Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Network-guided discovery and virtual perturbation identify clinically associated gene modules in acute myeloid leukemia

Yanming Li

yli8@kumc.edu

University of Kansas Medical Center

Peng Gao

Jilin Provincial People's Hospital

Jinfeng Li

Jilin Provincial People's Hospital

Yawen Guo

Jilin Provincial People's Hospital

Kun Fu

Jilin Provincial People's Hospital

Tao Feng

Jilin Provincial People's Hospital

Xuemei Wang

Jilin Provincial People's Hospital

Article

Keywords:

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11749320/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** conflict of interest to disclose.

Network-guided discovery and virtual perturbation identify clinically associated gene modules in acute myeloid leukemia

Yanming Li[1,2,*], Peng Gao[3], Jinfeng Li[3], Yawen Guo[3], Kun Fu[3], Tao Feng[3], Xuemei Wang[3]

1. Department of Biostatistics & Data Science, University of Kansas Medical Center, Kansas City, KS 66160, USA.
2. The University of Kansas Cancer Center, Kansas City, KS 66160, USA.
3. Department of Hematology, Jilin Provincial People's Hospital, Changchun, Jilin 130021, China.
* Corresponding author

Correspondence: Yanming Li, PhD, Department of Biostatistics & Data Science, University of Kansas Medical Center, 3901 Rainbow Boulevard, Kansas City, KS 66160, USA. Email: yli8@kumc.edu.

Running title: AML gene modules and virtual perturbation

Abstract: 190 words | Main text: 3309 words | Figures: 6 | Tables: 2

Article type: Article

Competing interests: The authors declare no competing interests.

Abstract

Acute myeloid leukemia (AML) expression programs reflect leukemic differentiation and the marrow environment. We investigated whether network-guided analysis identifies clinically associated gene modules whose connections contribute to prediction. Network-based linear discriminant analysis modeled one-year survival using 660 KEGG-pathway genes and standardized age in BEAT2. Frozen-model ablation and disconnection evaluated predictive contributions across seven additional cohorts. Public CRISPR screens, single-cell and spatial transcriptomics, and clinical analyses characterized the modules. The model selected 200 genes, forming seven KEGG predictive regulatory modules (KPRMs) and seven singletons. Disconnecting KPRM 1 network-recruited connections reduced mean evaluation-cohort area under the ROC curve (AUC) by 0.0145 after score-dispersion matching, exceeding matched gene-set and edge references (adjusted reference-rank probabilities, 0.0277 and 0.0140). KPRM 1 showed stronger AML-relative CRISPR dependency than pathway- and baseline-dependency-matched gene sets against other hematologic cancers ($P = 0.00340$; $q = 0.0238$). Cellular analyses localized KPRM 1 to myeloid compartments. Higher overall gene scores were associated with lower induction-remission odds in BEAT2 (odds ratio per standard deviation, 0.616; 95% CI, 0.473–0.804) and GSE146173 (0.532; 0.369–0.768). Network-guided discovery and virtual perturbation identified clinically associated AML modules with experimental dependency evidence, prioritizing gene programs for functional investigation.

Keywords: acute myeloid leukemia; netLDA; gene modules; virtual perturbation; CRISPR dependency; treatment response

1. 1 Introduction

Acute myeloid leukemia (AML) is heterogeneous in its genetics, differentiation states, and treatment outcomes. Molecular risk classification has improved disease characterization, but survival and treatment response vary within established categories [1, 2]. Expression profiles complement genetic annotations by capturing leukemic programs and the marrow environment. Identifying coordinated gene programs could clarify these clinical associations and prioritize biological targets.

PI3K–Akt, JAK–STAT, MAPK, NF-κB, and Wnt signaling connect survival, differentiation, and immune-associated processes. Functional genomic profiling links AML molecular features with drug sensitivity [3], while single-cell studies resolve malignant hierarchies and associated immune populations [4]. These findings motivate investigation of signaling programs across malignant and nonmalignant compartments.

Genes with weak marginal outcome associations can contribute to prediction through correlated expression [5, 6, 7]. We asked whether network-recruited genes define clinically associated AML modules and whether fitted connections contribute information beyond individual expression effects. We applied network-based linear discriminant analysis (netLDA) [6, 7] to five KEGG pathways in BEAT2 and examined the modules across additional AML cohorts.

Frozen-model ablation and network disconnection distinguished expression from connection contributions. Public CRISPR dependency screens supplied an experimental fitness endpoint, and single-cell, spatial, and clinical analyses characterized module context. KPRM 1 was the principal biological example; the TP53-containing module provided a complementary clinical case study.

2 Materials and methods

2.1 Study design and cohorts

BEAT2 was the development cohort for an age-adjusted one-year survival classifier. Connected components defined KEGG predictive regulatory modules (KPRMs), characterized through predictive perturbation, CRISPR screens, and functional and clinical analyses. The frozen classifier was evaluated in seven accession- or platform-defined cohorts; GSE146173 and GSE106291 contain overlapping patients. The workflow is shown in Supplementary Figure S16.

The frozen model remained the biological reference; patient-grouped refits assessed robustness. Cohort-specific Cox analyses [8, 9, 10] and comparator signatures are detailed in Supplementary Information Section A5.

2.2 Expression preprocessing and endpoint definition

Expression and clinical data came from Beat AML (BEAT2) [11], TCGA-LAML [12], and GEO [13] datasets

GSE146173, GSE37642, GSE12417, and GSE106291 [14, 15, 16, 17]. GPL570 and GPL96 subsets of GSE37642/GSE12417 were analyzed separately (Table 1).

Death at or before day 365 defined short survival; follow-up beyond day 365 established long survival. Earlier censoring was excluded. Eligibility yielded 571 BEAT2 specimens and 135, 241, 134, 409, 152, 77, and 245 specimens in the seven evaluation cohorts. Excluding three BEAT2 specimens with missing age left 568 specimens from 513 patients. Clinical-response analyses used separate eligibility criteria (Supplementary Information Section A4).

Cohort-specific identifiers linked clinical and expression records. GSE146173 age was recovered from GEO; TCGA survival identifiers were harmonized at patient level. Inclusion audits were retained.

Expression inputs comprised supplied BEAT2 normalized values, TCGA-LAML Xena Toil expected counts, log2(count + 1) values for GSE146173, processed series matrices for GSE37642/GSE12417, and the GSE106291 RNASeq worksheet. Gene-symbol mapping retained the highest-variance feature for duplicate symbols (Supplementary Information Section A4).

Candidates came from PI3K–Akt (hsa04151), JAK–STAT (hsa04630), MAPK (hsa04010), NF-κB (hsa04064), and Wnt (hsa04310) pathways [18]. Annotations retrieved October 1, 2026 yielded 865 genes; common measurement and BEAT2 finite-value/variance filtering retained 660. Expression was standardized within each cohort before endpoint filtering; nonfinite standardized columns were set to zero. netLDA estimated additional fitting-partition transformations.

Age used BEAT2 fitting-partition standardization, retained externally. Out-of-fold age and gene components were standardized and combined by penalized logistic reweighting without changing the gene network (Supplementary Information Section A1).

[Table 1 about here]

2.3 Network fitting and module definition

The gene network was estimated from BEAT2 expression, with standardized age entered separately. The biology-informed and ML-derived hub union contained 69 unique genes, including 57 available candidates ranked by absolute standardized survival-class contrast. Genes entering through their network neighborhoods were termed network-recruited genes.

BEAT2 cross-validation selected $\tau = 20$ hubs, $\nu = 200$ total genes, and absolute correlation threshold $\alpha = 0.50$. Default searches considered ten strongest connected neighbors per expansion (nb = 10) through three layers (d = 3); varying these defaults did not change predictions. Generalized inverses of local covariance blocks formed the operator Ω.

For selected genes j and k, normalized inverse-covariance weights from the saved operator were

$$w_{jk} = \frac{-\Omega_{jk}}{\sqrt{\Omega_{jj}\Omega_{kk}}}$$

Edges with absolute weights $>10^{-8}$ defined the fitted association graph. Components containing ≥2 genes defined KPRMs; isolated genes were singletons. Fractional KEGG membership assigned 1/k to each pathway for genes annotated to k pathways.

Each gene coefficient was separated into gene-specific diagonal and off-diagonal network contributions:

$$\beta_j = \Omega_{jj}\delta_j + \sum_{k \neq j}\Omega_{jk}\delta_k$$

δ denotes the standardized short-minus-long class contrast. Coefficient signs were compared with marginal contrasts. Connected recruited genes with marginal Benjamini–Hochberg (BH) false discovery rate (FDR) ≥0.05 [19] and nonzero network contribution were annotated as hidden co-regulator candidates (Supplementary Information Section A1).

We used five outer and five inner patient-grouped folds, keeping each patient's specimens together. Ranking, scaling, and network fitting were repeated within fitting partitions. Inner out-of-fold scores trained age–gene weights; outer held-out scores assessed discrimination. Patient-level scores averaged specimen scores; stratified patient bootstraps provided intervals. Five outer networks were compared with reference modules (Supplementary Information Section A13).

2.4 Predictive perturbation and virtual knockout

Gene ablation removed midpoint-centered expression contributions, retaining other coefficients. Network disconnection zeroed off-diagonal Ω entries incident to target genes and reconstructed coefficients from frozen class contrasts, retaining diagonal terms. Age effects and reweighting remained fixed.

Targets included all genes, hub and network-recruited subsets, and module components. Dispersion matching restored perturbed gene scores to their BEAT2 standard deviation. Baseline minus perturbed AUC measured loss, summarized by cohort and unweighted seven-cohort mean. Cohort intervals used 1,000 paired, stratified patient bootstraps.

Gene-set references matched size and hub composition, adding edge-burden matching for disconnection. Edge references matched counts and endpoint types; positive-definite references were assessed separately. Upper-tail rank probabilities were BH-adjusted across 21 ablation or 28 disconnection comparisons within each reference background and scaling specification (Supplementary Information Section A7).

