甲状腺癌患者微生物为中介诱导炎癌转化的机制研究

设计方案

1. 甲状腺癌表达谱数据获取
   我们从 TCGA 数据库下载人类甲状腺癌 RNAseq 表达谱数据，数据包含 60498 个 ENSEMBL 基因 ID，504 个肿瘤样本，60 个癌旁组织样本。TCGA 转录组数据均为 RSEM count 数据，并经过 TPM 标准化。考虑到对照样本量较少，样本不平衡可能在差异比较分析过程中引入 bias, 因此我们纳入了 279 个 GTEX 甲状腺正常组织样本。最终一共获得 504 个癌症组织样本和 339 个正常对照组织样本。

2. 微生物丰度定量分析
   我们从 TCMbio 数据库(https://microbiomex.sdu.edu.cn/) 获取 TCGA 甲状腺癌患者的菌群丰度数据，该数据包含多个微生物分层数据，包括 genus，class, family, order, phylum, species 等，我们以 genus 作为菌群分类标签，和转录组数据取交集后，最终获取肿瘤组织和癌旁组织样本对应的菌群丰度数据。

3. 菌群相关炎症因子/免疫细胞关联分析
   我们从 GSEA 官网获取人类炎症因子相关基因集 HALLMARK_INFLAMMATORY_RESPONSE，该数据集包含 200 个炎症应答相关基因。我们利用 Pearson 相关性算法平台每个炎症因子基因和菌群丰度之间的相关性，根据 P 小于 0.05 且 |R|>0.5 作为显著性阈值。我们分别在肿瘤组织样本和正常组织中评价菌群丰度和炎症因子基因之间的相关性。然后进一步筛选出在肿瘤组织中表现出特异性强相关的菌群-炎症因子关系对。

   我们利用 CIBERSORT 算法中评估了 22 种经典的免疫细胞浸润分数。然后我们同样分别在肿瘤组织和正常组织中评价菌群丰度和每种免疫细胞的相关性，筛选出在肿瘤组织中表现出特异性强相关的菌群-免疫细胞关系对。

4. 基于菌群-炎症因子-免疫细胞的炎癌标记物筛选
   我们获取了在肿瘤组织中和菌群丰度特异性相关的炎症因子和免疫细胞。这些炎症因子和免疫细胞很可能是在肿瘤环境中受到菌群特异性调控作用影响继而出现的免疫调控反应。因此我们首先通过 interaction network 构建菌群，炎症因子，免疫细胞，基因之间的相关性网络，筛选出同时参与调控菌群丰度， 炎症因子和免疫细胞的基因作为炎癌标记物。

5. 上下游调控通路分析

为了进一步从生物学调控水平解释菌群诱导炎症因子和免疫细胞释放，进而形成特异性肿瘤免疫维环境的分子机制，我们利用 clusterProfiler R package 对炎癌标记物基因进行 gene ontology , KEGG 和 Reactome 功能富集分析。根据 P 值小于 0.05 筛选出显著相关的功能通路。我们进一步利用 CBNplot 算法对 signature 上下游调控通路进行富集分析，识别出 signature 在生物学机制上的调控路径。

6. 炎癌分子模型搭建
   为了从炎癌标记物中提取具有早期诊断意义的分子标识，我们首先利用 Lasso 算法进行特征选择，获取在肿瘤组织和正常组织中具有显著表达水平差异的炎癌标记物。然后利用 random forest 算法构建炎癌分子模型实现对肿瘤组织和正常组织的精准区分。通过 ROC 曲线和混淆矩阵评价模型效能。

7. 炎癌分子模型外部数据验证
   我们从 GEO 数据库获取独立数据集(GSE50901，GSE27155，GSE151181，GSE5090)对炎癌分子模型进行验证。四套数据通过 combat 算法实现批次效应处理后，包含 X 个甲状腺癌患者和 Y 个正常对照组织样本。我们在独立验证数据上通过炎癌分子模型进行区分，并评价该模型在验证数据上的精度。

