

基于机器学习法初步探讨二型糖尿病发病机理

陈文豪^{1,2,3}, 郑允文^{1,2,3*}

¹ 五邑大学药学与食品工程学院 广东江门

² 华南生物医药大动物模型研究院 广东江门

³ 广东省医学大动物模型重点实验室 广东江门

【摘要】本研究旨在整合单细胞转录组学与系统生物学方法，系统解析二型糖尿病（T2D）胰岛的细胞与分子病理特征。通过对 GSE221156 数据集的高质量单细胞进行整合分析，构建了 T2D 胰岛的高分辨率转录图谱，发现 T2D 中 β 细胞和 δ 细胞比例显著降低，且 β 细胞内部功能亚群失衡。通过跨数据集差异分析获得 409 个 T2D β 细胞中共同失调的差异表达基因，并利用加权基因共表达网络分析 (hdWGCNA) 鉴定出与疾病进展紧密相关的核心模块 (red 与 turquoise 模块)。结合 LASSO 回归、XGBoost 与 SVM-RFE 三种机器学习算法，从关键模块中筛选并验证了一个由 CALR、IAPP、LAMTOR5、SCGN 组成的四基因核心集合。基于此构建的 T2D 风险预测模型具有良好的判别能力 ($AUC=0.732$)，其可视化列线图工具为风险评估提供了直观依据。本研究从单细胞尺度揭示了 T2D 胰岛细胞组成、亚群稳态及基因调控网络的多层次紊乱，鉴定出的核心基因模块为理解 T2D 发病机制及开发新型诊断标志物提供了新线索。

【关键词】二型糖尿病；单细胞转录组测序；机器学习；生物标志物

【收稿日期】2026 年 5 月 5 日 **【出刊日期】**2026 年 6 月 2 日 **【DOI】**10.12208/j.ijmd.20260048

A preliminary exploration of the pathogenesis of type 2 diabetes using machine learning methods

Wenhao Chen^{1,2,3}, Yunwen Zheng^{1,2,3*}

¹ School of Pharmacy and Food Engineering, Wuyi University, Jiangmen, Guangdong

² South China Institute of Large Animal Models for Biomedicine, Jiangmen, Guangdong

³ Guangdong Provincial Key Laboratory of Medical Large Animal Models, Jiangmen, Guangdong

【Abstract】 This study aims to integrate single-cell transcriptomics with systems biology approaches to systematically elucidate the cellular and molecular pathological characteristics of islets in type 2 diabetes (T2D). Through an integrated analysis of high-quality single-cell data from the GSE221156 dataset, we constructed a high-resolution transcriptional atlas of T2D islets, revealing a significant reduction in the proportion of β -cells and δ -cells in T2D, as well as an imbalance in functional subpopulations within β -cells. Through cross-dataset differential analysis, we identified 409 differentially expressed genes that are commonly dysregulated in T2D β -cells. Using weighted gene co-expression network analysis (hdWGCNA), we identified core modules (the red and turquoise modules) closely associated with disease progression. Combining three machine learning algorithms—LASSO regression, XGBoost, and SVM-RFE—we screened and validated a four-gene core set comprising CALR, IAPP, LAMTOR5, and SCGN from the key modules. The T2D risk prediction model constructed based on this set demonstrated good discriminatory ability ($AUC = 0.732$), and its visualized contour plot tool provided an intuitive basis for risk assessment. This study reveals, at the single-cell level, multilevel dysregulation in the composition, subpopulation homeostasis, and gene regulatory networks of T2D pancreatic islet cells. The identified core gene modules provide new insights for understanding the pathogenesis of T2D and developing novel diagnostic biomarkers.

【Keywords】 Type 2 diabetes; Single-cell transcriptomics; Machine learning; Biomarkers

*通讯作者：郑允文

二型糖尿病 (type 2 diabetes, T2D) 是一种以胰岛素抵抗和胰岛 β 细胞功能进行性衰退为核心的慢性代谢性疾病^[1]。传统研究手段难以解析胰岛内高度的细胞异质性及复杂的细胞间相互作用^[2]。近年来, 单细胞转录组测序 (scRNA-seq) 技术的发展, 使得在单细胞分辨率下绘制组织转录图谱、解析细胞异质性与细胞间通讯成为可能^[3]。同时, 机器学习算法为从高维、复杂的组学数据中稳健地筛选生物标志物和构建预测模型提供了有力工具^[4]。本研究整合多套公共单细胞数据集, 系统构建 T2D 胰岛高分辨率转录图谱, 结合多种机器学习算法筛选疾病相关核心基因并构建预测模型, 并通过虚拟敲除技术从系统层面验证其功能, 旨在为深入理解 T2D 发病机制及发现潜在诊疗靶点提供新线索。