Complementary scTenifoldKnk/scTenifoldNet virtual knockout [20, 21] reused one regulatory network from 700 malignant CD14 cells and 2,000 genes for three targets and nine controls (Supplementary Information Section A10.1).

2.5 Functional characterization

Gene set variation analysis (GSVA) [22] scored KPRMs and 50 MSigDB Hallmark sets [23]; ESTIMATE [24] summarized immune/stromal expression. Cohort-specific regressions adjusted for standardized age and frozen risk group. Hallmark analyses were repeated after removing KPRM genes. BH families covered 350 Hallmark or 21 ESTIMATE tests per cohort/specification.

STRING v12.0 [25] functional and physical networks used confidence thresholds 0.400/0.700 without added proteins. Verified mappings covered 658 candidate and all 200 selected genes. Connectivity was compared with 10,000 sets matched for size or additionally exact KEGG membership and degree bins. Seven-module BH adjustment was applied for each network, threshold, and background (Supplementary Information Section A6).

2.6 Cellular and spatial characterization

BoneMarrowMap-projected [26, 27] GSE116256 [4] and adult GSE185381 [28] cells included diagnosis-stage AML and healthy controls. Cells passed mapping criteria, assignment probability ≥0.50, valid annotations, and nonzero expression. UCell [29] scored full, hub, and recruited signatures within the common transcriptomic universe (maxRank = 1,500; coverage ≥70%). Author labels distinguished malignant, AML-associated nonmalignant, and healthy cells.

Sample-level means within each state and malignancy category required ≥20 cells and received equal weight. Paired tests required ≥3 complete pairs; Welch tests required at least three nonoverlapping samples per group. State-versus-rest

tests used each sample's other malignant states. BH families covered all modules and states within each study, signature component, and comparator, separately for fine-state and broad-lineage analyses. Replication required concordant direction and FDR <0.05 in both studies.

CellChat [30] networks used 12 shared populations and sample-level recurrence tests (Supplementary Information Section A10; Supplementary Figure S7).

Spatial analyses used four AML marrow specimens from GSE255217 and GSE279576 [31, 32], frozen UCell signatures, and robust cell type decomposition (RCTD) [33] compartment weights. Moran's I [34] used 999 within-component permutations; compartment associations used 1,999 graph-spectral surrogates, with BH correction per specimen/family. Depth adjustment, thinning, and reference sensitivity assessed technical dependence (Supplementary Information Sections A8 and A10).

2.7 Response and drug-sensitivity analyses

Frozen gene scores were related to complete remission or complete remission with incomplete hematologic recovery (CR/CRi) versus refractory disease by age and sex-adjusted logistic regression in BEAT2 (306 adult diagnosis-stage patients) and GSE146173 (214 patients). Overall, module, and component effects were reported per cohort-specific primary-endpoint SD, with BH correction by analysis family.

Twenty repeats of stratified five-fold cross-validation compared clinical models with and without frozen scores on identical patients. BEAT2 included 298 study-inclusion-day-zero patients, including 286 with European LeukemiaNet (ELN) 2017 risk. GSE146173 response slopes were estimated within that cohort (Supplementary Information Section A12).

Specimen-matched BeatAML screens [11] evaluated eight drugs using viability AUC (higher values indicate lower sensitivity). Models adjusted for age, sex, and source, with heteroskedasticity-consistent HC3 confidence intervals; sensitivities added differentiation scores and ELN. BH families covered 64 drug–score tests per specification.

Independent Finnish VenEx/Helsinki data [35] included 69 patients with venetoclax drug-sensitivity scores (DSS) and 68 with clinical response. Common differentiation signatures supported adjusted comparisons. Eight GSE271406 patients [36] supplied exploratory single-cell response analyses (Supplementary Information Sections A10 and A12).

2.8 TP53 module case study

TP53-module scores covered all 44 genes, 43 neighbors, or ATM/PTEN neighbors annotated to p53 signaling (hsa04115; accessed October 17, 2026). TP53 expression was unweighted. Scores used frozen coefficients and BEAT2 diagnosis-reference standardization.

Cox models adjusted for age and sex in BEAT2 and age externally; response models adjusted for age and sex. Additional models included TP53 alteration status. BH families covered four separate scores per cohort/endpoint (Supplementary Information Section A3).

2.9 Experimental CRISPR evidence

DepMap 24Q4 [37] Chronos scores [38] measured knockout fitness loss. Averaging cell lines within PatientID yielded 26 AML, 1,093 other-cancer, and 103 other-hematologic units. Welch module contrasts used seven-module BH correction per comparator. Benchmarks used 19,999 sets matched for size, exact KEGG membership, and non-AML dependency mean/variability. Upper-tail Monte Carlo tests included ties and plus-one correction, followed by seven-module BH adjustment.

Contributor exclusion and alternative backgrounds assessed robustness. GSE221578 target-versus-control comparisons used target-excluded UCell signatures and TMM/limma pseudobulk models [39, 40] (Supplementary Information Section A11).

3 Results

3.1 Network-guided discovery identifies seven AML modules

The model selected 20 hub and 180 network-recruited genes. Seven KPRMs contained 193 genes (64, 51, 44, 17, 12, 3, and 2); seven genes were singletons. Anchors were EIF4B, RAC3, GNB1, EDAR, COL6A1, PTK2, and CACNB2. Modules spanned pathway boundaries; KPRM 1 included all five pathways (Table 2; Figure 1).

Patient-grouped refits formed five to eight multi-gene components. KPRM 1 retained 81.9% of reference genes on average (best-component mean Jaccard, 0.606). EIF4B, PAK1, SIAH1, CSNK1E, and IFNGR2 recurred in all five refits. KPRM 5 had 95.0% mean retention; complete boundaries varied (Supplementary Table S5; Supplementary Figure S15).

[Table 2 about here]

[Figure 1 about here]

Network contributions reversed 83 of 200 coefficient signs relative to marginal contrasts (41.5%; Supplementary Figure S5). Frozen age-adjusted netLDA groups differed in survival in six of seven evaluation cohorts (Supplementary Information Section A5; Supplementary Figure S6).

Patient-grouped nested validation yielded AUC 0.746 (95% CI, 0.703–0.787) after reweighting, versus 0.647 for the original rule and 0.717 for its age component (paired difference versus age, 0.029; 95% CI, 0.014–0.044; Supplementary Table S4).

3.2 KPRM 1 connections contribute to discrimination

The perturbation implementation reproduced frozen-model predictions to numerical precision, enabling comparison of expression and connection contributions.

Ablating all genes, hubs, or recruited genes reduced mean evaluation AUC by 0.0271, 0.0206, and 0.0034. Corresponding raw disconnection losses were 0.0653, −0.0034, and 0.0643, attenuating to 0.0066, −0.0018, and 0.0063 after dispersion matching (Figure 2A).

KPRM 1 had the largest complete-module disconnection loss: 0.0570 before and 0.0144 after dispersion matching, versus 0.0017 after gene ablation. Disconnecting its 62 network-recruited genes produced a matched loss of 0.0145 (Figure 2B). AUC decreased in five of seven cohorts; intervals excluded zero in GSE37642/GPL96 (0.0264; 95% CI, 0.0095–0.0432) and GSE12417/GPL96 (0.0351; 95% CI, 0.0074–0.0620; Figure 2C).

The complete-module loss exceeded matched gene-set references (mean, 0.0036; central 95% range, −0.0003–0.0073; q = 0.0277) and positive-definite edge references (mean, 0.0039; range, 0.0012–0.0066; q = 0.0140). Network-recruited disconnection likewise exceeded both backgrounds (q = 0.0277 and 0.0140; Figure 2D). None of 21 gene-ablation comparisons exceeded its matched background after adjustment.

Single-cell virtual knockout identified SRGN after PAK1 perturbation, but SRGN also responded to six of nine controls. No module gene showed a significant downstream response (Supplementary Information Section A10.1).

[Figure 2 about here]

3.3 CRISPR screens support KPRM 1 dependency

DepMap covered 195 selected genes. KPRM 1 showed greater dependency in AML than other cancers (mean gene-effect difference, −0.0338; 95% CI, −0.0460 to −0.0215; $q = 2.21 \times 10^{-5}$) and other hematologic cancers (−0.0246; 95% CI, −0.0377 to −0.0114; q = 0.00389; Figure 3A). The contrast persisted after removing the strongest one, three, or five contributors; all 62 leave-one-gene-out contrasts retained direction (Figure 3D). MYB in KPRM 2 also showed greater AML dependency than other hematologic cancers (q = 0.00183; Figure 3C).

Against other hematologic cancers, KPRM 1 exceeded the pathway- and baseline-dependency-matched expectation (observed strength, 0.02479; expected, 0.00135; P = 0.00340; seven-module q = 0.0238; Figure 3B). It was the only module meeting this adjusted threshold. Size-only and baseline-matched sensitivities supported the result. The all-other-cancer benchmark had P = 0.0107 and q = 0.0749 (Supplementary Figure S10).

GSE221578 identified no FDR-significant module or selected-gene expression responses across patients following CDK6, JAK2, or FLT3 perturbation (Supplementary Information Section A11; Supplementary Figures S11–S12).

[Figure 3 about here]

3.4 Modules connect signaling and immune programs

In BEAT2, 291 of 343 estimable age- and risk-adjusted KPRM–Hallmark associations had $q < 0.05$ (Figure 4A). KPRM 1 was associated with IL6–JAK–STAT3 signaling (standardized $\beta = 0.886$; $q = 2.46 \times 10^{-175}$), and KPRM 2 with apoptosis ($\beta = 0.846$; $q = 1.52 \times 10^{-143}$). Both were positive and significant across all eight cohorts (maximum q, 5.46×10^{-9} and 1.56×10^{-13}).

