Hans Journal of Biomedicine 生物医学, 2026, 16(3), 576-597
Published Online May 2026 in Hans. https://www.hanspub.org/journal/hjbm
https://doi.org/10.12677/hjbm.2026.163060

Hans汉斯

调控网络重构揭示急性髓系白血病细胞群体的调控特征

叶一帆*，卢　婷*，赵　哲，莫少康，高瀛岱#，颜光玕#

中国医学科学院血液病医院(中国医学科学院血液学研究所)，血液与健康全国重点实验室，国家血液系统疾病临床医学研究中心，细胞生态海河实验室，天津医学健康研究院，天津

收稿日期：2026年4月13日；录用日期：2026年5月21日；发布日期：2026年5月29日

摘　要

急性髓系白血病(acute myeloid leukemia, AML)由具有干细胞样和原始细胞样特征的不同细胞群体构成，但这些细胞群体之间差异的调控基础仍未得到充分阐明。在本研究中，我们整合了公开发表的群体水平(bulk)染色质可及性数据和基因表达数据，涵盖前白血病造血干细胞(preleukemic hematopoietic stem cells, pHSC)、白血病干细胞(leukemia stem cells, LSC)以及AML原始细胞(AML blast)群体，从而系统比较了染色质可及性、预测的转录因子(transcription factor, TF)活性以及转录因子–靶基因(TF-target gene, TF-TG)的调控连接模式。研究结果表明，不同细胞群体在全局染色质可及性及转录因子组成层面总体上高度共享，而差异性可及性主要富集于远端调控元件。与此同时，转录因子活性呈现出具有结构性的、细胞群体偏倚的变化模式，而非整体一致的变化。更为重要的是，调控网络分析揭示，在转录因子身份相对保守的情况下，转录因子–靶基因调控连接模式发生了广泛改变，并在调控相互作用层面表现出显著分化和重构。综合上述多层次分析结果，本研究支持如下模型：基于这些细胞群体所推断的AML相关细胞状态之间的调控差异，主要来源于转录因子–靶基因调控连接的重构，而非转录因子组成的更替。总体而言，本研究为理解AML细胞群体间的调控组织方式及其变异提供了一个系统性的分析框架。

关键词

急性髓系白血病，染色质可及性，基因调控网络，转录因子活性

Network Regulatory Rewiring Characterizes AML Cell Populations

Yifan Ye*, Ting Lu*, Zhe Zhao, Shaokang Mo, Yingdai Gao#, Guangyu Yan#

*共同第一作者。
#通讯作者。

文章引用: 叶一帆, 卢婷, 赵哲, 莫少康, 高瀛岱, 颜光玕. 调控网络重构揭示急性髓系白血病细胞群体的调控特征[J]. 生物医学, 2026, 16(3): 576-597. DOI: 10.12677/hjbm.2026.163060

Chinese Academy of Medical Sciences & Peking Union Medical College (Institute of Hematology & Blood Diseases Hospital), State Key Laboratory of Experimental Hematology, National Clinical Research Center for Blood Diseases, Haihe Laboratory of Cell Ecosystem, Tianjin Institutes of Health Science, Tianjin

Received: April 13, 2026; accepted: May 21, 2026; published: May 29, 2026

Abstract

Acute myeloid leukemia (AML) comprises coexisting stem-like and blast-like cell populations, yet the regulatory basis underlying their differences remains incompletely understood. In this study, we integrated publicly available bulk chromatin accessibility and gene expression datasets across preleukemic hematopoietic stem cells, leukemia stem cells, and AML blast populations to systematically compare chromatin accessibility, inferred transcription factor activity, and transcription factor-target gene regulatory connectivity. We found that global chromatin accessibility and transcription factor repertoires were broadly shared across populations, while differential accessibility was mainly localized to distal regulatory elements. In parallel, transcription factor activity exhibited structured, population-biased patterns rather than uniform shifts. Importantly, regulatory network analysis revealed extensive changes in transcription factor-target gene connectivity, with substantial divergence at the level of regulatory interactions despite relative conservation of transcription factor identity. Together, these multi-layer analyses support a model in which regulatory variation across AML-related cell states inferred from these populations is more strongly associated with reorganization of TF-target regulatory connectivity than with replacement of transcription factor composition. Overall, this study provides a systematic framework for understanding regulatory organization and its variation across AML cell populations.

Keywords

Acute Myeloid Leukemia, Chromatin Accessibility, Gene Regulatory Networks, Transcription Factor Activity

1. 引言

急性髓系白血病(acute myeloid leukemia, AML)是一种高度异质性的血液系统恶性肿瘤，其特征是在单个患者体内共存多种细胞群体，这些群体在整体上反映出与不同分化阶段相关的细胞状态[1]。当前主流模型认为，AML 呈现类似正常造血过程的层级结构，其中具有自我更新能力的白血病干细胞(leukemia stem cells, LSC)维持疾病的持续传播，而分化程度更高的原始细胞(blast)则表现出有限的自我更新潜能[1] [2]。尽管该模型主要来源于功能性移植实验，这一框架在塑造当前对 AML 细胞组织结构的认识方面发挥了重要作用。

然而，越来越多的研究证据表明，在基因表达和染色质可及性层面，AML 相关细胞群体并未完全分离，而是表现出显著的重叠[3] [4]。这些观察结果对严格的层级模型提出了挑战，并提示不同细胞状态之间的调控差异可能并非由明确的离散边界所界定。在这一背景下，转录程序如何在部分重叠的细胞群体之间被建立并维持，仍有待进一步阐明。

在本研究中，我们对“细胞群体(cell populations)”与“细胞状态(cell states)”进行区分。细胞群体是

指通过实验定义或注释获得的细胞集合，例如通过流式细胞分选(fluorescence-activated cell sorting, FACS)获得或来源于数据集标注的类别(如 pHSC、LSC 及 AML 原始细胞)，构成本研究分析的基本单位。相比之下，细胞状态则是基于染色质可及性、转录因子(transcription factor, TF)活性及基因调控网络结构所推断的生物学或调控构型。尽管本研究的分析在细胞群体层面展开，但其解释旨在刻画潜在的细胞状态及其相互关系。因此，这些推断得到的细胞状态应被视为对统一定义细胞群体的调控层面解释，而非普遍意义上的离散生物学实体。

染色质可及性为解决上述问题提供了一个可行的研究切入点。转座酶可及染色质测序(assay for transposase-accessible chromatin using sequencing, ATAC-seq)能够在全基因组范围内刻画染色质开放状态，从而捕获顺式调控活性及潜在的转录因子结合信息[5]。在正常造血及白血病过程中，远端调控元件，尤其是增强子(enhancers)的动态调控，与细胞身份维持及状态转变密切相关[6]。将这些信息与转录输出相结合，有助于系统性解析不同细胞状态之间的调控组织方式。

在本研究中，我们整合多个 AML 队列的 bulk ATAC-seq 及 RNA 测序(RNA sequencing，RNA-seq)数据，对已注释的前白血病造血干细胞(preleukemic hematopoietic stem cells, pHSC)、白血病干细胞(LSC)及 AML 原始细胞群体进行调控特征比较。本研究进一步探讨：AML 相关细胞状态之间的变异，是否主要由全局染色质可及性的变化、推断的转录因子活性改变，或转录因子–靶基因(TF-target gene, TF-TG)调控连接关系的重构所驱动。通过比较不同细胞群体的调控结构，本研究旨在识别与 AML 细胞状态变异相关的候选调控特征。

2. 方法

2.1. 公共数据集收集与样本注释统一

从 Gene Expression Omnibus (GEO)数据库中收集多个 AML 队列的 bulk ATAC-seq 数据集(GSE74912、GSE256495、GSE150868 和 GSE197416)，并加以整合，用于分析 AML 相关造血细胞群体之间的调控变异。同时，从 GEO 数据集 GSE74246 获取 bulk RNA 测序(RNA sequencing, RNA-seq)数据，用于后续转录分析及调控网络分析。原始样本注释被统一整理为六类细胞群体标签：pHSC、LSC、Blast、Myeloblast、Bulk 及 primary AML cells。由于注释信息不完整或存在不一致，限制了跨数据集比较的可比性，因此在后续比较分析中排除了 Bulk 及 primary AML cells。基于一致的免疫表型特征及相似的全局染色质可及性模式，将 Blast 与 Myeloblast 群体合并为统一的 AML 原始细胞(AML blast)群体。最终构建的三类细胞群体分析框架(pHSC、LSC 及 AML blast)用于所有后续分析。

2.2. ATAC-seq 数据处理与质量控制

所有 ATAC-seq 样本均采用 BiasFreeATAC 流程[7]进行处理，该流程分别使用 fastp (v0.23.4)进行读段修剪，并使用 Bowtie2 (v2.2.5)进行比对。样本质量通过峰内 reads 比例(fraction of reads in peaks, FRiP)及转录起始位点(transcription start site, TSS)富集度进行评估。FRiP < 0.15 或 TSS 富集度 < 3 的样本被判定为失败样本并予以剔除。FRiP > 0.30 且 TSS 富集度 > 4.5 的样本被定义为高质量样本，其余未被判定为失败的样本定义为中等质量样本。后续分析仅保留中等及高质量样本。通过分析插入片段长度分布评估核小体相关的周期性特征，并基于单样本峰识别结果汇总可及染色质区域。基因组注释文件采用 Homo_sapiens.GRCh38.113.gtf。

2.3. 共识峰构建与可及性定量

为在统一特征空间中实现样本间比较，基于通过质量控制的 ATAC-seq 样本构建共识峰(consensus

peaks)。在每个注释细胞群体内，基于跨样本可重复出现的峰，对单样本峰进行合并，构建群体水平的共识峰集合。随后，将各细胞群体的共识峰集合进一步整合，形成覆盖所有群体的最终联合峰集合(union peak set)。基于该联合峰集合，采用 featureCounts (v2.0.8)对染色质可及性进行定量，构建峰 × 样本计数矩阵(peak-by-sample count matrix)。所得计数矩阵采用 DESeq2 (v1.46.0)进行归一化处理，并进一步进行方差稳定化转换(variance-stabilizing transformation, VST)。

2.4. 全局染色质可及性分析及方差分解

基于共识峰计数矩阵(consensus peak count matrix)，采用 Pearson 相关系数、欧氏距离及主成分分析(principal component analysis, PCA)对全局染色质可及性进行评估。分析基于变异性最高的峰(top variable peaks)进行，热图同时采用无监督聚类和基于细胞群体引导的样本排序进行构建。利用 PCA 分析样本在低维空间中的结构特征。为量化生物学因素与技术因素的贡献，采用线性模型对 PC1 及 PC2 进行方差分解，以患者来源(patient identity)、细胞群体(cell population)及数据集层面效应(dataset-level variation)作为解释变量。

2.5. 差异染色质可及性分析

在三类细胞群体(pHSC、LSC 及 AML 原始细胞(AML blast))构成的数据子集中，采用 DESeq2 进行差异染色质可及性分析。分析采用“一对其余(one-versus-rest)”比较框架，模型设计公式为~dataset + group，其中 group 表示当前比较的细胞群体相对于其他样本的对比。差异可及区域(differentially accessible regions, DARs)定义为校正后 p 值(adjusted p-value) < 0.05 且$|\log_2 \text{fold change}| > 1$ 的峰。为可视化分析，采用 limma (v3.62.1)从方差稳定化转换(variance-stabilizing transformation, VST)后的计数矩阵中去除与数据集相关的批次效应。显著 DARs 用于后续分析，包括热图绘制及功能注释。

2.6. 峰注释及基于区域的功能富集分析

采用 ChIPseeker (v1.42.1)结合 TxDb.Hsapiens.UCSC.hg38.knownGene 对 DARs 进行基因组注释。以峰到转录起始位点(transcription start site, TSS)的距离为依据，将其区分为启动子邻近区域与远端调控区域。功能富集分析采用 rGREAT (v2.6.0)进行，以 Gene Ontology 生物学过程(Gene Ontology biological process, GO:BP)作为基因集来源，并以共识峰集合(consensus peak set)作为背景。对可及性升高与降低的峰分别进行分析，并筛选显著富集条目用于后续可视化。