8. 炎症免疫治疗药物敏感性分析
   经过上述研究我们发现某些菌群具有特殊的调控作用，介导了炎症免疫反应，进而发展成区别于正常组织微环境的的分子和功能状态。如果在这过程中通过抗炎或免疫治疗药物进行干预，对于甲状腺癌治疗具有重要意义。因此我们利用 oncopredict 算法，评价了 CTRP2, GDSC1, GDSC2 三个数据库乳腺组织相关药物，过滤掉 IC50<1% 的药物后，最后一共分析了 151 种药物在不同亚群中的敏感性。然后我们比较了 151 种药物和菌群丰度，炎症因子，免疫细胞的相关性，最终筛选出能够有效阻断菌群诱导炎症因子和免疫细胞释放，形成肿瘤免疫维环境的药物。

分析结果

1. 菌群-炎症因子/免疫细胞关联分析
   我们比较了菌群丰度和炎症因子基因表达数据之间的相关性，发现在肿瘤组织和正常组织中大多数炎症因子都和菌群丰度呈现负相关，即当炎症因子相关基因高表达时，菌群丰度下降，如图 1 所示。

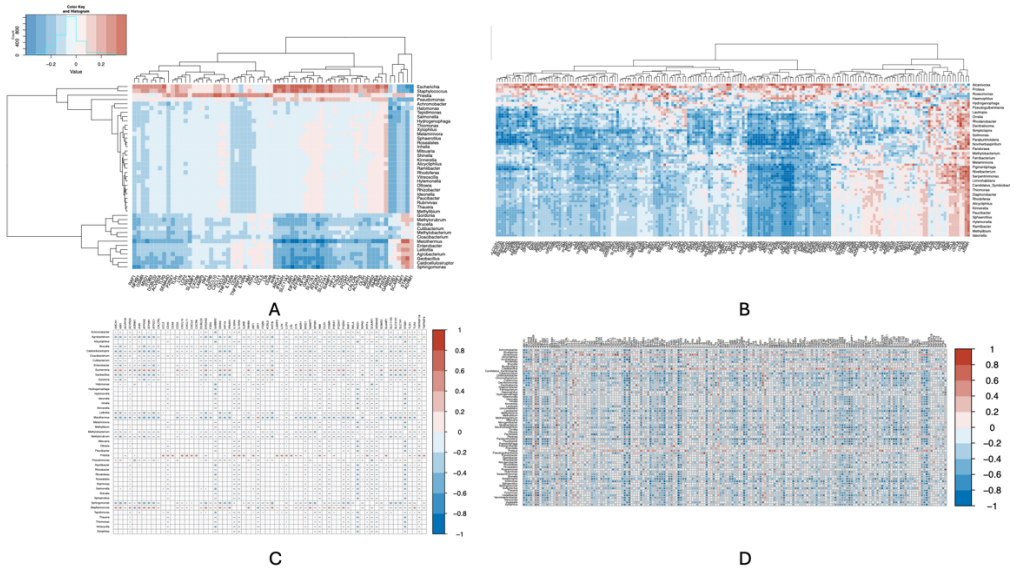


Figure 1 菌群-炎症因子相关性分析A 肿瘤组织中菌群和炎症因子相关性矩阵聚类B 正常组织中菌群和炎症因子聚类C 肿瘤组织中菌群和炎症因子的统计显著性D 正常组织中菌群和炎症因子的统计显著性。

这种负向调控关系很可能和菌群诱导的 Treg 抑制炎症反应或免疫检查点上调，抑制炎症因子释放有关。同时我们还发现，在肿瘤组织中体现出相关性的菌群-炎症因子关系对较少（Figure1A/C），而在正常组织中，显著相关的菌群-炎症因子关系对更多（Figure1B/D）。这有可能是因为在肿瘤微环境中基因表达呈现更高的异质性和多态性。这使得原本稳定相关的菌群-炎症因子关系，在肿瘤进展过程中小时。相反，我们也观测到一些在正常组织中并不相关的菌群-炎症因子关系，但是在肿瘤环境中体现出显著的相关性，例如 ABCA1 和 Agrobacterium。最终我们共获得 1250 个肿瘤组织特异性菌群-炎症因子关系对，涉及 66 个炎症因子基因和 43 个菌群（补充材料 Table S1）。

接下来我们 利用 CIBERSORT 算法对 22 种免疫细胞在乳腺癌肿瘤组织和正常组织中的浸润分数进行估计。然后比较菌群丰度和每种免疫细胞的相关性，如图 2 所示。

