Supplementary Materials for

Cross-sectional correlation network analysis identifies plasma proteins and metabolites associated with the pathogenesis and treatment of acute promyelocytic leukemia

Niu Qiao[*], Feng Liu[*#], Yizhu Lyu[*], Yuliang Zhang, Xiaolin Ma, Xiaojing Lin, Junyu Wang, Yinyin Xie, Ruihong Zhang, Jing Qiao, Hongming Zhu, Li Chen, Tong Yin, Hai Fang, Qiang Tian[#], Zhu Chen[#], and Saijuan Chen[#]

Affiliations:

Shanghai Institute of Hematology, State Key Laboratory of Medical Genomics, National Research Center for Translational Medicine at Shanghai, Ruijin Hospital Affiliated to Shanghai Jiao Tong University School of Medicine, Shanghai, China

[#]Corresponding authors: Feng Liu (lf12034@rjh.com.cn), Qiang Tian (tq12221@rjh.com.cn), Zhu Chen (zchen@stn.sh.cn), Saijuan Chen (sjchen@stn.sh.cn).

This PDF file includes:

Supplemental Methods

Supplemental Figures S1 to S6

Other Supplementary Materials for this manuscript include the following:

Supplemental Tables S1 to S10

Supplemental Methods

Clinical laboratory tests

A total of 88 clinical laboratory tests were collected as follows. (i) the percentage of bone marrow blast in primary APL patients; (ii) 20 routine blood tests, including white blood cell count, neutrophil percentage, lymphocyte percentage, monocyte percentage, eosinophil percentage, basophil percentage, neutrophil count, lymphocyte count, monocyte count, eosinophil count, basophil count, red blood count, haemoglobin, haematocrit, mean cell volume, mean cell haemoglobin, mean cell haemoglobin concentration, red cell distribution width, platelet count, and mean platelet volume; (iii) 6 coagulation-related tests, including activated partial thromboplastin time, prothrombin time, thrombin time, fibrinogen, fibrin degradation product, and D-dimer; (iv) 15 biochemical analyses including glucose, prealbumin, alanine aminotransferase, aspartate aminotransferase, alkaline phosphatase, gamma glutamyltransferase, total-value bilirubin, direct bilirubin, total protein, albumin, albumin-globulin, total bile acids, blood urea nitrogen, creatinine, and uric acid; (v) 12 blood lipid tests including triglyceride, total cholesterol, high-density lipoprotein, low-density lipoprotein, apolipoprotein A, apolipoprotein B, apolipoprotein E, lipoprotein A, lactate dehydrogenase, complement total, non-esterified fatty acids, and ferritin; (vi) 6 cytokines, such as IL-1β, IL-2R, IL-6, IL-8, IL-10, and tumor necrosis factor (TNF); (vii) 18 additional biochemical analyses as well as 10 endocrine tests.

Plasma proteome analysis

Plasma sample preparation

To effectively mine plasma proteins that are related to APL progression, we introduced Olink technology. According to the experimental design, all selected plasma samples were appropriately randomized. All samples were pooled as a mixed external control to assess potential variation between runs and plates. In addition, 4 Olink panels were pre-diluted so that the target proteins are in an optimal concentration range for detection. Samples are diluted according to the ratio in parentheses: Cardiometabolic (1:2025), Metabolism (1:10), CVD III (1:100), and Development (1:100) Panel.

Olink proximity extension assays

The main experimental steps include incubation, extension, and amplification, as well as detection. A total of 12 Olink panels were tested according to the manufacturer's instructions (Olink, Uppsala, Sweden). Briefly, 1 μL of plasma sample was incubated with a set of probes, and when a pair of matched antibody probes bind to their target protein, the corresponding DNA labels come into proximity and proximity-dependent DNA polymerization forms the PCR target sequence. The newly formed DNA molecules are then amplified and quantified by real-time PCR using the Fluidigm Biomark™ HD system.

Olink data quality control

We imported the raw image files into Fluidigm's real-time PCR Analysis software and used the Olink NPX Manager to export a Ct values matrix for downstream preprocessing. The Olink data is represented by an arbitrary unit called Normalized Protein eXpression (NPX), which is on a $\log_2$ scale. An estimate of 1 NPX represents a doubling of the concentration on the unknown original scale. NPX values were obtained by normalizing Ct values according to extension control, interplate control, and correction factor. Proteins with >60% missing frequency across all samples were filtered, leaving 991 unique proteins for downstream statistical analysis.

Differential proteins and functional enrichment analysis

P-values were calculated using a two-sided Wilcoxon Rank Sum test. The adjusted *P*-value was corrected by the Benjamini–Hochberg procedure. The difference between the average NPX values of the two sets of samples was used as the $\log_2$ fold change of the protein concentration. Proteins with an adjusted *P*-value < 0.05 and an $|\log_2 FC| > \log_2(1.5)$ were considered significantly differential proteins. As our proteomic analysis is based on targeted panels, the statistical significance of enriched pathways should be treated cautiously; however, relative enrichment is warranted. Functional Gene Ontology (GO) terms[1, 2] and KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways [3] were performed based on the Kolmogorov–Smirnov test using the clusterProfiler (version 3.16.1) R package [4]. *P*-values were adjusted by the Benjamini–Hochberg procedure and functional GO/KEGG terms were considered

enriched based on FDR < 0.05.