2.7. 基于 TOBIAS 的转录因子足迹分析

采用 TOBIAS (v0.17.3)在一个样本数量均衡的三组 ATAC-seq 子集中进行转录因子(transcription factor, TF)足迹分析，该子集包含 pHSC、LSC 及 AML 原始细胞(AML blast)群体，且各组样本数量相同(每组 18 个样本)。在进行偏倚校正(bias correction)后生成足迹信号，并在细胞群体层面进行汇总以用于比较分析。采用 BINDetect 模块结合脊椎动物 JASPAR 2026 motif 数据库评估差异 TF 活性。所得足迹得分用于后续 TF 层面的分析。

2.8. TF 特异性评分及 TF 层面足迹汇总

将 TOBIAS BINDetect 输出结果归并为每个 TF 对应的单一代表性 motif，用于后续分析。基于 pHSC、LSC 及 AML blast 群体中的平均足迹信号，计算 TF 特异性评分(TF specificity score)以量化群体偏倚的 TF 活性。该评分定义为某一 TF 在目标细胞群体中的足迹信号减去其在其他细胞群体中最高足迹信号的差

值。根据该评分，将每个 TF 归类至其相对活性最高的细胞群体，并据此进行排序，以筛选用于可视化的代表性群体偏倚调控因子。

2.9. 基于 Motif 的验证分析(HOMER 与 chromVAR)

采用独立的 motif 层面分析对 TOBIAS 推断的 TF 活性模式进行验证。使用 HOMER (v4.11)在差异 ATAC 峰上进行 motif 富集分析。同时，采用 chromVAR (v1.30.1)基于 JASPAR motif 数据库计算样本间的 motif 相关染色质可及性，数据组织于 SummarizedExperiment (v1.36.0)框架中。评估不同细胞群体之间的 motif 活性差异，并用于后续可视化分析。上述分析为群体相关的 TF 活性模式提供了独立的验证证据。

2.10. RNA-Seq 整合及差异表达分析

将来自 GEO 数据集 GSE74246 的 bulk RNA 测序(RNA sequencing, RNA-seq)数据统一至与 ATAC-seq 分析一致的三类细胞群体框架(pHSC、LSC 及 AML 原始细胞(AML blast))。构建基因层面的计数矩阵(gene-level count matrix)用于后续分析，并采用 edgeR 中的 filterByExpr 函数筛选低表达基因，保留在最小数量样本中具有足够表达量的基因，该阈值由文库大小(library size)及实验设计确定。差异表达分析采用 edgeR-voom/limma 流程，在“一对其余(one-versus-rest)”框架下进行，模型设计矩阵包括细胞群体(cell population)及数据集(dataset)作为协变量。显著上调基因定义为校正后 p 值(adjusted p-value) < 0.05 且 logFC > 1 的基因。采用 clusterProfiler (v4.14.0)基于 Gene Ontology 生物学过程(GO:BP)对上调基因进行功能富集分析，并以热图形式进行可视化。上述分析用于表征 AML 细胞群体之间的转录差异，并为后续调控网络分析提供支持。

2.11. 用于 PECA2 类网络推断的 ATAC 与 RNA 整合输入构建

为在细胞群体层面构建调控网络，分别在 pHSC、LSC 及 AML blast 群体内对 ATAC-seq 及 RNA-seq 数据进行聚合，并在群体层面进行整合。该设计能够捕获群体层面的共享调控特征，但不保留不同组学数据在样本层面的对应关系。对于 ATAC-seq 数据，在各细胞群体内进行峰识别，并合并得到共享的调控元件集合。对于 RNA-seq 数据，在每个细胞群体内对基因表达值取平均，构建群体水平的表达谱。上述聚合后的 ATAC 及 RNA 数据作为输入，用于后续调控网络推断。

2.12. 自定义 PECA2 类调控网络推断

采用基于 MATLAB (R2023b, MathWorks)实现的自定义 PECA2 类框架进行调控网络推断。在每个细胞群体中，整合染色质可及性、基因表达、转录因子(transcription factor, TF) motif 信息以及先验 TF-靶基因(TF-target gene, TF-TG)关系，以计算 TF-靶基因调控评分。负对照 TF-TG 对定义为来源于 PECA2 框架提供的预先构建的负对照集合中的 TF-基因配对，用于表示潜在的非调控关系。基于该负对照集合的评分分布构建经验背景。采用背景分布第 95 百分位数作为阈值筛选高置信度调控边。当无调控边通过该阈值时，保留评分最高的前 500 个 TF-TG 对用于后续分析。通过上述方法分别构建 pHSC、LSC 及 AML blast 群体的特异性调控网络。

2.13. 网络重叠、空模型比较、拓扑分析及 TF 重构量化

在细胞群体特异的 TF-TG 调控网络基础上进行比较分析。采用 Jaccard 相似性系数，在转录因子(TF)、靶基因(target gene, TG)及 TF-TG 调控连接三个层面量化 pHSC、LSC 及 AML blast 之间的重叠程度。为评估统计显著性，构建与真实网络规模匹配的随机网络，以估计空模型(null model)下的重叠分布。采用标准网络指标对网络拓扑结构进行表征。通过比较各转录因子在不同细胞群体中的靶基因集合，量化 TF

调控连接重构程度。重构得分(rewiring score)定义为 1 减去该 TF 在不同细胞群体中靶基因集合 Jaccard 相似性的平均值，评分越高表示重构程度越大。

3. 结果

3.1. 全局染色质可及性仅能部分区分 AML 细胞群体

为刻画 AML 相关造血细胞群体之间的全局表观遗传关系，我们整合了来自多个公开数据来源的 bulk ATAC-seq 数据，这些数据涵盖了经流式细胞分选(fluorescence-activated cell sorting, FACS)定义的前白血病造血干细胞(preleukemic hematopoietic stem cells, pHSC)、白血病干细胞(leukemia stem cells, LSC)以及原始细胞富集组分(blast-enriched fractions) (表 1；图 S1(A))。在该分析阶段，我们保留原始注释信息，以在数据整合前反映不同数据集中各自的分选策略。尽管分选标准及命名方式存在差异，这些细胞群体在整体上大致对应于从干细胞样状态向原始细胞富集白血病状态的谱系连续过程[8]，从而为跨群体比较染色质可及性特征及其相关调控属性提供了基础。

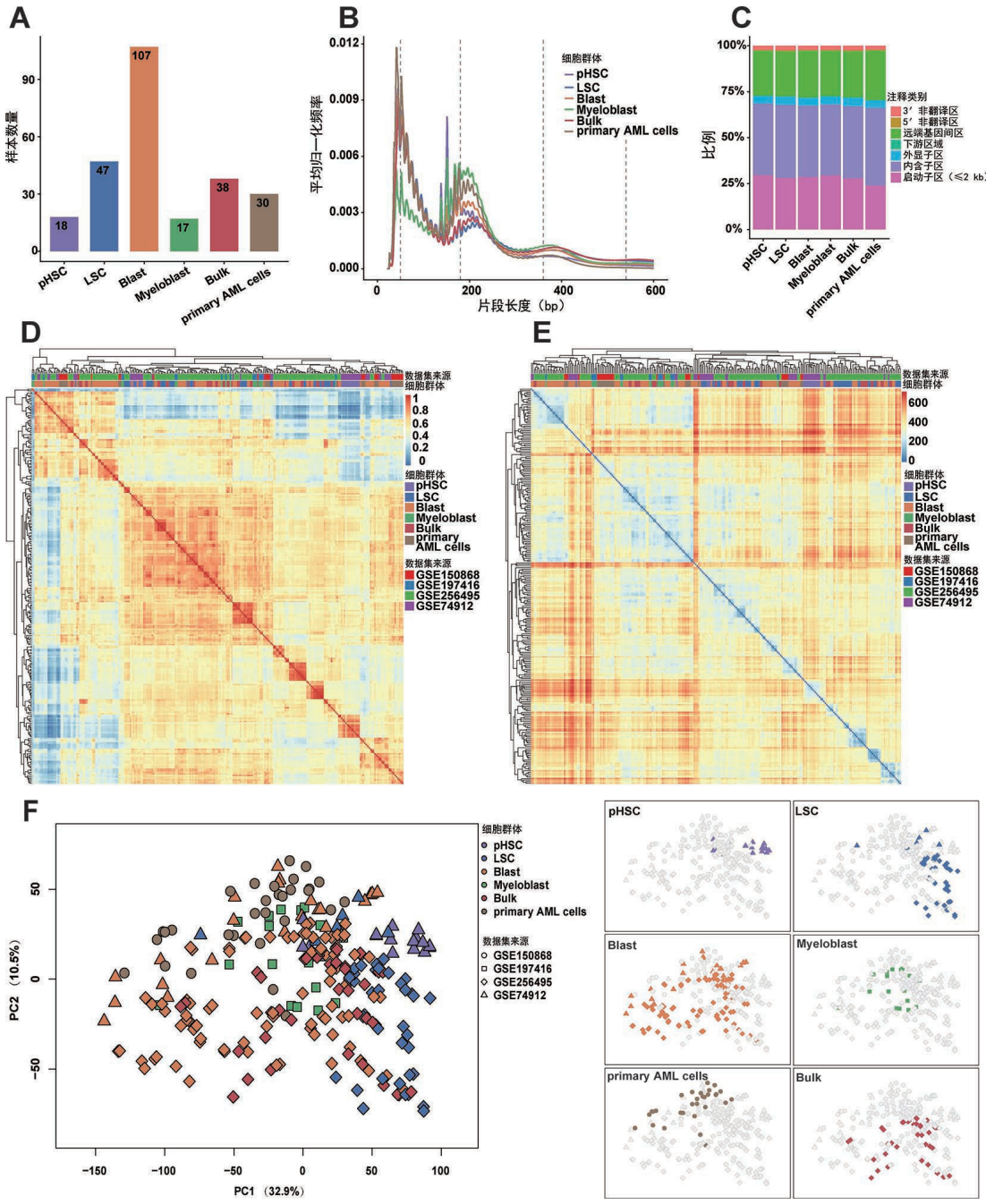
所有样本均采用统一的质量控制流程进行处理(见方法部分)。基于峰内 reads 比例(FRiP)及转录起始位点(TSS)富集度，大多数样本符合常用 ATAC-seq 质量标准[9] [10]，低质量样本已在后续分析中剔除(图 1(A)；图 S1(B)~(E))。整体片段长度分布呈现典型的核小体周期性特征(图 1(B)；图 S1(F))，支持整体数据质量的可靠性[5] [9]。在质量控制之后，我们对每个样本进行峰识别以界定候选开放染色质区域(见方法部分)。基因组注释结果显示，可及染色质区域主要分布于启动子邻近区、内含子区及远端基因间区(图 1(C))，其中相当比例定位于远端调控区域，提示顺式调控元件(如增强子)在 AML 表观遗传调控中具有重要作用[11] [12]。

随后，我们采用多种互补方法比较样本间的全局染色质可及性。为在统一特征空间中进行比较，我们构建了共识峰集合(见方法部分)。基于该集合进行的样本间相关性分析及欧氏距离分析均表明，同一细胞群体内样本具有更高相似性，同时不同群体之间仍存在较高接近性(图 1(D)，图 1(E)；图 S1(G)，图 S1(H))。因此，pHSC、LSC 及 blast 群体在染色质可及性上存在总体差异，但在全局层面未形成明确分离。

无监督主成分分析(PCA)进一步揭示了这一特征。在联合投影空间中，各样本呈连续分布，而非划分为离散簇(图 1(F)，左)。blast 样本主要分布于 PC1 负向区域，而 LSC 样本则偏向 PC1 正向区域；pHSC 样本集中于相对较窄的区域，并表现出较高的 PC2 取值，而其他细胞群体则在共享的 PCA 空间中呈更为分散的分布。在同一 PCA 投影中对各细胞群体进行高亮显示，可观察到各群体倾向于占据特定区域，但不同群体之间仍存在广泛重叠(图 1(F)，右)。结果表明，全局染色质可及性能够反映 AML 细胞群体间的结构性差异，但不足以将其分离为明确的离散表观遗传群体。

