Advances in Applied Mathematics 应用数学进展, 2021, 10(10), 3518-3531
Published Online October 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.1010371

Hans汉斯

基于马尔可夫随机场识别急性髓系白血病核心基因

许凌云[1]，王艺舒[2]，宋　凯[1]

[1]青岛大学数学与统计学院，山东　青岛
[2]北京科技大学数理学院，北京

收稿日期：2021年9月25日；录用日期：2021年10月18日；发布日期：2021年10月27日

摘　要

在大量的候选基因中筛选出关联急性髓系白血病(AML)的核心基因对于其治疗具有重要作用，鉴于肿瘤细胞高度异质性的特点，单细胞测序数据为AML生物标志物的筛选提供了可能。本研究基于马尔科夫随机场和高斯混合模型整合DNA甲基化和单细胞测序数据，将正常组和疾病组间基因互作网络的差异重连信息纳入马尔科夫随机场框架，通过重连网络和蛋白互作网络鉴定出CD86、TNF、GRAP2、FGFR1、IL18五个免疫相关的预后基因。基因GRAP2、FGFR1、TNF的高表达与预后较好有关，而DNA甲基化水平与预后较差有关，基因CD86、IL18的高表达与预后较差有关，而DNA甲基化水平与预后较好有关，CD86、TNF、GRAP2、FGFR1、IL18表达水平与DNA甲基化水平呈负相关；基因CD86表达水平与Monocytes、Macrophages M2免疫浸润水平显著正相关，与Mast cells resting、naïve B cell、naïve CD4 T cell、CD8 T cell、NK cell resting免疫浸润显著负相关，基因TNF、GRAP2、FGFR1反之。基因CD86、TNF、GRAP2、FGFR1、IL18可以作为AML发生、发展及免疫治疗中的预后生物标志物，为进一步研究提供依据。

关键词

马尔科夫随机场，差异网络，DNA甲基化，单细胞测序，急性髓系白血病

Identifying Core Genes of Acute Myeloid Leukemia Based on Markov Random Field

Lingyun Xu[1], Yishu Wang[2], Kai Song[1]

[1]School of Mathematics and Statistics, Qingdao University, Qingdao Shandong
[2]School of Mathematics and Physics, University of Science and Technology Beijing, Beijing

Received: Sep. 25th, 2021; accepted: Oct. 18th, 2021; published: Oct. 27th, 2021

文章引用: 许凌云, 王艺舒, 宋凯. 基于马尔可夫随机场识别急性髓系白血病核心基因[J]. 应用数学进展, 2021, 10(10): 3518-3531. DOI: 10.12677/aam.2021.1010371

Abstract

Screening out the core genes associated with acute myeloid leukemia (AML) from a large number of candidate genes plays an important role in its treatment. In view of the high heterogeneity of tumor cells, single-cell sequencing data provides the possibility to screen for AML biomarkers. The research is based on Markov random field and Gaussian mixture model to integrate DNA methylation and single-cell sequencing data, and incorporates the difference reconnection information of the gene interaction network between the normal group and the disease group into the Markov random field framework, through the reconnection network and protein interaction network identified five immune-related prognostic genes CD86, TNF, GRAP2, FGFR1, and IL18. High expression of genes GRAP2, FGFR1, TNF is related to better prognosis, while DNA methylation level is related to poor prognosis, high expression of genes CD86 and IL18 is related to poor prognosis, and DNA methylation level is related to better prognosis, CD86, TNF, GRAP2, FGFR1, IL18 expression levels are negatively correlated with DNA methylation levels; Gene CD86 expression level is significantly positively correlated with Monocytes and Macrophages M2 immune infiltration levels, and significantly negatively correlated with Mast cells resting, naïve B cell, naïve CD4 T cell, CD8 T cell, and NK cell resting immune infiltration levels. Genes TNF, GRAP2, FGFR1 are on the contrary. Genes CD86, TNF, GRAP2, FGFR1, IL18 can be used as prognostic biomarkers in the occurrence, development and immunotherapy of AML, providing a basis for further research.

Keywords

Markov Random Field, Differential Network, DNA Methylation, Single Cell Sequencing, Acute Myeloid Leukemia

1. 引言

急性髓系白血病(AML)是一种造血克隆性恶性肿瘤，其特征是造血干细胞和祖细胞不受控制地增殖，而不能分化为成熟细胞[1]。该病约占小儿白血病的30%和成人急性白血病的80%，具有发病快、病情发展迅速、控制难、易复发、预后差和发病率随着年龄的增加而上升等特征[2]，对人类健康造成严重危害。目前，AML通过化疗的方式可以达到一定疗效，但是仍有部分患者无法获得缓解，预后较差。因此，在大量的候选基因中筛选出关联疾病的预后核心基因对进一步提高预后的准确性评估具有举足轻重的意义与价值。

在过去的几十年中，人们基于传统意义上的bulk转录组对AML有了大量的研究，这有助于识别参与AML癌变和进展的关键基因及重要通道和功能。但是，肿瘤是一个复杂的恶性肿瘤、免疫细胞和基质细胞的混合体，具有瘤内异质性和瘤间异质性，整个肿瘤微环境中包括了促进肿瘤和抑制肿瘤的各种信号来调控肿瘤的生长和进化，bulk表达数据会将一些特异的细胞群体或细胞状态展示的信号掩盖，而这些特异性细胞群体在肿瘤的发生与发展中极为关键，除此之外，多种组学数据已经被应用于肿瘤的研究中，如单细胞技术层面的mRNA-蛋白质[3]、mRNA-染色质可及性[4]、mRNA-基因组[5]等，而如何整合各种不同的组学数据，发展相关的统计学方法，鉴定癌症预后基因，是一个亟需解决的问题。对此，

本文从单细胞的视角来探究 AML 的发展过程，有研究表明肿瘤的异质性主要体现在不同细胞类中表达水平的差异，而这种差异受表观因素的影响，如 DNA 甲基化、非编码 RNA、组蛋白修饰和染色质结构等，DNA 甲基化作为一种不仅影响基因表达而且可遗传又可逆的表观修饰，在癌症的形成过程中扮演着重要的作用。

马尔可夫随机场(MRF)模型已广泛用于图像分析，以解释观察到的像素强度的局部依赖性[6]并且还被应用于蛋白质的功能预测，通过蛋白质－蛋白质相互作用网络解释蛋白质功能的局部依赖性[7]，MRF 模型在基因组学也已应用于从蛋白质相互作用和基因表达数据中发现分子途径[8]，在 GWAS 研究中 MRF 模型还可以通过整合生物通道先验知识识别出潜在疾病基因[9]。在以往的大量研究中 MRF 模型已经被用于不同类型数据的整合，如：用于基因表达数据和 CHIP 芯片数据的空间正态混合模型[10]，用于 mRNA 微阵列数据的 Gamma-Gamma 模型和 MRF [11]，通过结合基因表达和蛋白质相互作用数据对基因进行优先排序[12]。在本文中，我们基于 MRF 和高斯混合模型整合 DNA 甲基化和单细胞测序数据，通过基因间的相互作用网络实现基因疾病关联状态的预测，DNA 甲基化信号值衡量不同网络下同一对基因间的重连系数，重连系数作为边的权重构造正常疾病两种状态下加权共表达先验网络 $p(X)$，不同细胞分类下单细胞 RNA 差异表达 $y_i=\Phi^{-1}(1-p_i)$ 服从高斯混合分布 $p(Y|X)$ 为似然函数，相比与同一网络中和疾病关联基因的邻近基因更可能和疾病相关，动态网络更能反映在不同表型下基因间相互作用的方式发生了怎样的改变，有助于我们更加精准地识别疾病关联基因。我们的模型主要有以下三点优势：1) 同时考虑到肿瘤异质性特点和表观影响因素；2) MRF 网络结构构造的动态网络很好地呈现不同表型下基因互作方式；3) MRF 具有良好的计算框架。