Plasma metabolome analysis

Chemical materials for untargeted metabolomics

MS-grade water (W6-4), methanol (A456-4), and acetonitrile (A955-4) were purchased from Fischer Scientific (Morris Plains, NJ, USA). MS-grade acetic acid (45754-100ML-F) was purchased from Sigma Aldrich (St. Louis, MO, USA). Analytical grade chemical standards were purchased [carnitine C2:0-d3 (DLM-754-PK), carnitine C16:0-d3 (DLM-1263-0.01), stearic acid C18:0-d3 (DLM-1154-PK), cholic acid CA-d4 (DLM-2611-PK), phenylalanine-d5 (DLM-1258-1)] from Cambridge Isotope Laboratories (Tewksbury, MA USA) and prepared in methanol solution to measure the limits of detection (LODs) and linearity. The 5 standards were dissolved into the precipitation solution with a volume ratio of acetonitrile:methanol = 1:1. The concentration of the 5 standards are 0.16 μg/mL, 0.15 μg/mL, 2.5 μg/mL, 1.85 μg/mL, and 3.61 μg/mL for carnitine C2:0-d3, carnitine C16:0-d3, stearic acid C18:0-d3, cholic acid CA-d4, and phenylalanine-d5, respectively.

Plasma sample preparation

An aliquot of 50 μL plasma was precipitated by adding 150 μL precipitating reagent and vortexed for 1 minute. The mixture was then placed in a −20 °C refrigerator for 1 hour to improve protein precipitation efficiency. Precipitated protein was subsequently removed by centrifugation (15,000 rpm, 15 min) at 4 °C. Then, 120 μL of the supernatant was transferred into the screw tube with an insert vial, ready for further Liquid chromatography(LC)–mass spectrometry (MS) analysis, or stored at −80 °C for later use. Mixed quality controls (QC, equal mixture of all the samples) were also processed to analyze the repeatability of samples under the same processing method. During sampling, a QC sample is inserted into every 10 samples to monitor the repeatability of the analysis process.

LC-MS chromatographic conditions

Reversed-phase high-performance liquid chromatographic (HPLC)-mass spectrometry (MS) separation was performed using a reversed-phase column (ACQUITY UPLC® HSS T3 1.8μm, 2.1×100mm) purchased from Waters (Milford, MA, USA). Mobile phases for HPLC consisted of water with 0.1%

formic acid (phase A) and acetonitrile (phase B). Metabolites were eluted from the column at a flow rate of 0.3 mL/min. A linear 1%–99% phase B gradient was applied over 1.5–16.5 min. The oven and cooler temperatures were set to 8 °C and 40 °C, respectively. The sample injection volume was 2 μL.

LC-MS data acquisition

Metabolic extracts were analyzed by RPLC–MS in both positive and negative ionization modes. The ExionLC™ AD coupled with TripleTOF® 6600+ system was operated in information-dependent acquisition (IDA) scan mode for data acquisition (acquisition range of m/z 40–1,000 for positive mode, and m/z 50–1,000 for negative mode). The MS/MS spectra of the QC sample were acquired under different collision energy (35 ± 15 CE) of the top 15 candidate ions. The source conditions were as follows: Curtain Gas flow (CUR) = 40, Temperature (TEM) = 550 ℃, Ion Source Gas 1 (GS1) = 60, Ion Source Gas 2 (GS2) = 60, IonSpray Voltage Floating (ISVF) = 5500 for positive mode, and 4500 for negative mode, Declustering Potential (DP) = 60V, Accumulation Time = 0.035s. The resulting mass spectra were exported into SCIEX OS software (Redwood City, CA, USA) for further processing.

Metabolic feature detection and retention time correction

The raw MS metabolic data (wiff format) were converted to mzXML and mgf formats by using ProteoWizard software (version 3.0.210) [5]. We set the Peak Picking msLevel parameter equal to TRUE to reduce the file size after conversion. For all QCs and samples, raw data were then processed for peak detection, retention time correction, and chromatogram alignment by xcms R package [6, 7]. Parameter settings for xcms processing of our reverse-phase LC-MS data were as follows: centWave for feature detection (Δ m/z = 15 ppm, snthr = 6, minimum peak width = 5 s, and maximum peak width = 40 s); obiwarp settings for retention-time correction (profStep = 1); and parameters for chromatogram alignment, including mzwid = 0.015, minfrac = 0.5, and bw = 5. The relative quantification of metabolite features was based on EIC (extracted ion chromatogram) areas.