为进一步解析 PCA 结构的来源，我们对主成分的方差贡献进行了分析。结果显示，在 PC1 中，患者来源(38.6%)及细胞群体(31.8%)解释了主要变异，而数据集层面效应贡献较小(6.0%)。相比之下，PC2 主要受到患者来源(51.8%)及数据集层面效应(40.4%)的影响，而细胞群体的贡献较低(图 S1(I))。对样本组成的进一步分析表明，不同细胞群体分布于多个患者及研究来源中，而多数患者仅对应单一细胞群体，不同研究之间缺乏充分的匹配样本。综合上述结果表明，在整合数据集中，患者背景、研究来源及细胞群体之间存在一定程度的耦合关系，因此全局染色质可及性信号不应被简单解释为仅反映细胞群体差异。

总体而言，全局染色质可及性能够捕获 AML 相关 pHSC、LSC 及 blast 群体之间具有生物学意义但不完全的差异。整合 ATAC-seq 数据所反映的图谱同时包含群体相关结构及显著的个体差异，提示需要引入更精细的调控特征以进一步解析细胞状态特异性的调控程序。



(A) 质量控制后纳入分析的各细胞群体 bulk ATAC-seq 样本数量，包括前白血病造血干细胞(preleukemic hematopoietic stem cells, pHSCs)、白血病干细胞(leukemia stem cells, LSCs)、blast 组分、myeloblast 组分、Bulk AML 细胞及原代 AML 细胞群体；(B) 各细胞群体的 ATAC-seq 片段长度分布，显示典型的核小体周期性特征；(C) 不同细胞群体中可及染色质区域的基因组注释。大多数峰分布于启动子邻近区、内含子区及远端基因间区；(D) 基于共识峰矩阵(consensus peak matrix)计算的所有样本染色质可及性 Pearson 相关性热图。样本按细胞群体及数据集来源标注；(E) 基于与(D)相同样本及特征空间计算的欧氏距离热图。样本按细胞群体及数据集来源标注；(F) 基于高变异共识峰的整合 ATAC-seq 样本主成分分析(principal component analysis, PCA)。左图中样本按细胞群体着色，并按数据集来源区分形状；右图为相同投影下分别高亮各细胞群体，展示其在共享染色质可及性空间中的分布。

Figure 1. Integrated ATAC-seq profiling defines the global chromatin accessibility landscape across AML-associated cell populations

图 1. 整合 ATAC-seq 分析揭示 AML 相关细胞群体的全局染色质可及性图谱

Table 1. Summary of GEO datasets and FACS-based cell population definitions
表 1. GEO 数据集及基于 FACS 的细胞群体定义汇总

GEO	测序类型	原始注释	FACS 分选定义	合并后的细胞群体
GSE74912	ATAC-seq	pHSC	Lin^- $CD34^+$ $CD38^-$ $TIM3^-$ $CD99^-$	pHSC
		LSC	Lin^- $CD34^+$ $CD38^-$ $TIM3^+$ $CD99^+$	LSC
		Blast	CD45 中等表达；SSC 高表达；非 LSC 组分	AML blast
GSE256495		LSC	$CD34^+$ $CD38^-$ $CD99^+$ $TIM3^+$	LSC
		Blast	CD45 中等表达；SSC 高表达；非 $CD34^+$ $CD38^-$	AML blast
		Bulk	未说明	排除*
GSE150868		Primary AML cells	未说明	排除*
GSE197416		Myeloblast	CD45/SSC 分选；选择 $CD33^+$和/或 $CD34^+$细胞；排除 $CD3^+$T 细胞及 $CD14^+$单核细胞	AML blast
GSE74246	RNA-seq	pHSC	Lin^- $CD34^+$ $CD38^-$ $TIM3^-$ $CD99^-$	pHSC
		LSC	Lin^- $CD34^+$ $CD38^-$ $TIM3^+$ $CD99^+$	LSC
		Blast	CD45 中等表达；SSC 高表达；非 LSC 组分	AML blast

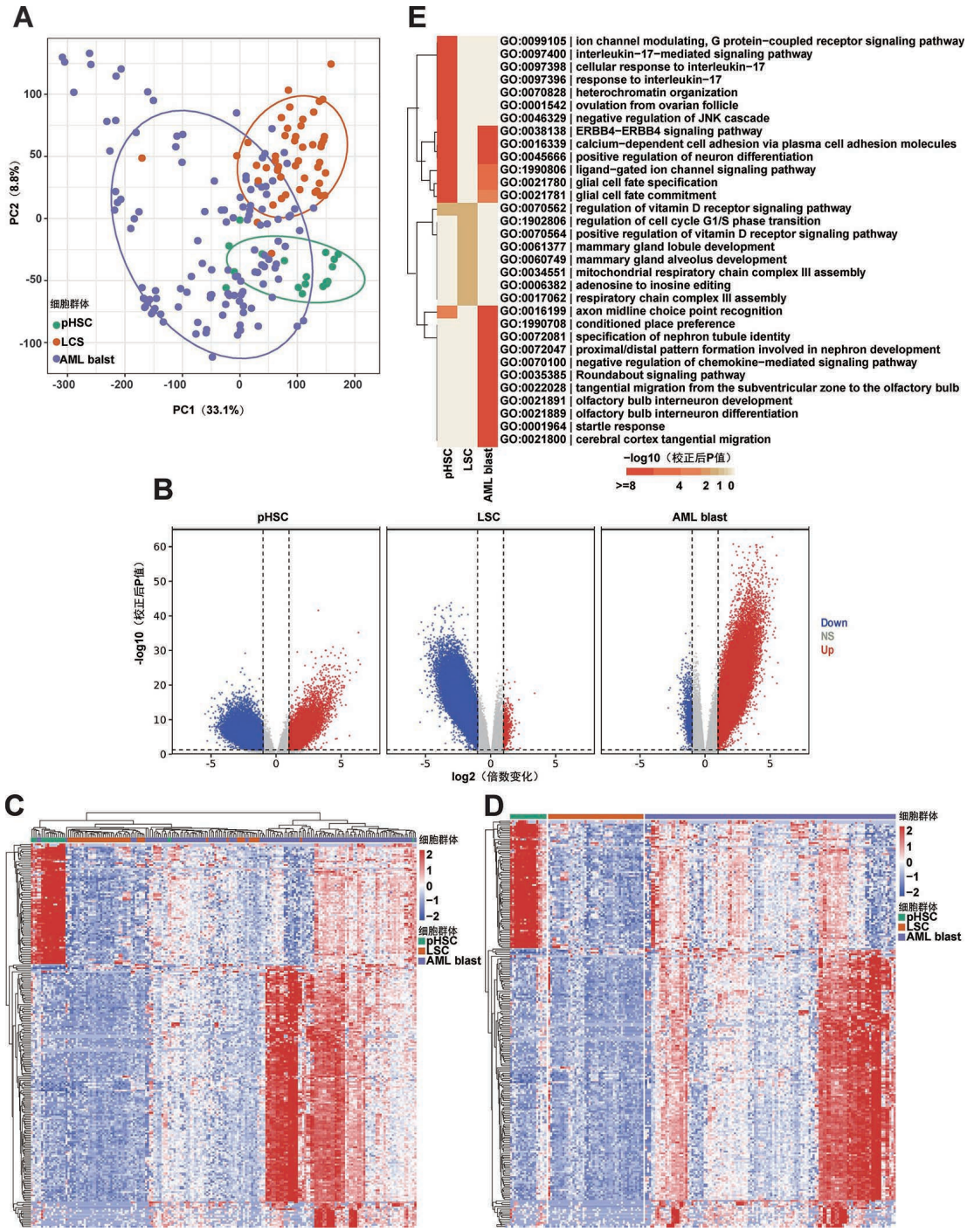
注：*由于不同数据集之间细胞分选注释不一致或信息不足，相关样本予以排除。

3.2. AML 细胞群体间差异性染色质可及性主要富集于远端调控元件

鉴于全局染色质可及性仅能在一定程度上区分不同 AML 细胞群体，我们进一步将后续分析聚焦于整合数据集中具有较高可比性的细胞群体。原始数据集中共包含六类注释细胞群体(pHSC、LSC、Blast、Myeloblast、Bulk 及 primary AML cells)。其中，由于分选信息不完整或不同数据集间缺乏一致性，Bulk 及 primary AML 细胞被排除。尽管 Blast 与 Myeloblast 群体基于不同分选策略定义，但两者在经典髓系标志物水平上表现出较高一致性(Myeloblast 依据 $CD45^+$、$CD3^-$及 CD34/CD33 组合定义；Blast 依据 CD45、侧向散射特征及排除 LSC 样群体定义)，且在主成分分析中呈现一致的分布位置[4] [13] [14]。因此，这两类群体被合并为统一的 AML 原始细胞(AML blast)群体。最终构建的三类细胞群体(pHSC、LSC 及 AML blast)作为后续分析的参照框架，而非被视为严格离散的 AML 细胞状态[4] [15]-[17]。

基于高变异染色质区域的主成分分析显示，三类细胞群体在低维空间中发生位置偏移，但仍存在部分重叠(见方法部分，图 2(A))。与此一致，基于距离的分析亦表明不同细胞群体之间并未完全分离，即使在按细胞群体分组后仍然如此(图 S2(A)，图 S2(B))。上述结果表明，即使在统一为最具可比性的三类细胞群体之后，全局染色质可及性仍不足以清晰区分 AML 细胞群体。

在此基础上，我们采用“一对其余(one-versus-rest)”框架并结合批次校正(见方法部分)，对各细胞群体进行差异性染色质可及性分析。差异可及区域(differentially accessible regions, DARs)在不同细胞群体中呈现出不同模式(图 2(B))：可及性升高区域在 AML blast 中最多，在 LSC 中最少，而在 pHSC 中处于中间水平。基于最显著差异区域的聚类分析能够在群体层面区分三类细胞群体(图 2(C)，图 2(D))，但在样本层面仍存在一定程度的重叠。值得注意的是，显著差异区域($padj < 0.05, |\log_2 FC| > 1$)主要位于远离转录起始位点(TSS)的区域，并富集于内含子区及远端基因间区(图 S2(C)，图 S2(D))，表明 AML 细胞群体间的染色质差异主要来源于远端顺式调控元件[4]。



(A) 基于高变异共识峰(consensus peaks)的主成分分析(principal component analysis, PCA)，展示 pHSC、LSC 及 AML 原始细胞(AML blast)样本在染色质可及性空间中的分布，表现为位置偏移但部分重叠；(B) 火山图展示在“一对其余(one-versus-rest)”框架下识别的各细胞群体差异可及区域(differentially accessible regions, DARs)。可及性升高的峰以红色表示，可及性降低的峰以蓝色表示，非显著峰以灰色表示；(C) 基于按行标准化(row-scaled)的可及性值并采用无监督聚类构建的最显著差异可及区域热图，展示所有样本的整体模式。样本按细胞群体进行标注；(D) 与(C)相同的差异可及区域热图，但样本按细胞群体排序，以突出群体相关的可及性模式；(E) 各细胞群体中可及性升高区域的 Gene Ontology (GO)生物学过程富集分析。颜色表示富集显著性。

Figure 2. Chromatin accessibility differences across AML cell populations are concentrated at distal regulatory elements
图 2. AML 细胞群体间染色质可及性差异主要富集于远端调控元件

对可及性升高区域进行功能富集分析显示，不同细胞群体具有不同的调控功能倾向(图 2(E))。在 pHSC 中，可及性升高区域富集于信号应答相关通路，包括 IL-17 信号通路及 G 蛋白偶联受体信号通路，同时涉及染色质相关过程。在 LSC 中，富集通路主要与细胞周期及代谢过程相关，包括 G1/S 期转换调控。在 AML blast 中，富集通路主要涉及细胞迁移、细胞间相互作用及发育过程，包括白细胞迁移及趋化因子信号通路。