2. 材料与方法

2.1. 加权状态网络构建

2.1.1. 基因共表达网络

基因—基因交互作用可以通过一个无向网络图 $G=(v,\varepsilon)$ 来呈现，网络中的每一个点代表基因，记该点集为 $v=\{i:i=1,\cdots,n\}$，点之间的连线(边)表示两个基因间的相关关系，记所有边的集合为 $\varepsilon=\{(i,i'):i\sim i'\}$，如果基因 i 和基因 i'存在某种已知的关系，例如：共同调控同一生物过程，记 $i\sim i'$，基因间相互作用的强弱由皮尔森相关系数衡量，ρ_{ij}、ρ'_{ij} 分别为疾病组和正常组下基因 i 和基因 j 的皮尔森相关系数，x_i 表示基因的疾病关联状态，$X=(x_1,\cdots,x_n)$ 为所有基因的关联标签集。

$$x_i=\begin{cases}x_i=1, \text{ 基因}i\text{与疾病相关联}\\ x_i=0, \text{ 基因}i\text{与疾病不相联}\end{cases}$$

每一个基因有 1 或 0 两种可能标签，因此在所有基因关联状态都被考虑的情况下总共有 2^n 个特征网络。

2.1.2. 差异重连系数

Fisher 变换是用于相关系数假设检验的方法，对共表达网络中基因间的相关系数执行 Fisher 变换 $F(\rho)=\frac{1}{2}\ln\frac{1+\rho}{1-\rho}$，在推断基因是否发生重连问题中更具有统计力[13]。原假设疾病—对照组基因间的 Pearson 相关系数不存在差异，在此假设问题下 $F(\rho)$ 服从均值 $\frac{1}{2}\ln\frac{1+\rho}{1-\rho}$ 标准差 $\frac{1}{\sqrt{N-3}}$ 的正态分布，

$$F(\rho)\sim N\left(\frac{1}{2}\ln\frac{1+\rho}{1-\rho},\frac{1}{N-3}\right) \tag{1}$$

经 Fisher 变换后的重连系数满足以下概率密度：

$$DR_{ij}=p\left(\left|X\right|\le\left|\frac{F\left(\rho_{ij}\right)-F\left(\rho'_{ij}\right)}{\sqrt{\frac{1}{N_1-3}+\frac{1}{N_2-3}}}\right|\right),\ X\sim N\left(0,1\right) \tag{2}$$

其中 N_1， N_2 为疾病组和对照组基因数。

2.2. 统计模型

2.2.1. 单细胞测序数据的高斯混合模型

单细胞测序数据经预处理后，我们选取 $i\times j$ 差异表达基因的单细胞测序表达矩阵，其中 $i\left(i=1,\cdots,n\right)$ 表示基因， $j\left(j=1,\cdots,m\right)$ 表示样本(单个细胞)，在本文中，我们考虑两种状态下不同细胞分类间基因表达的差异， $y_i=\Phi^{-1}\left(1-p_i\right)$ 对应于单细胞分类下差异 p 值的标准正态转化， Φ 是标准正态累积分布函数，假设 y_i 在给定的关联状态下条件独立， y_i 在两种关联状态下服从高斯混合分布[14]：

$$\begin{aligned}&y_i\left|x_i=0\sim N\left(0,1\right)\right.\\&y_i\left|x_i=1\sim N\left(\mu_i,\sigma_i^2\right)\right.\end{aligned} \tag{3}$$

其中， μ_i 和 σ_i^2 共轭先验为：

$$\begin{aligned}&\mu_i\left|\sigma_i^2\sim N\left(\bar{\mu},\frac{\sigma_i^2}{a}\right)\right.\\&\sigma_i^2\sim\text{Inverse Gamma}\left(\frac{\nu}{2},\frac{\nu d}{2}\right)\end{aligned} \tag{4}$$

y_i 概率密度函数可表示成：

$$f\left(y_i\left|x_i\right.\right)=\left(1-x_i\right)f_0\left(y_i\right)+x_i f_1\left(y_i\right) \tag{5}$$

$$f\left(Y\left|X\right.\right)=\prod_i f\left(y_i\left|x_i\right.\right) \tag{6}$$

其中 $f_0\left(y_i\right)$ 是零密度函数，与疾病没有关联的基因差异表达 p 值满足标准正态分布概率密度函数； $f_1\left(y_i\right)$ 是备择密度函数，与疾病相关的基因差异表达 p 值所满足均值 μ_i 、方差 σ_i^2 的正态分布。如何推断隐变量 x_i 状态是我们最为关注的问题，在下文中，我们构建了能够纳入基因间相互作用的马尔科夫随机场框架进而推断隐变量 x_i 状态。

2.2.2. 马尔科夫随机场建模先验概率

鉴于马尔科夫随机场自身所具有的两个特性：一方面，它包含网络结构，该结构可以解释长距离的依赖关系，因此能够将基因间相互作用关系纳入依存概率衡量的动态网络中，对整个研究问题而言，不仅考虑了基因－基因之间某种特定的关联，而且可以捕获疾病—对照间网络重构的差异特征，进一步识别更多关联状态；另一方面，可通过马尔可夫随机场局部结构特点建立良好的计算框架。因此，利用马尔可夫随机场定义 X 的先验概率图，首先，我们基于这样一个认识：网络中相邻连接的基因往往有相同的关联状态，那么在这 2^n 个可能会出现的网络中我们需要确定出网络中具有相同关联状态的基因连在一起的可能性。其次，我们不再单单局限于这一个网络，结合临床因素将网络二分成疾病和正常两个状态下的共表达网络[15]，假设疾病 VS 正常异常发生重连的基因更大可能地关联疾病，DNA 甲基化数据计算不同状态下同一对基因相关系数差异的显著性水平 p 值，若 $p\ge0.05$ 则不显著基因没有发生重连，若

$p<0.05$ 则基因发生重连且 p 值定义为该基因在不同状态下的差异重连度 DR_{ij}， DR_{ij} 以边的权重作为计算差异网络后验概率的另一个重要特征。在这，我们采用最邻近 Gibbs 抽样得到如下形式:

$$p\left(X\middle|\Phi\right)\propto\exp\left(-\alpha\sum_{i=1}^{n}I\left(x_i=1\right)+\beta_1\sum_{e_{ij}}DR_{ij}I\left(x_i=1\right)I\left(x_i'=1\right)-\beta_2\sum\nolimits_{e_{ij},DR_{ij}>\delta}DR_{ij}I\left(x_i=-1\right)I\left(x_i'=-1\right)\right)\tag{7}$$

(α,β_1,β_2) 是马尔可夫网络参数，先验概率 $p(X)$ 由三部分组成：1) 没有与任何基因相连独立于基因网络的单个基因，该类基因不伴随有网络信息，这些基因在关联疾病的情况下对 $p(X)$ 的影响程度为 α；2) 直接相连的基因均关联疾病，该类基因对 $p(X)$ 贡献程度记为 β_1； 3) 直接相连的基因不关联疾病，该类基因对 $p(X)$ 贡献程度为 β_2。由整个网络的联合概率密度函数可以得出某个基因 i 关联疾病的条件概率密度：

$$\frac{p\left(x_i=1\middle|X/x_i;\Phi\right)}{p\left(x_i=-1\middle|X/x_i;\Phi\right)}=\frac{\exp\left(F\left(DR_{ij};\Phi\right)\right)}{1+\exp\left(F\left(DR_{ij};\Phi\right)\right)}$$

其中，

$$F\left(DR_{ij};\Phi\right)=h+\beta_1\sum\nolimits_{e_{ij}}DR_{ij}-\beta_2\sum\nolimits_{e_{ij},DR_{ij}>\delta}DR_{ij}\tag{8}$$

2.2.3. 模型参数和后验概率估计

模型中存在两个超参数集：马尔科夫随机场中的网络参数 $\Phi=(\alpha,\beta_1,\beta_2)$、高斯混合模型参数 $\Theta=(\bar{\mu},a,\nu,d)$，为了估计这两个参数集 Φ 、 Θ，我们采用最初为解决统计图形问题由 Besag [6]提出的一种算法—ICM 算法，该算法采用下列步骤对参数和隐变量进行迭代更新：

1) 在不考虑重连信息的前提下，通过抽样高斯混合模型得到隐变量的初始估计值 $\hat{X}$

2) 最大化观测数据的概率密度函数 $f\left(Y\middle|\hat{X}\right)$，求得高斯混合模型的参数 Θ 的估计值 $\hat{\Theta}$