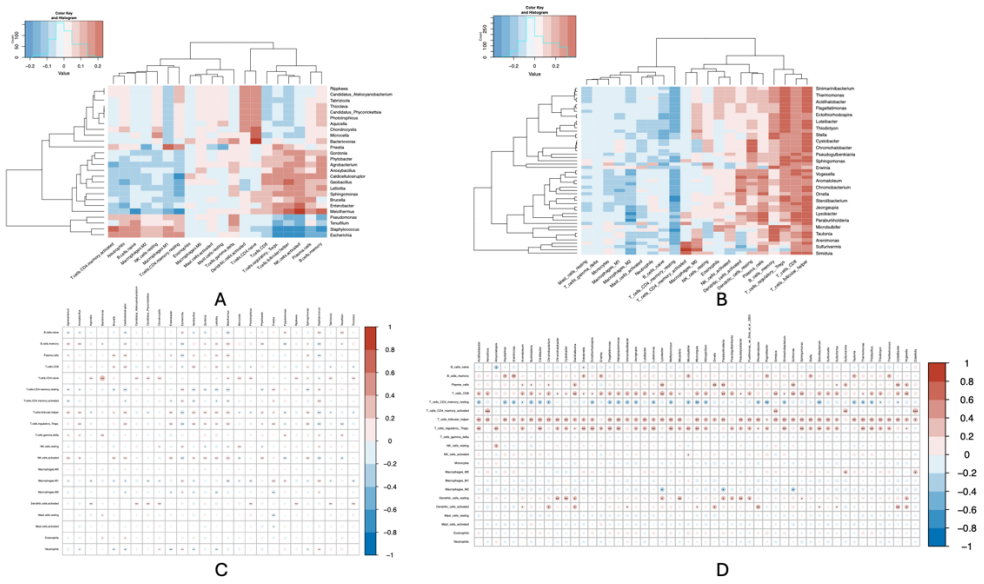


Figure 2 菌群-免疫细胞相关性分析A 肿瘤组织中菌群和免疫细胞相关性矩阵聚类 B 正常组织中菌群和免疫细胞聚类 C 肿瘤组织中菌群和免疫细胞的统计显著性 D 正常组织中菌群和免疫细胞的统计显著性。

图 2A/C 为肿瘤组织中显著相关的菌群和免疫细胞，图 2B/D 为正常组织中显著相关的菌群和免疫细胞。我们同样对肿瘤组织和正常组织进行比较分析，最终发现 180 个在肿瘤组织中特异性的菌群-免疫细胞关系对，涉及 19 个免疫细胞和 26 个菌群（补充材料 Table S2）。

2. 基于菌群-炎症因子-免疫细胞的分子亚型分析

通过菌群和炎症因子/免疫细胞关联分析，我们最终收集到 66 个炎症因子，19 个免疫细胞，以及 56 个菌群。我们以炎症因子，免疫细胞，菌群为特征，对乳腺癌肿瘤组织样本进行无监督聚类。通过 NMF 聚类算法，我们发现当聚类个数为 3 时为最优解，如图 3 A 所示。可以看出 RSS 曲线在 rank=3 后下降趋缓，Silhouette 在 rank=3 处达峰值，Cophenetic 在 rank=3 时最高，因此最终选择 rank=3 作为最佳亚群数。图 3B 为聚类个数从 2 到 5 分别对应的热图结果。

为了更直观的比较不同分子亚型的临床特征差异，我们在 ComplexHeatmap 聚类过程中补充了样本的临床信息，如图 3C 所示。首先我们发现 cluster2 中几乎所有炎症因子都体现出优势的高表达状态，因此我们标记 cluster2 为 high-cytokines 组。Cluster1 和 cluster3 均有部分 cytokines 高表达，但是二者表达趋势是相反的。这提示了 cluster1 和 cluster3 在菌群感染诱导炎症免疫过程存在显著差异的生物学过程。同时 cluster3 的免疫细胞浸润水平明显高于 cluster1，因此我们标记 cluster3 为 high-immune 组。Cluster1 标记为 low-immune 组。