1 材料与方法

1.1 数据获取与预处理

本研究的单细胞转录组数据来源于美国国立生物技术信息中心的基因表达综合数据库 (GEO)。我们获取了已发表的人尸体胰岛 scRNA-seq 数据集 GSE221156。作为本研究的核心发现集。所有数据分析均在 R 4.1.0 环境中进行。使用 Seurat (4.0 版本) 软件包进行标准预处理流程^[5]。首先进行严格的质控, 剔除每个细胞中检测到的基因数少于 10 或多于 5000 的细胞, 并排除线粒体基因占比超过 10% 的细胞, 以确保后续分析的可靠性。随后, 对每个样本的基因表达矩阵进行标准化 (Log Normalize), 并筛选表达变异度最高的 2000 个基因用于下游分析。为消除不同样本或数据集间的技术批次效应, 我们应用 Harmony 算法对整合后的数据进行校正。最后, 对校正后的数据执行主成分分析, 选取贡献度显著的前 15 个主成分进行 UMAP 非线性降维和基于图论的细胞聚类。

1.2 细胞类型注释与比例分析

基于 UMAP 降维和聚类结果, 我们首先使用 SingleR 软件包, 通过比对 Human RNA seq Data 参考数据集, 对各个细胞簇进行自动的细胞类型注释。随后, 结合胰岛组织的经典细胞类型标记基因进行人工校验与最终确认: 胰岛素 (INS) 标记 β 细胞, 胰高血糖素 (GCG) 标记 α 细胞, 生长抑素 (SST) 标记 δ 细胞, 胰多肽 (PPY) 标记 PP 细胞。此外, 还根据特征基因鉴定了腺泡细胞、导管细胞、内皮细胞、巨

噬细胞和星状细胞。细胞注释完成后, 分别统计并比较健康对照组与 T2D 组中各类细胞的比例差异。

1.3 β 细胞亚群分析与差异表达基因筛选

从整体细胞中提取出被注释为 β 细胞的群体, 使用 Seurat 包重新进行标准化、降维和聚类, 以探索 β 细胞内部的异质性亚群, 并比较这些亚群在健康和 T2D 状态下的比例变化。为了获得稳健的 T2D 相关差异表达基因, 我们在 GSE101207、GSE153855、GSE221156 三个数据集中, 分别对健康与 T2D 的 β 细胞进行差异表达分析。采用非参数的 Wilcoxon 秩和检验, 筛选标准为平均 $\log_2$ 倍变化绝对值大于 0.25 且经过 Bonferroni 校正后的 p 值小于 0.05。将三个数据集中鉴定出的差异表达基因取交集, 得到在 T2D 中 β 细胞中一致失调的共同差异表达基因集, 用于后续分析。

1.4 加权基因共表达网络分析

为了从系统层面识别与 T2D 病理状态协同变化的基因模块, 我们采用 hdWGCNA 方法对 T2D 样本的 β 细胞表达数据进行分析^[6]。首先, 选取合适的软阈值功率以确保构建的网络符合无尺度分布特征^[7]。然后, 使用动态树切割算法将基因划分到不同的共表达模块中。计算每个模块的特征基因 (module eigengene, ME), 并分析 ME 与疾病表型 (T2D vs 健康) 之间的相关性。选取与 T2D 状态最显著相关的模块作为关键模块, 提取该模块内的枢纽基因 (即模块内连接度最高的基因)。最后, 将这些枢纽基因与上述共同差异表达基因集取交集, 得到一个候选基因集合, 作为后续机器学习筛选的输入。

1.5 多策略机器学习筛选核心基因

为避免单一算法的偏差并提高筛选结果的稳健性, 我们采用了多步骤、多算法的集成式机器学习流程来筛选核心基因。流程如下:

初步关联分析: 对候选基因集中的基因进行单因素逻辑回归分析, 筛选出与 T2D 状态显著相关的基因。

LASSO 回归降维: 使用 glmnet 包执行最小绝对收缩和选择算子回归。通过 10 折交叉验证确定最优的正则化参数 λ , 将回归系数压缩至零, 从而进一步筛选出对模型贡献最大的变量。

多算法特征选择: 并行使用三种原理不同的特征选择算法对 LASSO 筛选后的基因进行重要性评

估与排序:

Boruta 算法: 一种基于随机森林的包装法, 通过创建“影子特征”来判定真实特征的重要性是否显著高于随机噪声。

XGBoost: 一种梯度提升决策树算法, 根据特征在构建所有树过程中的累积增益来评估其重要性。

核心基因确定: 选取在三种筛选方法 (包括 LASSO) 中被共同识别或排名靠前的基因, 确定为最终的 T2D 相关核心基因。

1.6 预测模型构建与验证

以最终确定的核心基因的表达水平作为特征, 此外, 在独立的验证数据集 (GSE153855 和 GSE221156) 中, 使用 Wilcoxon 秩和检验验证这些核心基因在健康与 T2D 样本间的表达差异是否与发现集一致, 并使用列线图预测核心基因对疾病的影响概率。

