Advances in Applied Mathematics 应用数学进展, 2025, 14(4), 783-797
Published Online April 2025 in Hans. https://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2025.144206

Hans汉斯

基于广义矩量法的自适应套索惩罚线性混合模型用于高维多组学数据的预测分析

王　乐

青岛大学数学与统计学院，山东　青岛

收稿日期：2025年3月26日；录用日期：2025年4月21日；发布日期：2025年4月28日

摘　要

在现代精准医学的探索中，对疾病风险的准确预测至关重要。高维多组学数据为此类预测研究提供了前所未有的资源，但其高维性和复杂的内部关系给分析带来了重大挑战。我们提出了一种基于广义矩估计框架方法(MpLMMGMM-AL)的自适应Lasso惩罚线性混合模型，用于使用高维多组学数据预测表型。我们的方法采用自适应Lasso作为惩罚函数，利用随机效应部分的核函数捕获不同组学数据层的各种类型的预测效应，并有效地选择预测组学区域及其相应的效应。通过大量的仿真，我们证明了MpLMMGMM-AL可以同时考虑大量变量，并有效地从各自的组学层中选择具有预测能力的变量。将该方法应用于公开数据集TCGA中的乳腺癌数据，并与MpLMMGMM进行了性能比较。

关键词

自适应Lasso，惩罚线性混合模型，广义矩法，高维数据，风险预测

An Adaptive Lasso Penalized Linear Mixed Model with Generalized Method of Moments for Prediction Analysis on High-Dimensional Multi-Omics Data

Le Wang

School of Mathematics and Statistics, Qingdao University, Qingdao Shandong

Received: Mar. 26th, 2025; accepted: Apr. 21st, 2025; published: Apr. 28th, 2025

文章引用: 王乐. 基于广义矩量法的自适应套索惩罚线性混合模型用于高维多组学数据的预测分析[J]. 应用数学进展, 2025, 14(4): 783-797. DOI: 10.12677/aam.2025.144206

Abstract

In the exploration of modern precision medicine, an accurate prediction of the disease risk is crucial. For such predictive research, high-dimensional multi-omics data provide unprecedented resources, however, their high dimensionality and intricate internal relationships pose significant analytical challenges. We propose an adaptive Lasso penalized linear mixed model under a generalized method of moments estimation framework (MpLMMGMM-AL) for predicting phenotypes using high-dimensional multi-omics data. Our approach employs adaptive Lasso as the penalty function, utilizes kernel functions in the random effects part to capture various types of predictive effects across different omics data layers, and effectively selects predictive omic regions and their corresponding effects. Through extensive simulations, we demonstrate that MpLMMGMM-AL can simultaneously consider a large number of variables and effectively choose variables with predictive power from their respective omics layers. Our method is applied on a breast cancer data from the publicly available dataset TCGA, and the performance is compared with MpLMMGMM.

Keywords

Adaptive Lasso, Penalized Linear Mixed Models, Generalized Method of Moments, High-Dimensional Data, Risk Prediction

1. 引言

表型预测是生物信息学和统计学领域的一个重要研究课题。随着高通量生物技术的最新进展和大规模精准医疗计划的启动，越来越多的多组学数据可用于反映了疾病的各个方面，如基因组学，转录组学，甲基组学，表观基因组学，蛋白质组学和代谢组学。这些高维和多层组学数据为全面研究预测因素在疾病风险预测中的作用提供了前所未有的机会[1]。在不同分子水平上对众多预测因素进行联合建模，同时考虑它们之间错综复杂的相互联系，是建立准确预测模型的关键一步[2]。虽然高维多层组学数据提供了基础信息，但其极高的维数给联合分析带来了计算挑战。现有的综合方法通常侧重于特定的基因或途径，主要针对检测疾病相关变量方面。同时，现有的方法阐明了潜在的疾病病因，但仅限于对有限数量的变量(如特定途径)进行建模，不能直接应用于预测分析。多组学数据面临的挑战包括：高维遗传数据中的大量噪声[3]，预测因素与表型之间的复杂关系[4] [5]，以及高计算成本[5]。这些因素极大地限制了现有模型的预测准确性。因此，迫切需要寻找一种能够从高维多组学数据中选择预测因子，提高预测精度的预测模型。

多年来，线性混合模型(LMM)及其扩展已被用于高维数据的预测分析。基因组最佳线性无偏预测(gBLUP)方法是 LMM 框架中最早的方法之一。最初由 Harris 等人引入[6]，用于预测牛的产奶量，后来扩展用于预测人类特征[7]。gBLUP (基因组最佳线性无偏预测)假设每个遗传变异的作用是加性的，它们的效应大小遵循相同的正态分布。gBLUP 相当于一个随机效应项的 LMM，其中方差－协方差结构编码了假设的线性加性关系[8]。gBLUP 只估计与随机效应项相关的一个参数，使其计算效率很高。尽管它易于实现，但 gBLUP 中的建模假设过于简单，导致人们试图放松这些假设。来自不同区域的遗传变异，如 eqtl、内含子 snp 和编码区，可以产生不同的效应大小[4]。因此，gBLUP 已经扩展到 MultiBLUP，其

中基因组被划分为多个区域(例如，基于基因或途径注释)。每个区域由一个随机效应项和它自己的方差参数来建模。多核线性混合模型(MKLMM)通过使用再现核希尔伯特空间中的核函数估计每个基因组区域的遗传相似性矩阵来扩展 MultiBLUP，这允许考虑每个基因组区域内的非线性效应[5]。

基于 LMM 的方法通过遗传相似矩阵对各种变异的遗传效应进行编码，显著降低了基因组数据的维数，使其对高维基因组数据的分析具有吸引力。类似的概念可以应用于高维多组学数据的建模，其中遗传相似性被组学相似性取代[9]。然而，在存在大量噪声的情况下，基于线性混合模型(LMM)的预测能力可能会受到限制。研究表明，在遗传相似性估计过程中排除噪声可以增强模型的可解释性，同时提高预测模型的鲁棒性和准确性。在目标函数中添加 L1 惩罚是减少噪声影响的常用方法。此外，对于惩罚 LMM，获得限制最大似然估计(REML)的计算代价很高。广义矩量法(GMM)作为一种惩罚 LMM 方差分量估计的替代方法[8]。在基因组数据预测分析中，Wang 和 Wen 开发了一种带有广义矩估计方法(pLMMGMM)的惩罚线性混合模型[8]。对于多组学数据的预测分析，Wang 和 Wen 提出了一种带有 GMM 估计的惩罚 LMM (MpLMMGMM) [10]。现有方法采用 L1 惩罚，即 Lasso 来选择预测标记，其中 Lasso 对每个系数施加相同的惩罚。然而，对于小偏差和更好的稀疏性，应该对大系数应用较小的惩罚，对小系数应用较大的惩罚。