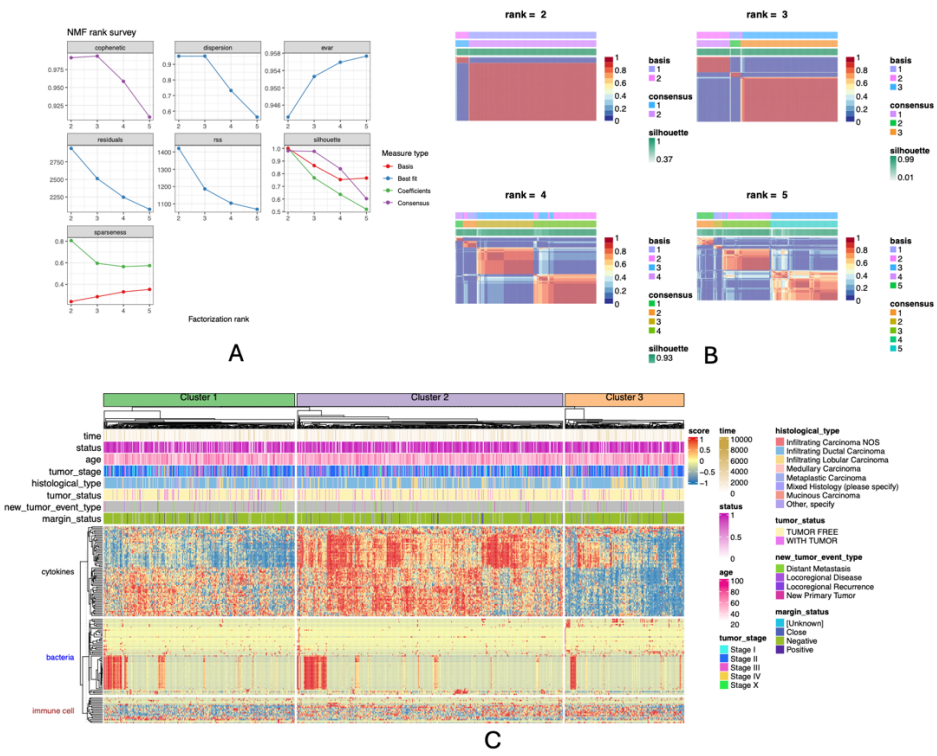


Figure 3 分子亚型分析A NMF 最佳聚类个数评估. B 2-5 聚类效果热图. C ComplexHeatmap 聚类热图结果

3. 分子亚型特异性 signature 识别

   我们利用 limma 算法分别对三个分子亚型进行差异比较，识别出每个亚型特异性的 signature。这些 signature 可认为是每个亚型特异性的分子标记物。我们分别比较了三组亚型和其他样本之间的差异，结果如图 4A-C 所示。均值分布图乳图 4D-F 所示。我们对三组亚型特异性高表达基因进行比较，通过韦恩图识别出各自亚型特异性的 signature。最后我们识别出 345 个 low-immune 组特异性 signature，1122 个 high-cytokine 组特异性 signature，786 个 high-immune 组特异性 signature，结果如图 4G 所示。亚型特异性 signature 在三组亚型中的表达分布如图 4H 所示。