3) 通过最大化下列伪似然函数求得网络参数 Φ 的估计值 $\hat{\Phi}$

$$\begin{aligned}l\left(\hat{X};\Phi\right)&=\prod\nolimits_{i=1}^{n}p_i\left(\hat{X}_i\middle|\hat{X}_{N_i};\Phi\right)\\&=\frac{\exp\left(-\alpha I\left(\hat{x}_i=1\right)\beta_1\sum_{e_{ij}}DR_{ij}I\left(\hat{x}_i=1\right)I\left(\hat{x}_i'=1\right)-\beta_2\sum_{e_{ij},DR_{ij}>\delta}DR_{ij}I\left(\hat{x}_i=-1\right)I\left(\hat{x}_i'=-1\right)\right)}{\exp\left(-\alpha+\beta_1\sum_{e_{ij}}DR_{ij}\right)I\left(\hat{x}_i'=1\right)-\exp\left(\beta_2\sum_{e_{ij},DR_{ij}>\delta}DR_{ij}\right)I\left(\hat{x}_i'=-1\right)}\end{aligned}\tag{9}$$

4) 基于两组参数和隐变量的初始值 $\hat{\Theta}$ 、 $\hat{\Phi}$ 、 $\hat{X}$ 执行 ICM 算法循环更新疾病关联标签 $\hat{X}$

特别地，对于 $i(i=1,\cdots,n)$， X_i 基于下式更新：

$$p\left(X_i\middle|Y,X_{I/i}\right)\propto f\left(Y_i\middle|X_i;\hat{\Theta}\right)p_i\left(X_i\middle|\hat{X}_{N_i};\hat{\Phi}\right)\tag{10}$$

5) 继续到第二步直到 $\max_{\theta\in(\Theta,\Phi)}\dfrac{\left|\theta^{k+1}-\theta^{k}\right|}{\left|\theta^{k+1}\right|}$

在算法达到收敛后，ICM 算法得到的参数估计值作为模型中的参数，基于(10)式采用 Gibbs 抽样抽取 M 次潜在变量，进而估计隐变量的后验概率 $p_i=p\left(x_i=0\middle|Y\right)$， $p_{(i)}$ 是 p_i 的升序排列，在是否关联疾病的推断问题中采用基于后验概率的 FDR [16]定义。

H_{i0}：基因 i 与疾病不关联；

H_{i1}：基因 i 与疾病相关联。

基于后验概率当 $k=\max\left\{m:\frac{1}{m}\sum_{i=1}^{m}p_i\le\alpha\right\}$ 时，该研究中设置 $\alpha=0.05$，拒绝所有原假设 $H_{(i)}$，$i=1,\cdots,k$。

2.3. 免疫相关的重连基因预后评估

在免疫相关的重连基因中，为了验证基因表达水平和甲基化水平的独立预后价值，通过 Kaplan-Meier 生存曲线和 TCGA 队列的对数秩检验评估基因转录表达和 DNA 甲基化对总生存时间(OS)的影响，学生 t 检验和多重假设校正(错误发现率，FDR)用于识别基因表达水平及甲基化水平高低对预后生存的差异，所有分析均使用 R4.0.3 软件包进行。急性髓系白血病患者 OS 相关的基因在 UCSC Xena 数据库(GDC TCGA Acute Myeloid Leukemia (LAML))中得以验证，UCSC Xena 数据库中的基因表达 RNAseq 和 DNA 甲基化评估了表达及甲基化对 OS 的潜在影响。

2.4. 急性髓系白血病数据

急性髓系白血病模型中的甲基化和单细胞表达数据集在 GEO 数据库下载，对应编号为 GSE116256、GSE58477，基因集验证中所用数据来源于 UCSC Xena 数据库，包含 DNA 甲基化、基因表达及临床数据。

3. 结果

3.1. 单细胞测序和甲基化数据预处理

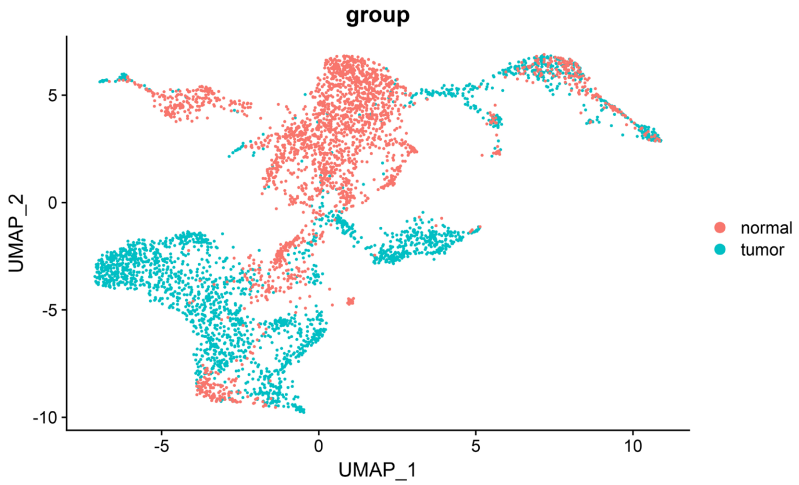
GEO 数据库下载急性髓系白血病单细胞测序数据(GSE116256)和甲基化数据(GSE58477)。在单细胞测序数据集(GSE116256)中，获取 4 个急性髓系白血病患者和 4 个与之匹配健康供体的 4941 个细胞测序样本，每个细胞测序深度为 27,899 个基因，形成 27,899*4941 的单细胞表达矩阵。首先，使用 Seurat 包[17]对表达矩阵进行数据标准化、归一化、线性降维，其次，对细胞聚类分型、非线性降维及细胞类型注释如图 1(a)，图 1(b)，最后，得到不同细胞簇类间 16,308 个差异表达基因。在甲基化数据集(GSE58477)中，由 Illumina HumanMethylation450K 芯片平台取样具有 485,512DNA cg 位点的 62 个急性髓系白血病患者和 10 个与之匹配健康供体。首先，使用 ChAMP 包[18]对甲基化信号矩阵质控、标准化、归一化处理，然后，探针注释且对对应于一个探针的多个 cg 位点甲基化水平取均值，最后，正常组 VS 疾病组差异分析，在 $adj.p.Val<0.05$ 且 $|logFC|>0.1$ 的条件下筛选出具有唯一 symbol 的 6825 个差异甲基化基因。

3.2. 模型在急性髓系白血病数据中的应用

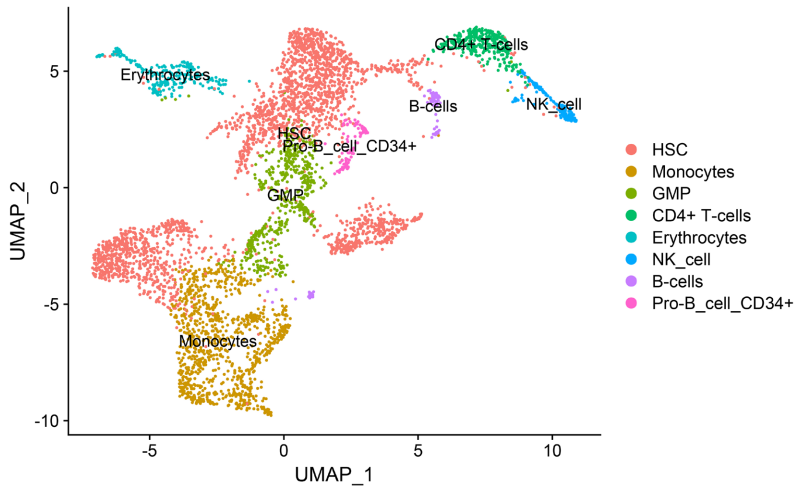
基于单细胞测序和甲基化数据集，将差异表达和差异甲基化的 3603 个交集基因纳入差异网络中，甲基化信号矩阵根据样本表型分成正常组和疾病组两个矩阵，两个状态下交集基因间相关性矩阵 ρ、ρ'，相关系数经费希尔转化及差异重连系数的定义进而衡量出不同状态下网络中基因互相作用的差异，重连特征作为权重确定先验信息，根据马尔可夫随机场框架计算出网络中每一个基因关联疾病的后验概率，将后验概率从小到大排序，基于后验概率定义 FDR 选取使得后验概率平均值小于显著性水平的最大的前 K 个基因，最后从差异网络中选出显著关联疾病的 1320 个优先级基因。图 2 展示了这些基因的功能富集(GO)和通道富集(KEGG)，通过功能富集分析($p<0.05$)可以发现大多数基因参与细胞间通讯及分子功能调节、细胞内信号传导及细胞表面受体信号通路、免疫反应生物过程，细胞组成主要为膜结合囊泡、细胞外囊泡、细胞外器、胞外外泌体，分子功能表现为大分子复合物结合、酶结合、蛋白质复合物结合等，通路富集分析表明基因富集于癌症相关途径、转录失调、趋化因子信号通路、T 细胞受体信号通路。