为了应对这些挑战，我们开发了一个具有自适应 Lasso 惩罚的线性混合模型，命名为 MpLMMGMM-AL，用于高维多组学数据的预测分析。MpLMMGMM-AL (i)识别预测性生物标志物，(ii)允许每个基因组区域的多个核来解释各种类型的遗传效应，以及(iii)考虑多组学数据之间复杂的内部/内部关系。在接下来的章节中，我们将详细介绍 MpLMMGMM-AL 方法。随后，我们将通过仿真研究比较其与现有方法 MpLMMGMM 的预测精度。此外，我们将使用乳腺癌数据集说明其应用。

2. 材料与方法

MpLMMGMM-AL 通过协方差矩阵的核化和广义矩量法的惩罚方法，扩展了多组学数据分析的惩罚线性混合模型。在本节中，我们简要概述了 LMM 和我们的多组学数据分析模型，并提出了具有 GMM 估计的自适应 Lasso 惩罚线性混合模型。

2.1. 各区域多组学数据整合

假设我们有一个包含 n 个人的数据集。设 Y 为 $n\times1$ 结果向量，X_d 为人口统计变量(如年龄和性别)的 $n\times P_d$ 矩阵。我们将基因组划分为 R 组，根据不同的标准(如基因和通路注释)定义，并使用 O_i 来表示 i-th 组中所有预测因子的联合预测效果。我们将结果建模为:

$$Y = X_d\beta_d + \sum_{i=1}^{R} O_i + \varepsilon,\ \varepsilon \sim N\left(0, \sigma_0^2 I_n\right) \tag{1}$$

特别地：

$$Y = X_d\beta_d + \sum_{i=1}^{R}\sum_{j\in S_i} m_j^i + \varepsilon,\ \varepsilon \sim N\left(0, \sigma_0^2 I_n\right) \tag{2}$$

这里 $m_j^i \sim N\left(0, K_j^i\sigma_{ij}^2\right)$ 和 S_i 区域 i 是否考虑了所有组学效应的集合(例如基因组数据的边际预测效应以及基因组和甲基化之间的相互作用)。

在不丢失一般性的情况下，我们使用基因注释来定义集合，并且只考虑基因表达、基因组学和甲基化数据。我们将基因表达数据的预测效应视为固定效应。生物系统受到多个层面的调控，涉及组学数据功能层之间和功能层内部的相互作用。假设我们将区域 m 的不同组学数据(如基因组学和甲基化数据)视为 J_1，层内相互作用为 J_2，层间相互作用为 J_3。随后，我们的模型可以写成：

$$Y = X_d\beta_d + \sum_{i=1}^{R} E_i\gamma_i + \sum_{i=1}^{R}\sum_{j=1}^{J_1} o_j^i + \sum_{i=1}^{R}\sum_{j'}^{J_2} W_{j'}^i + \sum_{i=1}^{R}\sum_{j''}^{J_3} B_{j''}^i + \varepsilon \tag{3}$$

这里 E_i 是 $n\times p_i$ 基因表达数据维数矩阵， γ_i 是它们的效应； o_j^i 表示区域 i 内 jth 组学层所有预测变量的综合预测效应； $W_{j'}^i$ 表示区域 i 内 j'th 层内相互作用； $B_{j''}^i$ 表示区域 i 内 j''th 区域的层间交互作用。

类似于 Hai 和 Wen 两位作者提出的 BLMM [11]，我们使用随机效应项来模拟这些联合效应为 $o_j^i \sim N\left(0, K_{o,j}^i\sigma_{o,i,j}^2\right)$， $W_{j'}^i \sim N\left(0, K_{w,j'}^m\sigma_{w,i,j'}^2\right)$， $B_{j''}^i \sim N\left(0, K_{b,j'}^m\sigma_{b,i,j''}^2\right)$。对于组学数据的每个边际预测效应(即 o_j^i)，我们采用线性核函数，定义为 $K_{o,j}^i\left(Z_{ij}\right) = \left(\frac{1}{\sqrt{p_{ij}}}Z_{ij}^k\right)^{\mathrm{T}}\left(\frac{1}{\sqrt{p_{ij}}}Z_{ij}^l\right)$。这里 p_{ij} th 组学层是变异的数量， Z_{ij}^k 和 Z_{ij}^l th 分别为个体 k 和 l 的 jth 组学数据向量。对于层内相互作用，我们考虑具有 2 个自由度的多项式核 $K_{w,j'}^i\left(Z_{ij'}\right) = \left(\left(\frac{1}{\sqrt{p_{ij'}}}Z_{i,j'}^k\right)^{\mathrm{T}}\left(\frac{1}{\sqrt{p_{ij'}}}Z_{ij'}^l\right)\right)^2$。用于捕获层间交互的内核定义为 $K_{b,j''}^i = K_{j_1''}^i \circ K_{j_2''}^i$， $\circ$表示哈达玛积运算， $K_{j_1''}^i$、 $K_{j_2''}^i$ 分别为组学层 j_1'' 和 j_2'' 的协方差矩阵[12]。

2.2. 带 GMMs 估计器的自适应套索惩罚线性混合模型

许多复杂疾病的根本原因往往事先不为人所知，因此，分析中包括的大量区域可能难以预测，特别是在高维数据的情况下。此外，综合证据表明，并非所有的遗传变异和区域都具有预测能力[4]。因此，在分析中纳入所有组学层及其潜在的相互作用可以减轻预测因素的影响，并导致次优性能。因此，变量选择对于预测模型的稳健性和准确性具有重要意义[2]。由于所提出的 LMM 框架的高度灵活性，它可以很容易地适应各种场景，其变量选择过程涉及固定和随机效应。例如，在模型(3)中，具有预测表达水平的基因的选择涉及到固定效应的选择(即 $\gamma_i \neq 0$)，CpG 位点和遗传变异的选择需要随机效应的选择(即 $\sigma_{o,i,j}^2 \neq 0$， $\sigma_{w,i,j'}^2 \neq 0$， $\sigma_{b,i,j''}^2 \neq 0$)。

在目标函数中添加 L1 惩罚是同时进行变量选择和参数估计的常用方法。在这种情况下，Lasso 对每个系数施加相同的惩罚。然而，对于小偏差和改进的稀疏性，应该对大系数应用小的惩罚，对小系数应用大的惩罚。