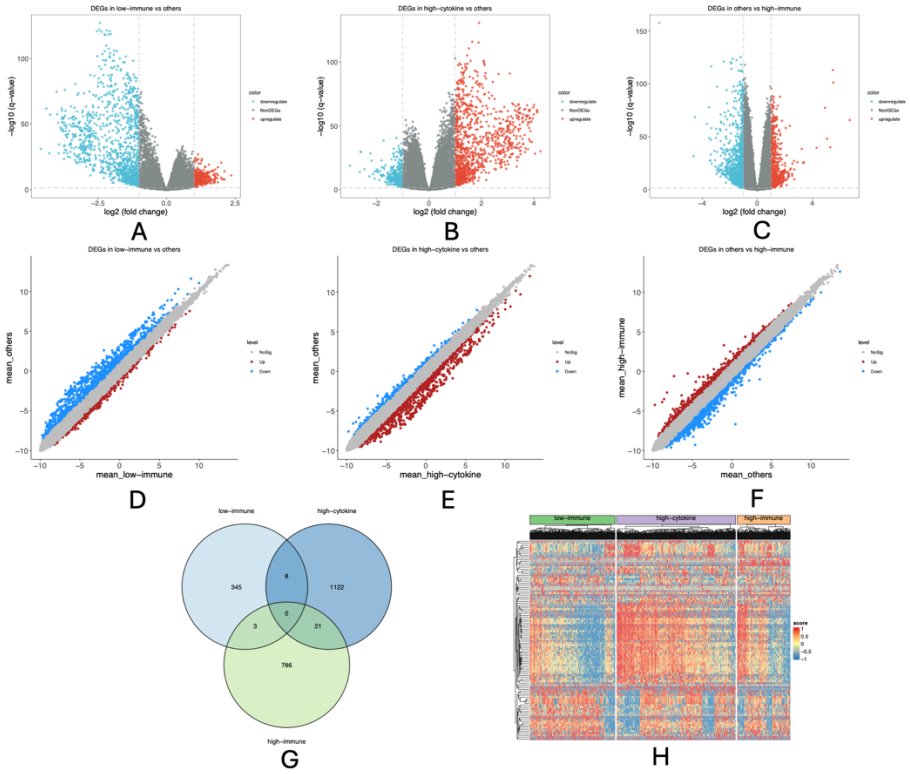


Figure 4 不同亚型之间差异比较分析

4. 上下游调控通路分析

   我们识别出了每个新亚型特异性对应的分子特征。为了进一步从生物学过程上解释每个分子亚型之间的差异，我们对所有亚型的 signature 基因进行功能富集分析，如图 5 所示。

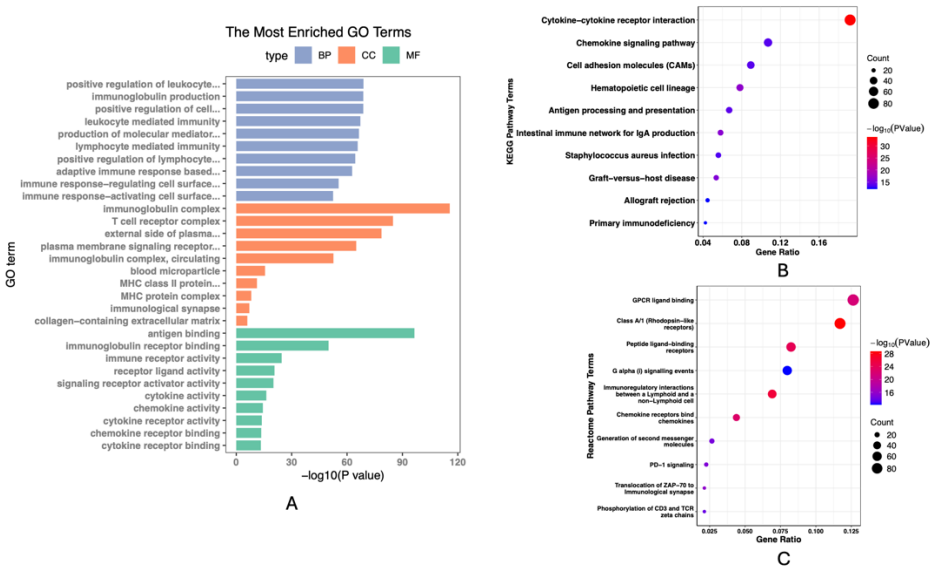


Figure 5 功能富集分析A Gene Ontology 富集分析(BP, CC, MF). B KEGG pathway 富集分析. C Reactome pathway 富集分析

图 5A 展示了 Gene Ontology 富集分析结果，蓝色为 BP， 橙色为 CC, 绿色为 MF。可以看出在 Gene Ontology 中亚型 signature 主要集中在调控炎症免疫反应相关生物学过程。尤其是白细胞介导的抗炎免疫过程（Figure5A）。白细胞的激活和肿瘤微环境细菌感染显著相关。KEGG 通路富集分析提示亚型特异型 signature 主要参与调控炎症因子互作以及其他免疫相关通路（Figure5B）。Reactome 通路分析同样提示 signature 基因主要参与调控炎症免疫相关通路（Figure5C）。