Removing KPRM genes from Hallmark sets preserved 98.5% of association directions across 2,793 estimable paired comparisons; the median absolute coefficient change was 0.008 (Figure 4B–C). KPRM 1 complement, KPRM 2 IL2–STAT5, KPRM 3 protein-secretion, KPRM 4 allograft-rejection, and KPRM 6 myogenesis associations remained significant in all cohorts.

KPRM 1 was associated with ImmuneScore and StromalScore in every cohort (β ranges, 0.440–0.822 and 0.175–0.795). KPRM 4 ImmuneScore and KPRM 6 StromalScore associations were also positive throughout (all $q \leq 0.000203$ and ≤ 0.035; Figure 4D).

The high-confidence STRING functional network contained 1,164 edges, including 330 within modules and 787 between modules; 86.0% linked two network-recruited genes (Figure 4E).

Connectivity exceeded pathway- and degree-matched expectations (1,164 versus 943.0 edges; 1.23-fold; $P = 1.00 \times 10^{-4}$). KPRMs 1–5 were individually enriched (ratios, 1.47–4.23; $q = 0.00140$–0.0192; Figure 4F). The physical network was enriched overall (356 versus 302.9 edges; $P = 0.0121$), with module-level evidence concentrated in KPRM 1 (2.13-fold; $q = 0.000700$; Supplementary Information Section A6).

[Figure 4 about here]

3.5 Myeloid localization across cellular and spatial data

Filtering retained 16,499 cells from 17 primary sample units and 124,141 cells from 30 replication units. All 193 module genes were measured in both datasets. Across 13 matched malignant states, KPRM 1 enrichment extended through progenitor and differentiated myeloid populations (Figure 5A–D).

Malignant-minus-nonmalignant CD14 differences were positive in five primary pairs (0.0122 UCell units; P = 0.0302; q = 0.0705) and six replication pairs (0.0162; P = 0.0134; q = 0.0940; Figure 5E), without meeting joint FDR replication.

Broad-monocyte contrasts retained direction (Figure 5F). Within malignant cells, monocyte enrichment relative to other lineages was strongest in replication (difference, 0.0277; n = 13; q = 0.0033; Figure 5G). Replication CD14 scores exceeded other malignant states in 12 of 13 sample units (q = 0.0147).

[Figure 5 about here]

Communication analysis identified 89 of 96 recurrent interactions in replication, including incoming MIF signaling, but no module–communication correlation met joint FDR criteria (Supplementary Figure S7).

In four spatial specimens (4,817 spots), KPRM 1 enrichment correlated with monocytic weights in BM1 and BM2 after depth adjustment ($\rho = 0.189$ and 0.212, respectively) and thinning ($\rho = 0.130$ and 0.109, respectively; all $q \leq 0.00700$). AML11 and AML12 associations were nonsignificant. KPRM 5 retained thinned clustering in BM1 and BM2 (both q = 0.00700; Supplementary Figure S17).

3.6 Clinical response associations

Higher overall gene scores were associated with lower CR/CRi odds in BEAT2 (211 responders/95 refractory; OR per SD, 0.616; 95% CI, 0.473–0.804; q = 0.00190) and GSE146173 (156/58; OR, 0.532; 95% CI, 0.369–0.768; q = 0.00308; Figure 6A). Restriction to 298 BEAT2 day-zero patients retained the association (OR, 0.587; 95% CI, 0.445–0.774).

KPRMs 2, 3, 6, and 7 passed response BH correction in BEAT2; none did in the GSE146173 primary endpoint. Both hub and recruited components were separately associated in both cohorts. In GSE146173, hubs added information beyond recruited contributions (likelihood-ratio P = 0.0110; q = 0.0154); the reciprocal addition was nonsignificant (P = 0.0956). No individual gene passed BH correction (Supplementary Figure S20). Patient-level hub and network-recruited score distributions are shown in Figure 6C.

[Figure 6 about here]

Adding the overall score to age and sex increased mean cross-validated remission AUC from 0.605 to 0.663 in BEAT2 and 0.699 to 0.744 in GSE146173. Among 286 ELN-complete BEAT2 patients, adding the score changed age and sex/ELN AUC from 0.718 to 0.719 (Supplementary Information Section A12).

The TP53-module score was associated with shorter survival in BEAT2 (HR, 1.84; 95% CI, 1.60–2.11) and TCGA-LAML (1.46; 1.20–1.79), and lower remission odds in BEAT2 (OR, 0.688; 95% CI, 0.531–0.890; q = 0.00884) and GSE146173 (0.699; 0.519–0.944; q = 0.0264). The BEAT2 response association persisted after TP53-alteration adjustment (Supplementary Figure S18).

Among 217 venetoclax-screened patients, adjustment for differentiation state and ELN reversed the KPRM 1 viability-AUC coefficient from 24.91 to −17.39 per SD (95% CI, −29.18 to −5.60; q = 0.0369), with all patient-omission coefficients negative. Independent Finnish DSS replication was not established (0.66; 95% CI, −4.82 to 6.14; P = 0.811; n = 69; Supplementary Figures S14 and S19).

4 Discussion

Network-guided analysis identified pathway-spanning AML modules with complementary clinical, cellular, and experimental dependency evidence. KPRM 1 combined a recurring gene core, predictive connection contributions, and excess AML-relative dependency under matched backgrounds. These findings prioritize a coordinated program extending beyond individually associated genes.

KPRM 1 disconnection reduced discrimination more than expression ablation and exceeded matched gene-set and positive-definite edge references after dispersion matching. This extends local-network learning [5, 6, 7] by identifying dependence structure relevant to prediction. Attenuation of raw losses after matching shows that perturbation effects also depend on score scale.

Public CRISPR screens supplied an experimental fitness endpoint for the computational module. Excess AML-relative dependency persisted against other hematologic cancers after pathway and baseline matching and exclusion of its strongest contributors, supporting a distributed program. Direct tests of particular relationships remain needed: virtual knockout responses overlapped controls, and analyzed perturbation-expression data showed no FDR-significant module response [20, 21].

Functional and cellular findings placed KPRM 1 in an inflammatory myeloid context [4, 26, 28]. Hallmark associations persisted after removing shared genes; STRING connectivity exceeded pathway- and degree-matched expectations. Single-cell and spatial localization [32] extended across malignant and nonmalignant compartments. Malignant specificity did not meet joint replication criteria, while ESTIMATE scores represented immune/stromal expression programs [24].

Survival-derived gene scores were associated with remission in both bulk cohorts and improved discrimination beyond age and sex, with negligible addition beyond ELN in BEAT2. The TP53 module further illustrated clinical information carried by expression neighborhoods. These results motivate biological stratification within established risk groups rather than replacement of molecular risk assessment.

Differentiation adjustment reversed the KPRM 1 venetoclax association [11], distinguishing its association with myeloid state from residual drug sensitivity. The Finnish analysis did not establish replication; cellular composition, drug assays, and combination-treatment endpoints differed [35]. The conditional association warrants matched follow-up.

Patient-grouped validation supported reweighting and a recurring KPRM 1 core, although boundaries varied. It retained the established candidate universe and model settings; earlier cohort-wide preprocessing remained part of development. Evaluation-population overlap, treatment-timing uncertainty, and limited spatial replication affect generalization. Supplementary comparator analyses were descriptive, with an outcome-selected retained ML panel and cohort-specific Cox refits.

Targeted perturbation of recurrent KPRM 1 genes should measure AML fitness and module responses. Independently assembled clinical cohorts could assess whether these programs refine response stratification within molecular risk groups.

netLDA identified clinically associated AML modules and highlighted KPRM 1 through predictive disconnection, matched CRISPR dependency, and myeloid localization, providing specific programs for functional follow-up.

Acknowledgements and declarations

Funding. This study received no external funding.

Ethics approval and consent to participate. This study involved secondary analyses of publicly available, de-identified datasets. No participants were recruited and no new human samples were collected.

Consent for publication. Not applicable. The analyses used publicly available, de-identified data and report no identifiable participant-level information.

Data availability. Public data were obtained from GEO (GSE146173, GSE37642, GSE12417, GSE106291, GSE116256, GSE185381, GSE255217, GSE279576, GSE271406, and GSE221578; https://www.ncbi.nlm.nih.gov/geo/), Beat AML [11], TCGA-LAML [12], BoneMarrowMap [26, 27], DepMap [37], and the Finnish study [35]. Supporting results are supplied in the supplementary materials. Study-specific scripts, frozen coefficients, module membership, and essential processed summaries are available from the corresponding author on reasonable request.

Code availability. netLDA software is available at https://github.com/lyqglyqg/netLDA [41].

Competing interests. The authors declare that they have no competing interests.

Authors' contributions. **Y. Li**: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **P. Gao**: Conceptualization, Investigation, Project administration, Validation, Visualization, Writing – original draft, Writing – review & editing. **J. Li, Y. Guo, K. Fu, T. Feng, and X. Wang**: Investigation, Validation, Writing – original draft, Writing – review & editing.

Acknowledgements. The first author, Y. Li, dedicates this work to the memory of his mother, Linyi Zhao, who passed away in 2025. Her illness and passing provided a deeply personal motivation for this research.

References

1. Döhner H, Wei AH, Appelbaum FR, et al. Diagnosis and management of AML in adults: 2022 recommendations from an international expert panel on behalf of the ELN. Blood. 2022;140(12):1345–1377. doi:10.1182/blood.2022016867.

2. Papaemmanuil E, Gerstung M, Bullinger L, et al. Genomic classification and prognosis in acute myeloid leukemia. N Engl J Med. 2016;374(23):2209–2221. doi:10.1056/NEJMoa1516192.

3. Tyner JW, Tognon CE, Bottomly D, et al. Functional genomic landscape of acute myeloid leukaemia. Nature. 2018;562:526–531. doi:10.1038/s41586-018-0623-z.