因此，我们用自适应 Lasso 惩罚代替 Lasso 惩罚。虽然 REML 被广泛用于估计线性混合模型(Linear Mixed Models, LMM)的参数[4] [7] [13]，但其计算成本很高，特别是对于具有大量随机效应的 LMM。事实上，在 REML 和最大似然估计(MLE)中，由于计算负担的原因，无法考虑大量的随机效应。因此，根据 Wang 和 Wen 提出的类似思路，我们建议继续使用广义矩量法(Generalized Method of Moments, GMM)估计模型参数[14]。因此，模型(3)的目标函数可表示为：

$$\begin{aligned}\left(\hat{\beta}_d, \hat{\gamma}, \hat{\sigma}^2\right) = \arg\min_{\beta_d,\gamma,\sigma^2} \frac{1}{2}\left\| ZZ^{\mathrm{T}} - \sum_{i=1}^{R}\sum_{j=1}^{J_1} K_{o,j}^i\sigma_{o,i,j}^2 - \sum_{i=1}^{R}\sum_{j'}^{J_2} K_{w,j'}^i\sigma_{w,i,j'}^2 - \sum_{i=1}^{R}\sum_{j''=1}^{J_3} K_{b,j''}^i\sigma_{b,i,j''}^2 - \sigma_0^2 I_n \right\|_F^2 \\ + \lambda_1\left(\sum_{i=1}^{R}\sum_{j=1}^{J_1}\omega_{o,i,j}\sigma_{o,i,j}^2 + \sum_{i=1}^{R}\sum_{j'}^{J_2}\omega_{w,i,j'}\sigma_{w,i,j'}^2 + \sum_{i=1}^{R}\sum_{j''=1}^{J_3}\omega_{b,i,j''}\sigma_{b,i,j''}^2\right) + \lambda_2\sum_{i=1}^{R}\omega_i\left|\gamma_i\right|\end{aligned} \tag{4}$$

这里 $\gamma = \left(\gamma_1, \cdots, \gamma_R\right)$， $Z = Y - X_d\beta_d - \sum_{i=1}^{R} E_i\gamma_i$， $\sigma^2 = \left(\sigma_0^2, \sigma_{O,11}^2, \cdots, \sigma_{o,R,J_1}^2, \sigma_{w,11}^2, \cdots, \sigma_{w,R,J_2}^2, \sigma_{b,11}^2, \cdots, \sigma_{b,R,J_3}^2\right)$， λ_1 和 λ_2 分别为随机效应和固定效应的非负正则化参数。 $\omega = \left(\omega_{o,1,1}, \cdots, \omega_{o,R,J_1}, \omega_{w,1,1}, \cdots, \omega_{w,R,J_2}, \omega_{b,1,1}, \cdots, \omega_{b,R,J_3}\right)$ 是自适应权重， $\omega = 1/\left|\tilde{\phi}\right|$，用 $\tilde{\phi}$ 表示 ϕ 的初始的 $\sqrt{n}$ 一致估计量(例如最大似然估计量)。对于固定和随机效应，如果我们不希望对特定参数进行变量选择，我们将相应的权重设置为零(例如，如果我们打算包

括所有人口统计变量进行预测，则将这些人口统计变量对应的自适应权重设置为零)。

我们采用迭代方法来估计随机效应(即：σ^2)和固定效应(即：γ、β_d)中的参数。在迭代步骤 $t+1$ 中，我们首先将随机效应项更新为：

$$\begin{aligned}\hat{\sigma}^{2,t+1} = \arg\min_{\sigma^2} \frac{1}{2}\left\| Z_t Z_t^{\mathrm{T}} - \sum_{i=1}^{R}\sum_{j=1}^{J_1} K_{o,j}^{i}\sigma_{o,i,j}^{2} - \sum_{i=1}^{R}\sum_{j'}^{J_2} K_{w,j'}^{i}\sigma_{w,i,j'}^{2} - \sum_{i=1}^{R}\sum_{j''=1}^{J_3} K_{b,j''}^{i}\sigma_{b,i,j''}^{2} - \sigma_0^2 I_n \right\|_F^2 \\ + \lambda_1\left(\sum_{i=1}^{R}\sum_{j=1}^{J_1}\omega_{o,i,j}\sigma_{o,i,j}^{2} + \sum_{i=1}^{R}\sum_{j'}^{J_2}\omega_{w,i,j'}\sigma_{w,i,j'}^{2} + \sum_{i=1}^{R}\sum_{j''=1}^{J_3}\omega_{b,i,j''}\sigma_{b,i,j''}^{2}\right), \lambda_1 > 0.\end{aligned} \tag{5}$$

这里 $Z_t = Y - X_d\beta_d^t - \sum_{i=1}^{R} E_i\gamma_i^t$；给定步骤 $t+1$ 中随机效应的参数估计，我们将与固定效应相关的参数更新为

$$\left(\hat{\beta}_d^{t+1}, \hat{\gamma}^{t+1}\right) = \arg\max_{\beta_d,\gamma} -\frac{1}{2}\log\left|\Sigma_{t+1}\right| - \frac{1}{2}Z^{\mathrm{T}}\Sigma_{t+1}^{-1}Z - \lambda_2\sum_{i=1}^{R}\omega_i\left|\gamma_i\right|, \lambda_2 > 0. \tag{6}$$

这里 $\Sigma_{t+1} = \sum_{i=1}^{R}\sum_{j=1}^{J_1} K_{o,j}^{i}\sigma_{o,i,j}^{2,t+1} + \sum_{i=1}^{R}\sum_{j'}^{J_2} K_{w,j'}^{i}\sigma_{w,i,j'}^{2,t+1} + \sum_{i=1}^{R}\sum_{j''=1}^{J_3} K_{b,j''}^{i}\sigma_{b,i,j''}^{2,t+1} + \sigma_0^{2,t+1} I_n$。与依赖 REML 估计的惩罚 LMM 相比，我们的目标函数在每次迭代中更容易优化。与现有的 LMM 只能考虑有限数量的随机效应相比[8]，我们的方法允许联合考虑大量区域(即随机效应)，并有效识别具有预测能力的区域。与对每个系数应用相同惩罚的常见$L1$ 惩罚不同，我们的方法可以对不同的系数施加不同的惩罚。

让 $Y_a = \left(Y_p, Y\right)$，这里 Y_p 是 $n_p \times 1$ 要预测的结果向量。给定参数估计，如：$\sigma^2, \beta_d, \gamma$，$Y_a$ 的方差可以直接推导为 $\Sigma_{Y_a} = \sum_{i=1}^{R}\sum_{j=1}^{J_1} K_{o,i,j}^{i}\sigma_{o,i,j}^{2} + \sum_{i=1}^{R}\sum_{j'=1}^{J_2} K_{w,i,j}^{i}\sigma_{w,i,j}^{2} + \sum_{i=1}^{R}\sum_{j''=1}^{J_3} K_{b,i,j}^{i}\sigma_{b,i,j}^{2} + \sigma_0^2 I_n$，$Y_a$ 的方差可以进一步写成：

$$\Sigma_{Y_a} = \begin{pmatrix} \Sigma_{pp} & \Sigma_{po} \\ \Sigma_{op} & \Sigma_{oo} \end{pmatrix}$$