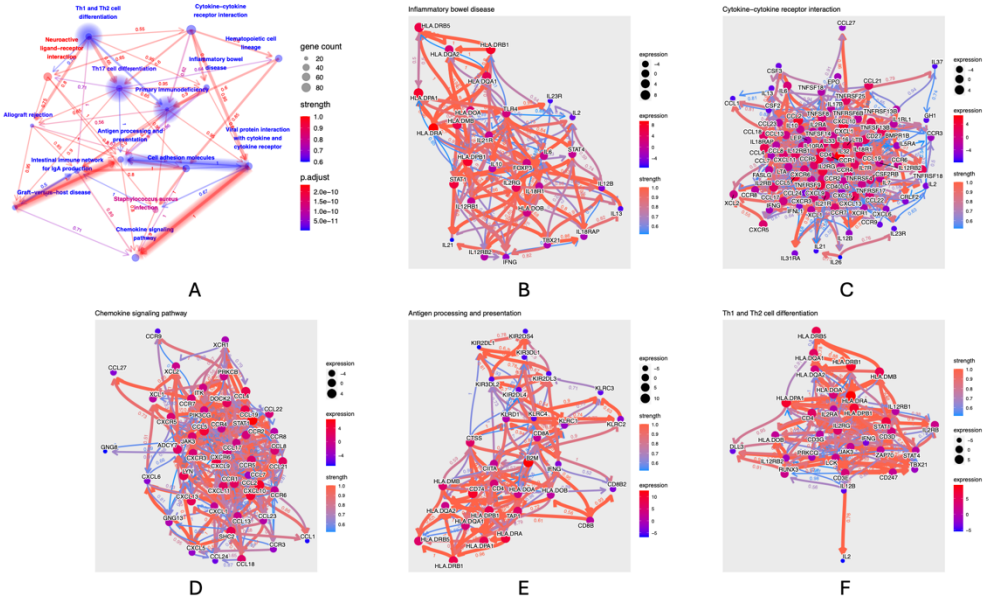


Figure 6 CBNplot 因果推断分析

我们进一步利用 CBNplot 算法对 signature 基因上下游调控关系进行因果推断，结果如图 6 所示。图 6A 为显著通路 cross talk 通讯互作网络。节点大小代表功能通路包含的基因个数，边的颜色代表互作强度，节点颜色代表通路富集的显著性。我们发现强度最高的互作关系来自 viral protein infection with cytokine and cytokine receptor to chemokine signaling pathway，以及 Graft-versus-host disease to antigen processing and presentation。同时我们深入探查了 5 个关键通路内部基因上下游调控作用关系，如图 6B-F 所示。通过上下游因果调控关系推断，我们发现 XX 通路中 XX 基因对 XX 呈现显著的调控作用。（该段落内容可以结合后续湿实验分子功能实验的结果。如无实验补充，则无需深度探讨分子机制。）

5. 预后分析

为了进一步评价不同分子亚型之间是否存在显著的预后水平差异，我们首先用单因素 cox 回归算法筛选出和预后显著相关的基因。然后利用 Lasso 特征选择算法对预后相关基因进行筛选，结果如图 6 A-B 所示。最后利用 Lasso 筛选的特

征构建多因素 cox 回归模型，特征基因的贡献通过森林图展示如图 6C 所示。列线图用于个体化预测，通过各变量的贡献分数计算总分，进而预测目标事件，结果如图 6D 所示。校准曲线评估模型的预测准确性，横轴为预测概率，纵轴为实际观察到的比例，如图 6E 所示。图中校准曲线接近对角线，代表预测结果和实际结果相符。三组分子亚型的生存曲线如图 6 G 所示，P 值小于 0.0001，尤其是 high-immune 组，预后明显好于其他两个亚型。