4. van Galen P, Hovestadt V, Wadsworth MH II, et al. Single-cell RNA-seq reveals AML hierarchies relevant to disease progression and immunity. Cell. 2019;176(6):1265–1281.e24. doi:10.1016/j.cell.2019.01.031.

5. Li Y, Hong HG, Ahmed SE, Li Y. Weak signals in high-dimensional regression: detection, estimation and prediction. Appl Stoch Models Bus Ind. 2019;35(2):283–298. doi:10.1002/asmb.2340.

6. Li Y, Hong HG, Li Y. Multiclass linear discriminant analysis with ultrahigh-dimensional features. Biometrics. 2019;75(4):1086–1097. doi:10.1111/biom.13065.

7. Li Y. A local-network guided linear discriminant analysis for classifying lung cancer subtypes using individual genome-wide methylation profiles. In: Arai K, Bhatia R, Kapoor S, editors. Proceedings of the Future Technologies Conference (FTC) 2019. Advances in Intelligent Systems and Computing. Vol. 1069. Cham: Springer; 2020. p. 676–687. doi:10.1007/978-3-030-32520-6_50.

8. Cox DR. Regression models and life-tables. J R Stat Soc Series B. 1972;34(2):187–202. doi:10.1111/j.2517-6161.1972.tb00899.x.

9. Harrell FE Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. JAMA. 1982;247(18):2543–2546. doi:10.1001/jama.1982.03320430047030.

10. Schoenfeld D. Partial residuals for the proportional hazards regression model. Biometrika. 1982;69(1):239–241. doi:10.1093/biomet/69.1.239.

11. Bottomly D, Long N, Schultz AR, et al. Integrative analysis of drug response and clinical outcome in acute myeloid leukemia. Cancer Cell. 2022;40(8):850–864.e9. doi:10.1016/j.ccell.2022.07.002.

12. Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. N Engl J Med. 2013;368(22):2059–2074. doi:10.1056/NEJMoa1301689.

13. Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res. 2002;30(1):207–210. doi:10.1093/nar/30.1.207.

14. Bamopoulos SA, Batcha AMN, Jurinovic V, et al. Clinical presentation and differential splicing of SRSF2, U2AF1 and SF3B1 mutations in patients with acute myeloid leukemia. Leukemia. 2020;34(10):2621–2634. doi:10.1038/s41375-020-0839-4.

15. Li Z, Herold T, He C, 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. 2013;31(9):1172–1181. doi:10.1200/JCO.2012.44.3184.

16. Metzeler KH, Hummel M, Bloomfield CD, et al. An 86-probe-set gene-expression signature predicts survival in cytogenetically normal acute myeloid leukemia. Blood. 2008;112(10):4193–4201. doi:10.1182/blood-2008-02-134411.

17. Herold T, Jurinovic V, Batcha AMN, et al. A 29-gene and cytogenetic score for the prediction of resistance to induction treatment in acute myeloid leukemia. Haematologica. 2018;103(3):456–465. doi:10.3324/haematol.2017.178442.

18. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res. 2023;51(D1):D587–D592. doi:10.1093/nar/gkac963.

19. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B. 1995;57(1):289–300. doi:10.1111/j.2517-6161.1995.tb02031.x.

20. Osorio D, Zhong Y, Li G, et al. scTenifoldKnk: an efficient virtual knockout tool for gene function predictions via single-cell gene regulatory network perturbation. Patterns. 2022;3(3):100434. doi:10.1016/j.patter.2022.100434.

21. Osorio D, Yu X, Zhong Y, et al. scTenifoldNet: a machine learning workflow for constructing and comparing transcriptome-wide gene regulatory networks from single-cell data. Patterns. 2020;1(9):100139. doi:10.1016/j.patter.2020.100139.

22. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. BMC Bioinformatics. 2013;14:7. doi:10.1186/1471-2105-14-7.

23. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1(6):417–425. doi:10.1016/j.cels.2015.12.004.

24. Yoshihara K, Shahmoradgoli M, Martínez E, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun. 2013;4:2612. doi:10.1038/ncomms3612.

25. Szklarczyk D, Kirsch R, Koutrouli M, et al. The STRING database in 2023: protein–protein association networks and functional enrichment analyses for any sequenced genome of interest. Nucleic Acids Res. 2023;51(D1):D638–D646. doi:10.1093/nar/gkac1000.

26. Zeng AGX, Iacobucci I, Shah S, et al. Single-cell transcriptional atlas of human hematopoiesis reveals genetic and hierarchy-based determinants of aberrant AML differentiation. Blood Cancer Discov. 2025;6(4):307–324. doi:10.1158/2643-3230.BCD-24-0342.

27. Zeng A. AML Single Cell RNA-seq Datasets (Zeng et al. 2025) [dataset]. Version 1. Zenodo; 2026. doi:10.5281/zenodo.18528442.

28. Lasry A, Nadorp B, Fornerod M, et al. An inflammatory state remodels the immune microenvironment and improves risk stratification in acute myeloid leukemia. Nat Cancer. 2023;4:27–42. doi:10.1038/s43018-022-00480-0.

29. Andreatta M, Carmona SJ. UCell: robust and scalable single-cell gene signature scoring. Comput Struct Biotechnol J. 2021;19:3796–3798. doi:10.1016/j.csbj.2021.06.043.

30. Jin S, Guerrero-Juarez CF, Zhang L, et al. Inference and analysis of cell–cell communication using CellChat. Nat Commun. 2021;12:1088. doi:10.1038/s41467-021-21246-9.

31. NCBI Gene Expression Omnibus. GSE255217: Spatial Transciptomics of human trephine BM samples from AML patients and healthy donors [dataset]. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE255217. Accessed October 18, 2026.

32. Dasdemir E, Veletic I, Ly C, et al. Integrative spatial multi-omics reveal niche-specific inflammatory signaling and differentiation hierarchies in AML. iScience. 2026;29(1):114289. doi:10.1016/j.isci.2025.114289.

33. Cable DM, Murray E, Zou LS, et al. Robust decomposition of cell type mixtures in spatial transcriptomics. Nat Biotechnol. 2022;40:517–526. doi:10.1038/s41587-021-00830-w.

34. Moran PAP. Notes on continuous stochastic phenomena. Biometrika. 1950;37(1–2):17–23. doi:10.1093/biomet/37.1-2.17.

35. Vänttinen I, Saad J, Ruokoranta T, et al. Drivers of clinical resistance to venetoclax and hypomethylating agents in acute myeloid leukemia and strategies for improving efficacy. HemaSphere. 2026;10(1):e70282. doi:10.1002/hem3.70282.

36. Gui G, Bingham MA, Herzog JR, et al. Single-cell spatial transcriptomics reveals immunotherapy-driven bone marrow niche remodeling in AML. Sci Adv. 2025;11(28):eadw4871. doi:10.1126/sciadv.adw4871.

37. DepMap Consortium. DepMap 24Q4 Public. Version 1. Figshare; 2024. doi:10.25452/figshare.plus.27993248.v1.

38. Dempster JM, Boyle I, Vazquez F, et al. Chronos: a cell population dynamics model of CRISPR experiments that improves inference of gene fitness effects. Genome Biol. 2021;22:343. doi:10.1186/s13059-021-02540-7.

39. Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. 2010;11:R25. doi:10.1186/gb-2010-11-3-r25.

40. Ritchie ME, Phipson B, Wu D, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015;43(7):e47. doi:10.1093/nar/gkv007.

41. Li Y. netLDA software. GitHub. https://github.com/lyqglyqg/netLDA. Accessed 22 October 2026.

Figure legends

Figure 1. netLDA-defined KPRMs and their KEGG pathway composition. (A–G) Seven non-overlapping KEGG predictive regulatory modules (KPRMs) identified from the age-adjusted BEAT2 gene network, containing 64, 51, 44, 17, 12, 3, and 2 genes, respectively. (H) Seven singleton genes. Orange nodes represent hub genes, and teal nodes represent network-recruited genes. Larger outlined nodes mark representative hub genes with the greatest absolute standardized marginal expression difference within each module. Solid and dashed edges indicate positive and negative normalized inverse-covariance weights, respectively; edge widths represent absolute weights scaled within each module. Edges with absolute weights $>10^{-8}$ are shown. Pie charts summarize membership in the five candidate KEGG pathways, with genes annotated to multiple pathways contributing equally across their pathway assignments.

Figure 2. Predictive contributions of genes and KPRM network connections in AML. (A) AUC losses following gene ablation, network disconnection, and standard-deviation (SD)-matched disconnection for all 200 selected genes, 20 retained hub genes, and 180 network-recruited genes. Circles represent the seven external cohorts; diamonds indicate their equally weighted mean. (B) Mean external AUC losses for complete-module perturbations of the seven KEGG predictive regulatory modules (KPRMs); n denotes module size. Gene ablation removes midpoint-centered expression contributions while retaining the other coefficients. Network disconnection removes the relevant off-diagonal entries of the BEAT2 network matrix, preserves its diagonal, and reconstructs gene coefficients using frozen BEAT2 marginal contrasts. SD matching rescales the perturbed gene score to its original BEAT2 SD. Age standardization and age–gene reweighting remain fixed. (C) Cohort-specific AUC losses after SD-matched disconnection of KPRM 1 network-recruited connections, with 95% paired, outcome-stratified patient-bootstrap confidence intervals (1,000 replicates). (D) Observed KPRM 1 complete-module and network-recruited effects (red diamonds) compared with matched gene-set and positive-definite edge references. Grey circles show reference means; bars span the central 95% of reference effects. Gene-set references match set size, hub/network-recruited composition, and edge burden within 10%; edge references match edge counts and hub/network-recruited endpoint types exactly. q(rank) denotes upper-tail reference-rank probabilities adjusted by the Benjamini–Hochberg method across 28 planned module/component comparisons within each reference background and scaling variant. Positive AUC loss indicates reduced discrimination. All evaluations use the seven non-BEAT2 cohorts and the frozen fitted classifier.