3.3. 蛋白互作网络构建及生存相关免疫特征筛选

图 3(a)展示了 string 数据库中差异网络优先级基因通过 cytoscape 构建的蛋白互作网络，cytohubba 筛选出蛋白互作网络 100 个基因。ImmPort 数据库获得了免疫相关基因(IRG)列表，将免疫相关基因与差异网络优先级基因取交集，共有 177 个发生重连的免疫相关基因。为了探索基因的预后特征，对这些基因做批量生存分析，选取与预后显著相关的 IRG50 个基因，图 3(b)，图 3(c)展示了预后显著相关的 IRG 基因表达水平和甲基化水平热图结果。蛋白互作网络中核心基因与预后显著相关的 IRG 取交集筛选出 5 个具有预后特征的生物标志物，这五个预后相关的免疫核心基因是 CD86、TNF、GRAP2、FGFR1、IL18，从 TCGA 数据库下载急性髓系白血病的表达数据、甲基化数据及临床特征验证了核心基因预后相关性，图 4 展示了五个预后相关的免疫核心基因生存分析图，通过生存分析结果可以看出基因 GRAP2、FGFR1、TNF 的高表达与预后较好有关，而 DNA 甲基化水平与预后较差有关，基因 CD86、IL18 的高表达预后较差有关，而 DNA 甲基化水平与预后较好有关，CD86、TNF、GRAP2、FGFR1、IL18 表达水平与 DNA 甲基化水平呈负相关。

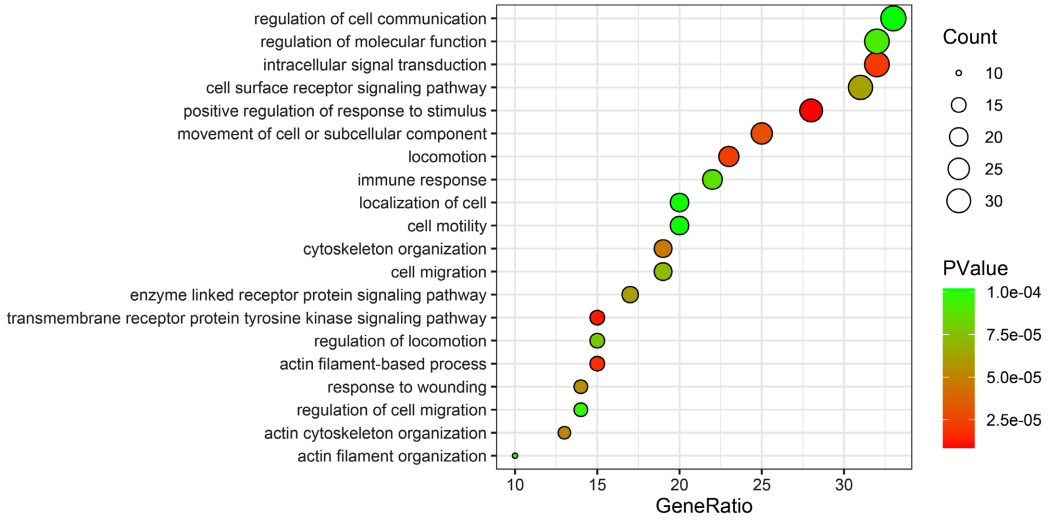


(a)

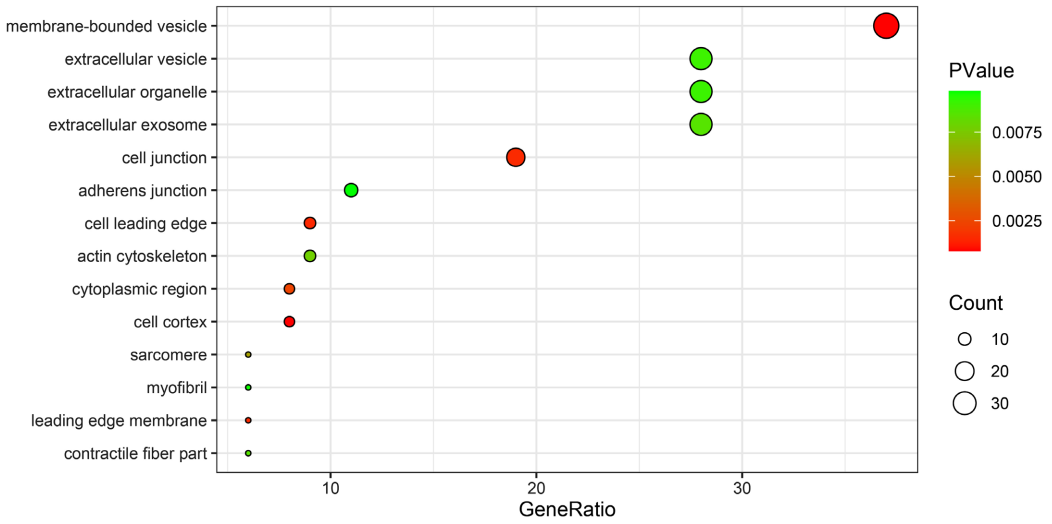


(b)

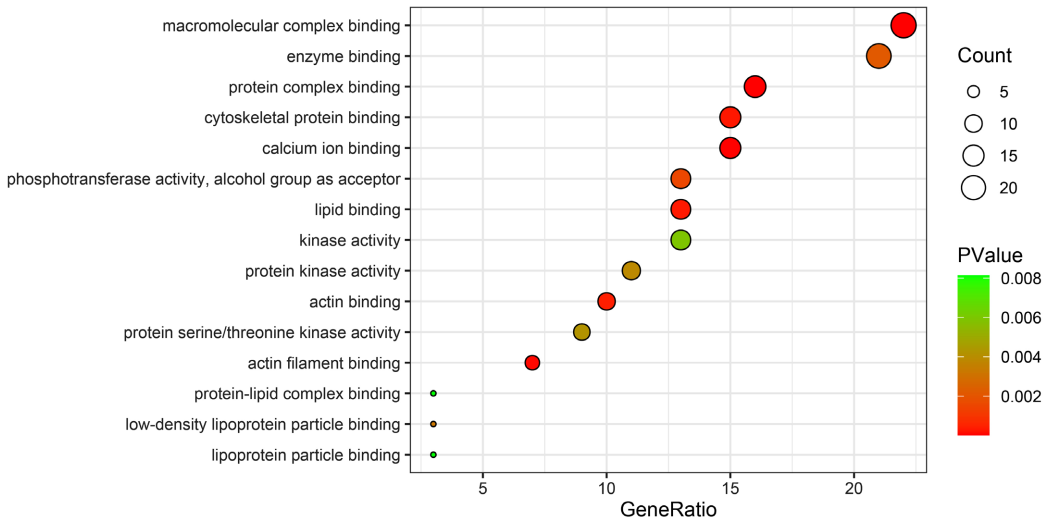
Figure 1. (a) Cell classification and distribution of samples in different states; (b) Cell cluster annotation
图 1. (a) 细胞分类及不同状态下样本的分布；(b) 细胞簇类注释


regulation of cell communication
regulation of molecular function
intracellular signal transduction
cell surface receptor signaling pathway
positive regulation of response to stimulus
movement of cell or subcellular component
locomotion
immune response
localization of cell
cell motility
cytoskeleton organization
cell migration
enzyme linked receptor protein signaling pathway
transmembrane receptor protein tyrosine kinase signaling pathway
regulation of locomotion
actin filament-based process
response to wounding
regulation of cell migration
actin cytoskeleton organization
actin filament organization
10
15
20
25
30
GeneRatio
Count
10
15
20
25
30
PValue
1.0e-04
7.5e-05
5.0e-05
2.5e-05