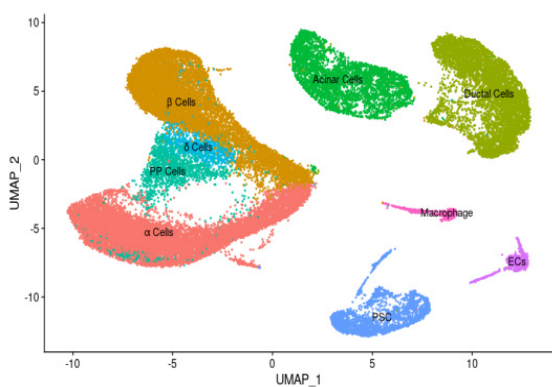
2 结果

2.1 T2D 胰岛单细胞转录图谱揭示细胞组成失衡 β 细胞异质性

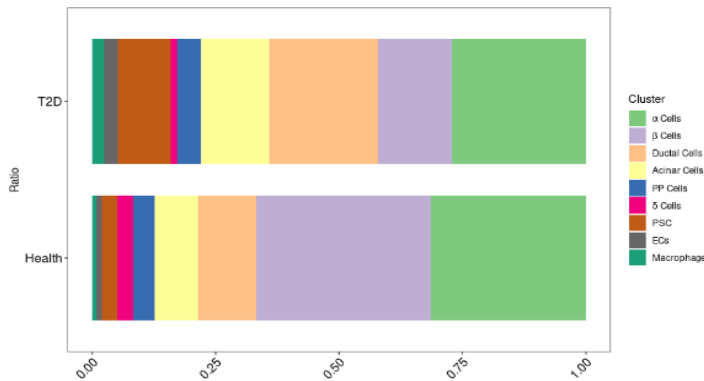
本研究整合了 GSE221156 数据集的单细胞数据,

经过严格质控和批次校正后, 共保留了 32, 322 个高质量单细胞用于构建胰岛转录图谱。UMAP 降维与聚类分析将细胞划分为 24 个簇, 通过经典标记基因, 我们成功注释出 9 种主要的胰岛细胞类型: β 细胞、 α 细胞、 δ 细胞、PP 细胞、腺泡细胞、导管细胞、内皮细胞、巨噬细胞和星状细胞 (图 1a)。细胞比例分析揭示了一个重要的病理变化, 与健康对照组相比, T2D 患者胰岛中 β 细胞和 δ 细胞的比例均显著降低 (图 1b), 这说明 T2D 的发病不仅涉及胰岛素分泌细胞 (β 细胞) 的减少, 还与分泌生长抑素、具有旁分泌调控功能的 δ 细胞减少有关, 反映了胰岛内分泌网络的整体失调^[8]。

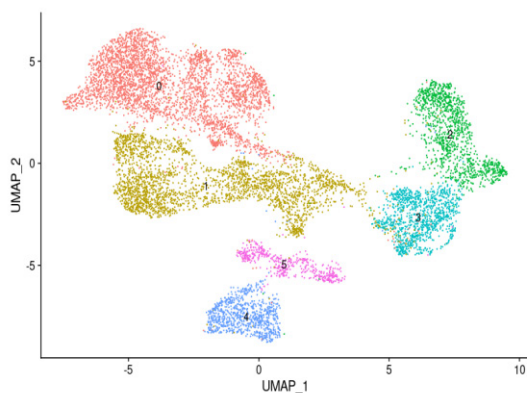
对 β 细胞进行亚群再聚类分析, 鉴定出 3 个具有不同转录特征的亚群 (图 1c), 证实了 β 细胞群体的功能异质性^[9]。比较发现, T2D 样本中这些亚群的比例构成发生了显著改变 (图 1d), 表明疾病状态破坏了 β 细胞内部的稳态平衡。为筛选稳定的 T2D 相关分子特征, 我们在三个独立数据集中进行了差异表达分析, 并取交集获得了 409 个在 T2D β 细胞中一致失调的共同差异表达基因 (图 1e)。