Metabolic feature identification

Metabolite identification was performed using three approaches. For in-house database searching, all 45 QCs were firstly quantitatively processed against our in-house metabolite library, which contains

650 chemical standards compound lists. Parameter settings for quantitative processing were as follows: integration algorithm = MQ4, XIC width = 0.02 Da, RT half window = 30 seconds, minimum peak width = 3 points, and S/N integration threshold = 3. A smart confirmation search was used for the qualitative library search. We selected the top 5 peaks and manually inspected the retention time for all in-house compounds based on the m/z, MS1, and MS2 results. After that, all metabolomic features of all samples were extracted with a unique mass/charge ratio and retention time identified by QCs, and then aligned, quantified, and manually corrected with the SCIEX OS software (Redwood City, CA, USA). For public database searching, using the metabolic features table generated by xcms software, the metabolic features, and MS/MS spectra were matched according to their accurate masses (±25 ppm), and RT values (±30 s). The generated MS1/MS2 pairs were automatically searched in the public databases by using metID R package (version 0.4.1)[8]: HMDB (http://www.hmdb.ca/), MoNA (http://mona.fiehnlab.ucdavis.edu/), and MassBank (http://www.massbank.jp/). All above libraries gives an MSI[9] level 2 identification. The MS/MS spectra similarity score cutoff was set as 0.5. Thirdly, the metabolic peaks with MS/MS spectra that were not matched in public databases were analyzed by the structural similarities algorithm MetDNA [10], which gives an MSI[9] level 4 identification.

Metabolic data processing

We combined all the above positive and negative results, and filtered the metabolites according to the following criteria: (i) for a given metabolic feature identified in both positive and negative modes by metID, we preferentially retain substances with a high mz.match.score; (ii) for a given metabolic feature identified in both positive and negative modes by metDNA, we preferentially retain substances with a high score; (iii) when a given metabolic feature was matched differently between different methods, we choose the matching based on the identification level: AB Sciex > metID > MetDNA. We then removed all metabolites that were quantified in less than 30% of the samples and replaced the missing value with 1/5 of the minimum positive value of each feature. To minimize the impact of batch effects in our bulk metabolomics data, we applied systematic error removal using random forest (SERRF) algorithm[11] for batch corrections.

Differential metabolites and pathway enrichment analysis

The processed matrix was extracted and imported into limma (version 3.48.0) R package [12] for calculating significantly altered metabolic features/compounds. We define significant differential metabolites as absolute fold change > 1.5 and adjusted *P*-value < 0.05. We utilized MetaboAnalystR (version 3.0.3) [13] to perform the metabolite set enrichment analysis as well as metabolic pathway analysis on metabolites [2], relatively enriched metabolic pathways (*P*-value < 0.1) were extracted. The Enrichment ratio is computed by the number of hits within a particular metabolic pathway divided by the expected number of hits. The larger the enrichment ratio, the bigger degree of the enrichment.

Supplemental References

1. Ashburner M*, et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25-29 (2000).
2. Chong J*, et al.* MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. *Nucleic Acids Res* **46**, W486-W494 (2018).
3. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* **28**, 27-30 (2000).
4. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284-287 (2012).
5. Chambers MC*, et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nat Biotechnol* **30**, 918-920 (2012).
6. Gowda H*, et al.* Interactive XCMS Online: simplifying advanced metabolomic data processing and subsequent statistical analyses. *Anal Chem* **86**, 6931-6939 (2014).
7. Huan T*, et al.* Systems biology guided by XCMS Online metabolomics. *Nature Methods* **14**, 461-462 (2017).
8. Shen X*, et al.* metID: a R package for automatable compound annotation for LC-MS-based data. *Bioinformatics*, (2021).
9. Sumner LW*, et al.* Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* **3**, 211-221 (2007).
10. Shen X*, et al.* Metabolic reaction network-based recursive metabolite annotation for untargeted metabolomics. *Nat Commun* **10**, 1516 (2019).
11. Fan S*, et al.* Systematic Error Removal Using Random Forest for Normalizing Large-Scale Untargeted Lipidomics Data. *Anal Chem* **91**, 3590-3596 (2019).
12. Ritchie ME*, et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res* **43**, e47 (2015).
13. Chong J, Yamamoto M, Xia J. MetaboAnalystR 2.0: From Raw Spectra to Biological Insights. *Metabolites* **9**, (2019).