对可及性降低区域的分析显示出互补模式(图 S2(E))。在 LSC 中，可及性降低区域富集于白细胞迁移、趋化及炎症反应相关通路，与其在免疫及微环境相关调控中的参与降低一致。在 AML blast 中，与核受体相关的调控元件可及性降低，提示分化相关转录程序的活性减弱[18]-[20]。此外，一些在 pHSC 中不活跃的发育及迁移相关通路，在不同细胞群体之间表现出差异性调控。总体而言，这些结果表明 AML 细胞群体之间呈现出协调的染色质可及性变化模式，其中大部分差异信号集中于远端调控区域。

3.3. AML 细胞群体间染色质可及性差异伴随转录因子活性的结构化变化

鉴于差异性染色质可及性主要富集于远端调控元件，我们进一步探讨这些染色质差异是否与上游转录因子(transcription factor, TF)活性在 pHSC、LSC 及 AML 原始细胞(AML blast)群体之间的变化相关。

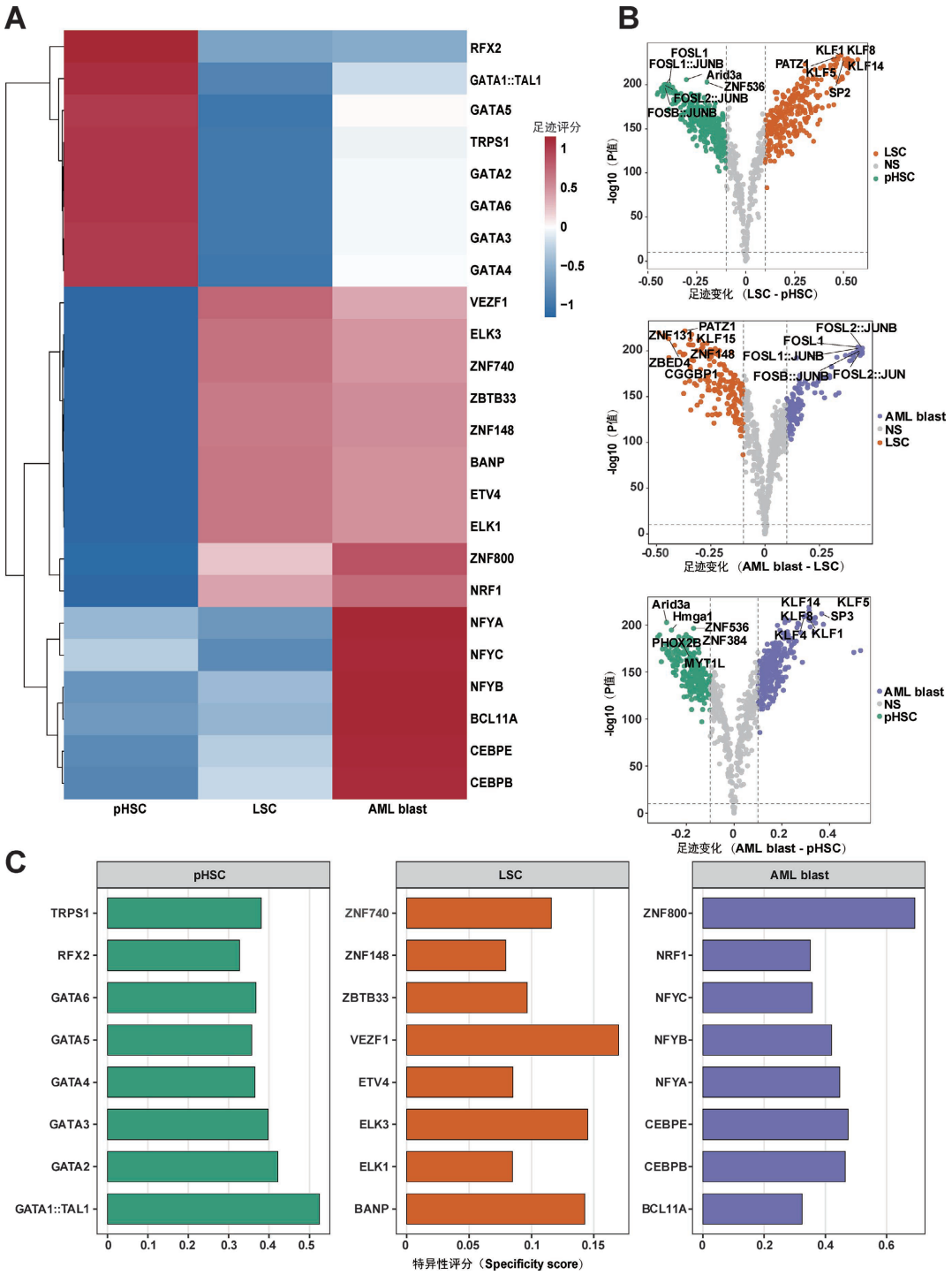
为此，我们基于整合的 bulk ATAC-seq 数据，采用 TOBIAS 方法进行足迹分析(foot printing) [21]，推断转录因子结合活性，并在不同细胞群体之间进行比较(见方法部分)。在主要转录因子家族层面观察到清晰且具有生物学意义的差异模式(图 3(A))。在 pHSC 中，足迹信号富集于 GATA 家族转录因子(如 GATA1/2/3 及相关模块)，与其在造血干/祖细胞维持及早期谱系调控中的作用一致[22] [23]。在 LSC 中，ETS 家族转录因子(如 ELK1、ETV4)足迹信号更为显著，反映其与细胞增殖、信号转导及应激反应相关的调控过程[24] [25]。在 AML blast 中，CEBP 及 NFY 家族转录因子足迹信号显著富集，这些因子分别与髓系分化及细胞周期调控相关，与白血病原始细胞增强的增殖及转录活性一致[26] [27]。

两两比较的转录因子足迹变化进一步支持上述模式(图 3(B))。相较于 pHSC，LSC 中 GATA 家族转录因子活性降低，而 ETS 家族转录因子活性升高；在涉及 AML blast 的比较中，可观察到 CEBP 及 NFY 等转录因子活性增强。这些变化沿 pHSC-LSC-AML blast 轴呈现方向性趋势，而非在不同细胞群体间随机分布，其中 LSC 表现出中间型的转录因子活性特征。

为系统量化转录因子在不同细胞群体中的偏倚性，我们基于 TOBIAS 计算的平均足迹信号定义了转录因子特异性评分(TF specificity score)，定义为某一转录因子在目标细胞群体中的足迹信号减去其在其他群体中最高信号的差值。根据该评分，将每个转录因子归类至其相对活性最高的细胞群体，并筛选出具有代表性的群体偏倚转录因子(见方法部分，图 3(C))。与上述趋势一致，GATA 相关转录因子在 pHSC 中富集，ETS 相关转录因子在 LSC 中富集，而 CEBP 及 NFY 相关转录因子在 AML blast 中富集。上述结果表明，AML 细胞群体间的染色质可及性差异表现为具有结构化、群体偏倚的转录因子活性组合，而非全局一致性变化。

为验证上述观察结果的稳健性，我们进一步采用独立的基于 motif 的方法进行验证。Homer 等[28]分析在不同细胞群体中识别到差异性的 motif 富集模式，包括 GATA、CEBP 及 AP-1 家族等转录因子 motif，表现出群体特异性富集趋势，与足迹分析结果一致(见方法部分，图 S3(A))。此外，chromVAR [29]分析基于差异性峰集合计算 motif 相关可及性变化，也观察到一致的转录因子活性差异(见方法部分，图 S3(B))。无监督及按细胞群体排序的 deviation 热图均在样本层面重现了上述差异模式(图 S3(C)，图 S3(D))。上述独立分析共同支持上述 TF 活性差异的稳健性。

综上所述，AML 细胞群体间的染色质可及性差异伴随着系统性且具有群体偏倚的转录因子活性变化，这些变化沿干细胞样至原始细胞样状态呈现方向性趋势，表明不同细胞群体对应于具有结构化、群体偏倚的转录因子活性组合。



(A) 代表性转录因子(transcription factor, TF)在 pHSC、LSC 及 AML 原始细胞(AML blast)群体中的足迹得分(footprint score)热图，显示推断的 TF 活性在不同细胞群体中的偏倚分布；(B) 转录因子足迹差异的两两比较。每个点代表一个 TF motif；正值表示在标注细胞群体中推断的 TF 活性增强，负值表示在比较细胞群体中推断的 TF 活性增强；(C) 基于 TF 特异性评分(TF specificity score)排序的各细胞群体偏倚转录因子。评分越高，表示该 TF 在对应细胞群体中推断活性的相对增强程度越强。

Figure 3. Structured, distinct transcription factor activity patterns accompany chromatin accessibility differences across AML cell populations

图 3. AML 细胞群体间染色质可及性差异伴随转录因子活性的结构化特异性模式

3.4. AML 细胞群体间推断的调控连接差异超越了共享的转录因子组成

上述分析表明，AML 细胞群体在染色质可及性及推断的转录因子(transcription factor, TF)活性层面存在差异，但这些差异在 pHSC、LSC 及 AML 原始细胞(AML blast)之间总体呈连续变化且部分重叠。为进一步明确 AML 细胞状态之间的调控差异主要体现于转录输出层面还是调控结构层面，我们整合 RNA 测序(RNA sequencing, RNA-seq)数据与染色质可及性数据，从基因表达及转录因子–靶基因(TF-target gene, TF-TG)调控连接两个层面进行联合分析。该分析旨在将染色质及转录因子层面的观察置于更完整的调控框架中，而非将推断的网络结构作为孤立证据。