Figure 3. Experimental CRISPR dependency of AML KPRMs. (A) Patient-unit module contrasts against other cancers and other hematologic cancers. Positive values indicate stronger dependency in AML; bars show 95% confidence intervals. (B) KPRM 1 compared with 19,999 candidate-gene sets matched for size, exact KEGG membership and non-AML dependency mean and variability, using the hematologic comparator. (C) Mean gene effects for ten specified genes; more negative values indicate greater dependency. Lines connect group means. (D) KPRM 1 contrasts after removing its top one, three or five dependency contributors; the top five were RPS6, CDK6, GRB2, CFLAR and EIF4B. P-values in D are nominal sensitivity statistics. Significant q-values are labeled in A–C; complete results are supplied separately. Related cell lines were averaged within PatientID, yielding 26 AML and 103 other hematologic units.

Figure 4. Functional characterization of KPRMs. (A) Age- and risk-adjusted associations between KPRM and Hallmark GSVA scores in BEAT2. Displayed Hallmarks comprise the three largest absolute coefficients per module, supplemented by the strongest remaining associations to reach 18. (B) The strongest absolute BEAT2 association per module after removing all KPRM genes from Hallmark sets, shown across eight cohorts. (C) Original and overlap-removed coefficients for 2,793 estimable pairs; the dashed line denotes equality. (D) Associations with ESTIMATE ImmuneScore and StromalScore. Colors in A, B, and D indicate standardized regression coefficients. Stars indicate BH-adjusted P-values: * <0.05, ** <0.01, *** <0.001; correction is within cohort and analysis family, including 21 tests for ESTIMATE. Complete-case sample sizes were 566, 135, 241, 134, 409, 152, 77, and 245 in the displayed cohort order. (E) STRING functional associations at confidence ≥0.700. Node colors indicate modules; circles denote hub genes and squares network-recruited genes. Node size represents degree, edge width represents confidence, and line style distinguishes within-module, between-module, and singleton-associated edges. (F) Observed/expected connectivity ratios from 10,000 permutations of 658 mapped candidates; all 200 selected genes were mapped. Blue circles indicate size matching and orange triangles additional exact KEGG-membership and candidate-degree-bin matching. The dashed line marks unity. Stars use the same thresholds; the overall row uses permutation P-values and module rows use BH-adjusted P-values across seven modules separately for each background. Zero ratios indicate no internal edges. Module sizes were 64, 51, 44, 17, 12, three, and two genes. Full statistical definitions are provided in Supplementary Information Section A6.

Figure 5. Single-cell characterization of KEGG predictive regulatory modules in AML. (A) Existing BoneMarrowMap-projected UMAP coordinates for the primary dataset, GSE116256, colored by the supplied broad-lineage annotations. Each dot represents a retained cell. (B) Projected UMAPs colored by author-annotated malignancy status for GSE116256 and the adult subset of GSE185381; the nonmalignant category includes AML-associated microenvironment cells and healthy controls. (C) Sample-balanced mean UCell scores for seven KPRMs in malignant cells across 13 matched fine states with at least three eligible sample units in each dataset. (D) KPRM_001 scores in six selected shared broad lineages; dark blue indicates the primary dataset and vermilion indicates replication in panels D–G. Points represent sample-level means; vertical marks indicate dataset means. (E) KPRM_001 scores in malignant and nonmalignant CD14 monocytes from complete within-sample pairs (primary, n = 5; replication, n = 6). Lines connect matched observations. (F) Malignant-minus-nonmalignant effects for fine CD14 monocytes and pooled broad monocytes. (G) Within-sample differences between malignant monocytes and the unweighted mean of other eligible malignant broad lineages. Error bars in F and G are 95% t confidence intervals. Eligibility required at least 20 cells per sample/state/status. P denotes a two-sided paired t-test in E; q denotes the Benjamini–Hochberg FDR within each analysis family across all seven modules. KPRM_001 underwent exploratory follow-up.

Figure 6. Bulk gene scores and induction response in AML. (A) Age- and sex-adjusted odds ratios for CR/CRi versus refractory disease in BEAT2 (306 diagnosis-stage specimens: 211 CR/CRi and 95 refractory) and GSE146173 (214 samples: 156 CR/CRi and 58 refractory). Each score was fitted separately; points and bars show odds ratios per cohort-specific primary-endpoint SD and 95% confidence intervals. (B) Linear and three-degree-of-freedom natural-spline response curves in GSE146173, evaluated at median age and the female reference category; shading shows pointwise 95% confidence intervals. (C) GSE146173 patient-level contributions from 20 hub genes and 180 network-recruited genes. Components were standardized using their primary-endpoint reference means and SDs. Dots represent individual samples; boxes show medians and interquartile ranges, with whiskers extending to observations within 1.5 interquartile ranges. These distributions are unadjusted; adjusted separate and joint component estimates appear in Supplementary Figure S20B. Higher scores indicate the original shorter-survival direction. CR/CRi, complete remission or complete remission with incomplete hematologic recovery. BEAT2 response analyses used the survival-model development cohort; specimen-to-induction episode linkage was not confirmed.

Table 1 Characteristics of the eight bulk AML classification cohorts

A Source and classification sample counts

Cohort	Role	Source samples	One-year eligible	Age available	Short survival	Long survival
BEAT2	Development	649	571	568	258	310
TCGA-LAML	External	161	135	135	57	78
GSE146173	External	246	241	241	88	153
GSE37642 / GPL570	External	136	134	134	61	73
GSE37642 / GPL96	External	417	409	409	217	192
GSE12417 / GPL96	External	163	152	152	86	66
GSE12417 / GPL570	External	79	77	77	30	47
GSE106291	External	250	245	245	85	160

B Characteristics of age-available classification samples

Cohort	Age, years median (IQR)	Male sex n/available (%)	Deaths over recorded follow-up n/N (%)
BEAT2	61 (46–71)	320/568 (56.3)	396/568 (69.7)
TCGA-LAML	58 (45–67)	NR	93/135 (68.9)
GSE146173	58 (46–66)	NR	149/241 (61.8)
GSE37642 / GPL570	59.5 (45–66)	NR	98/134 (73.1)
GSE37642 / GPL96	57 (43–66)	NR	308/409 (75.3)
GSE12417 / GPL96	57 (46–67)	NR	103/152 (67.8)
GSE12417 / GPL570	62 (54–69)	NR	47/77 (61.0)
GSE106291	58 (46–66)	118/245 (48.2)	150/245 (61.2)

Short- and long-survival counts and Panel B summaries refer to samples with an available age. One-year-eligible counts precede the age requirement. Three BEAT2 samples lacked age; all other eligible samples had age available.

Short survival denotes death at or before day 365; long survival denotes follow-up beyond 365 days. All stored class labels agreed with these rules. Death counts in Panel B cover the recorded follow-up, not only the first year.

Recorded zero ages in two BEAT2 samples were retained to reproduce the existing classification population. Recorded ages below 18 occur in 18 BEAT2 samples (including the two zero ages) and one GSE12417/GPL96 sample.

Counts refer to specimens, not necessarily distinct patients. BEAT2 has 568 age-available classification samples from 513 distinct patient identifiers.

NR, not reported in the submitted clinical file. Sex percentages use records with reported sex. Sex was available for BEAT2 and GSE106291; the other six submitted clinical tables had absent or entirely missing sex fields.

GSE146173 and GSE106291 contain overlapping samples and are displayed as separate accession-level analyses. Their counts should not be added to estimate a unique-patient total. IQR, interquartile range; AML, acute myeloid leukemia.

Table 2 Molecular composition of the seven KPRMs identified by netLDA

A Module membership and fitted gene contributions

Module	Genes, n	Retained hub genes	Network-recruited genes, n	Largest absolute network-recruited coefficients
KPRM 1	64	EIF4B, PAK1 (n = 2)	62	SIAH1 (−0.651); CSNK1E (+0.540); IFNGR2 (+0.440)
KPRM 2	51	RAC3, LPAR2 (n = 2)	49	BRCA1 (+0.461); MAPK3 (+0.425); PPARD (+0.381)
KPRM 3	44	GNB1, PIK3R1, ROCK2 (n = 3)	41	DAXX (+0.750); CRK (+0.462); SOS1 (−0.318)
KPRM 4	17	EDAR, FLT3LG, LAT (n = 3)	14	IL23A (+0.258); DUSP4 (−0.221); WNT1 (−0.217)
KPRM 5	12	COL6A1 (n = 1)	11	TCF7 (−0.790); RASGRP1 (+0.397); ZAP70 (−0.319)
KPRM 6	3	PTK2 (n = 1)	2	GNG11 (+0.135); CCND1 (−0.093)
KPRM 7	2	CACNB2 (n = 1)	1	FGF16 (+0.036)