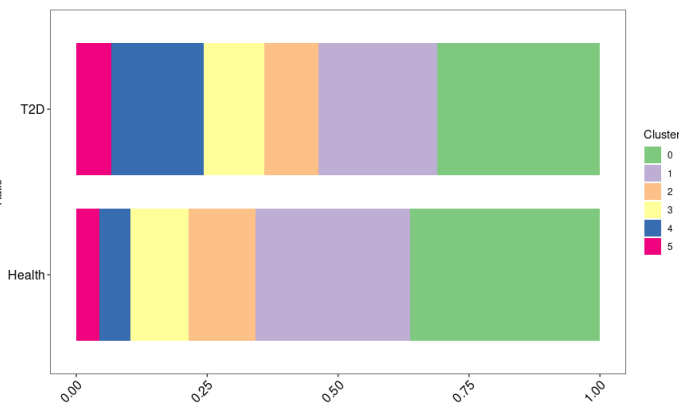
a. 细胞注释结果



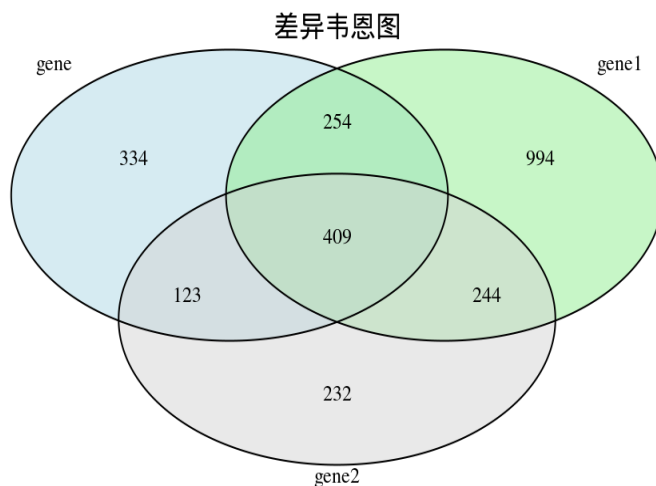
b. 不同组别的细胞比例分布



c. β 细胞亚群



d. β 细胞亚群比



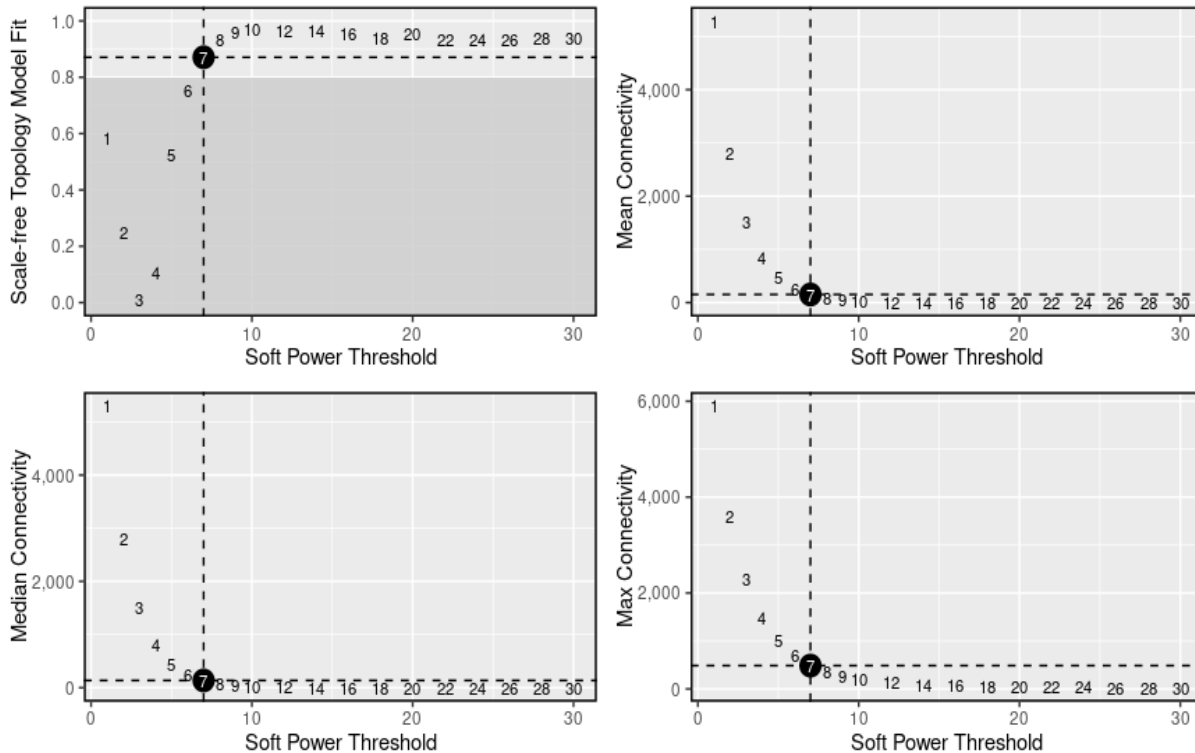
e. 三组数据差异基因交集

图 1 细胞注释及其分布情况

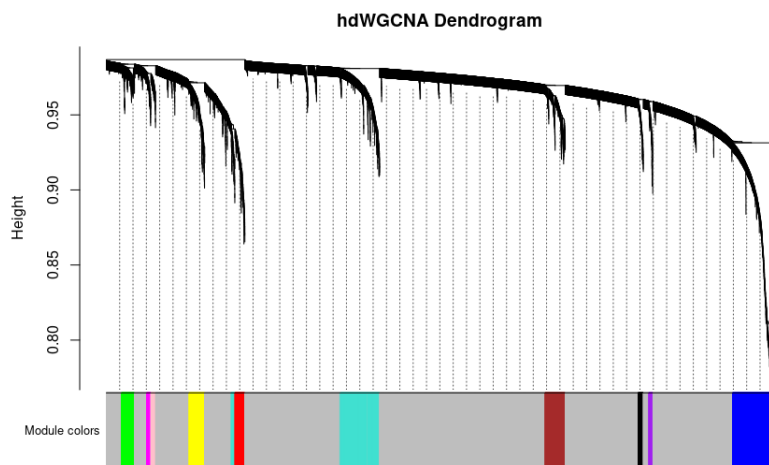
2.2 表达网络分析揭示 T2D 相关基因模块

进一步, 我们利用 hdWGCNA 对 T2D β 细胞的转录组进行了共表达网络分析^[10]。网络构建的软阈值设定为 7 (图 2a), 最终鉴定出 10 个共表达模块 (图 2b)。模块-性状相关性分析显示, 各模块在两组中表现存在显著差异 (图 2c)。接着, 我们对这