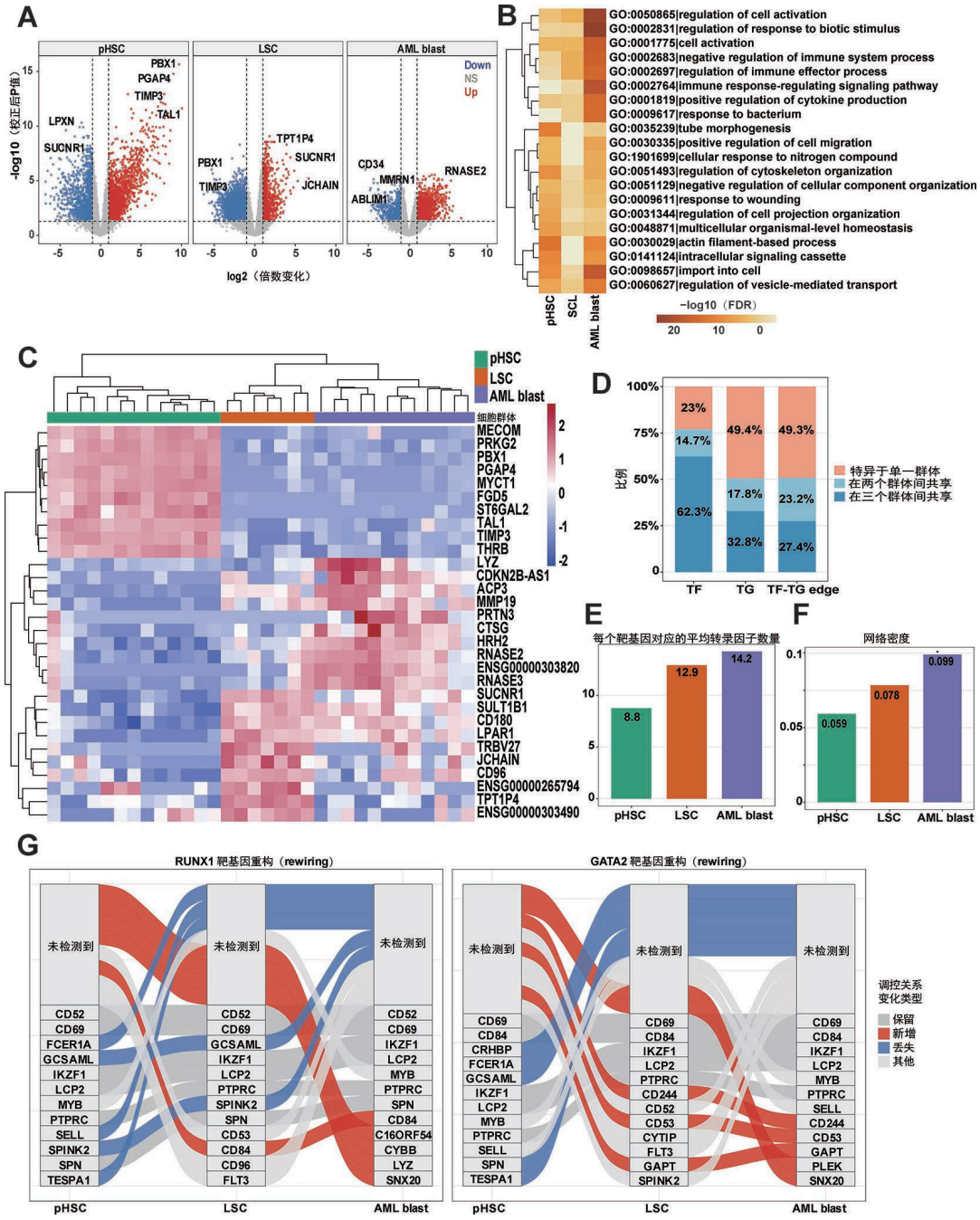
在与 ATAC-seq 分析一致的细胞群体框架下，我们构建了 pHSC、LSC 及 AML blast 的批量 RNA-seq 表达谱，并进行差异表达分析(见方法部分)。结果显示，三类细胞群体在转录水平上存在显著差异(图 4(A))。代表性基因呈现出明确的群体偏倚表达模式：与干细胞及祖细胞相关的调控因子(如 PBX1、TAL1)在 pHSC 中高表达，并在 LSC 及 AML blast 中逐渐降低；而与髓系分化及效应功能相关的基因(如 RNASE2)在 AML blast 中上调[30] [31]。LSC 在这些基因模块中表现出中间型表达模式。对上调基因的功能富集分析进一步支持上述差异(图 4(B))：pHSC 相关基因富集于细胞激活及外界刺激响应相关通路，而 AML blast 则富集于免疫反应及细胞迁移相关过程。与此一致，代表性标志基因的表达能够在一定程度上区分不同细胞群体，但仍保留部分重叠(图 4(C))。上述结果表明，AML 细胞群体不仅在染色质可及性及转录因子活性层面存在差异，其下游转录输出亦呈现出系统性变化，其中 LSC 表现出介于 pHSC 与 AML blast 之间的中间型表达模式。

为直接刻画调控结构，我们进一步整合染色质可及性与基因表达数据，采用 PECA2 方法构建 TF-TG 调控网络(见方法部分) [32]。网络组成的比较显示，不同细胞群体之间的转录因子集合具有较高重叠(图 4(D))，提示上游调控因子组成相对保守。相比之下，靶基因层面的重叠程度较低，而在 TF-TG 连接层面的重叠最低。因此，在本研究的整合分析框架下，AML 细胞群体之间最显著的差异体现在调控连接关系，而非转录因子本身的组成。结合前述远端调控元件富集及转录因子活性偏倚模式，这一结果支持如下模型：AML 细胞状态之间的调控差异主要源于部分共享调控框架内 TF-TG 连接的重构，而非调控因子的替代。

为评估上述模式是否来源于生物学结构而非随机效应，我们构建了与真实网络规模匹配的随机网络进行对照分析(见方法部分，图 S4(A))。在随机网络中，转录因子及靶基因的重叠迅速趋于饱和，而 TF-TG 连接的重叠始终较低。相比之下，真实网络表现为中等程度的转录因子重叠、较低的靶基因重叠以及高于随机预期的 TF-TG 连接重叠。该结果表明，观测到的网络重叠模式无法由随机连接解释，而更符合在共享调控框架内存在结构化调控差异的情形。

对高置信度调控连接的进一步分析揭示，不同细胞群体在网络拓扑结构上亦存在系统性差异(见方法部分，图 4(E)，图 4(F))。包括每个靶基因所对应的平均转录因子数目及网络整体密度等指标，在 pHSC、LSC 及 AML blast 之间均存在差异，提示在转录因子组成相对稳定的情况下，调控组织方式发生了改变。值得注意的是，相同转录因子在不同细胞群体中连接不同的靶基因，而同一靶基因则由不同的转录因子组合调控，表明推断的调控关系在不同细胞群体之间发生了广泛重构。

进一步地，我们在单个转录因子层面对调控重构进行量化，通过比较其在不同细胞群体中的靶基因集合相似性定义重构评分(rewiring score) (见方法部分)。结果识别出一系列高度重构的转录因子，包括 HOX 家族、GATA1/TAL1、CEBPA/CEBPD、WT1、STAT5A 及 IRF 家族成员(图 S4(B))。这些转录因子在不同细胞群体中对应的靶基因集合发生显著变化，提示其调控作用更多依赖于连接关系的改变，而非固定的靶基因程序。



(A) 火山图展示在“一对其余(one-versus-rest)”框架下识别的 pHSC、LSC 及 AML 原始细胞(AML blast)差异表达基因。表达上调基因以红色表示，下调基因以蓝色表示，非显著基因以灰色表示；(B) 各细胞群体中上调基因的 Gene Ontology(GO)生物学过程富集分析。颜色表示富集显著性；(C) 基于按行标准化(row-scaled)的表达值绘制的代表性标志基因热图，展示 pHSC、LSC 及 AML blast 之间的群体相关转录差异。样本按细胞群体标注；(D) 在转录因子(transcription factor, TF)、靶基因(target gene, TG)及 TF-TG 调控连接三个层面，展示推断调控网络中共享组分与群体特异组分的比例；(E) 三类细胞群体中每个靶基因对应的平均转录因子数量；(F) pHSC、LSC 及 AML blast 中推断调控网络的密度；(G) Sankey 图展示代表性转录因子在不同细胞群体中的靶基因连接重构(rewiring)。不同颜色的流表示推断 TF-TG 调控关系中的保留、新增、丢失或其他变化类型。

Figure 4. Inferred regulatory network rewiring reveals differences across AML cell populations beyond shared transcription factor repertoires

图 4. 推断的调控网络重构揭示 AML 细胞群体间超越共享转录因子组成的调控差异

为在个体层面展示调控重构，我们选取代表性转录因子进行具体分析(图 4(G))。例如，GATA2 在 pHSC 中主要调控与造血干/祖细胞功能相关的基因[33]，而在 LSC 及 AML blast 中，其靶基因集合部分保留，同时伴随新的靶基因获得及原有靶基因丢失。RUNX1 亦表现出类似的靶基因组成变化模式[34]。这些结果表明，转录因子–靶基因关系在不同细胞群体间发生显著改变，符合沿 pHSC-LSC-AML blast 轴的动态调控重构过程。为进一步结合生物学背景，我们还分析了在造血过程中具有明确功能的转录因子(图 S4(C))。其中，HOXA11 在 AML blast 中获得大量新的调控连接，而 GATA1 则在从 pHSC 到 AML blast 过程中逐渐丧失其调控连接，提示正常调控程序的逐步减弱[35]。

综上所述，AML 细胞群体在转录因子组成上具有较高共享性，但在 TF-TG 调控连接层面存在显著差异。本研究结果表明，AML 相关细胞状态之间的调控差异更主要体现为 TF-TG 调控连接的重构，而非转录因子组成的改变。

4. 讨论

急性髓系白血病(acute myeloid leukemia, AML)正逐渐被认识为一种由多种恶性细胞群体共同构成的疾病，这些细胞群体在分化状态及功能特征上部分重叠，而非彼此完全分离的独立单元[3] [4] [36]。本研究结果与这一认识相一致，并进一步提示，pHSC-LSC-AML blast 这一框架更适合被理解为一个简化的调控轴，而非严格离散的细胞分类体系。既往研究通常依据分化阶段或功能特征对 AML 细胞群体进行划分，并假设这些类别对应于相对独立的调控程序[37] [38]。在本研究中，通过整合基于流式细胞分选(fluorescence-activated cell sorting, FACS)定义的 pHSC、LSC 及 blast 群体的批量 ATAC-seq 及 RNA 测序(RNA sequencing, RNA-seq)数据，我们发现这些细胞状态之间的调控差异仅部分体现在全局染色质可及性层面，而更多关联于远端顺式调控元件、具有群体偏倚的转录因子(transcription factor, TF)活性以及转录因子–靶基因(TF-target gene, TF-TG)调控连接的变化。综合来看，这些结果表明，AML 细胞状态之间的调控变异更主要来源于在部分共享调控框架内的选择性调控连接重构(selective rewiring)，而非转录因子组成的替代。

本研究的一个重要发现是，全局染色质可及性在区分 AML 细胞状态方面的能力有限。尽管 pHSC、LSC 及 blast 群体在相关性分析、距离分析及主成分分析中表现出一定的群体相关结构，但仍存在广泛重叠。这一结果与既往研究中 AML 相关细胞群体在转录组或表观基因组空间中未能清晰分离的观察一致[4] [39]，提示这些细胞状态之间的调控边界可能并不像传统层级模型所假设的那样明确。值得注意的是，区分不同细胞群体的染色质差异主要集中于远端调控元件。差异性可及区域主要富集于内含子区及远端基因间区，提示 AML 细胞状态之间的调控差异主要通过以增强子为核心的调控结构进行编码，而非局限于启动子邻近区域。这一观察与远端顺式调控元件在造血系统细胞身份维持及状态可塑性中的重要作用相一致[4] [40] [41]。在这一背景下，本研究结果进一步表明，AML 细胞状态间的调控变异更多体现为对特定顺式调控元件(如增强子)的选择性使用，而非单纯的基因表达水平改变。

对转录因子活性及调控网络的联合分析进一步支持上述观点。推断的转录因子活性在 pHSC-LSC-AML blast 轴上呈现出结构化差异，其中 pHSC 富集 GATA 相关活性，LSC 富集 ETS 相关活性，而 AML blast 则富集 CEBP 及 NFY 相关活性。这些模式亦得到独立 motif 分析的支持，表明 AML 细胞状态对应于转录因子活性组合的协调变化，而非全局一致性的调控改变。值得注意的是，LSC 在该轴上表现出中间型特征，既保留部分干细胞相关调控特征，同时又获得部分与 blast 相关的转录输出特征。

本研究多层次分析中最突出的发现是调控网络层面的显著分化。尽管不同 AML 细胞状态之间的转录因子组成具有较高共享性，但在靶基因层面重叠程度较低，而在 TF-TG 调控连接层面的重叠最低。这一结果表明，AML 细胞状态之间的调控差异更主要体现为转录因子–靶基因连接关系的重构，而非转录因子本身的更替。然而，这些调控连接变化的驱动机制仍有待进一步阐明，可能涉及分化状态差异、突变背景、

微环境影响、表观遗传记忆或细胞组成混杂等多种因素。在单个转录因子层面，同时存在靶基因的保留、获得及丢失，进一步支持调控连接的选择性重构，而非简单的整体激活或抑制。识别出的高度重构转录因子(如 GATA、TAL1、HOX、CEBP、WT1、STAT5A 及 IRF 家族成员)进一步提示，转录因子的功能在 AML 中具有显著的情境依赖性，不能仅依据转录因子的存在与否进行推断。在这一过程中，转录因子表达水平、辅因子可用性、染色质状态及更高阶调控结构均可能共同决定相同转录因子在不同细胞群体中的调控输出。

总体而言，本研究结果对 AML 细胞群体组织模式的理解提供了新的视角。我们并未否定功能层级结构的存在，而是表明 pHSC、LSC 及 AML blast 之间的分子差异较经典模型所暗示的更为连续。更广泛地，这些结果提示，相较于传统基于特征的分析框架，调控网络结构可作为理解 AML 细胞状态变异的重要补充维度。本研究支持如下模型：AML 的异质性并非主要由新调控因子的引入所驱动，而更主要源于在部分重叠的恶性细胞状态之间，以增强子为核心的转录调控网络发生选择性调控连接重构。