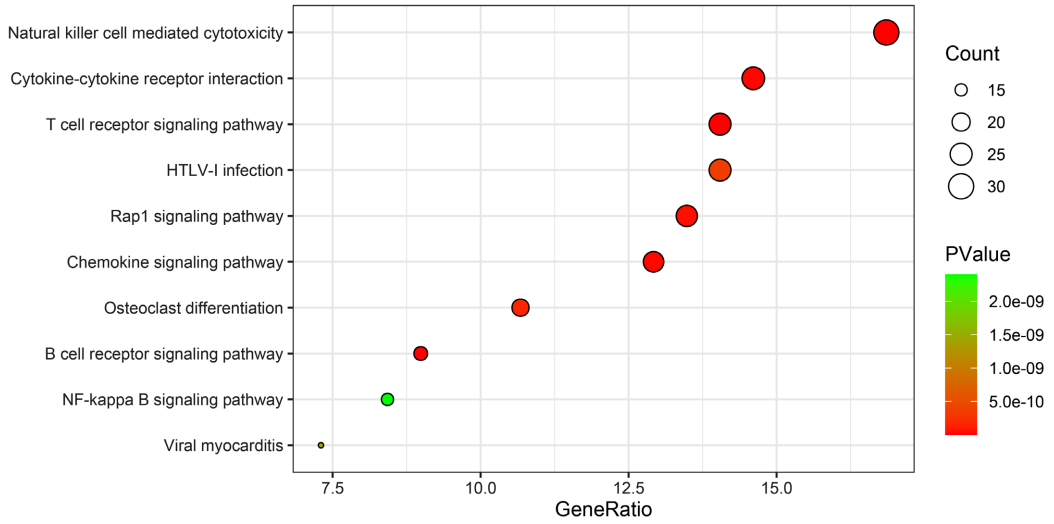
(a)


membrane-bounded vesicle
extracellular vesicle
extracellular organelle
extracellular exosome
cell junction
adherens junction
cell leading edge
actin cytoskeleton
cytoplasmic region
cell cortex
sarcomere
myofibril
leading edge membrane
contractile fiber part
10
20
30
GeneRatio
PValue
0.0075
0.0050
0.0025
Count
10
20
30

(b)

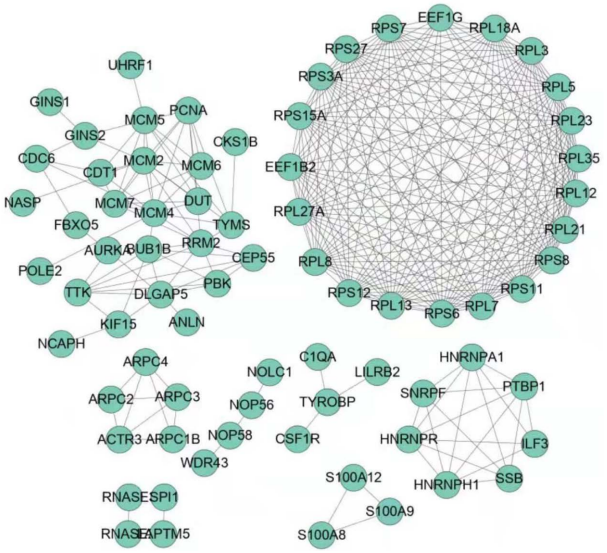

macromolecular complex binding
enzyme binding
protein complex binding
cytoskeletal protein binding
calcium ion binding
phosphotransferase activity, alcohol group as acceptor
lipid binding
kinase activity
protein kinase activity
actin binding
protein serine/threonine kinase activity
actin filament binding
protein-lipid complex binding
low-density lipoprotein particle binding
lipoprotein particle binding
5
10
15
20
GeneRatio
Count
5
10
15
20
PValue
0.008
0.006
0.004
0.002

(c)

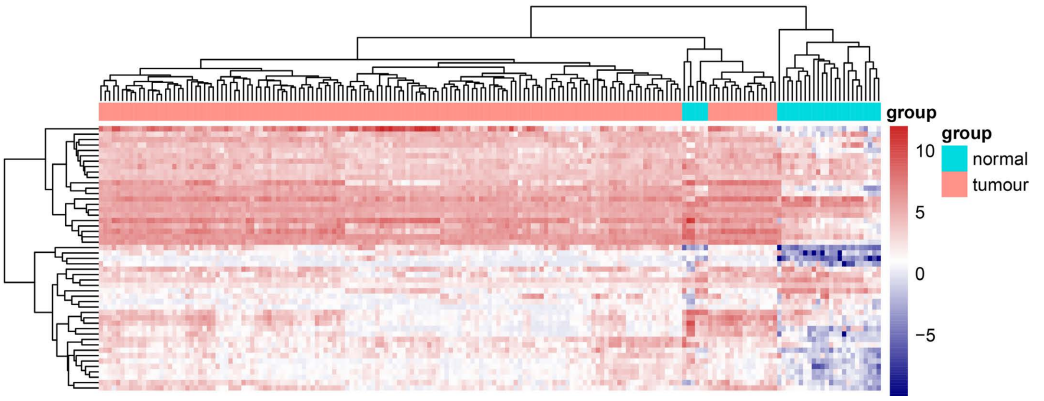


(d)

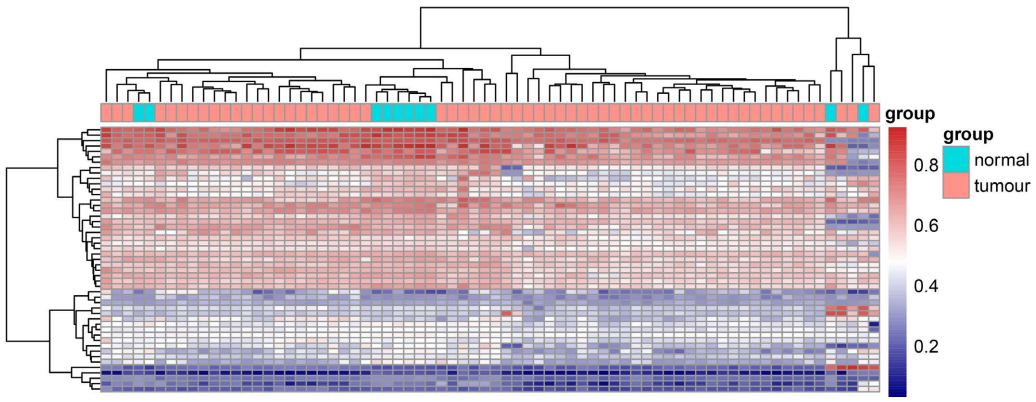
Figure 2. Network priority gene GO, KEGG enrichment analysis (a) BP; (b) CC; (c) MF; (d) Pathway enrichment
图 2. 网络优先级基因 GO、KEGG 富集分析 (a) 生物过程；(b) 细胞组成；(c) 分子功能；(d) 通路富集



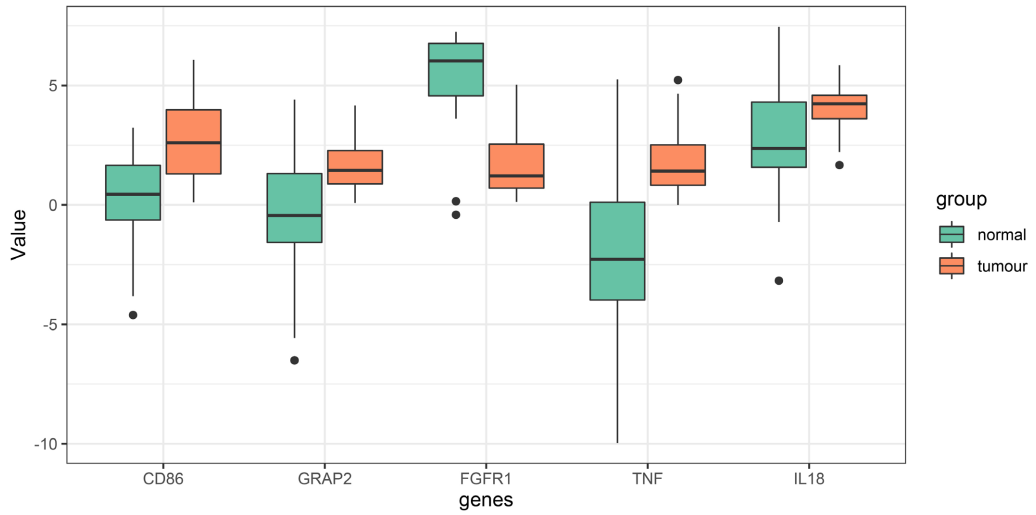
(a)