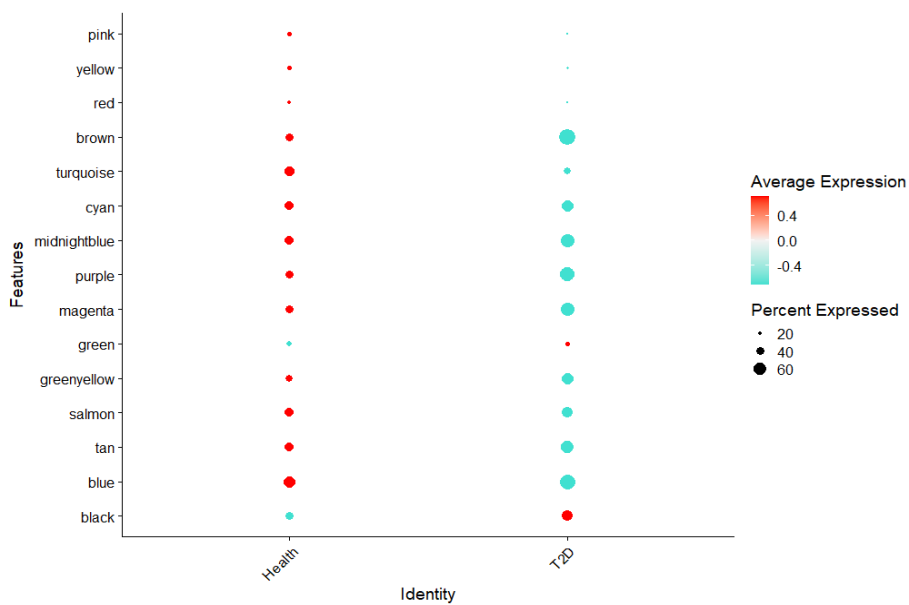
些模块进行了模块保存分析, 筛选出与疾病进展最相关的模块。结果显示, red 模块和 turquoise 模块在模块整体质量评估和基因保留率的分析中均大于 10, 符合我们的筛选便准。于是, 我们将与这两个模块中的枢纽基因与前述 409 个共同差异表达基因取交集 (图 3e), 用于后续的机器学习筛选。