本研究亦存在若干局限。首先，基于批量 ATAC-seq 及 RNA-seq 的数据分析无法解析细胞群体内部的异质性，可能掩盖亚克隆差异或过渡性调控状态[39] [42]。其次，不同研究来源的数据往往仅包含单一细胞群体样本，尽管已通过数据整合和方差分解控制批次效应，患者来源、数据集差异及细胞群体之间仍可能存在一定程度的混杂[43]。第三，转录因子足迹分析及调控网络推断均为间接方法，主要依赖染色质可及性、motif 信息及统计关联，而缺乏对转录因子结合或功能扰动的直接验证[44]。此外，RNA-seq 与 ATAC-seq 在样本层面缺乏严格匹配，限制了跨组学数据整合的精度。在可获得的有限匹配样本中，网络推断受样本量制约，难以得到稳健结果；同时，由于公开数据缺乏全面的分子背景信息，不同分子亚型的 AML 患者分层分析亦难以开展。尽管存在这些限制，整合多个独立数据集及多层次验证结果显示，本研究所推断的转录因子－靶基因调控重构模式具有稳健性，为未来在具备更多匹配多组学及分子背景信息的数据中进一步验证提供了基础。未来研究仍需结合高分辨率单细胞多组学数据及功能实验，以更直接地解析这些调控关系。

5. 结论

本研究系统比较了前白血病造血干细胞(preleukemic hematopoietic stem cells, pHSC)、白血病干细胞(leukemia stem cells, LSC)及 AML 原始细胞(AML blast)群体在染色质可及性、转录因子(transcription factor, TF)活性以及调控网络结构方面的差异。研究结果表明，尽管不同细胞群体在转录因子组成及全局染色质可及性模式上总体相似，其调控差异主要体现在远端顺式调控元件、群体偏倚的转录因子活性以及转录因子－靶基因(transcription factor-target gene, TF-TG)调控连接层面。

综合上述结果，本研究支持如下模型：AML 相关细胞状态之间的调控变异主要体现在推断的 TF-TG 调控连接层面，而非转录因子组成差异。该分析框架在一定程度上修正了对 AML 细胞群体组织模式的传统认识，并为后续更直接解析 AML 异质性的调控机制提供了理论基础。

致　谢

作者感谢颜光玕实验室全体成员在研究过程中提供的讨论与支持。

基金项目

本研究由国家重点研发计划(项目编号：2022YFA1106100)和国家自然科学基金(项目编号：32370600)资助。文章处理费(APC)由作者承担。

参考文献

[1] Döhner, H., Weisdorf, D.J. and Bloomfield, C.D. (2015) Acute Myeloid Leukemia. *New England Journal of Medicine*, **373**, 1136-1152. https://doi.org/10.1056/nejmra1406184

[2] Bonnet, D. and Dick, J.E. (1997) Human Acute Myeloid Leukemia Is Organized as a Hierarchy That Originates from a Primitive Hematopoietic Cell. *Nature Medicine*, **3**, 730-737. https://doi.org/10.1038/nm0797-730

[3] van Galen, P., Hovestadt, V., Wadsworth II, M.H., Hughes, T.K., Griffin, G.K., Battaglia, S., *et al*. (2019) Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease Progression and Immunity. *Cell*, **176**, 1265-1281.e24. https://doi.org/10.1016/j.cell.2019.01.031

[4] Corces, M.R., Buenrostro, J.D., Wu, B., Greenside, P.G., Chan, S.M., Koenig, J.L., *et al*. (2016) Lineage-Specific and Single-Cell Chromatin Accessibility Charts Human Hematopoiesis and Leukemia Evolution. *Nature Genetics*, **48**, 1193-1203. https://doi.org/10.1038/ng.3646

[5] Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. and Greenleaf, W.J. (2013) Transposition of Native Chromatin for Fast and Sensitive Epigenomic Profiling of Open Chromatin, DNA-Binding Proteins and Nucleosome Position. *Nature Methods*, **10**, 1213-1218. https://doi.org/10.1038/nmeth.2688

[6] Mulet-Lazaro, R. and Delwel, R. (2023) From Genotype to Phenotype: How Enhancers Control Gene Expression and Cell Identity in Hematopoiesis. *HemaSphere*, **7**, e969. https://doi.org/10.1097/hs9.0000000000000969

[7] Zhang, H., Lu, T., Liu, S., Yang, J., Sun, G., Cheng, T., *et al*. (2021) Comprehensive Understanding of Tn5 Insertion Preference Improves Transcription Regulatory Element Identification. *NAR Genomics and Bioinformatics*, **3**, lqab094. https://doi.org/10.1093/nargab/lqab094

[8] Thomas, D. and Majeti, R. (2017) Biology and Relevance of Human Acute Myeloid Leukemia Stem Cells. *Blood*, **129**, 1577-1585. https://doi.org/10.1182/blood-2016-10-696054

[9] Corces, M.R., Trevino, A.E., Hamilton, E.G., Greenside, P.G., Sinnott-Armstrong, N.A., Vesuna, S., *et al*. (2017) An Improved Atac-Seq Protocol Reduces Background and Enables Interrogation of Frozen Tissues. *Nature Methods*, **14**, 959-962. https://doi.org/10.1038/nmeth.4396

[10] Miskimen, K.L.S., Chan, E.R. and Haines, J.L. (2017) Assay for Transposase-Accessible Chromatin Using Sequencing (ATAC-Seq) Data Analysis. *Current Protocols in Human Genetics*, **92**, 20.4.1-20.4.13. https://doi.org/10.1002/cphg.32

[11] Corces, M.R., Granja, J.M., Shams, S., Louie, B.H., Seoane, J.A., Zhou, W., *et al*. (2018) The Chromatin Accessibility Landscape of Primary Human Cancers. *Science*, **362**, eaav1898. https://doi.org/10.1126/science.aav1898

[12] Bell, C.C., Fennell, K.A., Chan, Y., Rambow, F., Yeung, M.M., Vassiliadis, D., *et al*. (2019) Targeting Enhancer Switching Overcomes Non-Genetic Drug Resistance in Acute Myeloid Leukaemia. *Nature Communications*, **10**, Article 2723. https://doi.org/10.1038/s41467-019-10652-9

[13] Gorczyca, W., Sun, Z., Cronin, W., Li, X., Mau, S. and Tugulea, S. (2011) Immunophenotypic Pattern of Myeloid Populations by Flow Cytometry Analysis. In: *Methods in Cell Biology*, Elsevier, 221-266. https://doi.org/10.1016/b978-0-12-385493-3.00010-3

[14] Ally, F. and Chen, X. (2024) Acute Myeloid Leukemia: Diagnosis and Evaluation by Flow Cytometry. *Cancers*, **16**, Article 3855. https://doi.org/10.3390/cancers16223855

[15] Dick, J.E. (2005) Acute Myeloid Leukemia Stem Cells. *Annals of the New York Academy of Sciences*, **1044**, 1-5. https://doi.org/10.1196/annals.1349.001

[16] Shlush, L.I., Zandi, S., Mitchell, A., Chen, W.C., Brandwein, J.M., Gupta, V., *et al*. (2014) Identification of Pre-Leukaemic Haematopoietic Stem Cells in Acute Leukaemia. *Nature*, **506**, 328-333. https://doi.org/10.1038/nature13038

[17] Corces-Zimmerman, M.R., Hong, W., Weissman, I.L., Medeiros, B.C. and Majeti, R. (2014) Preleukemic Mutations in Human Acute Myeloid Leukemia Affect Epigenetic Regulators and Persist in Remission. *Proceedings of the National Academy of Sciences*, **111**, 2548-2553. https://doi.org/10.1073/pnas.1324297111

[18] Magliulo, D., Simoni, M., Caserta, C., Fracassi, C., Belluschi, S., Giannetti, K., *et al*. (2023) The Transcription Factor HIF2α Partakes in the Differentiation Block of Acute Myeloid Leukemia. *EMBO Molecular Medicine*, **15**, e17810. https://doi.org/10.15252/emmm.202317810

[19] Sanchez, P.V., Glantz, S.T., Scotland, S., Kasner, M.T. and Carroll, M. (2014) Induced Differentiation of Acute Myeloid Leukemia Cells by Activation of Retinoid X and Liver X Receptors. *Leukemia*, **28**, 749-760. https://doi.org/10.1038/leu.2013.202

[20] Pan, P. and Chen, X. (2020) Nuclear Receptors as Potential Therapeutic Targets for Myeloid Leukemia. *Cells*, **9**, Article 1921. https://doi.org/10.3390/cells9091921

[21] Bentsen, M., Goymann, P., Schultheis, H., Klee, K., Petrova, A., Wiegandt, R., *et al*. (2020) Atac-Seq Footprinting Unravels Kinetics of Transcription Factor Binding during Zygotic Genome Activation. *Nature Communications*, **11**, Article No. 4627. https://doi.org/10.1038/s41467-020-18035-1

[22] Rodrigues, N.P., Tipping, A.J., Wang, Z. and Enver, T. (2012) GATA-2 Mediated Regulation of Normal Hematopoietic Stem/Progenitor Cell Function, Myelodysplasia and Myeloid Leukemia. *The International Journal of Biochemistry & Cell Biology*, **44**, 457-460. https://doi.org/10.1016/j.biocel.2011.12.004

[23] Hewitt, K.J., Johnson, K.D., Gao, X., Keles, S. and Bresnick, E.H. (2016) The Hematopoietic Stem and Progenitor Cell Cistrome. In: *Current Topics in Developmental Biology*, Elsevier, 45-76. https://doi.org/10.1016/bs.ctdb.2016.01.002

[24] Ciau-Uitz, A., Wang, L., Patient, R. and Liu, F. (2013) ETS Transcription Factors in Hematopoietic Stem Cell Development. *Blood Cells*, *Molecules*, *and Diseases*, **51**, 248-255. https://doi.org/10.1016/j.bcmd.2013.07.010

[25] Kar, A. and Gutierrez-Hartmann, A. (2013) Molecular Mechanisms of ETS Transcription Factor-Mediated Tumorigenesis. *Critical Reviews in Biochemistry and Molecular Biology*, **48**, 522-543. https://doi.org/10.3109/10409238.2013.838202

[26] Pabst, T. and Mueller, B.U. (2009) Complexity of CEBPA Dysregulation in Human Acute Myeloid Leukemia. *Clinical Cancer Research*, **15**, 5303-5307. https://doi.org/10.1158/1078-0432.ccr-08-2941

[27] Ly, L.L., Yoshida, H. and Yamaguchi, M. (2013) Nuclear Transcription Factor Y and Its Roles in Cellular Processes Related to Human Disease. *American Journal of Cancer Research*, **3**, 339-346.