Figure S1

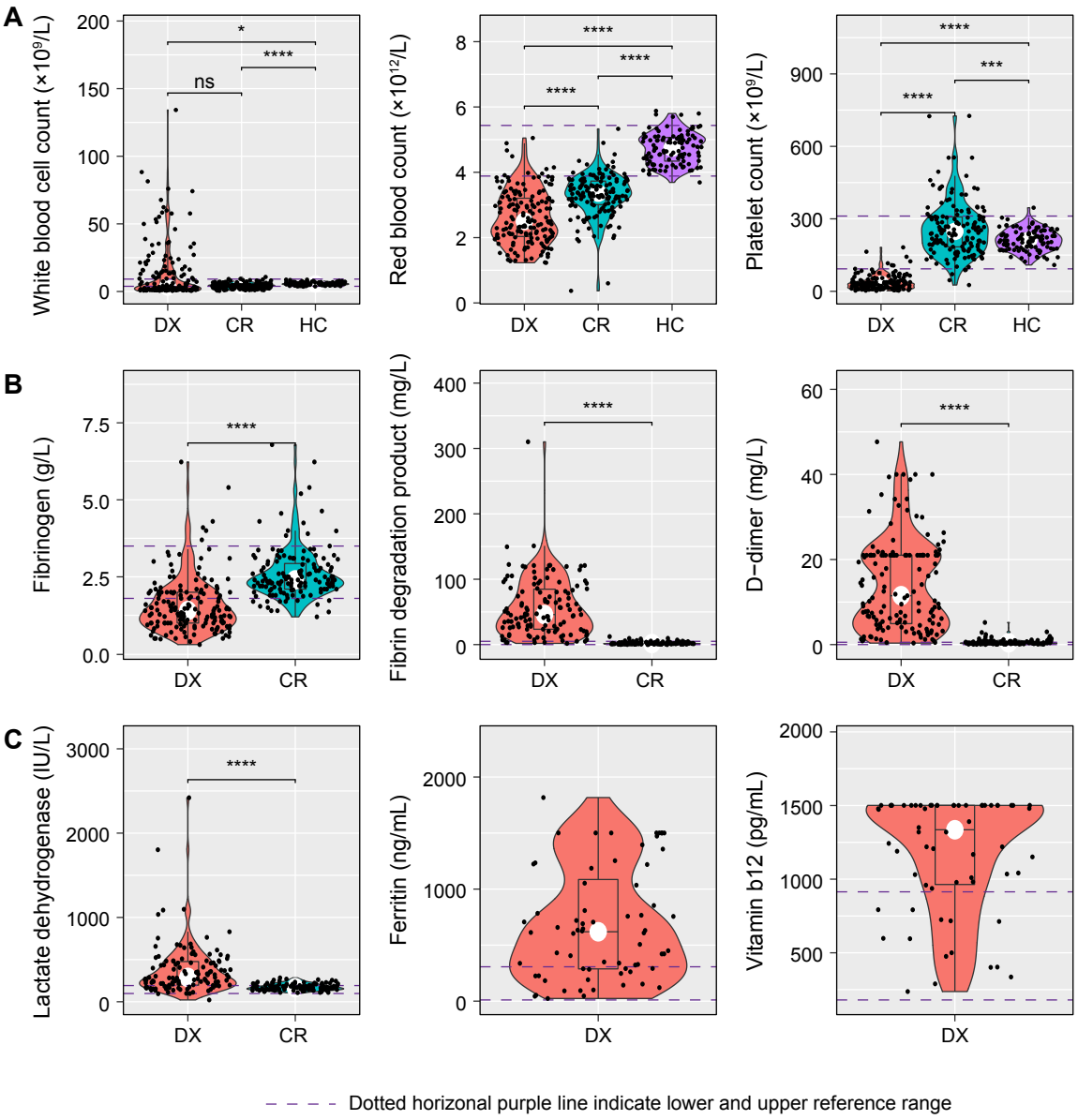


Figure S1. Clinical and laboratory characteristics of primary APL.

(**A**) Violin plot of a subset of routine blood tests, including white blood cell count, red blood cell count, and platelet counts. (**B**) Violin plot of a subset of coagulation tests, including fibrinogen, fibrin degradation product, and d-dimer. (**C**) Violin plot of lactate dehydrogenase, ferritin, and vitamin b12. DX, primary APL; CR, complete remission; HC, healthy controls. *P*-value is calculated using the Wilcoxon test and corrected for multiple-hypothesis testing with Benjamini-Hochberg method. * $P_{adj} < 0.05$, ** $P_{adj} < 0.01$, *** $P_{adj} < 0.001$, **** $P_{adj} < 0.0001$; ns, not significant.