B KEGG pathway membership within each module

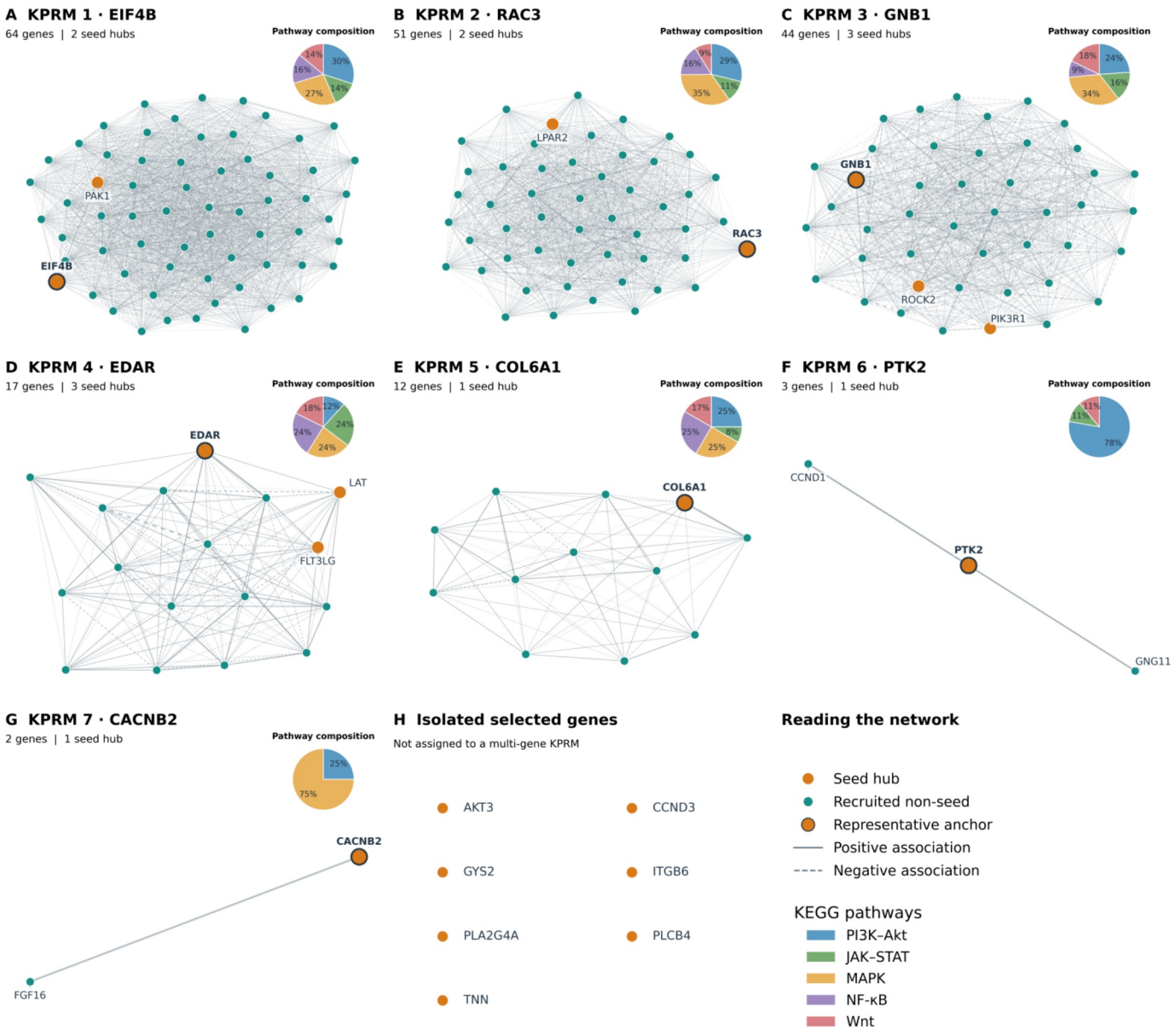
Module	PI3K–Akt (n)	JAK–STAT (n)	MAPK (n)	NF-κB (n)	Wnt (n)
KPRM 1	25	11	23	13	11
KPRM 2	17	7	23	11	6
KPRM 3	15	10	20	6	10
KPRM 4	4	4	6	4	3
KPRM 5	4	2	3	3	2
KPRM 6	3	1	0	0	1
KPRM 7	1	0	2	0	0

The seven KPRMs contain 193 of the 200 selected genes: 13 retained hub genes and 180 network-recruited genes. The remaining seven genes are singleton components: ITGB6, CCND3, GYS2, PLA2G4A, PLCB4, AKT3, TNN. All seven singletons are retained hub genes.

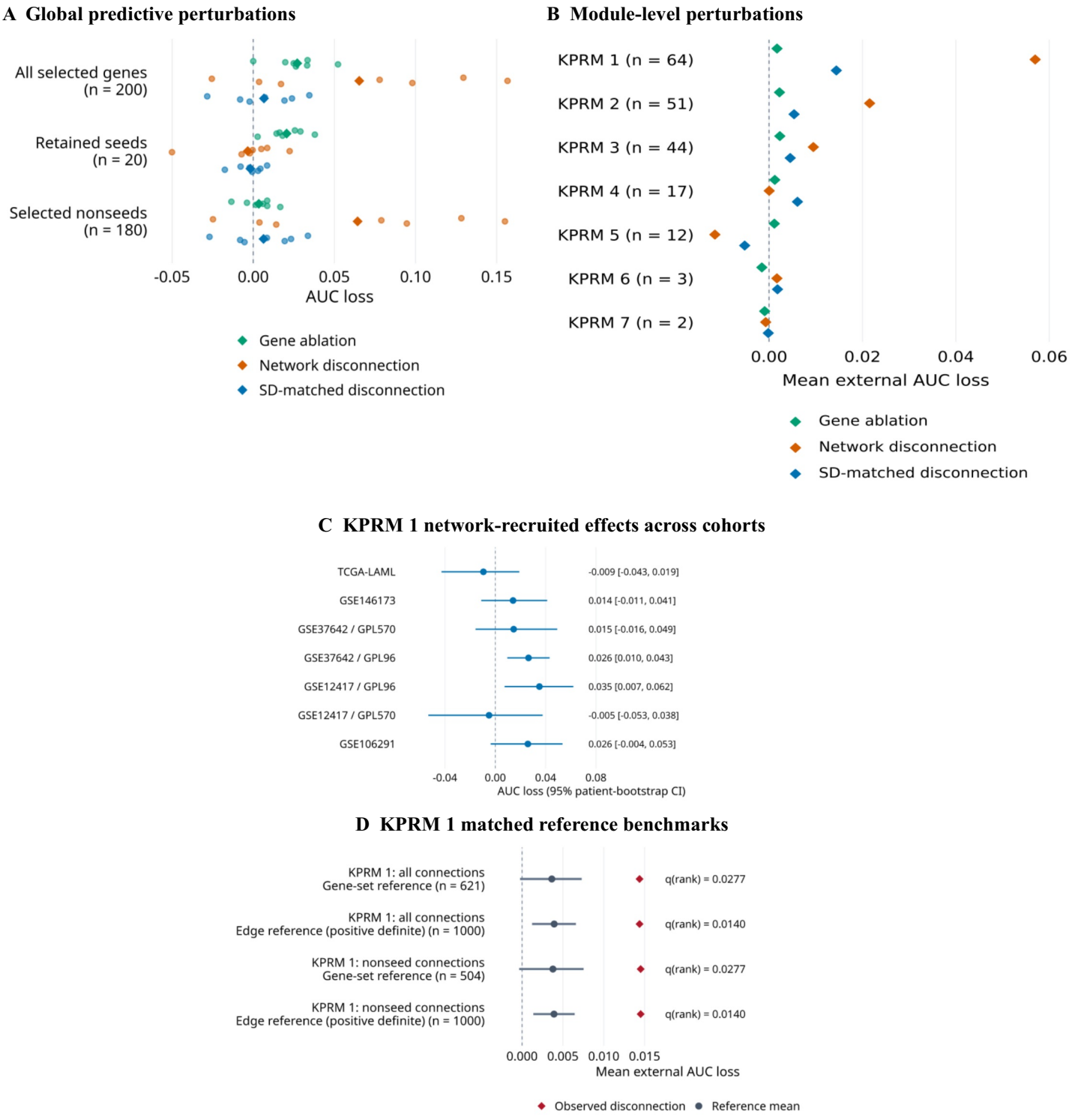
Panel A lists up to three network-recruited genes per module ranked by the absolute frozen netLDA coefficient; parentheses contain signed coefficients. Retained hub genes are taken from the final fitted model. Network-recruited status is distinct from marginal statistical significance.

Panel B reports numbers of genes assigned to each of the five candidate KEGG pathways. A gene can belong to multiple pathways, so row sums can exceed the module size. These counts describe membership and are not pathway-enrichment tests.

Module membership was verified against connected components of the saved inverse-covariance network using normalized off-diagonal weights with absolute value $>10^{-8}$. Coefficients and hub labels were checked against the raw fitted model. KPRM, KEGG predictive regulatory module.

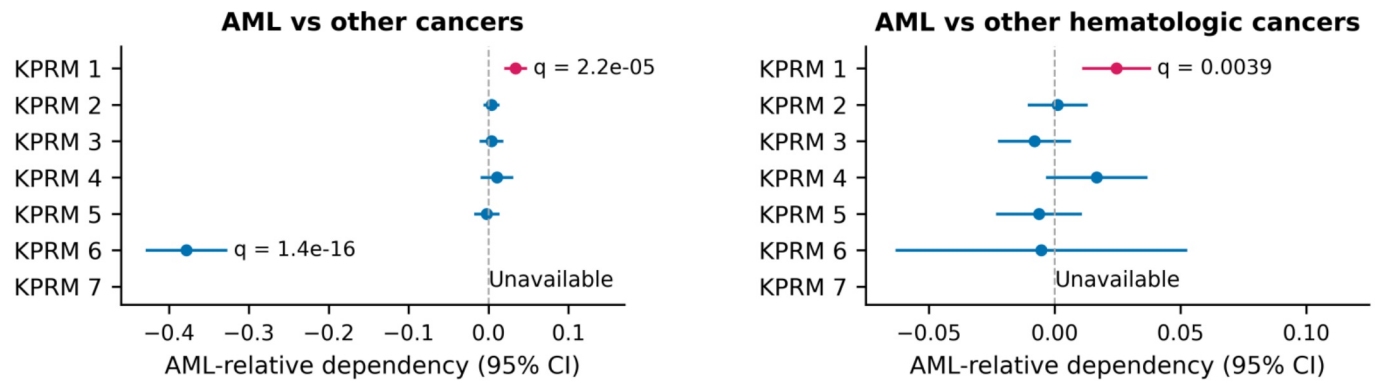

Seven distinct netLDA gene-network modules
Age-adjusted BEAT2 model • 193 genes in seven modules • 7 isolated genes
A KPRM 1 · EIF4B
64 genes | 2 seed hubs
Pathway composition
14%
30%
16%
14%
27%
PAK1
EIF4B
B KPRM 2 · RAC3
51 genes | 2 seed hubs
Pathway composition
9%
29%
16%
11%
35%
LPAR2
RAC3
C KPRM 3 · GNB1
44 genes | 3 seed hubs
Pathway composition
18%
24%
9%
16%
34%
GNB1
ROCK2
PIK3R1
D KPRM 4 · EDAR
17 genes | 3 seed hubs
Pathway composition
18%
12%
24%
24%
24%
EDAR
LAT
FLT3LG
E KPRM 5 · COL6A1
12 genes | 1 seed hub
Pathway composition
17%
25%
25%
8%
25%
COL6A1
F KPRM 6 · PTK2
3 genes | 1 seed hub
Pathway composition
11%
11%
78%
CCND1
PTK2
GNG11
G KPRM 7 · CACNB2
2 genes | 1 seed hub
Pathway composition
25%
75%
CACNB2
FGF16
H Isolated selected genes
Not assigned to a multi-gene KPRM
AKT3
CCND3
GYS2
ITGB6
PLA2G4A
PLCB4
TNN
Reading the network
Seed hub
Recruited non-seed
Representative anchor
Positive association
Negative association
KEGG pathways
PI3K–Akt
JAK–STAT
MAPK
NF-κB
Wnt