[28] Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., *et al*. (2010) Simple Combinations of Lineage-Determining Transcription Factors Prime CIS-Regulatory Elements Required for Macrophage and B Cell Identities. *Molecular Cell*, **38**, 576-589. https://doi.org/10.1016/j.molcel.2010.05.004

[29] Schep, A.N., Wu, B., Buenrostro, J.D. and Greenleaf, W.J. (2017) ChromVAR: Inferring Transcription-Factor-Associated Accessibility from Single-Cell Epigenomic Data. *Nature Methods*, **14**, 975-978. https://doi.org/10.1038/nmeth.4401

[30] Ficara, F., Murphy, M.J., Lin, M. and Cleary, M.L. (2008) Pbx1 Regulates Self-Renewal of Long-Term Hematopoietic Stem Cells by Maintaining Their Quiescence. *Cell Stem Cell*, **2**, 484-496. https://doi.org/10.1016/j.stem.2008.03.004

[31] Hu, X., Ybarra, R., Qiu, Y., Bungert, J. and Huang, S. (2009) Transcriptional Regulation by TAL1: A Link between Epigenetic Modifications and Erythropoiesis. *Epigenetics*, **4**, 357-361. https://doi.org/10.4161/epi.4.6.9711

[32] Duren, Z., Chen, X., Xin, J., Wang, Y. and Wong, W.H. (2020) Time Course Regulatory Analysis Based on Paired Expression and Chromatin Accessibility Data. *Genome Research*, **30**, 622-634. https://doi.org/10.1101/gr.257063.119

[33] Peters, I.J.A., de Pater, E. and Zhang, W. (2023) The Role of GATA2 in Adult Hematopoiesis and Cell Fate Determination. *Frontiers in Cell and Developmental Biology*, **11**, Article 1250827. https://doi.org/10.3389/fcell.2023.1250827

[34] Ichikawa, M., Asai, T., Chiba, S., Kurokawa, M. and Ogawa, S. (2004) Runx1/AML-1 Ranks as a Master Regulator of Adult Hematopoiesis. *Cell Cycle*, **3**, 720-722. https://doi.org/10.4161/cc.3.6.951

[35] Takahashi, S., Komeno, T., Suwabe, N., Yoh, K., Nakajima, O., Nishimura, S., *et al*. (1998) Role of GATA-1 in Proliferation and Differentiation of Definitive Erythroid and Megakaryocytic Cells *in Vivo*. *Blood*, **92**, 434-442. https://doi.org/10.1182/blood.v92.2.434

[36] Levine, J.H., Simonds, E.F., Bendall, S.C., Davis, K.L., Amir, E.D., Tadmor, M.D., *et al*. (2015) Data-Driven Phenotypic Dissection of AML Reveals Progenitor-Like Cells That Correlate with Prognosis. *Cell*, **162**, 184-197. https://doi.org/10.1016/j.cell.2015.05.047

[37] Saygin, C., Hu, E., Zhang, P., Sher, S., Lozanski, A., Doong, T., *et al*. (2021) Genomic Analysis of Cellular Hierarchy in Acute Myeloid Leukemia Using Ultrasensitive LC-FACSEQ. *Leukemia*, **35**, 3406-3420. https://doi.org/10.1038/s41375-021-01295-1

[38] Velten, L., Story, B.A., Hernández-Malmierca, P., Raffel, S., Leonce, D.R., Milbank, J., *et al*. (2021) Identification of Leukemic and Pre-Leukemic Stem Cells by Clonal Tracking from Single-Cell Transcriptomics. *Nature Communications*, **12**, Article No. 1366. https://doi.org/10.1038/s41467-021-21650-1

[39] Fan, H., Wang, F., Zeng, A., Murison, A., Tomczak, K., Hao, D., *et al*. (2023) Single-Cell Chromatin Accessibility Profiling of Acute Myeloid Leukemia Reveals Heterogeneous Lineage Composition Upon Therapy-Resistance. *Communications Biology*, **6**, Article No. 765. https://doi.org/10.1038/s42003-023-05120-6

[40] Bhagwat, A.S., Lu, B. and Vakoc, C.R. (2018) Enhancer Dysfunction in Leukemia. *Blood*, **131**, 1795-1804. https://doi.org/10.1182/blood-2017-11-737379

[41] McKeown, M.R., Corces, M.R., Eaton, M.L., Fiore, C., Lee, E., Lopez, J.T., *et al*. (2017) Superenhancer Analysis Defines Novel Epigenomic Subtypes of Non-Apl AML, Including an RARα Dependency Targetable by SY-1425, a Potent and Selective RARα Agonist. *Cancer Discovery*, **7**, 1136-1153. https://doi.org/10.1158/2159-8290.cd-17-0399

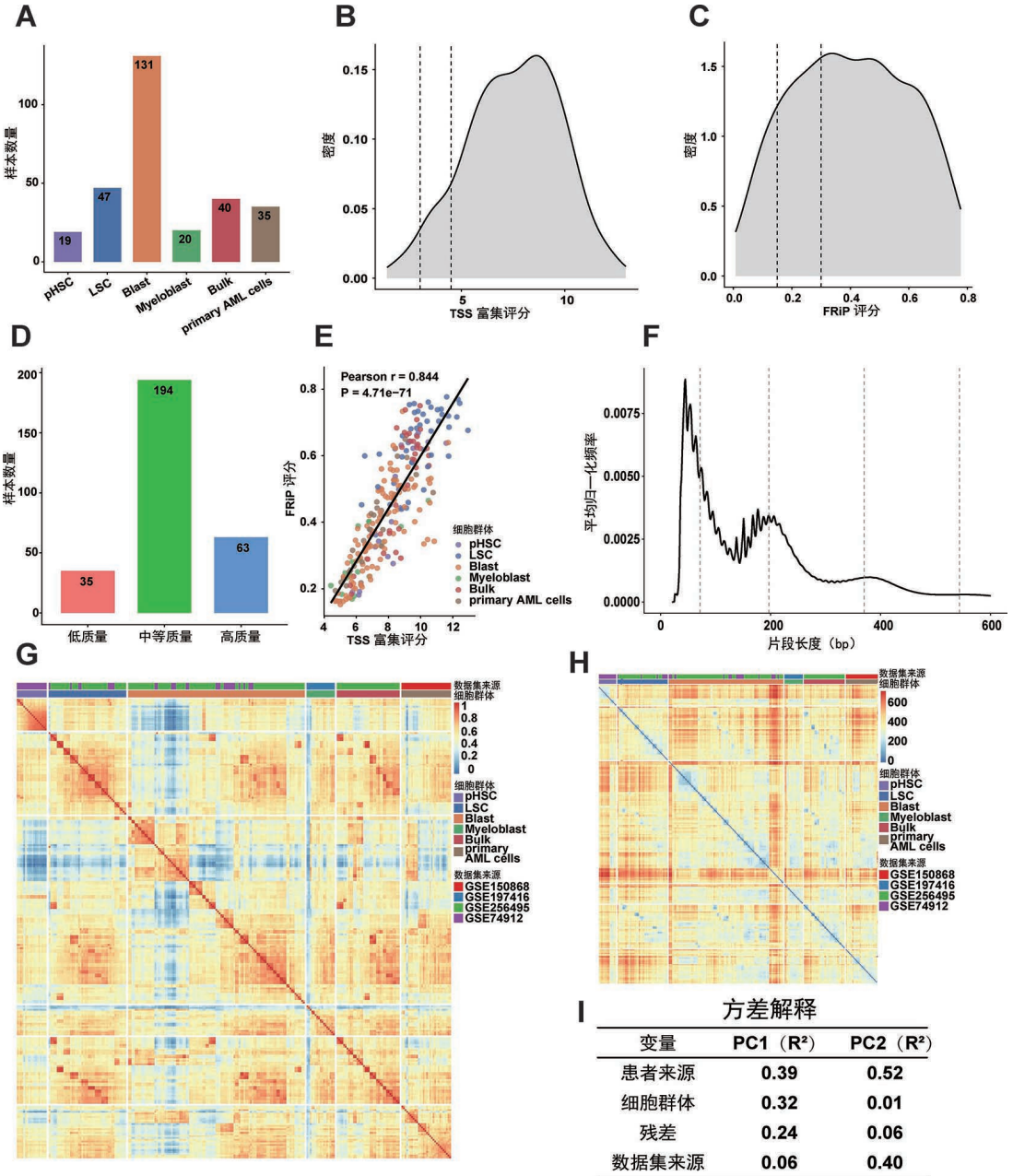
[42] Ediriwickrema, A., Gentles, A.J. and Majeti, R. (2023) Single-Cell Genomics in AML: Extending the Frontiers of AML Research. *Blood*, **141**, 345-355. https://doi.org/10.1182/blood.2021014670

[43] Yu, Y., Zhang, N., Mai, Y., Ren, L., Chen, Q., Cao, Z., *et al*. (2023) Correcting Batch Effects in Large-Scale Multiomics Studies Using a Reference-Material-Based Ratio Method. *Genome Biology*, **24**, Article No. 201.

https://doi.org/10.1186/s13059-023-03047-z

[44] Grandi, F.C., Modi, H., Kampman, L. and Corces, M.R. (2022) Chromatin Accessibility Profiling by ATAC-Seq. *Nature Protocols*, **17**, 1518-1552. https://doi.org/10.1038/s41596-022-00692-9