Figure S2

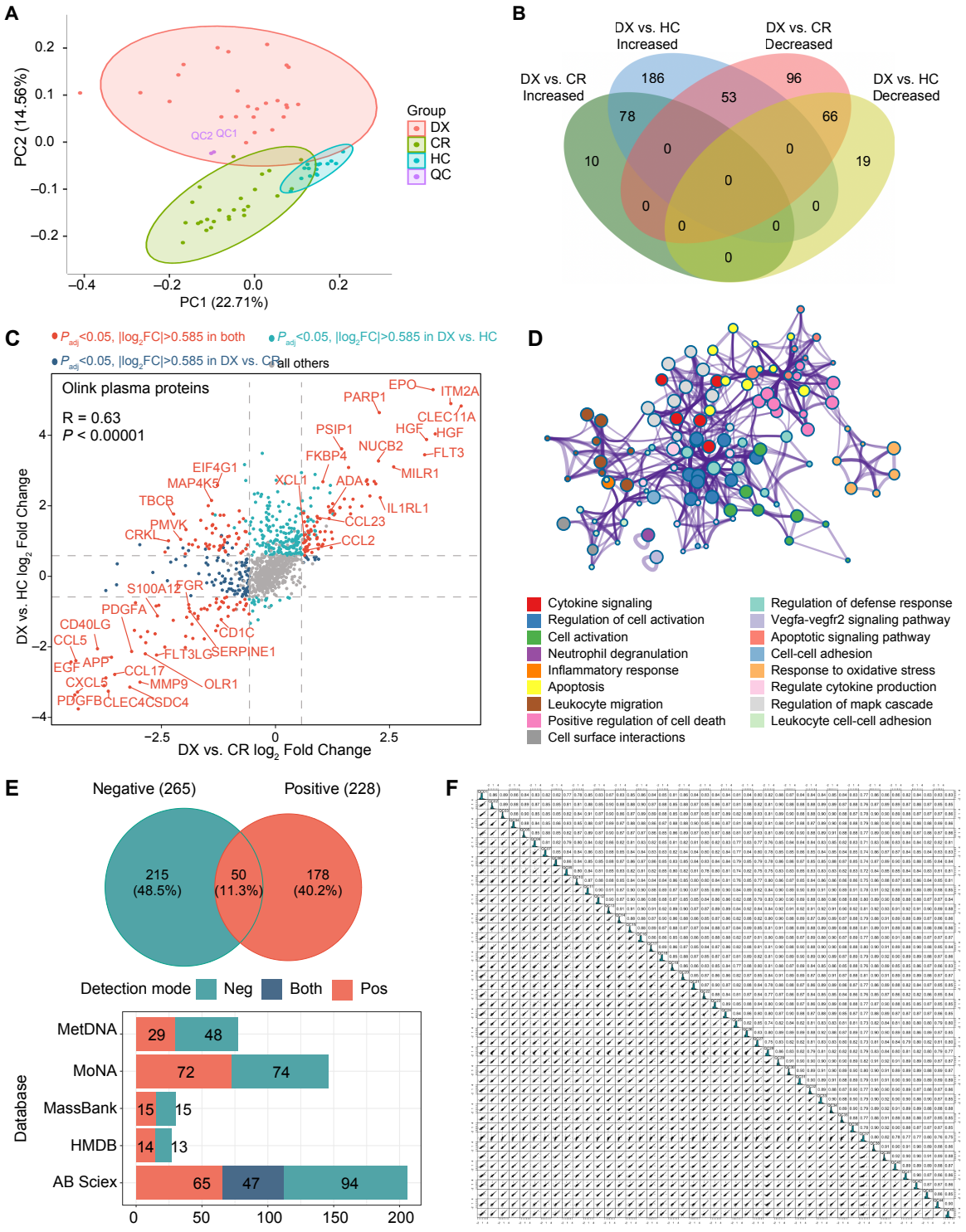


Figure S2. Quality assessments for plasma proteome and metabolome data.

(**A**) The PCA plot of different sample groups from the proteome layer. (**B**) Venn diagram of the intersection number of differential proteins between DX vs. CR and DX vs. HC comparisons. (**C**) Log_2

fold-change of plasm proteins in DX vs. CR (x-axis) and DX vs. HC (y-axis) comparisons. Significant changes at DX vs. CR (in blue), DX vs. HC (in cyan), and both levels (in red) are highlighted. (**D**) Significant GO enriched terms of PDPs in high-risk vs. standard-risk at DX stage. (**E**) The number of metabolites identified in untargeted HPLC-MS/MS analysis (upper panel) and the detailed number of metabolites identified by metDNA, MoNA, MassBank, HMDB public databases, and local AB Sciex library (lower panel). (**F**) The quantification repeatability of longitudinal 45 benchmark QC samples shows the robust and accurate metabolome platform. DX, primary APL; CR, complete remission; HC, healthy controls; PDP, proteins with different plasma levels; QC, quality controls. Neg, negative mode; Pos, positive mode.