其中 Σ_{pp} 和 Σ_{oo} 分别为测试样本和训练样本的方差，Σ_{po} 为测试样本和训练样本之间的协方差矩阵。利用多元正态分布的条件分布公式，检验样本的预测值可计算为：

$$Y_p = X_{d,p}\hat{\beta}_d + \sum_{i=1}^{R} E_{i,p}\hat{\gamma}_i + \Sigma_{po}\Sigma_{oo}^{-1}\left(Y - X_d\hat{\beta}_d - \sum_{i=1}^{R} E_i\hat{\gamma}_i\right) \tag{7}$$

其中 $X_{d,p}\left(X_d\right)$，$E_{i,p}, i \in (1,\cdots,R)\left(E_i\right)$ 分别表示人口学变量和测试(训练)样本中的基因表达水平。

3. 模拟研究

我们进行了广泛的仿真研究，以评估 MpLMMGMM-AL 的性能，并进一步将其与默认设置下的 MpLMMGMM 进行比较。在下面描述的所有模拟研究中，我们包括三种类型的组学数据：基因表达、甲基化和基因组数据。为了彻底评估我们的方法的性能，模拟数据集被设计为在相同类型的特征(例如，通路中基因的共表达水平)和不同数据类型(例如，甲基化影响启动子区域基因的表达)中表示现实的相关性。因此，我们利用 InterSIM 软件生成基因表达和甲基化数据。该软件以 TCGA 卵巢癌研究为基础，模拟了多种真实的内部/相互关联的数据类型[15]。由于 InterSIM 软件不模拟基因组数据，我们发现为了真实地模拟人类基因组，Consortium 等从 1000 基因组计划的全基因组测序数据中提取了所有的单核苷酸变异(SNV) [16]。根据提供的前 100 个基因的列表，我们在线搜索每个基因对应的详细数据信息(https://www.ensembl.org/index.html?ref=opentrons.ghost.io)。随后，我们使用模拟基因表达和甲基化数据选择位于基因组区域的 SNV。我们排除了没有单核苷酸变异(SNV)的基因。在模拟研究中使用的基因的详细信息在附录中提供(见表 S1)。

我们将 SNV 和 CpG 位点定位到基因区域，并基于因果基因模拟定量表型，其中 25%的 SNV 位于被指定为因果基因的因果基因上。对于下面描述的所有模拟，我们将前三个区域设置为关联区域，而将其余区域设置为噪声。我们假设样本量为 500，其中 70%的样本用于模型训练其余用于模型评估。预测精度采用 Pearson 相关性和均方根误差(RMSE)来衡量。对于我们提出的方法，我们还计算了从不同疾病模型中正确选择预测区域的概率。

3.1. 方案一：噪音区的影响

积累的证据表明，从多个组学数据源收集的过多变量引入了噪声。为了评估它们的影响，我们指定了三个区域作为关联区域，并逐渐将噪声区域的数量从 7 个增加到 97 个。在这种情况下，我们只考虑加性效应，模型的表述如下：

$$Y=\sum_{i=1}^{3}E_i\gamma_i+\sum_{i=1}^{3}o_i^G+\sum_{i=1}^{3}o_i^M+\varepsilon \tag{8}$$

我们使用上标来表示相关的组学数据，其中，E、G、M 分别表示基因表达，基因组学和甲基化数据。$o_i^j\sim N\left(0,K_i^j\sigma_i^{2,j}\right),j\in G,M$ 在区域 i 中表示基因组和甲基化数据的预测效果。对于每个给定数量的噪声区域，我们改变总遗传度，记为

$$h^2=\frac{\sum_{k=1}^{3}\sigma_k^2}{\sum_{k=1}^{3}\sigma_k^2+\sigma_e^2} \tag{9}$$

其中，σ_k^2 为基因组或甲基化数据的 k-th 类方差成分，σ_e^2 为残差方差，并允许不同区域对总遗传力的贡献不同。具体地说，对于第 r 个因果基因组区域，它对总遗传力的贡献比例为 $rh/6$，其中 $r=1,2,3$ [17]。如果模型涉及基因组和甲基化数据的随机效应，则总遗传力修改为 rh/12，范围为 20%~80%。样本量设置为 500，其中训练样本量为 350，测试样本量为 150，在蒙特卡罗模拟中重复实验 500 次。总遗传力为 0.6 的 RMSE 和 Pearson 相关性如图 1 所示。

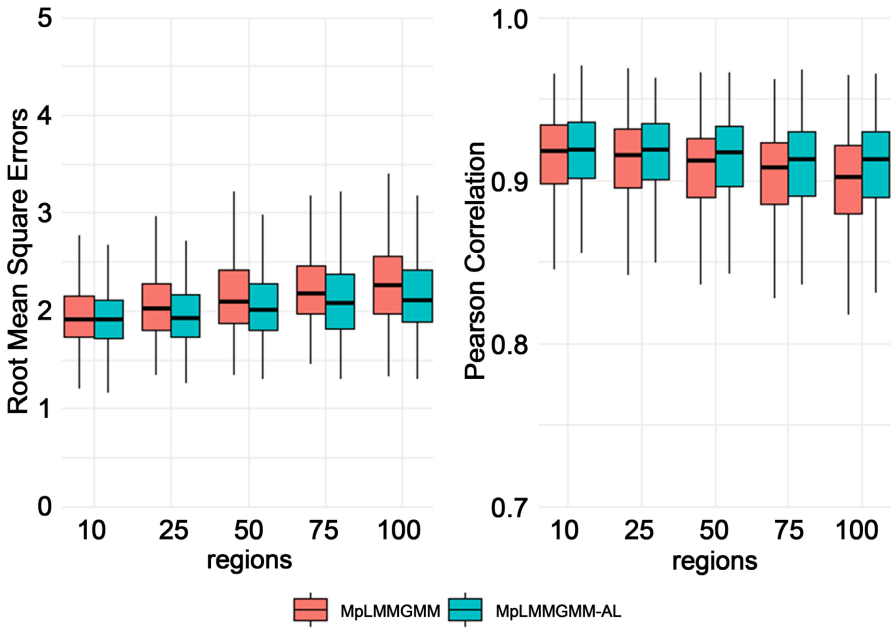

Figure 1. The impact of the number of noise regions (h = 0.6)
图 1. 噪声区域数的影响(h = 0.6)