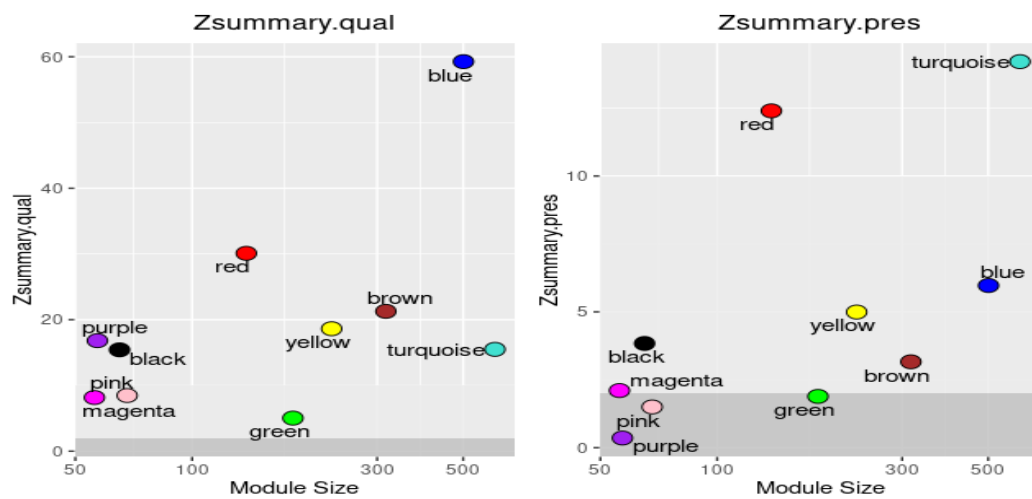
a. 软阈值的筛选



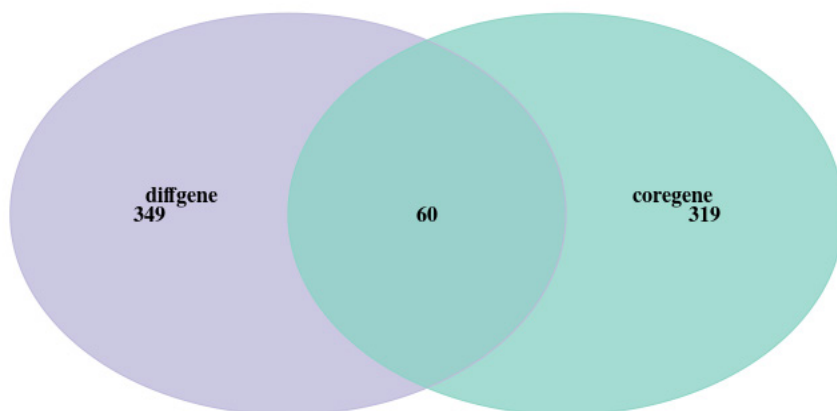
b. 基因共表达网络



c. 模块在不同组别的表达情况



d. 模块保留分析



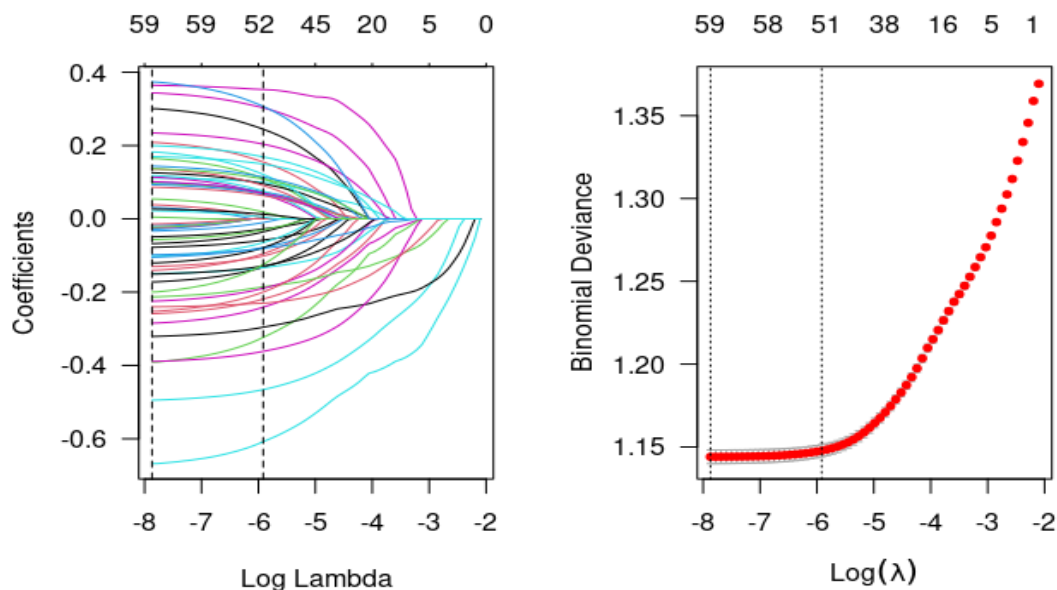
c. 差异基因与模块基因交集

图2 高维加权基因共表达网络分析

2.3 多步骤机器学习鉴定 T2D 核心基因并构建预测模型

为从候选基因集中筛选出最稳健、判别力最强的标志物,我们采用了多算法集成的策略^[11]。首先, LASSO 回归分析在最优 λ 值下保留了 51 个系数非零的基因 (图 3a)。随后,通过 XGBoost (图 3b) 和 SVM-RFE (图 3c) 两种独立算法对这些基因的重要性进行了一致性排序。综合三方法的结果,并经过在独立数据集中进行基因 wilcoxon 秩和检验的表达验证 (图 3d, e),通过比较与我们实验组的趋势,最终确定了一个由 CALR, IAPP, LAMTOR5, SCGN

四个基因组成的核心基因集合。基于这四个基因的表达水平,我们构建了列线图进一步评估核心基因组合的临床转化潜力,我们基于筛选出的特征基因构建了 T2D 风险预测列线图^[12] (图 3f)。如图所示, IAPP 和 LAMOTR 对总评分的贡献权重相对较高,表明这两个基因的表达水平在区分 T2D 与健康个体中具有较大的判别力,为 T2D 的风险评估提供了直观的量化工具。为了验证该预测工具的可靠性,我们评估了其底层逻辑回归模型的性能。结果显示,模型在测试集中的判别效能良好 ($AUC = 0.732$, 图 3g)。