Figure S3

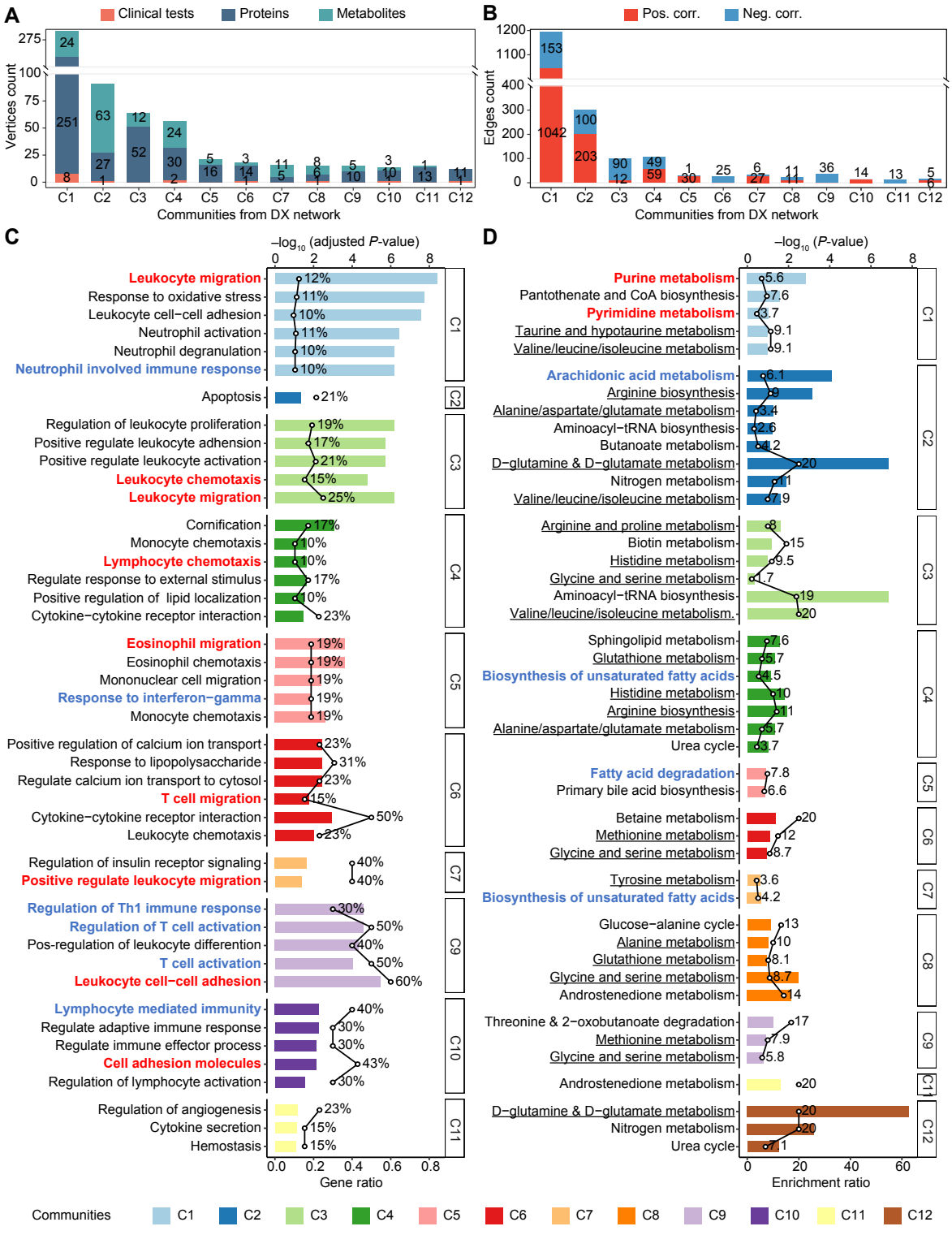


Figure S3. Functional enrichment results of differential proteins/metabolites of the top 12 APL perturbed communities.

(**A**) Distribution of vertex counts in the top 12 communities from the primary APL correlation networks.

(**B**) Distribution of edge counts in the top 12 communities from the primary APL correlation networks. (**C**) GO and KEGG functional enrichment results of plasma proteins in the top 12 communities (node degree ≥ 10 or vertex count ≥ 12) from the primary APL correlation network. Texts colored in bold-red represent leukocyte migration-related GO terms, and texts colored in bold-blue represent immunity-response-related GO terms. (**D**) MSEA results of plasma metabolites in the top 12 communities from the primary APL correlation network. MSEA, metabolite set enrichment analysis. Underlined texts represent amino acid metabolisms-related terms. Text colored in bold-blue indicates nucleotide metabolism-related terms. Text colored in bold-blue indicates lipid metabolism-related terms.