在所有考虑的场景中，MpLMMGMM-AL 的性能都优于 MpLMMGMM 方法。随着噪声量的增加，我们的方法的性能保持相对稳定，显示出对噪声的鲁棒性，可能归因于其参数估计的惩罚函数。随着噪声区域数量的增加，当遗传度等于 0.6 时选择关联区域的概率见表 1。

Table 1. As the number of noise regions increases, the opportunity to select associated regions (h = 0.6)
表 1. 随着噪声区域数量的增加，选择关联区域的机会(h = 0.6)

区域	基因表达数据		基因组数据		甲基化数据	
噪声数	灵敏度	特异性	灵敏度	特异性	灵敏度	特异性
10	0.9920	0.9449	0.8967	0.6894	0.2573	0.7429
25	0.9927	0.9403	0.8967	0.7043	0.2360	0.8605
50	0.9773	0.9553	0.8787	0.7483	0.2087	0.9082
75	0.9680	0.9459	0.8500	0.7728	0.2033	0.9076
100	0.9600	0.9624	0.8193	0.8045	0.1787	0.9301

当 h = 0.6 时，MpLMMGMM-AL 的平均灵敏度和特异性分别为 69%和 85%；虽然选择的敏感性和特异性可能受到噪声基因数量及其影响程度的影响，但 MpLMMGMM-AL 总体上显示出正确选择预测基因和滤除噪声的能力。

3.2. 情景 2：疾病影响模型

复杂的人类疾病表现在不同的分子水平上[16]，因此，我们在这组模拟中评估了疾病模型的影响。包括(i)只有个体组学数据对疾病风险的贡献，以及(ii)多重组学数据对风险的共同贡献。从 50 个随机选择的基因中提取遗传、基因表达和甲基化数据，并将其中 3 个基因设置为因果关系。

3.2.1. 结果受单层组学数据的影响

我们首先评估了 MpLMMGMM-AL 在只有单层基因组数据有助于疾病风险的情况下的性能。我们考虑了线性和非线性预测效应，排除了人口统计学变量，并将结果模拟为

$$Y = \sum_{i=1}^{3} I_i^E E_i \gamma_i + \sum_{i=1}^{3} o_i^G + \sum_{i=1}^{3} o_i^M + \varepsilon \tag{10}$$

如果区域 i 上的相关组学对结果有贡献，那么 I_i^E 是一个等于 1 的指标函数。我们考虑了四种类型的疾病模型，包括不同组学数据层，见表 S2。

表 2 总结了本文方法在单层组学数据模型下的平均 RMSE 和 Pearson 相关性。除了 Multi-omics 列中的 MpLMMGMM- AL 和 MpLMMGMM 外，MpLMMGMM 在转录组学列中相当于 Lasso。此后，它也可以被称为转录组学。MpLMMGMM-AL 类似于在套索之上构建的自适应套索。在基因组学和表观基因组学中，MpLMMGMM 与 pLMMGMM 相对应[8]。标记为 S1~S4 的行对应于模型 S1~S4，表 3 描述了在这些疾病模型中选择相关区域的可能性。

Table 2. The impact of single-layer omics data on disease models
表 2. 单层组学数据疾病模型的影响

模型	方法	多组学	转录组	基因组	甲基化
S1-RMSE	MpLMMGMM-AL	1.03	1.03	3.31	1.69
	MpLMMGMM	1.08	1.04	3.38	1.69

续表

S2-RMSE	MpLMMGMM-AL	1.40	1.48	1.39	1.47
	MpLMMGMM	1.37	1.46	1.37	1.46
S3-RMSE	MpLMMGMM-AL	1.04	1.12	1.33	1.02
	MpLMMGMM	1.02	1.12	1.31	1.02
S4-RMSE	MpLMMGMM-AL	1.51	2.04	2.05	2.03
	MpLMMGMM	2.04	2.03	2.05	2.04
S1-Pearson 相关性	MpLMMGMM-AL	0.951	0.952	0.0306	0.865
	MpLMMGMM	0.947	0.950	0.0149	0.865
S2-Pearson 相关性	MpLMMGMM-AL	0.311	0.0326	0.340	0.0301
	MpLMMGMM	0.344	0.0575	0.356	0.0277
S3-Pearson 相关性	MpLMMGMM-AL	0.521	0.432	0.0286	0.539
	MpLMMGMM	0.536	0.424	0.0300	0.559
S4-Pearson 相关性	MpLMMGMM-AL	0.124	0.0322	0.125	0.0319
	MpLMMGMM	0.106	0.00120	0.125	0.0239

Table 3. The opportunity to select associated regions under different disease models.
表 3. 在不同疾病模型下选择关联区域的机会

区域	基因表达数据		基因组数据		甲基化数据	
	特异性	灵敏性	特异性	灵敏性	特异性	灵敏性
S1	0.996	0.9523	-	0.7540	-	0.9844
S2	-	0.8710	0.8807	0.7751	-	0.9647
S3	-	0.8767	-	0.7935	0.2393	0.9093
S4	-	0.8671	0.7733	0.8071	-	0.9758
S5	-	0.9942	0.042	0.9787	0.006	0.9893
S6	0.9927	0.9537	0.8973	0.7367	-	0.9528
S7	0.99	0.9549	-	0.7980	0.23	0.9041
S8	-	0.8802	0.8667	0.7915	0.2133	0.9182
S9	0.988	0.9490	0.7733	0.7367	0.1707	0.9203

从表 2 中可以观察到，当只有转录组学数据对结果有贡献时(S1)，我们的方法与仅使用基因表达数据的方法相似。此外，当基因组数据仅显示非线性预测效应时(S4)，MpLMMGMM-AL 优于 MpLMMGMM。在这两种情况下，MpLMMGMM-AL 的平均敏感性和特异性分别为 88%和 89%。

3.2.2. 结果受多个组学数据的影响

当多种类型的组学数据单独或共同影响结果时，我们评估了 MpLMMGMM-AL 的性能。我们加入组学之间的相互作用($I_i o_i^O$)加入公式(10)，模拟结果为

$$Y = \sum_{i=1}^{3} I_i^E E_i \gamma_i + \sum_{i=1}^{3} o_i^G + \sum_{i=1}^{3} o_i^M + \sum_{i=1}^{3} I_i^O o_i^O + \varepsilon \tag{11}$$