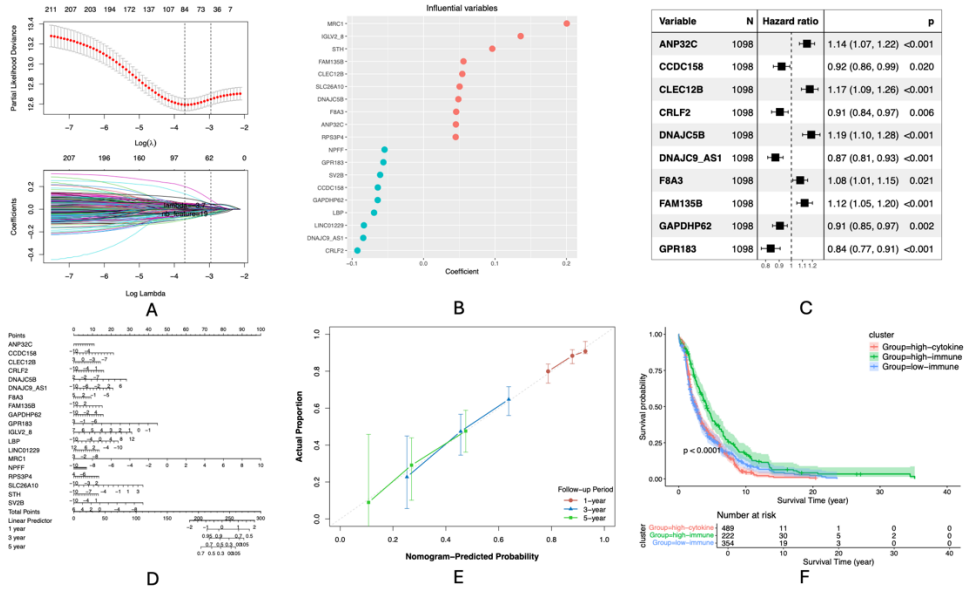


Figure 7 预后分析

6. TCGA-BRCA 分子亚型外部数据验证

我们从 GEO 数据库获取独立验证数据 GSE3494 对新亚型进行验证。首先我们在 TCGA cohort 上训练了一个随机森林模型，用于区分三个分子亚型。模型预测效能如图 8A-C 所示。 图 8A 分别 为预后基因经过特征选择后贡献度排序情况。图 6B 为根据显著特征构建随机森林模型后对三组亚型预测效能。我们发现该模型可以准确预测 high-cytokine 亚型(precision=0.87)。对其余两个亚型预测精度分别为 0.7 和 0.6。这代表我们构建的随机森林模型可以通过显著特征表达模式，从乳腺癌样本中准确快速筛选出 high-cytokine 亚型。图 8C 为 DCA 曲线，DCA 曲线测试模型在不同阈值下对应的净收益，图中结果显示阈值概率 >0.4 时（即认为干预的“必要概率”超过 40%），使用该模型指导决策能带来更高的净获益。而另外两个亚型，我们进一步构建二分类模型，准确度提高到 0.8 和 0.78（图 8D-E）。图 8E 为 ROC 曲线，综合精度为 0.87。然后我们用该模型在验证数据集上进行分类，获得三组亚型的样本标签。最后我们利用 KM 生存曲线验证三组亚型的生存水平差异如图 8F 所示，P 值小于 0.05，代表在验证数据上预测的 High-cytokine 亚型的样本生存预后和其他样本存在显著差异。