Figure S4

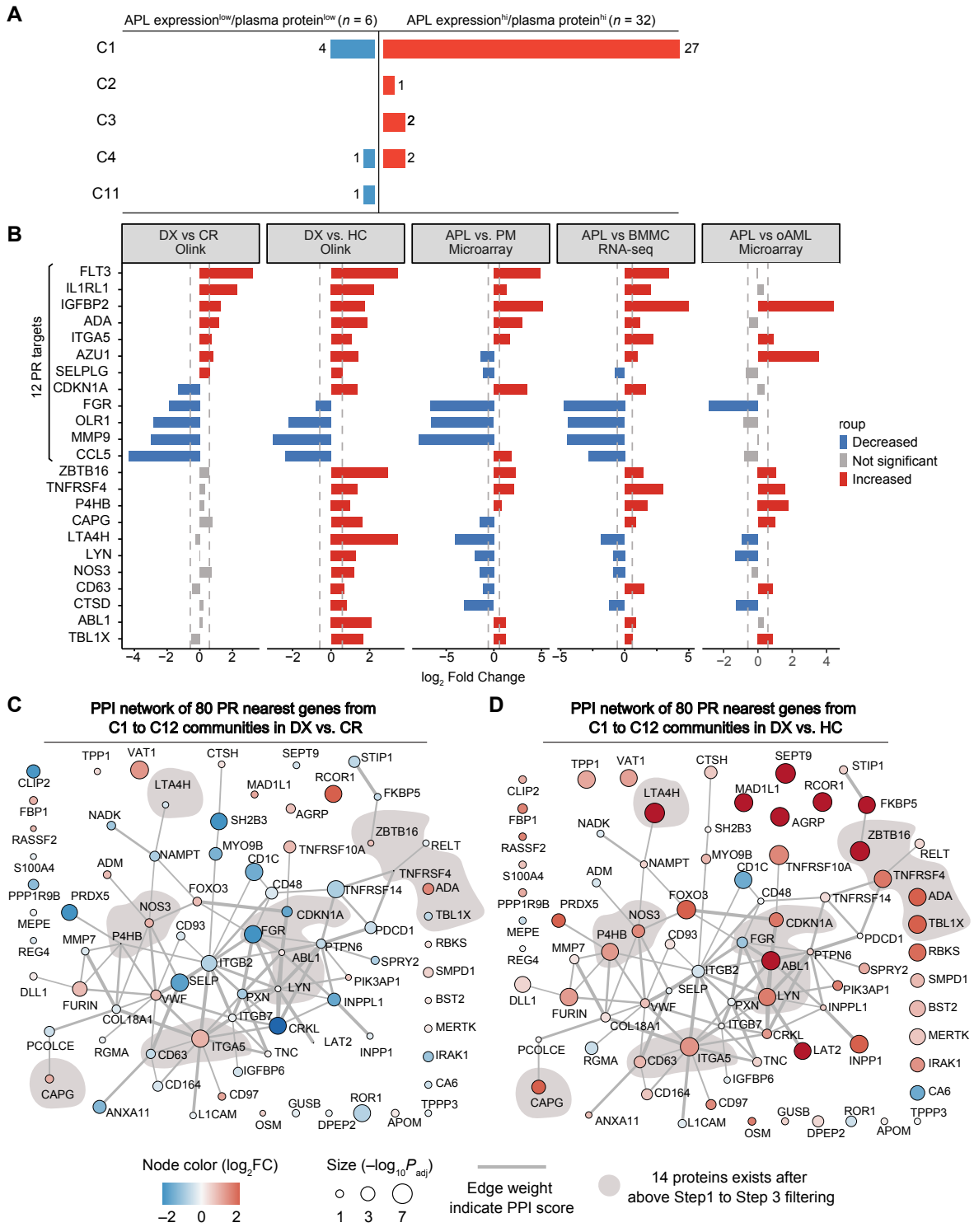


Figure S4. Detailed change trend of APL derived factors and PML/RARA targets.

(**A**) Detailed change trend of the 32 APL expression^{hi}/plasma protein^{hi} factors and 6 expression^{low}/plasma protein^{low} factors in each community. (**B**) Detailed change trend of 23 PML/RARA

(PR) ChIP-seq nearest genes in DX vs. CR (Olink data), DX vs. HC (Olink data), APL vs. PM (Microarray data), APL vs. BMMC (RNA-seq data), APL vs. oAML (Microarray data). (**C**) Protein-protein interaction (PPI) network of 80 PML/RARA targets from C1 to C12 community in DX vs. CR comparison. The node color intensities (red: increased, blue: decreased) depict the $\log_2$ fold-change between DX and CR. The node size indicates the $\log_{10}$ adjusted *P*-value between DX vs. CR. The edges widths were weighted by the interaction strengths from the STRING database. (**D**) PPI network of the 80 PML/RARA targets (were within C1 to C12 communities) in DX vs. HC comparison. The node color intensities (red: increased, blue: decreased) depict the $\log_2$ fold-change between DX and HC. The node size indicates the $\log_{10}$ adjusted *P*-value between DX vs. HC. The edges widths were weighted by the interaction strengths from the STRING database. DX, primary APL; CR, complete remission; HC, healthy controls; PM, normal promyelocytes; BMMC, bone marrow mononuclear cells; oAML, other types of AMLs.