a. LASSO 回归分析

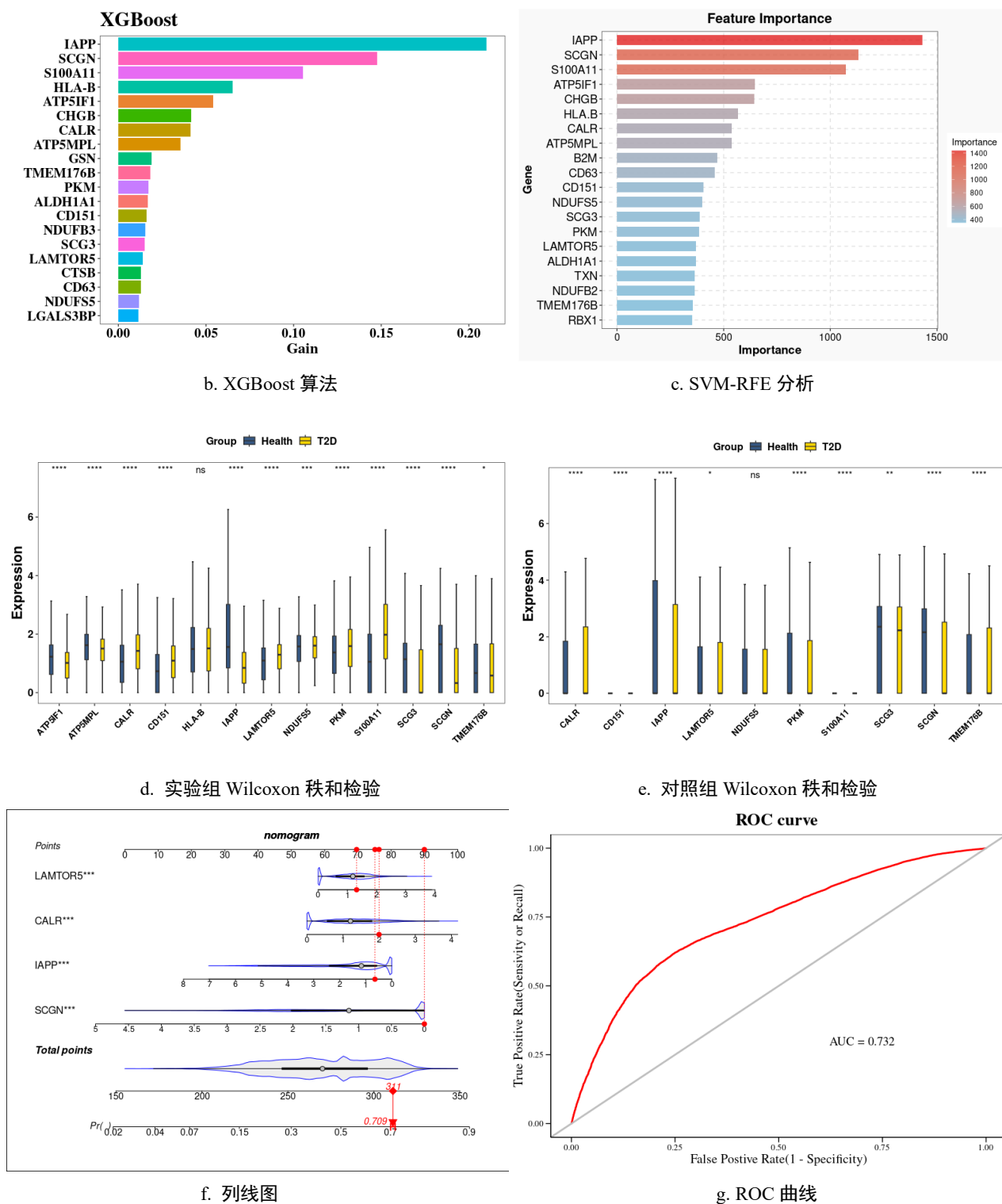


图3 多步骤机器学习结果

3 讨论

本研究通过整合单细胞转录组与机器学习, 系统描绘了 T2D 胰岛的多层次病理景观。我们不仅证实了 T2D 中 β 细胞和 δ 细胞的丢失^[13], 更重要的是

揭示了 β 细胞内部功能亚群的比例失调, 这提示 T2D 的病理进展伴随着 β 细胞功能特化的紊乱。从分子机制层面, 我们通过共表达网络分析鉴定出 red 与 turquoise 两个与 T2D 状态强相关的核心模块, 表

明存在协同调控的基因功能单元在疾病中发生紊乱。

通过多算法机器学习筛选, 我们鉴定出一个由 CALR、IAPP、LAMTOR5 和 SCGN 组成的四基因核心模块。该模块中的基因具有重要的生物学意义: IAPP(胰岛淀粉样多肽)是胰岛 β 细胞分泌激素^[14], 其异常折叠与聚集是 T2D 胰岛病理的标志之一; CALR(钙网蛋白)参与内质网应激与钙稳态调节, 而内质网应激是 β 细胞功能障碍的关键环节, 目前在 T2D 方面暂未有太多研究, 是本次研究的重点发现; LAMTOR5 是 mTORC1 信号通路溶酶体分支的关键组件, 与营养感知和代谢调控密切相关^[15]; SCGN(分泌粒蛋白)则在胰岛细胞中富集^[16], 与胰岛素分泌囊泡的成熟和释放有关。这提示该基因模块可能通过内质网应激、蛋白质稳态失衡、营养感知与分泌通路失调等多个相互关联的生物学过程, 共同驱动 T2D 胰岛 β 细胞的病理变化。

基于此四基因构建的预测模型(AUC=0.732)展现了良好的诊断潜力, 其可视化列线图工具为未来潜在的临床转化提供了基础。然而, 本研究仍存在一定局限性: 结论主要基于生物信息学分析, 需要在独立的临床队列以及通过细胞或动物实验进行功能验证; 模型未整合临床指标, 其实际应用价值需前瞻性研究评估。未来工作将深入探究这四个基因之间的调控关系及其在胰岛功能衰竭中的具体作用机制。

当然, 本研究也存在一定的局限性。所有分析均基于公共数据库的回顾性数据, 结论需要在更多的独立队列以及通过细胞、动物实验进行功能验证。研究主要聚焦于转录组层面, 未来需要整合蛋白质组、代谢组等多组学数据, 以更全面地描绘调控网络。此外, 构建的预测模型尚未纳入临床指标, 其临床应用价值有待进一步前瞻性研究评估。