这里为 o_i^O 区域 i 基因型与甲基化之间的相互作用效应且 $o_i^O \sim N\left(0, K_i^O \sigma_i^2\right)$，$K_i^O = K_i^G \circ K_i^M$ 是区域 i 的上位性效应，$K_i^G = Z_i^G \left(Z_i^G\right)^{\mathrm{T}} \Big/ p_i^G$、$K_i^M = Z_i^M \left(Z_i^M\right)^{\mathrm{T}} \Big/ p_i^M$，$Z_i^G$ 与 Z_i^M 分别表示区域 i 的基因组数据和甲基化数据。I_i^O 是一个指示函数，如果组学之间存在相互作用，则它为 1。在这个模拟中，我们主要考虑五种情况，S5：只有基因组学和甲基化数据之间的相互作用才会导致疾病风险($I_i^O \neq 0$)；S6：转录组学和基因组数据独立地促进疾病风险($I_i^E \neq 0$，$\sigma_i^{2,G} \neq 0$，K_i^G 是使用线性核计算的)；S7：转录组学和甲基化数据独立影响疾病风险($I_i^E \neq 0$，$\sigma_i^{2,M} \neq 0$，K_i^M 是使用线性核计算的)；S8：基因组和甲基化数据独立地影响疾病风险($\sigma_i^{2,G} \neq 0$，$\sigma_i^{2,M} \neq 0$，K_i^G，K_i^M 使用线性核计算)；S9：多种类型的组学数据共同促成了结果，并存在不同层组学数据之间的相互作用($I_i^E \neq 0$，$\sigma_i^{2,G} \neq 0$，$\sigma_i^{2,M} \neq 0$，K_i^G，K_i^M 使用线性核计算，$I_i^O \neq 0$)。与上述类似，详情见表 S2 在附录中。

表 4 总结了该方法在多种组学数据模型下的平均均方根误差和 Pearson 相关性。当多个组学数据层独立地贡献结果时(如 S5、S6 和 S7 中观察到的)，MpLMMGMM-AL 优于 MpLMMGMM。当不同类型的组学数据和不同组学数据层之间的相互作用共同影响结果时(如 S9 所示)，MpLMMGMM-AL 优于其他方法。此外，它比仅使用单层组学数据的模型性能更好。这表明，当多个组学数据层对结果有贡献时，对所有组学数据的联合分析可能有利于预测建模。为了评估选择性能，我们进一步计算了我们的方法的敏感性和特异性见表 3。在疾病模型 S5，S6，S7，S9 中，我们的方法平均灵敏度为 57%，特异性为 91%。

Table 4. The impact of multi-omics data on disease models
表 4. 多组学数据疾病模型的影响

模型	方法	多组学	转录组	基因组	甲基化
S5-RMSE	MpLMMGMM-AL	1.05	1.06	3.44	1.70
	MpLMMGMM	1.06	1.08	3.36	1.70
S6-RMSE	MpLMMGMM-AL	1.42	1.49	3.56	2.01
	MpLMMGMM	1.47	1.51	3.49	2.01
S7-RMSE	MpLMMGMM-AL	1.05	1.13	3.48	1.71
	MpLMMGMM	1.12	1.14	3.39	1.69
S8-RMSE	MpLMMGMM-AL	1.42	1.54	1.63	1.48
	MpLMMGMM	1.41	1.54	1.62	1.03
S9-RMSE	MpLMMGMM-AL	1.44	1.48	3.57	1.98
	MpLMMGMM	1.49	1.50	3.51	1.98
S5-Pearson 相关性	MpLMMGMM-AL	0.0323	0.0282	0.0325	0.0296
	MpLMMGMM	0.0323	0.0053	0.0297	0.0195
S6-Pearson 相关性	MpLMMGMM-AL	0.916	0.906	0.0553	0.824
	MpLMMGMM	0.909	0.904	0.0678	0.865
S7-Pearson 相关性	MpLMMGMM-AL	0.949	0.940	0.0188	0.861
	MpLMMGMM	0.941	0.938	0.0282	0.865
S8-Pearson 相关性	MpLMMGMM-AL	0.519	0.325	0.253	0.443
	MpLMMGMM	0.517	0.326	0.278	0.416
S9-Pearson 相关性	MpLMMGMM-AL	0.911	0.905	0.0330	0.823
	MpLMMGMM	0.903	0.904	0.0392	0.823

4. 实际数据应用

在本节中，我们将提出的方法应用于来自 TCGA (癌症基因组图谱)基因组数据库的乳腺癌数据。TCGA 提供了大量的数据，包括体细胞突变、拷贝数变异、DNA 甲基化、mRNA 表达，以及来自数千个与正常细胞基因组数据匹配的肿瘤的临床信息。TCGA 数据使不同癌症类型的比较和对比研究成为可能，旨在为精确肿瘤学提供见解。在我们的分析中，我们选择了乳腺癌的数据，重点关注乳腺癌中指标的检测。长期以来，雌激素受体(ER)表达被认为存在于三分之二的乳腺癌中[14]，但研究表明其患病率可能接近 70% [18]。

本研究共使用了 333 个样本，包括基因表达和体细胞突变状态。我们选择了 41 个先前报道的与乳腺癌相关的基因。详细信息见表 5。为了避免过拟合，我们使用 80%的样本训练 ER 模型，并基于剩余 20%的数据计算平均 Pearson 相关系数和 RMSE。这个过程要重复 200 次，以防止偶然发现。为了比较，我们还采用 MpLMMGMM 方法建立了预测模型。两种方法的预测精度如表 5 示。

Table 5. The prediction accuracy of ER
表 5. ER 的预测精度

	RMSE 均值	Pearson 相关性均值
MpLMMGMM-AL	0.09404	0.4078
MpLMMGMM	0.072553453	0.63809188

变量选择情况如表 6 示。可以观察到，与 MpLMMGMM 相比，我们的方法在 RMSE 均值和 Pearson 相关均值方面的表现略差。而在变量选择方面，综合固定效应和随机效应，MpLMMGMM-AL 下选择 ATM 和 CDH1 基因的概率超过 98%。值得注意的是，ATM 中的蛋白截断变异与总体乳腺癌风险相关，p 值小于 0.0001 [19]。Banerjee 等明确指出 CDH1 基因在预测各类乳腺癌的发病几率方面具有决定性作用，准确率超过 90% [20]。相比之下，MpLMMGMM 选择风险基因的概率不高。

Table 6. Prediction of ER using both genetic and variable methods (ER[a]: Probability of a gene being selected for ER under MpLMMGMM-AL; ER[b]: Probability of a gene being included in ER under MpLMMGMM, N/A: Not applicable)
表 6. 用基因和变量两种方法进行 ER 预测(ER[a]：MpLMMGMM-AL 下基因被 ER 选择的概率；ER[b]：MpLMMGMM 下基因入选 ER 的概率，N/A：无)