Predictive contributions of genes and KPRM connections

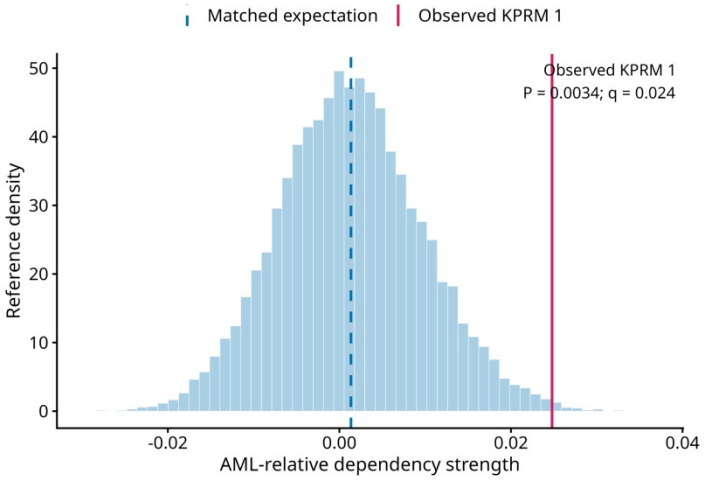


Experimental CRISPR evidence for AML KPRMs

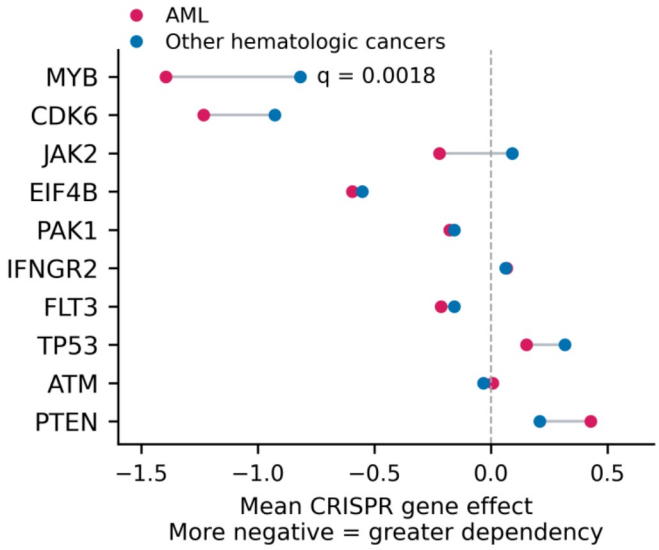
A Module dependency contrasts



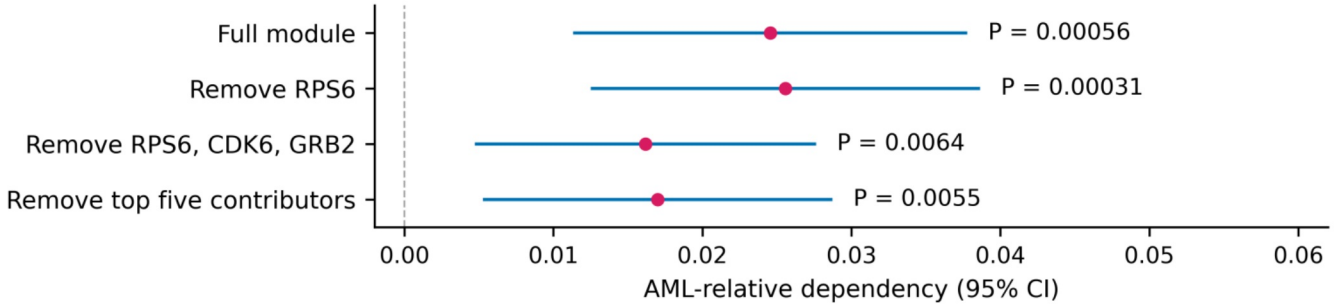
B Matched KPRM 1 reference

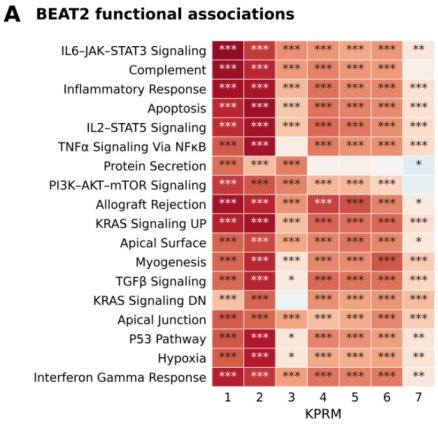


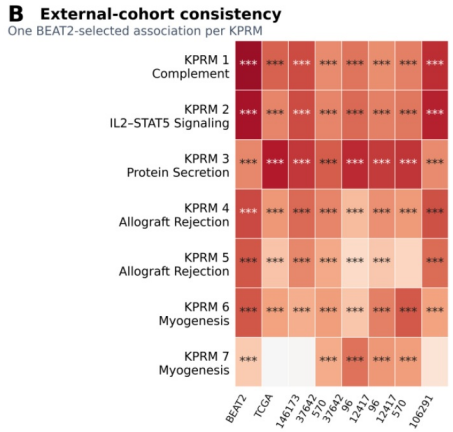
C Gene dependency effects

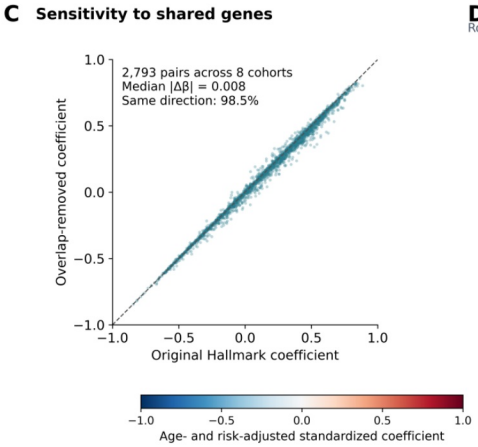


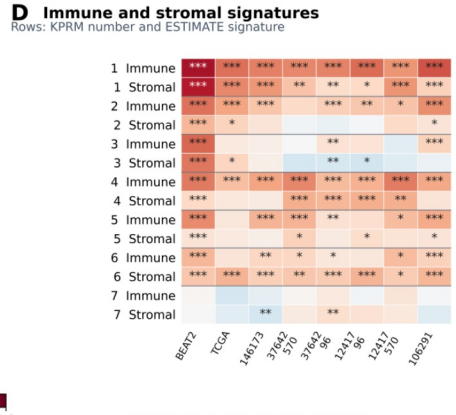
D KPRM 1 contributor sensitivity




A BEAT2 functional associations
IL6–JAK–STAT3 Signaling
Complement
Inflammatory Response
Apoptosis
IL2–STAT5 Signaling
TNFα Signaling Via NFκB
Protein Secretion
PI3K–AKT–mTOR Signaling
Allograft Rejection
KRAS Signaling UP
Apical Surface
Myogenesis
TGFβ Signaling
KRAS Signaling DN
Apical Junction
P53 Pathway
Hypoxia
Interferon Gamma Response
1 2 3 4 5 6 7
KPRM

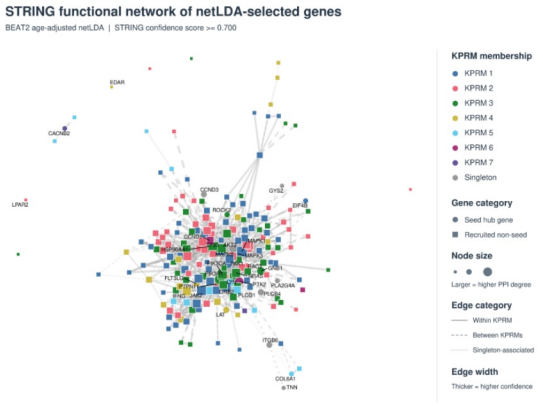

B External-cohort consistency
One BEAT2-selected association per KPRM
KPRM 1 Complement
KPRM 2 IL2–STAT5 Signaling
KPRM 3 Protein Secretion
KPRM 4 Allograft Rejection
KPRM 5 Allograft Rejection
KPRM 6 Myogenesis
KPRM 7 Myogenesis
BEAT2
TCGA
146173
37642 570
37642 96
12417 96
12417 570
106291


C Sensitivity to shared genes
2,793 pairs across 8 cohorts
Median |Δβ| = 0.008
Same direction: 98.5%
Overlap-removed coefficient
Original Hallmark coefficient
Age- and risk-adjusted standardized coefficient