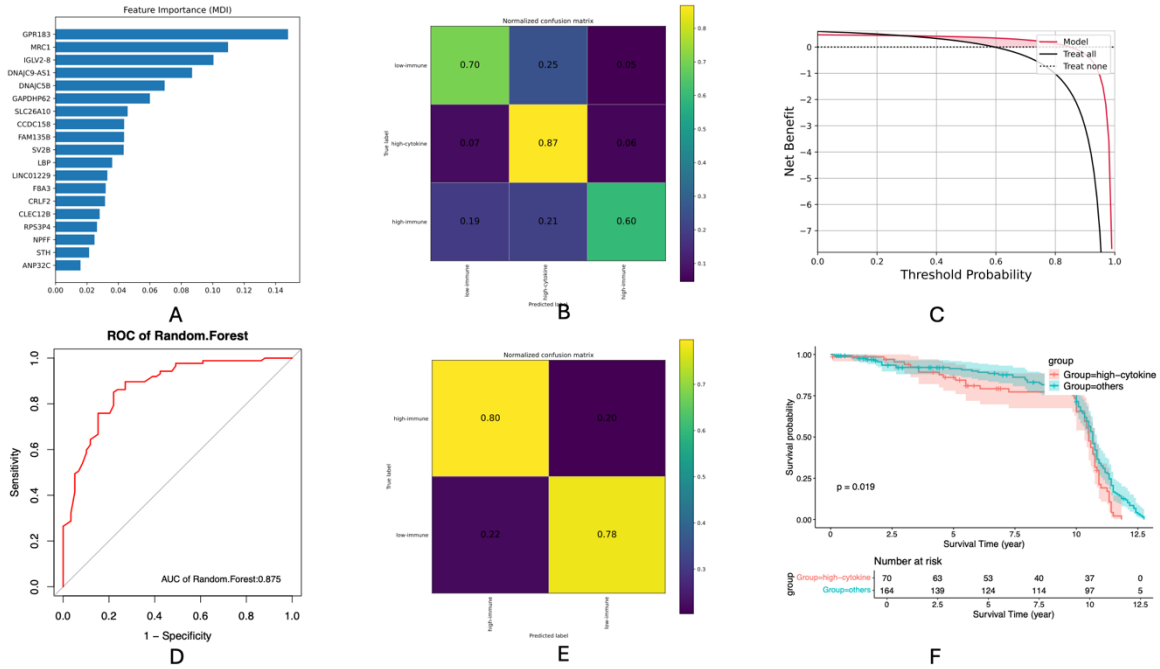


Figure 8 随机森林模型建模和效能评估

7. 炎症免疫治疗药物敏感性分析

我们利用 oncopredict 算法对 151 种药物进行敏感性分析，结果如图 9 所示。横轴为三个亚型，纵轴为药物敏感性。通过 anova 方差分析比较每种药物在三个亚型中的统计学差异，我们按照 P 值排序，取前 8 个结果进行展示。最后通过比较发现，药物 CAY10618（GPP78）在三组乳腺癌样本中呈现显著差异(Pvalue= p = 0.034)。CAY10618 尤其在 high-immune 组里具有更高的敏感性。GPP78 (CAY10618) 是一种有效的烟酰胺磷酸核糖转移酶 (Nampt) 抑制剂。GPP78 通过诱导自噬对肿瘤细胞细胞具有毒性，有抗癌和抗炎作用。因此临床上通过分子特征或免疫浸润水平筛选出 high-immune 患者，针对性给药 GPP78 可能会有显著高于其他亚型的效果。

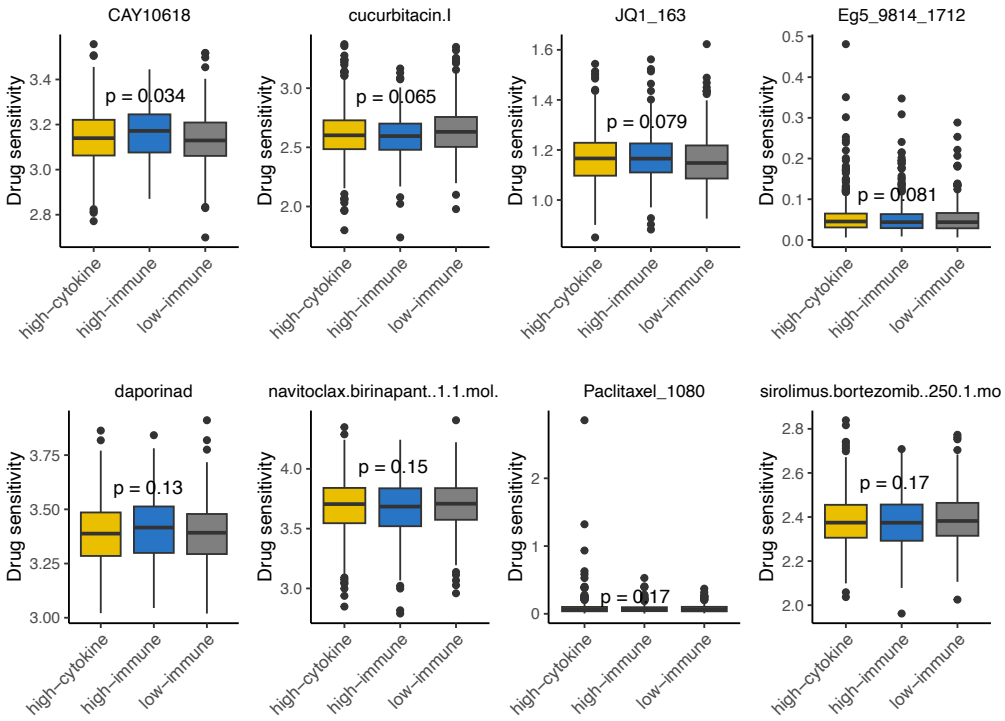


Figure 9 药物敏感性分析