附　录

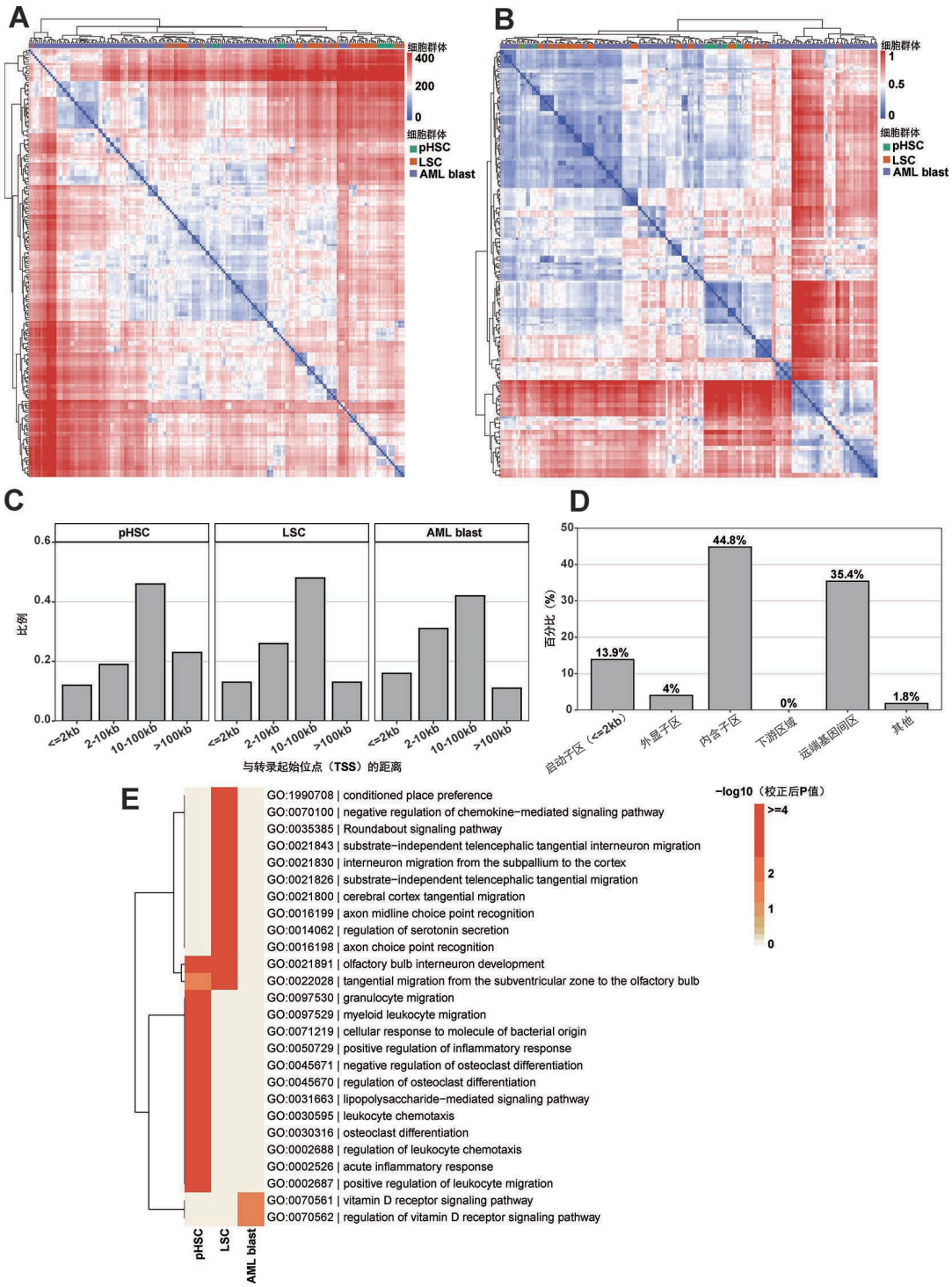


I 方差解释

变量	PC1（R^2）	PC2（R^2）
患者来源	0.39	0.52
细胞群体	0.32	0.01
残差	0.24	0.06
数据集来源	0.06	0.40

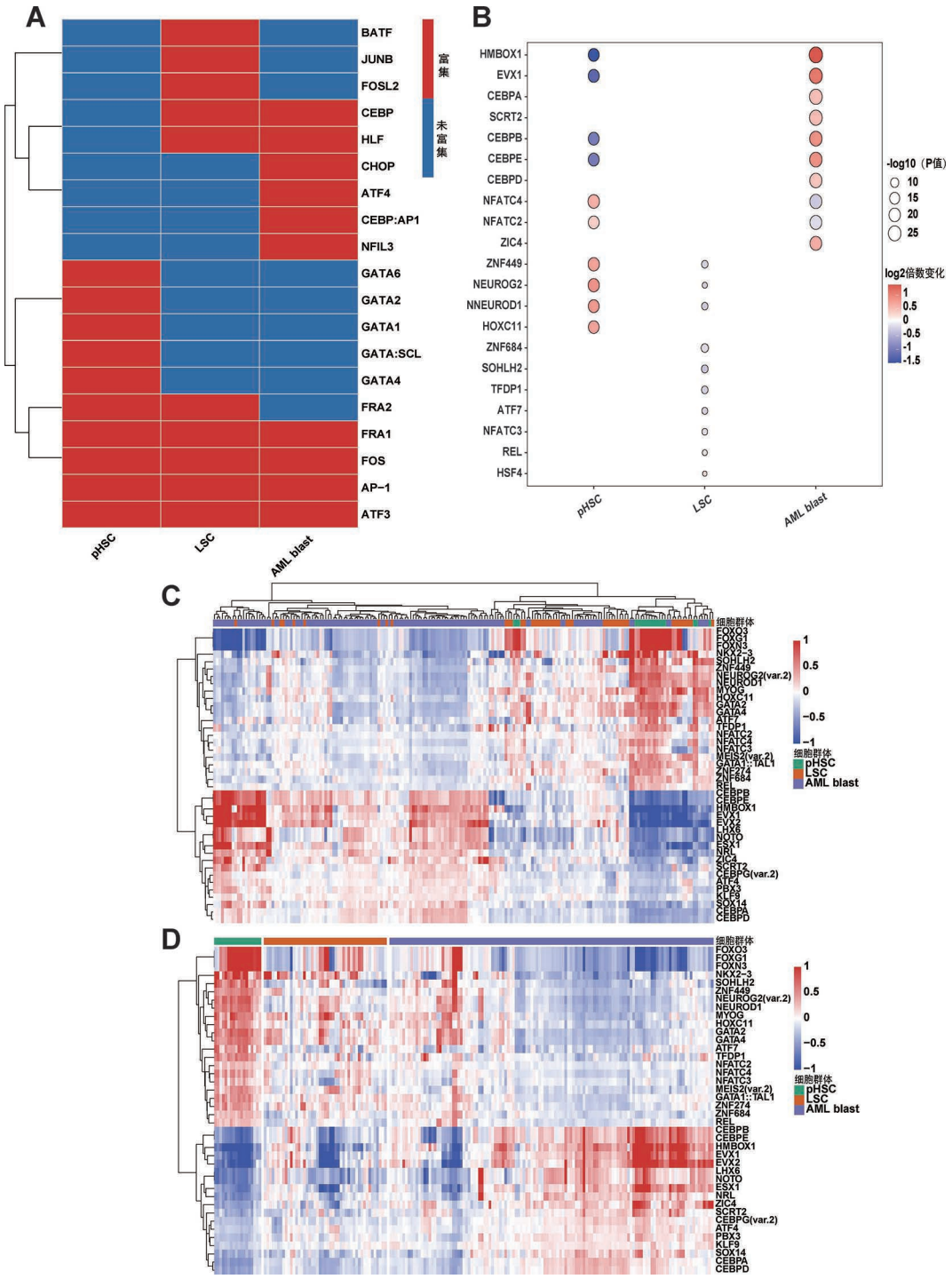
(A) 质量控制前从 pHSC、LSC、blast 组分、myeloblast 组分、Bulk 及原代 AML 细胞群体中收集的 ATAC-seq 样本数量；(B) 所有样本中转录起始位点(transcription start site, TSS)富集评分的分布。虚线表示用于样本质量分类的阈值；(C) 所有样本中峰内 reads 比例(fraction of reads in peaks, FRiP)评分的分布。虚线表示用于样本质量分类的阈值；(D) 基于 TSS 富集度与 FRiP 标准，将样本划分为不合格、高质量及中等质量的数量统计；(E) 样本中 FRiP 与 TSS 富集度之间的关系。每个点代表一个样本，并按细胞群体着色；(F) 所有通过质量控制样本的 ATAC-seq 片段长度分布，显示典型的核小体周期性特征；(G) 基于共识峰矩阵(consensus peak matrix)计算的所有通过质量控制样本的染色质可及性 Pearson 相关性热图，样本排序以突出细胞群体结构。样本按细胞群体及数据集来源标注；(H) 基于与(G)相同样本及特征空间计算的欧氏距离热图。样本按细胞群体及数据集来源标注；(I) 前两个主成分中患者来源、细胞群体、数据集及残差效应所解释的方差比例。

Figure S1. Quality assessment and global variance structure of the integrated ATAC-seq dataset
图 S1. 整合 ATAC-seq 数据集的质量评估及全局变异结构



(A) 基于高变异共识峰(consensus peaks)计算的 pHSC、LSC 及 AML 原始细胞(AML blast)样本的欧氏距离热图，采用无监督聚类。样本按细胞群体标注；(B) 基于与(A)相同样本及特征空间计算的欧氏距离热图，样本按细胞群体排序，以突出细胞群体相关结构；(C) 各细胞群体中差异可及区域到最近转录起始位点(transcription start site, TSS)的距离分布；(D) 三类细胞群体中差异可及区域的基因组注释。大多数区域分布于内含子区及远端基因间区；(E) 各细胞群体中可及性降低区域的 Gene Ontology (GO)生物学过程富集分析。颜色表示富集显著性。

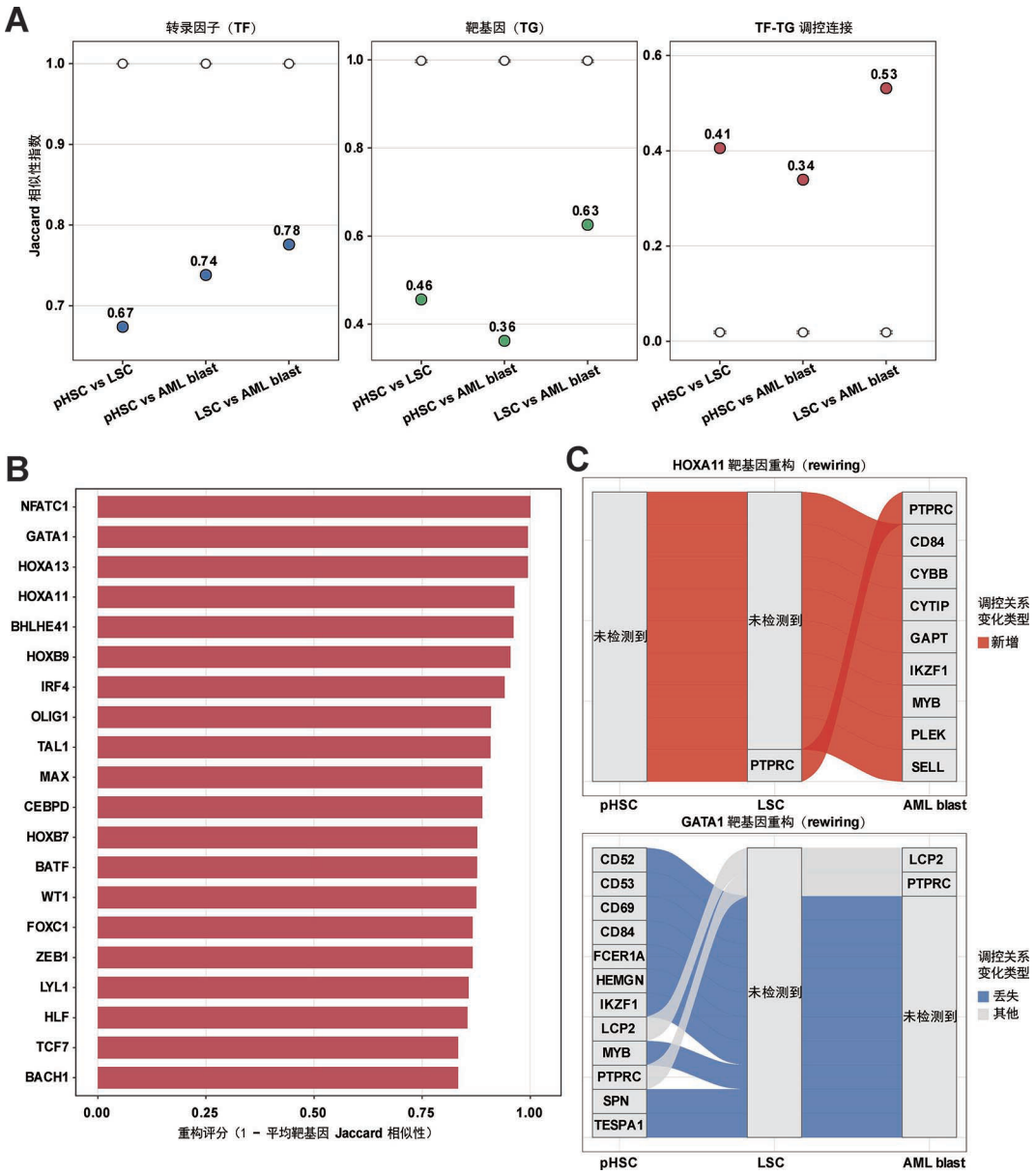
Figure S2. Differentially accessible regions across AML cell populations are enriched at distal regulatory elements
图 S2. 整合 AML 细胞群体间差异可及区域富集于远端调控元件



(A) 基于 HOMER 分析，在 pHSC、LSC 及 AML 原始细胞(AML blast)群体中富集的代表性转录因子(transcription factor, TF) motif 的二元热图(binary heatmap)；(B) 基于 chromVAR 推断的 pHSC、LSC 及 AML blast 群体之间差异 motif 活性的气泡图。点大小表示统计显著性，颜色表示 motif 相关可及性偏差的方向及幅度；(C) 基于代表性 motif 的 chromVAR 偏差得分(deviation score)在样本中的无监督聚类热图。样本按细胞群体标注；(D) 与(C)相同的 chromVAR 偏差得分矩阵热图，但样本按细胞群体排序，以突出细胞群体相关的 motif 活性模式。

Figure S3. Independent motif-based analyses support population-biased models of transcription factor activity across AML cell populations

图 S3. 基于 motif 的独立分析支持 AML 细胞群体间转录因子活性的群体偏倚模式



(A) pHSC、LSC 及 AML 原始细胞(AML blast)调控网络中转录因子(transcription factor, TF)、靶基因(target gene, TG)及 TF-TG 调控连接的两两 Jaccard 重叠。实心点表示实际观测值，空心点表示在规模匹配的随机网络中的期望值；(B) 基于重构评分(rewiring score)对转录因子进行排序。该评分定义为其在不同细胞群体中靶基因集合 Jaccard 相似性的平均值的补数，评分越高表示对应转录因子在调控网络中的靶基因连接重构程度越高；(C) Sankey 图展示代表性转录因子在不同细胞群体中的 TF-TG 连接重构。不同颜色的流表示推断 TF-TG 调控关系中的获得或丢失。

Figure S4. Quantitative analyses further support extensive inferred regulatory network rewiring across AML cell populations
图 S4. 定量分析进一步支持 AML 细胞群体间广泛的推断调控网络重构