysis using cumulative statistic calculation. *bioRxiv*:060012.
12. Wijetunga NA, et al. (2017) SMITE: An R/Bioconductor package that identifies network modules by integrating genomic and epigenomic information. *BMC Bioinformatics* 18(1):1–13.
13. R Core Team (2017) R: A language and environment for statistical computing. *R Found Stat Comput Vienna, Austria*.
14. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 30(1):207–10.
15. Wollscheid B, et al. (2009) Mass-spectrometric identification and relative quantification of N-linked cell surface glycoproteins. *Nat Biotechnol* 27(4):378–386.
16. Hofmann A, et al. (2010) Proteomic cell surface phenotyping of differentiating acute myeloid leukemia cells. *Blood* 116(13):1–3.