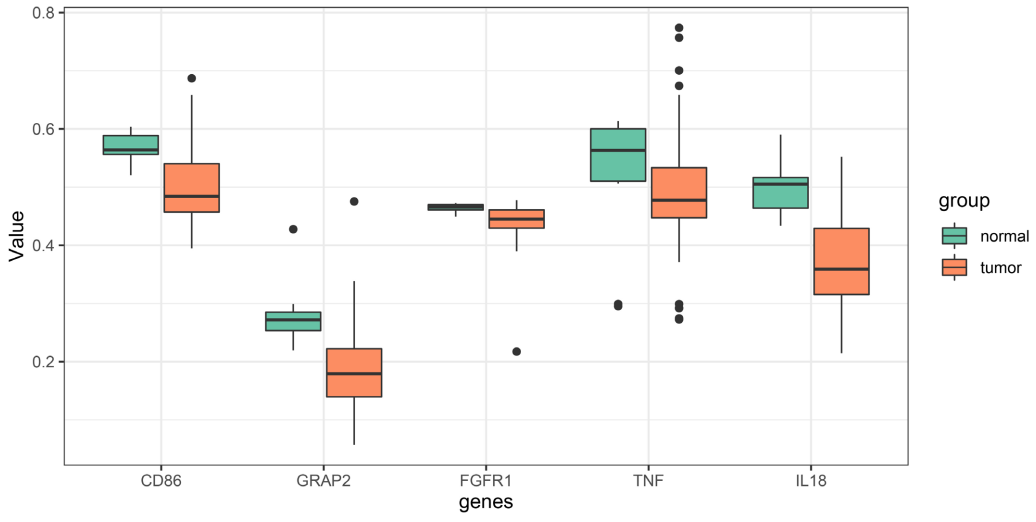
(b)



(c)



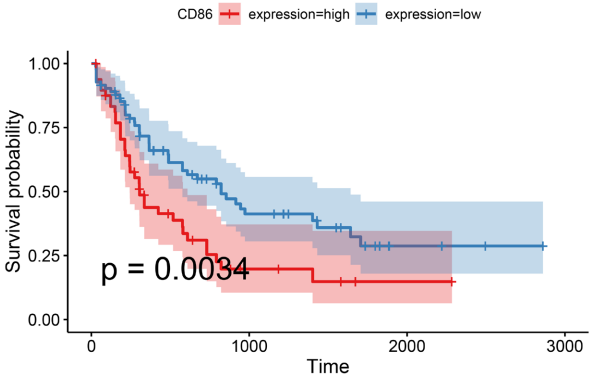
(d)



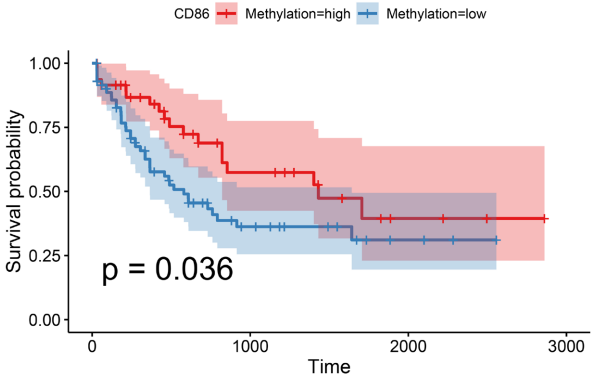
(e)

Figure 3. (a) Protein interaction network diagram of priority genes; (b) IRG gene expression levels that are significantly related to prognosis; (c) IRG gene methylation levels that are significantly related to prognosis; (d) Box plot of five core gene expression levels; (e) Box plot of methylation signal values of five core genes

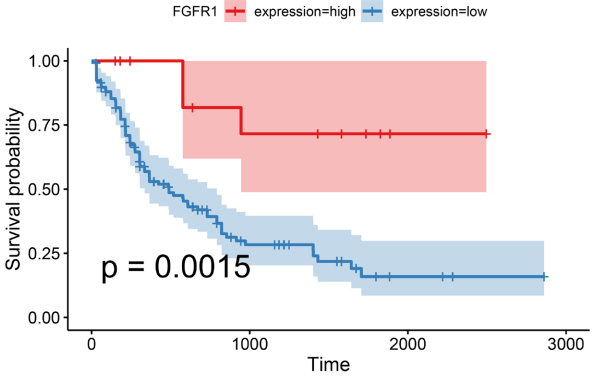
图 3. (a) 优先级基因的蛋白互作网络图；(b) 预后显著相关的 IRG 基因表达水平；(c) 预后显著相关的 IRG 基因甲基化水平；(d) 五个核心基因表达水平箱线图；(e) 五个核心基因甲基化信号值箱线图



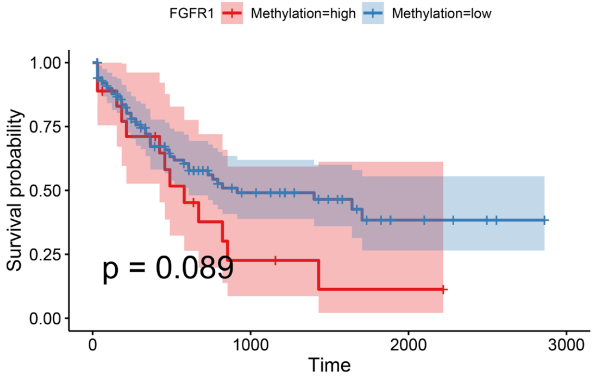
(a1)



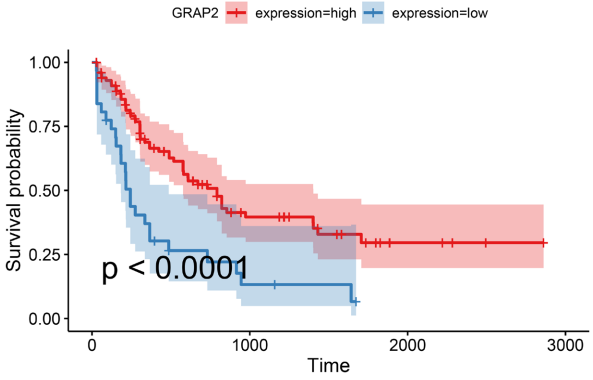
(a2)



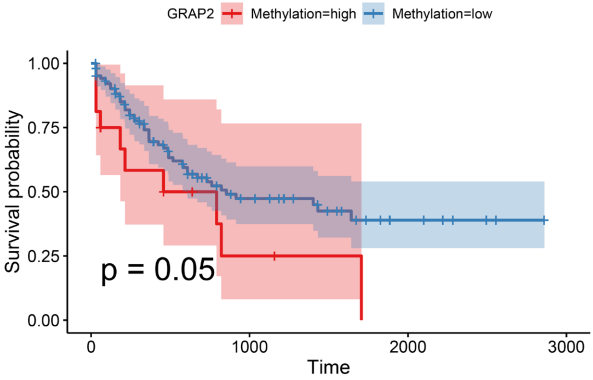
(b1)



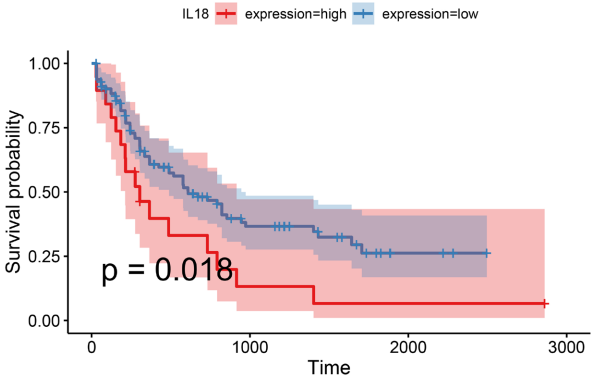
(b2)



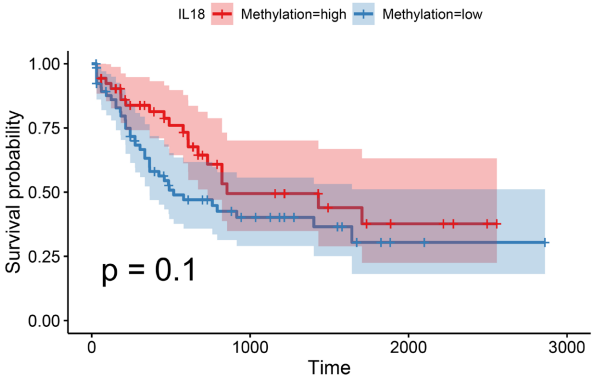
(c1)