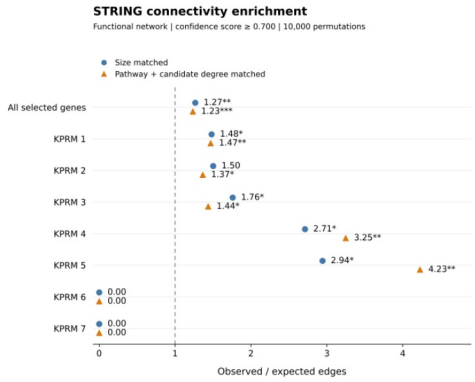

D Immune and stromal signatures
Rows: KPRM number and ESTIMATE signature
1 Immune
1 Stromal
2 Immune
2 Stromal
3 Immune
3 Stromal
4 Immune
4 Stromal
5 Immune
5 Stromal
6 Immune
6 Stromal
7 Immune
7 Stromal
BEAT2
TCGA
146173
37642 570
37642 96
12417 96
12417 570
106291
BH-adjusted P: * < 0.05 ** < 0.01 *** < 0.001

532

E

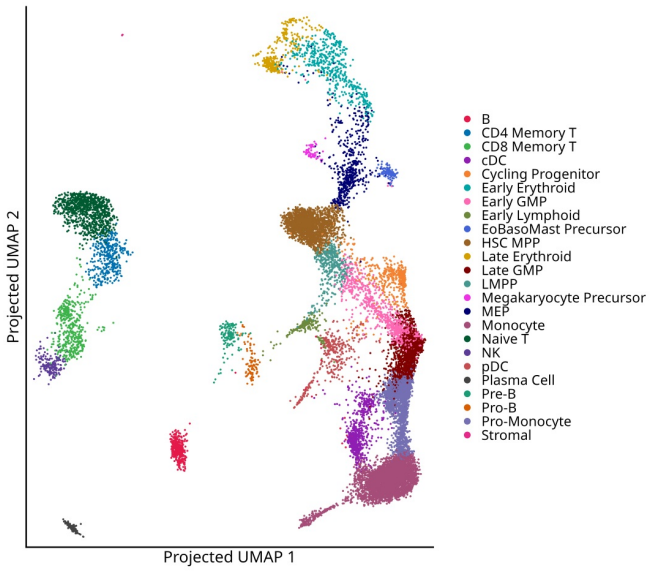

STRING functional network of netLDA-selected genes
BEAT2 age-adjusted netLDA | STRING confidence score >= 0.700
KPRM membership
KPRM 1
KPRM 2
KPRM 3
KPRM 4
KPRM 5
KPRM 6
KPRM 7
Singleton
Gene category
Seed hub gene
Recruited non-seed
Node size
Larger = higher PPI degree
Edge category
Within KPRM
Between KPRMs
Singleton-associated
Edge width
Thicker = higher confidence

F

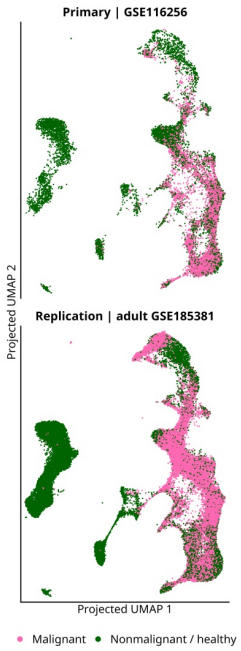

STRING connectivity enrichment
Functional network | confidence score ≥ 0.700 | 10,000 permutations
Size matched
Pathway + candidate degree matched
All selected genes
KPRM 1
KPRM 2
KPRM 3
KPRM 4
KPRM 5
KPRM 6
KPRM 7
1.27** 1.23***
1.48* 1.47**
1.50 1.37*
1.76* 1.44*
2.71* 3.25**
2.94* 4.23**
0.00 0.00
0.00 0.00
Observed / expected edges

533

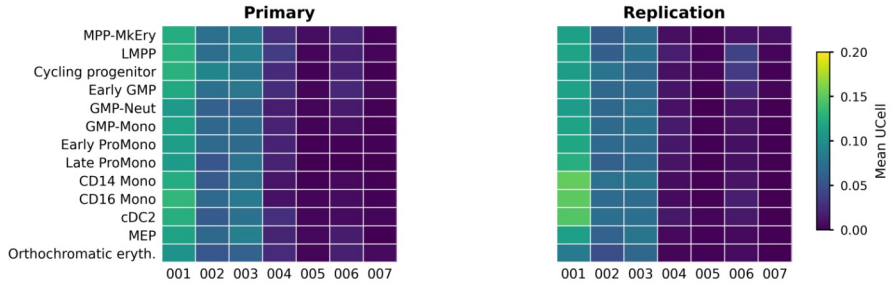
A Broad lineage annotations in the primary cohort



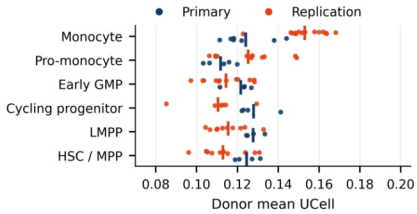
B Projected cellular landscapes by malignancy status



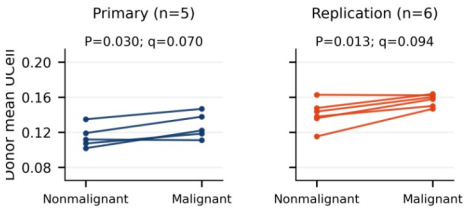
C KPRM localization in malignant cells



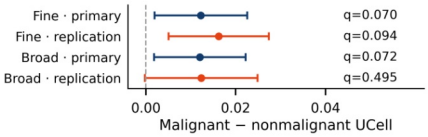
D KPRM_001 broad lineage enrichment



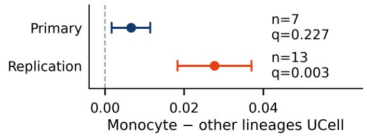
E CD14 monocyte paired contrasts



F Fine state and broad lineage sensitivity

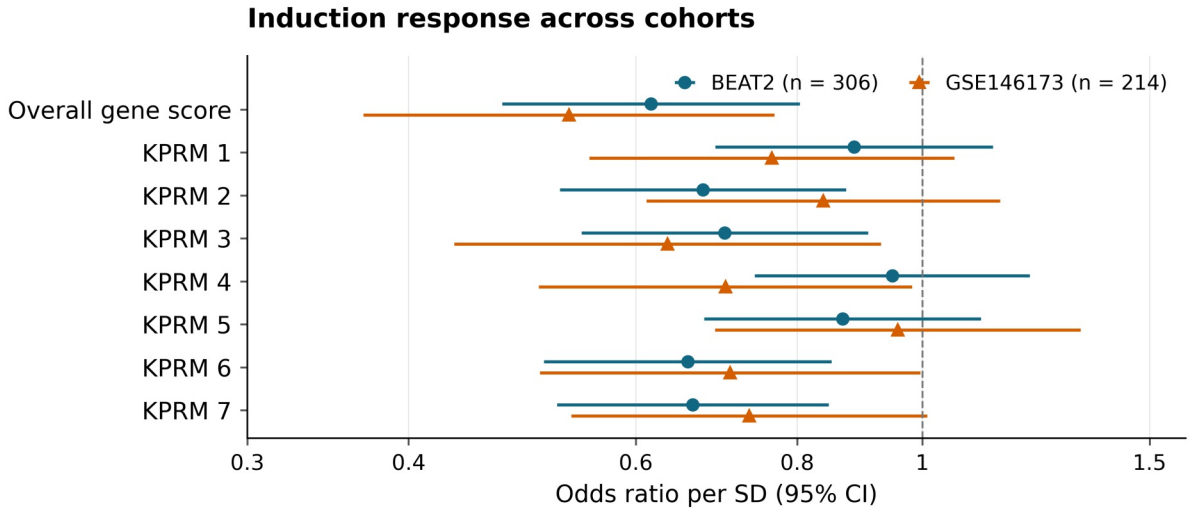


G Monocyte localization within donors

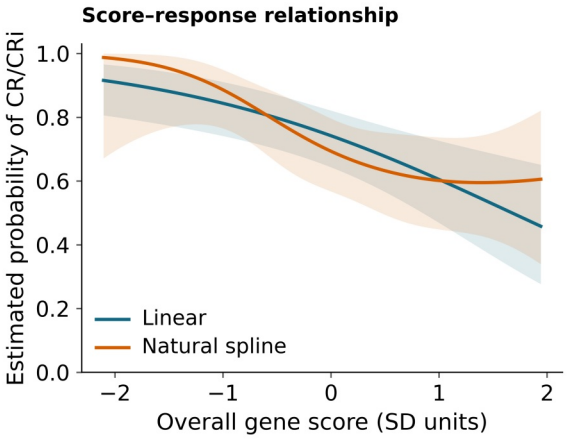


Bulk gene scores and treatment response in AML

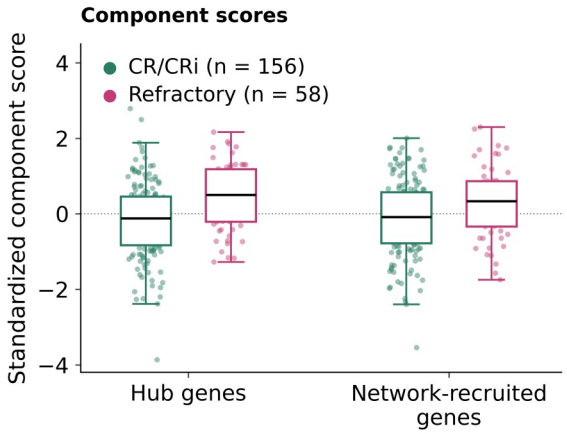
A



B



C



Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- AMLSupplementaryLeukemia.pdf