基因	转录组	体细胞突变	ER[a]	ER[b]
AKT1	Included	Included	0	0
ATM	Included	Included	0.999	0
ATRIP	Included	Included	0.5	0.5
BARD1	Included	N/A	0	0
BLM	N/A	Included	0.5	0.4
BRCA1	Included	Included	0	0
BRCA2	Included	Included	0	0
BRIP1	Included	Included	0.5	0.0025
CDH1	Included	Included	0.9878	0.005
CHEK2	Included	Included	0.3625	0.5
EPCAM	Included	N/A	0	0.5

续表

ERBB2	Included	Included	0	0.0025
FANCC	Included	Included	0	0
FANCD2	Included	Included	0	0
FANCM	Included	N/A	0.5	0.6
GEN1	Included	Included	0.495	0
HOXB13	Included	N/A	0.0025	0
MCPH1	Included	Included	0	0
MEN1	Included	Included	0.0075	0
MLH1	Included	Included	0.495	0
MSH2	Included	Included	0	0
MSH6	Included	Included	0	0.5
MUTYH	Included	Included	0	0.215
NBN	Included	Included	0	0
NF1	Included	Included	0.5	0
PALB2	Included	Included	0	0.015
PMS2	Included	Included	0	0
POLG	Included	N/A	0	0
PPM1D	Included	Included	0.3425	0
PTEN	Included	Included	0	0
RAD51B	N/A	Included	0	0
RAD51C	Included	Included	0	0
RAD51D	N/A	Included	0	0
RBBP8	Included	Included	0.5	0.1075
RECQL	Included	Included	0	0.0125
RINT1	Included	Included	0.3525	0
SERPINA3	Included	N/A	0.5	0.6
STK11	Included	Included	0.5	0
TEX15	Included	N/A	0	0.5
TP53	Included	Included	0.5	0.5525
XRCC2	Included	Included	0	0

5. 结论

在这项研究中，我们提出了一个自适应 Lasso 惩罚线性混合模型与 GMM 估计器用于多组学数据的预测分析。提出的 MpLMMGMM-AL 将多组学数据划分为多个区域，它可以根据各种标准来定义。它使用多个随机效应项对不同分子水平预测变量的累积预测效应进行建模，并通过使用多个核函数捕获线性和非线性预测效应。该方法采用惩罚项选择预测区域和组学层，利用 GMM 估计器加快计算速度。将自适应套索惩罚应用于不同大小的系数，以增强多组学数据的预测能力。通过利用来自公开数据集

TCGA 的乳腺癌数据，我们观察到我们的方法在预测 ER 水平方面的表现略低于 MpLMMGMM。然而，我们的方法在变量选择方面优于竞争方法。

随着噪声水平的上升，MpLMMGMM-AL 表现出一致和准确的预测性能，而 MpLMMGMM 的稳定性可能不如前者。基于平均灵敏度和特异性的可识别趋势，我们的方法显示出对噪声的鲁棒性，这是开发精确风险预测模型的关键属性。我们的模型，利用自适应套索，显著推进我们对疾病机制的理解。例如，当只有一个组学层可预测时，所提出的方法可以准确地从相应的组学层检测相关区域，从而达到与仅使用疾病相关组学层的模型(例如疾病模型 S1，S4)相似的预测精度水平。即使对于没有边际效应的模型(即疾病模型 S5)，我们的方法也显示出正确检测关联的能力，这与 MpLMMGMM 相当。

虽然我们的方法取得了更好的预测性能，但也有一些局限性。MpLMMGMM-AL 只关注持续的结果。为指数族(例如，二进制和泊松)的结果开发一般线性混合模型(LMM)框架将是有意义的。此外，我们的目标是通过引入有效的筛选规则，如顺序强规则和增强的对偶多面体投影规则，进一步降低计算成本。这将简化和加速计算，特别是对于具有大样本量的高维数据。这些方面将是我们今后研究的重点。总之我们开发了一种带有 GMM 估计器的自适应 Lasso 线性混合模型，用于多组学数据的风险预测分析。我们的方法显示了对噪声的鲁棒性，同时从多个组学层捕获预测标记，包括它们之间的相互作用。在预测性能方面，它在某些情况下优于 MpLMMGMM。

参考文献

[1] Boekel, J., Chilton, J.M., Cooke, I.R., Horvatovich, P.L., Jagtap, P.D., Käll, L., *et al*. (2015) Multi-Omic Data Analysis Using Galaxy. *Nature Biotechnology*, **33**, 137-139. https://doi.org/10.1038/nbt.3134

[2] Morris, J.S. and Baladandayuthapani, V. (2017) Statistical Contributions to Bioinformatics: Design, Modelling, Structure Learning and Integration. *Statistical Modelling*, **17**, 245-289. https://doi.org/10.1177/1471082x17698255

[3] Kirchner, H., Osler, M.E., Krook, A. and Zierath, J.R. (2013) Epigenetic Flexibility in Metabolic Regulation: Disease Cause and Prevention? *Trends in Cell Biology*, **23**, 203-209. https://doi.org/10.1016/j.tcb.2012.11.008

[4] Speed, D. and Balding, D.J. (2014) MultiBLUP: Improved SNP-Based Prediction for Complex Traits. *Genome Research*, **24**, 1550-1557. https://doi.org/10.1101/gr.169375.113

[5] Weissbrod, O., Geiger, D. and Rosset, S. (2016) Multikernel Linear Mixed Models for Complex Phenotype Prediction. *Genome Research*, **26**, 969-979. https://doi.org/10.1101/gr.201996.115

[6] Harris, B.L., Johnson, D.L. and Spelman, R.J. (2009) Genomic Selection in New Zealand and the Implications for National Genetic Evaluation. *Proceedings ICAR* 36*th Session*, Niagara, 16-20 June 2009, 325-330.

[7] Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., *et al*. (2010) Common SNPs Explain a Large Proportion of the Heritability for Human Height. *Nature Genetics*, **42**, 565-569. https://doi.org/10.1038/ng.608

[8] Wang, X. and Wen, Y. (2021) A Penalized Linear Mixed Model with Generalized Method of Moments for Complex Phenotype Prediction. *Bioinformatics*, **38**, 5222-5228. https://doi.org/10.1093/bioinformatics/btac659

[9] Li, J., Lu, Q. and Wen, Y. (2019) Multi-Kernel Linear Mixed Model with Adaptive Lasso for Prediction Analysis on High-Dimensional Multi-Omics Data. *Bioinformatics*, **36**, 1785-1794. https://doi.org/10.1093/bioinformatics/btz822

[10] Wang, X. and Wen, Y. (2022) A Penalized Linear Mixed Model with Generalized Method of Moments for Prediction Analysis on High-Dimensional Multi-Omics Data. *Briefings in Bioinformatics*, **23**, bbac193. https://doi.org/10.1093/bib/bbac193

[11] Hai, Y. and Wen, Y. (2020) A Bayesian Linear Mixed Model for Prediction of Complex Traits. *Bioinformatics*, **36**, 5415-5423. https://doi.org/10.1093/bioinformatics/btaa1023

[12] Hai, Y., Ma, J., Yang, K. and Wen, Y. (2023) Bayesian Linear Mixed Model with Multiple Random Effects for Prediction Analysis on High-Dimensional Multi-Omics Data. *Bioinformatics*, **39**, btad647. https://doi.org/10.1093/bioinformatics/btad647

[13] VanRaden, P.M. (2008) Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science*, **91**, 4414-4423. https://doi.org/10.3168/jds.2007-0980

[14] Allred, D.C., Harvey, J.M., Berardo, M. and Clark, G.M. (1998) Prognostic and Predictive Factors in Breast Cancer by Immunohistochemical Analysis. *Modern Pathology*, **11**, 155-168.