Figure S5

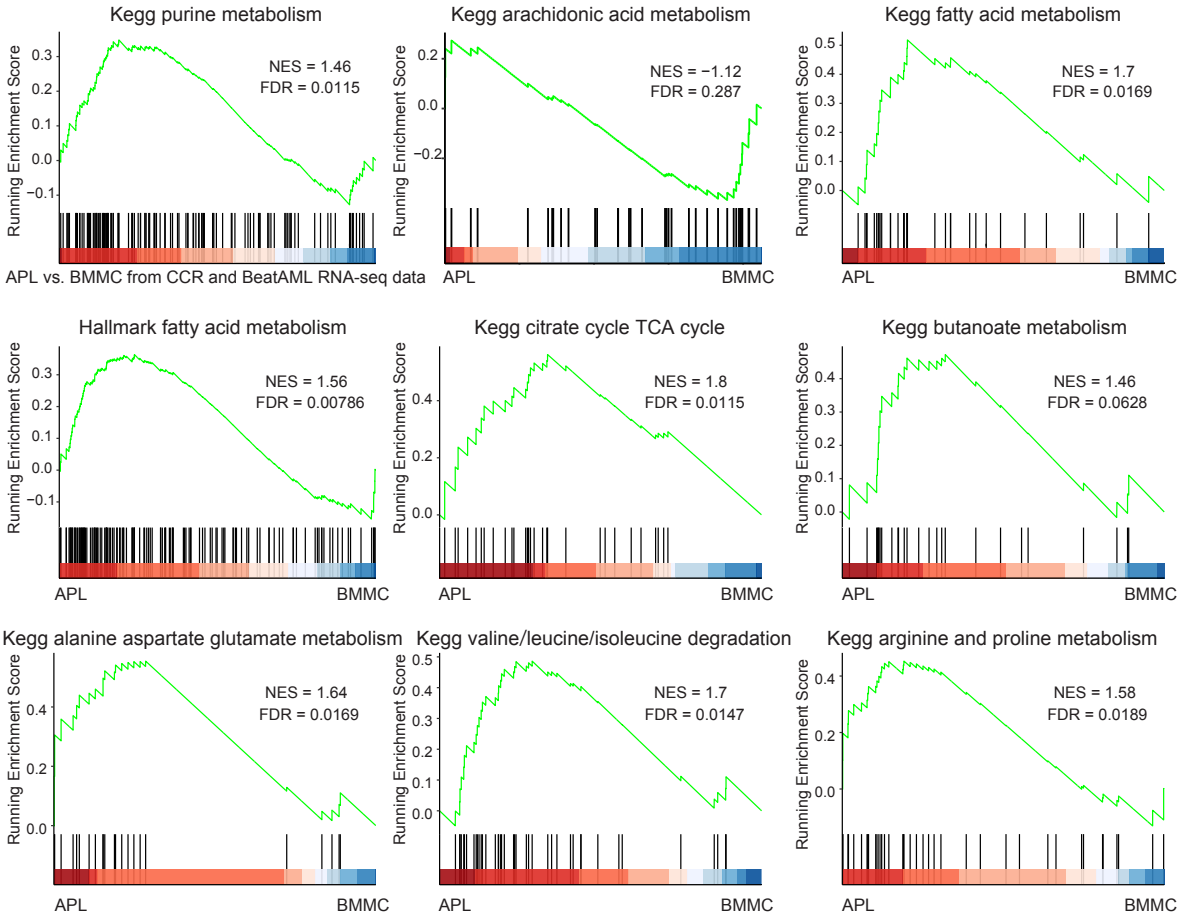
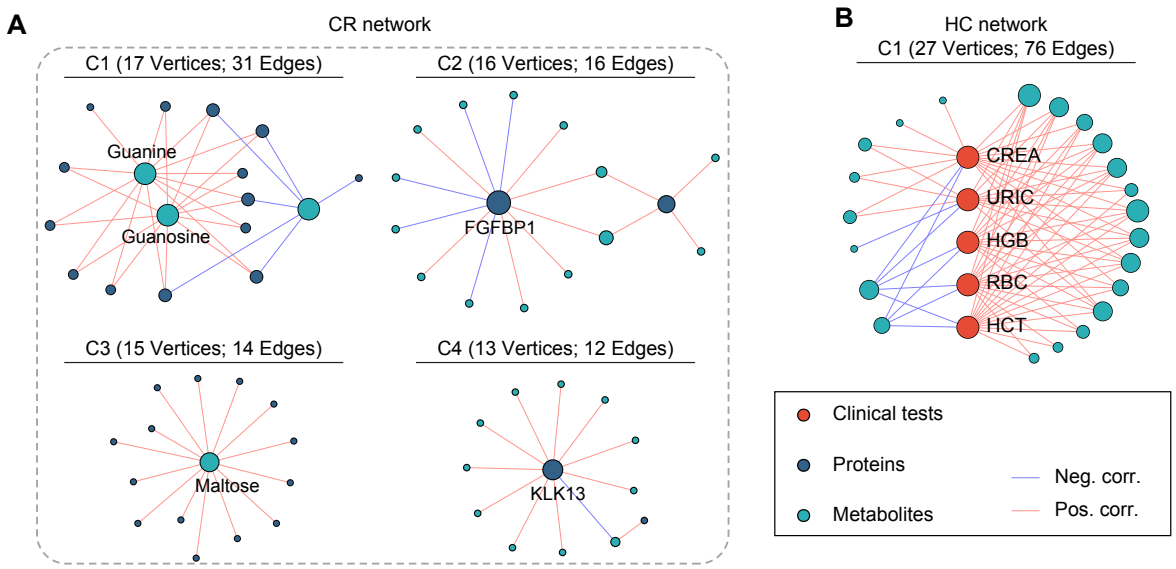


Figure S5. Representative GSEA plot of metabolomic terms in APL at diagnosis vs. normal healthy bone marrow mononuclear cells (BMMC) by RNA-seq data.

Figure S6



241

242 **Figure S6.** Representative top communities in CR and HC.

243 Representative top communities (hub node degree ≥ 10) in CR (**A**) and HC (**B**). CR, complete remission;

244 HC, healthy controls.

Table S1. Clinical characteristics of the entire APL and healthy control cohorts.

Table S2. Differentially proteins/genes in Olink DX vs. CR, DX vs. HC, CR vs. HC, APL vs. PM, APL vs. BMMC, and HR vs. SR comparisons.

Table S3. Enriched GO and KEGG terms for DX vs. CR proteins with different plasma levels (PDPs) clusters.

Table S4. Differentially metabolites in DX vs. CR, DX vs. HC, CR vs. HC, and HR vs. SR comparisons.

Table S5. Enriched KEGG and SMPDB metabolic pathways of differential metabolites for DX vs. CR, DX vs. HC, and CR vs. HC comparisons.

Table S6. Complete cross-sectional correlation network in primary APL (DX), complete remission (CR) APL, and healthy controls (HC) networks.

Table S7. All enriched GO and KEGG terms of plasma proteins in the top 12 communities from the primary APL correlation network.

Table S8. All enriched KEGG and SMPDB results of plasma metabolites in the top 12 communities from the primary APL correlation network.

Table S9. Community types and degrees of nodes in the primary APL (DX), complete remission (CR) APL, and healthy controls (HC) networks.

Table S10. Enriched GO terms for CR vs. HC proteins with different plasma levels (PDPs).