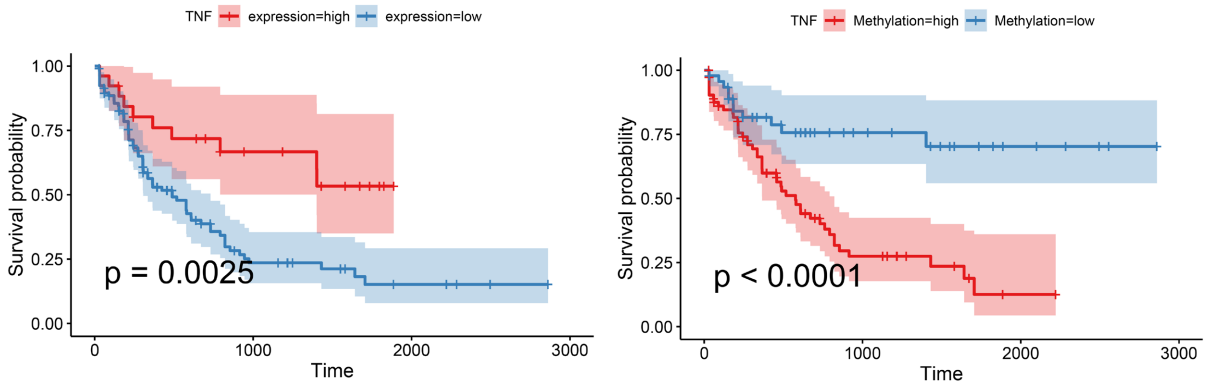
(c2)



(d1)



(d2)

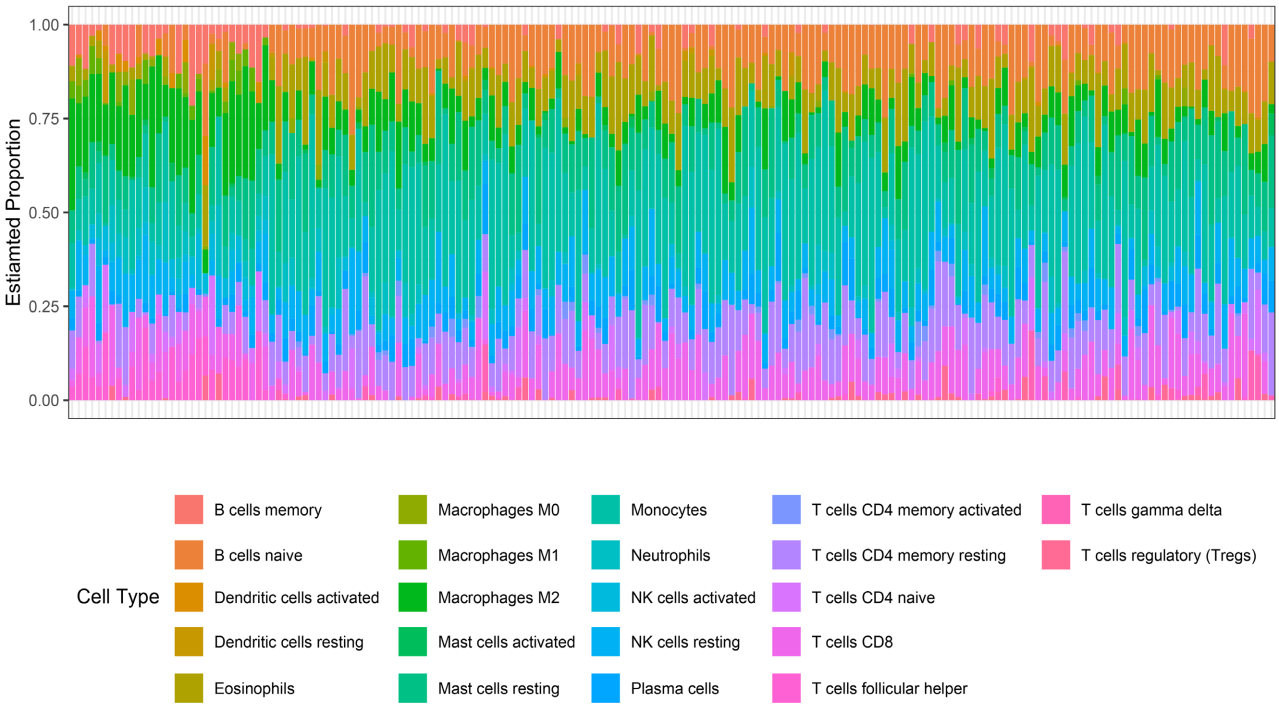


(e1) (e2)

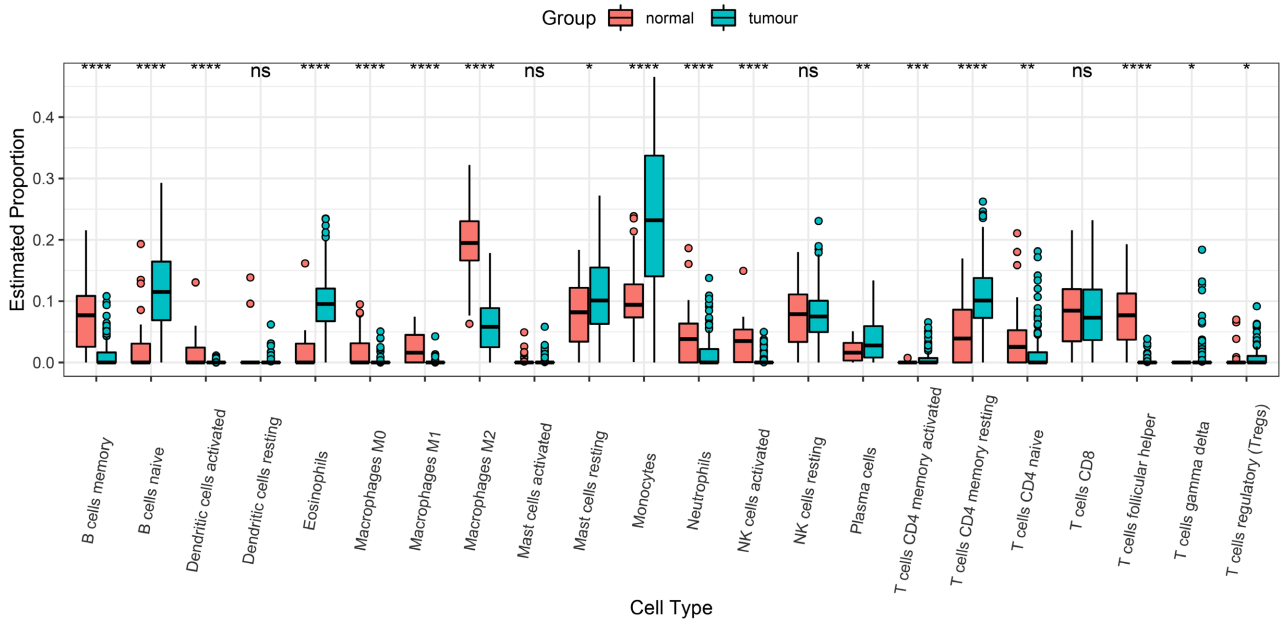
Figure 4. Five core genes prognostic survival analysis chart
图 4. 五个核心基因预后生存分析图