[15] Chalise, P., Raghavan, R. and Fridley, B.L. (2016) *InterSIM*: Simulation Tool for Multiple Integrative 'Omic Datasets'. *Computer Methods and Programs in Biomedicine*, **128**, 69-74. https://doi.org/10.1016/j.cmpb.2016.02.011

[16] The 1000 Genomes Project Consortium (2015) A Global Reference for Human Genetic Variation. *Nature*, **526**, 68-74. https://doi.org/10.1038/nature15393

[17] Wen, Y. and Lu, Q. (2020) Multikernel Linear Mixed Model with Adaptive Lasso for Complex Phenotype Prediction. *Statistics in Medicine*, **39**, 1311-1327. https://doi.org/10.1002/sim.8477

[18] Nadji, M., Gomez-Fernandez, C., Ganjei-Azar, P. and Morales, A.R. (2005) Immunohistochemistry of Estrogen and Progesterone Receptors Reconsidered: Experience with 5993 Breast Cancers. *American Journal of Clinical Pathology*, **123**, 21-27. https://doi.org/10.1309/4wv79n2ghj3x1841

[19] Breast Cancer Association Consortium (2021) Breast Cancer Risk Genes—Association Analysis in More than 113,000 Women. *New England Journal of Medicine*, **384**, 428-439. https://doi.org/10.1056/nejmoa1913948

[20] Banerjee, S., Sengupta, A., Ghosh, S.K. and Banerjee, R. (2024) CDH1 Gene as Biomarker Towards Breast Cancer Prediction. *Journal of Biomolecular Structure and Dynamics*. https://doi.org/10.1080/07391102.2024.2316770

附　录

Table S1. Detailed information on genes in simulation studies
表 S1. 模拟研究中基因的详细信息

基因	SNP 数	CPG 数	基因	SNP 数	CPG 数	基因	SNP 数	CPG 数
ACACA	5700	1	EGFR	4736	2	PCNA	208	3
AKT1	736	1	EIF4E	1088	2	PDK1	1364	2
ASNS	460	2	EIF4EBP1	580	2	PGR	2352	1
ATM	2904	5	EIF4G1	436	1	PIK3R1	1732	2
BAD	276	3	ERBB2	824	11	PRDX1	772	2
BAK1	188	2	ERBB3	512	2	PREX1	5640	2
BAP1	136	2	ETS1	2760	2	PRKAA1	700	1
BAX	184	2	FN1	1744	2	PRKCD	856	2
BCL2	4568	13	FOXO3	2348	2	PTEN	2264	7
BCL2L1	1024	2	GAB2	4308	2	PXN	1024	2
BECN1	400	2	GAPDH	140	2	RAD50	1620	2
BID	1188	1	GSK3A	212	2	RAD51	668	2
BRAF	3988	7	HSPA1A	32	3	RAF1	1872	2
BRCA2	1576	2	IRS1	1300	1	RB1	6348	19
CASP7	1228	1	ITGA2	2368	2	RPS6	92	2
CAV1	752	6	JUN	172	2	RPS6KA1	940	2
CCND1	308	16	KDR	1188	2	RPS6KB1	972	3
CCNE1	296	1	KIT	1760	1	SERPINE1	356	2
CCNE2	336	2	LCK	688	2	SHC1	216	3
CDH1	2200	7	MAP2K1	2256	2	SMAD1	1280	2
CDH2	4604	2	MAPK1	772	2	SMAD3	3020	1
CDH3	2088	2	MAPK14	1660	2	SMAD4	888	2
CDKN1A	236	6	MAPK9	1528	1	SQSTM1	1044	3
CDKN1B	732	2	MET	2088	2	SRC	1372	2
CHEK1	868	2	MYC	148	2	STAT3	1292	2
CHEK2	1212	4	MYH11	4648	2	STAT5A	588	2
CLDN7	68	1	NFKB1	1916	2	STK11	1524	2
COL6A1	900	2	NOTCH1	1588	2	STMN1	372	3
CTNNB1	1140	2	NRAS	220	1	SYK	2632	7
DVL3	304	2	PARK7	772	2	TGM2	984	1
TP53	548	3	XBP1	84	2	VHL	376	7
TSC1	1096	2	XRCC1	948	2	YWHAE	1680	2
TSC2	1284	4	XRCC5	2200	2			
TUBA1B	76	2	YWHAB	432	2			

Table S2. Simulation setup
表 S2. 模拟设置

模型	转录组	基因组	甲基化组	相互作用
S1	0.5			
	1			
	1.5			
S2		0.125		
		0.25		
		0.375		
S3			0.125	
			0.25	
			0.375	
S4		0.0556		
		0.1111		
		0.1667		
S5				0.0556
				0.1111
				0.1667
S6	0.5	0.125		
	1	0.25		
	1.5	0.375		
S7	0.5		0.125	
	1		0.25	
	1.5		0.375	
S8		0.125	0.125	
		0.25	0.25	
		0.375	0.375	
S9	0.5	0.0556	0.0556	0.0556
	1	0.1111	0.1111	0.1111
	1.5	0.1667	0.1667	0.1667

注意：有七个模拟模型：S1：只有转录组学数据对结果有贡献；S2：只有基因组学数据以加性方式对结果做出贡献；S3：只有甲基化数据以加性方式对结果有贡献；S4：只有基因组学数据中的成对相互作用对结果有贡献；S5：只有基因组学和甲基化数据之间的相互作用才会导致疾病风险；S6：转录组学和基因组数据独立导致疾病风险；S7：转录组学和甲基化数据依赖性地促进疾病风险；S8：基因组和甲基化数据独立导致疾病风险；S9：几种类型的组学数据和不同组学数据层之间的相互作用共同影响结果。