4 结论

本研究从单细胞尺度系统揭示了二型糖尿病(T2D)胰岛在细胞组成、功能亚群稳态及分子调控网络层面的多层次紊乱。首先, 我们构建的高分辨率胰岛转录图谱证实, T2D 中不仅存在 β 细胞和 δ 细胞的比例丢失, 还伴有 β 细胞内部功能亚群的失衡, 这共同导致了胰岛内分泌网络的整体失调。其次, 通过整合多数据集差异分析与共表达网络(hdWGCNA), 我们锁定了一个与 T2D 病理状态

紧密相关的关键基因模块, 并进一步利用多算法机器学习筛选出一个高度稳健的四个核心基因(CALR, IAPP, LAMTOR5, SCGN)。这些基因涉及内质网应激、激素分泌、营养感知等多个关键生物学过程。最后, 基于这些核心基因构建的 T2D 风险预测模型(AUC=0.732)及其可视化列线图展现出了良好的诊断潜力。综上所述, 本研究不仅从系统层面阐释了 T2D 胰岛的病理重塑规律, 更为深入理解其发病机制及开发新型诊断策略提供了重要的分子靶点与理论依据。

参考文献

- [1] MARTÍNEZ-LÓPEZ J A, LINDQVIST A, LOPEZ-PASCUAL A, et al. Single-cell mRNA-regulation analysis reveals cell type-specific mechanisms of type 2 diabetes [J]. Nat Commun, 2025, 16(1): 9475.
- [2] XIE X, WU C, DAO F, et al. scRiskCell: A single-cell framework for quantifying islet risk cells and their adaptive dynamics in type 2 diabetes [J]. Imeta, 2025, 4(4): e70060.
- [3] BAO K, CUI Z, WANG H, et al. Pseudotime Ordering Single-Cell Transcriptomic of β Cells Pancreatic Islets in Health and Type 2 Diabetes [J]. Phenomics, 2021, 1(5): 199-210.
- [4] XIE X, WU C, YANG Y, et al. Interpretable machine learning-guided single-cell mapping deciphers multi-lineage pancreatic dysregulation in type 2 diabetes [J]. Cardiovasc Diabetol, 2025, 24(1): 300.
- [5] HAO Y, HAO S, ANDERSEN-NISSEN E, et al. Integrated analysis of multimodal single-cell data [J]. Cell, 2021, 184(13): 3573-87.e29.
- [6] MORABITO S, REESE F, RAHIMZADEH N, et al. hdWGCNA identifies co-expression networks in high-dimensional transcriptomics data [J]. Cell Rep Methods, 2023, 3(6): 100498.
- [7] LANGFELDER P, HORVATH S. WGCNA: an R package for weighted correlation network analysis [J]. BMC Bioinformatics, 2008, 9: 559.
- [8] 陈康, 傅强, 卢建强. δ 2 T 细胞在二型糖尿病中的功能及临床研究 [J]. 国际检验医学杂志, 2020, 41(13):

- 1560-2+7.
- [9] WANG J, WEN S, CHEN M, et al. Regulation of endocrine cell alternative splicing revealed by single-cell RNA sequencing in type 2 diabetes pathogenesis [J]. *Commun Biol*, 2024, 7(1): 778.
- [10] MORABITO S, MIYOSHI E, MICHAEL N, et al. Single-nucleus chromatin accessibility and transcriptomic characterization of Alzheimer's disease [J]. *Nat Genet*, 2021, 53(8): 1143-55.
- [11] YU Y S, LEE D S, LIM J H, et al. Identifying potential biomarkers for type 2 diabetes in the adipose tissue of older adults via multiple machine learning algorithms [J]. *Sci Rep*, 2025, 15(1): 44904.
- [12] WU Q, WANG Z, FAN M, et al. Integrative Multi-Omics and Machine Learning Reveal Shared Biomarkers in Type 2 Diabetes and Atherosclerosis [J]. *Int J Mol Sci*, 2025, 27(1).
- [13] LAWLOR N, GEORGE J, BOLISSETTY M, et al. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes [J]. *Genome Res*, 2017, 27(2): 208-22.
- [14] XIN Y, KIM J, OKAMOTO H, et al. RNA Sequencing of Single Human Islet Cells Reveals Type 2 Diabetes Genes [J]. *Cell Metab*, 2016, 24(4): 608-15.
- [15] LI Z, QIAO Z, ZHENG W, et al. Network Cluster Analysis of Protein-Protein Interaction Network-Identified Biomarker for Type 2 Diabetes [J]. *Diabetes Technol Ther*, 2015, 17(7): 475-81.
- [16] SHARMA A K, KHANDELWAL R, KUMAR M J M, et al. Secretagoin Regulates Insulin Signaling by Direct Insulin Binding [J]. *iScience*, 2019, 21: 736-53.
- 版权声明:** ©2026 作者与开放获取期刊研究中心 (OAJRC) 所有。本文章按照知识共享署名许可条款发表。
<https://creativecommons.org/licenses/by/4.0/>

**OPEN ACCESS**