3.4. 预后核心基因表达水平与免疫浸润的相关性

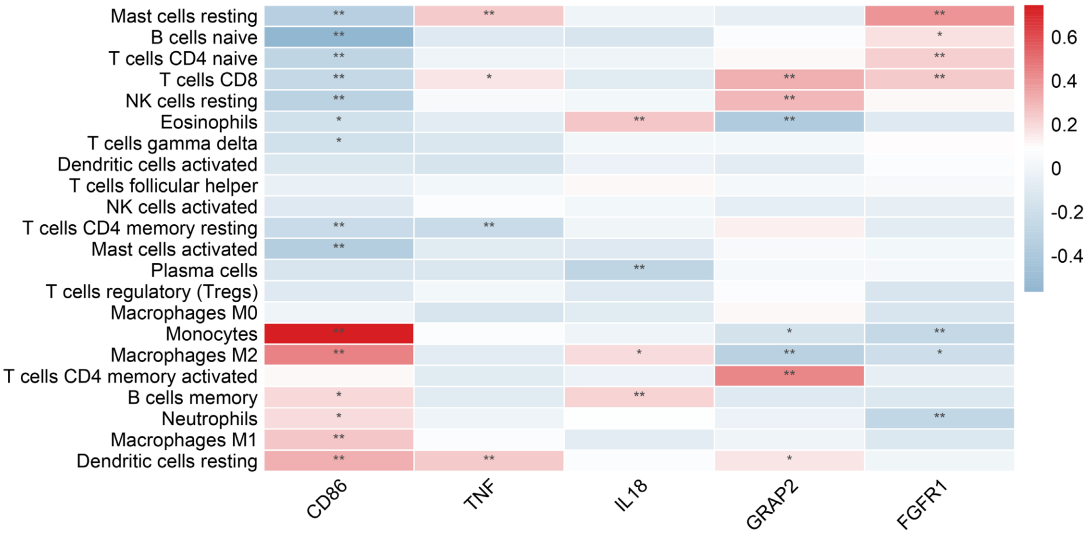
图 3(d)，图 3(e)五个核心基因的基因表达及甲基化信号值箱线图，展示了图 5(c)基因 CD86 的表达水平与单核细胞、巨噬细胞 M1、巨噬细胞 M2 浸润水平呈显著正相关，而与静息 Mast 细胞、naïve B 细胞、naïveCD4 T 细胞、CD8 T 细胞、静息 NK 细胞浸润水平显著负相关。基因 GRAP2 的表达水平与单核细胞、巨噬细胞 M2、嗜酸性粒细胞浸润水平呈负相关，与 CD8 T 细胞、静息 NK 细胞、激活 CD4 记忆 T 细胞浸润水平显著正相关。基因 FGFR1 的表达水平与单核细胞、巨噬细胞 M2、中性粒细胞呈负相关，与静息 Mast 细胞、naïve B 细胞、naïve CD4 T 细胞、CD8 T 细胞浸润水平呈正相关。



(a)



(b)



(c)

Figure 5. (a) Immune cell infiltration ratio; (b) The difference in the abundance of immune cells in the two states; (c) The correlation between five core prognostic genes and immune infiltration

图 5. (a) 免疫细胞浸润比例；(b) 免疫细胞丰度在两种状态下的差异；(c) 五个预后核心基因与免疫浸润的相关

4. 结论

总之，本文通过整合单细胞测序表达数据和 DNA 甲基化数据，利用马尔科夫随机场的网络结构特点构造疾病组和正常组间的差异重连网络，最后筛选出急性髓系白血病免疫相关的五个预后核心基因。基于 TCGA 数据库表达数据和甲基化数据验证了五个预后生物标志物表达水平与甲基化水平的相关性以及两者与预后的相关性，进一步分析了免疫相关的预后生物标志物表达水平与免疫细胞浸润的相关性，这使得免疫细胞丰度成为衡量预后特征的重要指标。

结果表明，基因 CD86、IL18 表达水平与静息 Mast 细胞、naïve B 细胞、naïve CD4 T 细胞、CD8 T

细胞、静息 NK 细胞免疫浸润减少和预后较差有关，基因 GRAP2、FGFR1、TNF 表达水平与预后较好和静息 Mast 细胞、naïve B 细胞、naïve CD4 T 细胞、激活 CD4 记忆 T 细胞、CD8 T 细胞、静息 NK 细胞免疫浸润增加有关，与单核细胞、巨噬细胞、粒细胞浸润减少有关，基因 CD86、TNF、GRAP2、FGFR1、IL18 可以作为 AML 发生、发展及免疫治疗中的预后生物标志物，为进一步研究提供依据。

参考文献

[1] Ferrara, F. and Schiffer, C.A. (2013) Acute Myeloid Leukaemia in Adults. *Lancet*, **381**, 484-495. https://doi.org/10.1016/S0140-6736(12)61727-9

[2] 林凡琳, 潘慧, 刘胜先, 陈晨, 王黎, 崔昌浩. 急性髓系白血病发病机制的研究进展[J]. 中国细胞生物学学报, 2018, 40(5): 850-856.

[3] Genshaft, A.S., Li, S., Gallant, C.J., Darmanis, S., Prakadan, S.M., Ziegler, C.G.K., *et al*. (2016) Multiplexed, Targeted Profiling of Single-Cell Proteomes and Transcriptomes in a Single Reaction. *Genome Biology*, **17**, Article No. 188. https://doi.org/10.1186/s13059-016-1045-6

[4] Cao, J., Cusanovich, D.A., Ramani, V., Aghamirzaie, D., Pliner, H.A., Hill, A.J., *et al*. (2018) Joint Profiling of Chromatin Accessibility and Gene Expression in Thousands of Single Cells. *Science*, **361**, 1380-1385. https://doi.org/10.1126/science.aau0730

[5] Rodriguez-Meira, A., Buck, G., Clark, S.A., Povinelli, B.J., Alcolea, V., Louka, E., *et al*. (2019) Unravelling Intratumoral Heterogeneity through High-Sensitivity Single-Cell Mutational Analysis and Parallel RNA Sequencing. *Molecular Cell*, **73**, 1292-1305. E8. https://doi.org/10.1016/j.molcel.2019.01.009

[6] Besag, J.E. (1986) On the Statistical Analysis of Dirty Pictures. *Journal of the Royal Statistical Society*: *Series B* (*Methodological*), **48**, 259-279. https://doi.org/10.1111/j.2517-6161.1986.tb01412.x

[7] Deng, M., Chen, T. and Sun, F. (2004) An Integrated Probabilistic Model for Functional Prediction of Proteins. *Journal of Computational Biology*, **11**, 463-475. https://doi.org/10.1089/1066527041410346

[8] Segal, E., Wang, H. and Koller, D. (2003) Discovering Molecular Pathways from Protein Interaction and Gene Expression Data. *Bioinformatics*, **19**, i264-i272. https://doi.org/10.1093/bioinformatics/btg1037

[9] Chen, M., Cho, J. and Zhao, H. (2011) Incorporating Biological Pathways via a Markov Random Field Model in Genome-Wide Association Studies. *PLoS Genetics*, **7**, Article ID: e1001353. https://doi.org/10.1371/journal.pgen.1001353

[10] Wei, P. and Pan, W. (2008) Incorporating Gene Networks into Statistical Tests for Genomic Data via a Spatially Correlated Mixture Model. *Bioinformatics*, **24**, 404-411. https://doi.org/10.1093/bioinformatics/btm612

[11] Wei, Z. and Li, H. (2007) A Markov Random Field Model for Network-Based Analysis of Genomic Data. *Bioinformatics*, **23**, 1537-1544. https://doi.org/10.1093/bioinformatics/btm129

[12] Ma, X., Lee, H., Wang, L. and Sun, F. (2007) CGI: A New Approach for Prioritizing Genes by Combining Gene Expression and Protein-Protein Interaction Data. *Bioinformatics*, **23**, 215-221. https://doi.org/10.1093/bioinformatics/btl569

[13] Hu, R., Qiu, X., Glazko, G., Klebanov, L. and Yakovlev, A. (2009) Detecting Intergene Correlation Changes in Microarray Analysis: A New Approach to Gene Selection. *BMC Bioinformatics*, **10**, Article No. 20. https://doi.org/10.1186/1471-2105-10-20

[14] Li, H., Xu, Z., Adams, T., Kaminski, N. and Zhao, H. (2020) A Markov Random Field Model for Network-Based Differential Expression Analysis of Single-Cell RNA-seq Data. https://doi.org/10.1101/2020.11.11.378976

[15] Hou, L., Chen, M., Zhang, C.K., Cho, J. and Zhao, H. (2014) Guilt by Rewiring: Gene Prioritization through Network Rewiring in Genome Wide Association Studies. *Human Molecular Genetics*, **23**, 2780-2790. https://doi.org/10.1093/hmg/ddt668

[16] Li, H. (2010) A hidden Markov Random Field Model for Genome-Wide Association Studies. *Biostatistics*, **11**, 139-150. https://doi.org/10.1093/biostatistics/kxp043

[17] Qiu, X., Hill, A., Packer, J., Lin, D., Ma, Y.-A. and Trapnell, C. (2017) Single-Cell mRNA Quantification and Differential Analysis with Census. *Nature Methods*, **14**, 309-315. https://doi.org/10.1038/nmeth.4150

[18] Yuan, T., Morris, T.J., Webster, A.P., Yang, Z., Beck, S., Feber, A, *et al*. (2017) ChAMP: Updated Methylation Analysis Pipeline for Illumina BeadChips. *Bioinformatics*, **33**, 3982-3984. https://doi.org/10.1093/bioinformatics/btx513