研究文章

# 基于样本特异性因果网络熵（SCNE）揭示疾病进展过程中的预恶化状态

钟佳媛 , 唐慧 , 黄子怡 , 柴华 , 凌菲 , 陈佩 , 和睿

佛山大学数学与大数据学院，中国佛山 528000。 华南理工大学生物与生物工程学院，中国广州 510640。 华南理工大学数学学院，中国广州 510640。

\*通信地址： fling@scut.edu.cn (F.L.); chenpei@scut.edu.cn (P.C.); scliurui@scut.edu.cn (R.L.)

这些作者对本研究做出了同等贡献。

复杂的疾病并不总是循序渐进的。相反，它们可能会经历被称为临界状态或临界点的突然转变，在那里发生明显的质变。检测这种关键的过渡或恶化前状态至关重要，因为它与严重的疾病恶化有关。然而，精确定位复杂疾病的恶化前状态仍然是一个障碍，尤其是在涉及高维数据和样本有限的情况下，传统的统计方法往往被证明是不够的。在这项研究中，我们引入了一种创新的定量方法，即样本特异性因果关系网络熵（SCNE），它可以推断出每个个体的样本特异性因果关系网络，并有效量化分子间因果关系的动态变化，从而捕捉复杂疾病的临界点或恶化前状态。我们通过数值模拟和各种实际数据集，包括结直肠癌上皮细胞退化（EPCD）的单细胞数据、流感感染数据以及癌症基因组图谱（TCGA）库中的三种不同肿瘤病例，证实了我们方法的准确性和有效性。与其他现有的六种单样本方法相比，我们提出的方法在识别临界信号或恶化前状态方面表现出卓越的性能。此外，通过分析信号生物标志物的功能，也能凸显计算发现的功效。

引用：Zhong J, Tang H, Huang Z, Chai H, Ling F, Chen P, Liu R. 《基于样本特异性因果网络熵（SCNE）揭示疾病进展过程中的预恶化状态》（Uncovering the Pre-Deterioration State during Disease Progression Based on Sample-Specific Causality Network Entropy (SCNE).Research 2024;7:Article 0368. https://doi.org/10.34133/research.0368

2023 年 12 月 10 日提交 2024 年 4 月 6 日接受 2024 年 4 月 29 日发表

Copyright © 2024 Jiayuan Zhong et al.独家授权科学技术评论出版社。不对美国政府作品原件提出权利主张。以知识共享署名许可协议 4.0 (CC BY 4.0) 分发。

# 导言

复杂的疾病通常是由环境或遗传因素引起的体内平衡改变造成的。大量实验和临床证据表明，复杂疾病的演变并不总是以渐进模式为特征，而是在达到临界过渡或临界点时，系统状态会发生突然的质变 。因此，撇开各种疾病在临床表现和生物机制上的特殊差异不谈，疾病的演变可分为三种不同的状态：相对正常的稳定状态、以复原力减弱和易感性增强为特征的恶化前状态，以及另一种稳定的恶化状态（图 1A）。恶化前状态是相对正常状态与恶化状态之间的临界点，在疾病进展过程中具有重要意义。如果复杂的疾病越过了这种恶化前的状态，病情就会迅速恶化，最终形成恶化后的状态。与不可逆转的恶化状态不同，确定恶化前的状态可以提供缓解进一步恶化的潜力，并利用合适的干预方法帮助控制疾病的进展。然而，准确定位复杂疾病的恶化前状态或临界点是一项艰巨的任务，因为在疾病恶化前，分子表达或临床表型的改变可能微乎其微[3]。

最近，一种被称为动态网络生物标志物（DNB） 的量化方法被用于利用一组在疾病进展过程中集体波动的分子来辨别临界点。DNB 方法在各种疾病和生物现象中的有效性显而易见，它已被用于识别疾病前状态 ，检测细胞命运转变 ，以及研究免疫检查点阻断 [9]。然而，评估统计条件需要多个样本，这就带来了挑战，因为在现实世界中获取每个个体的多样本数据具有挑战性，从而限制了 DNB 方法及其在生物研究和临床环境中的扩展。针对这一问题，人们开发了许多不同的单样本方法，包括单样本景观熵（SLE）[10]、单样本网络模块生物标志物（sNMB）[11]、基于单样本的隐马尔可夫模型（sHMM）[12]、时序网络流熵（TNFE）[13]、个性化动态网络生物标志物（PDNB）[14]和景观动态网络生物标志物（LDNB）[15]，以利用特定样本量化复杂疾病的危急程度。尽管如此，这些方法主要侧重于利用从大块 omics 数据中获得的临界状态的动态特征来识别临界转换的早期指标，但由于存在高噪声数据，特别是在处理单细胞数据时，仍然会遇到与鲁棒性有关的挑战。因此，人们迫切需要开发创新的单样本方法，这种方法既能适用于大样本数据，也能适用于单细胞表达数据，从而能够检测复杂疾病的恶化前状态，并预测与疾病进展有关的关键分子。

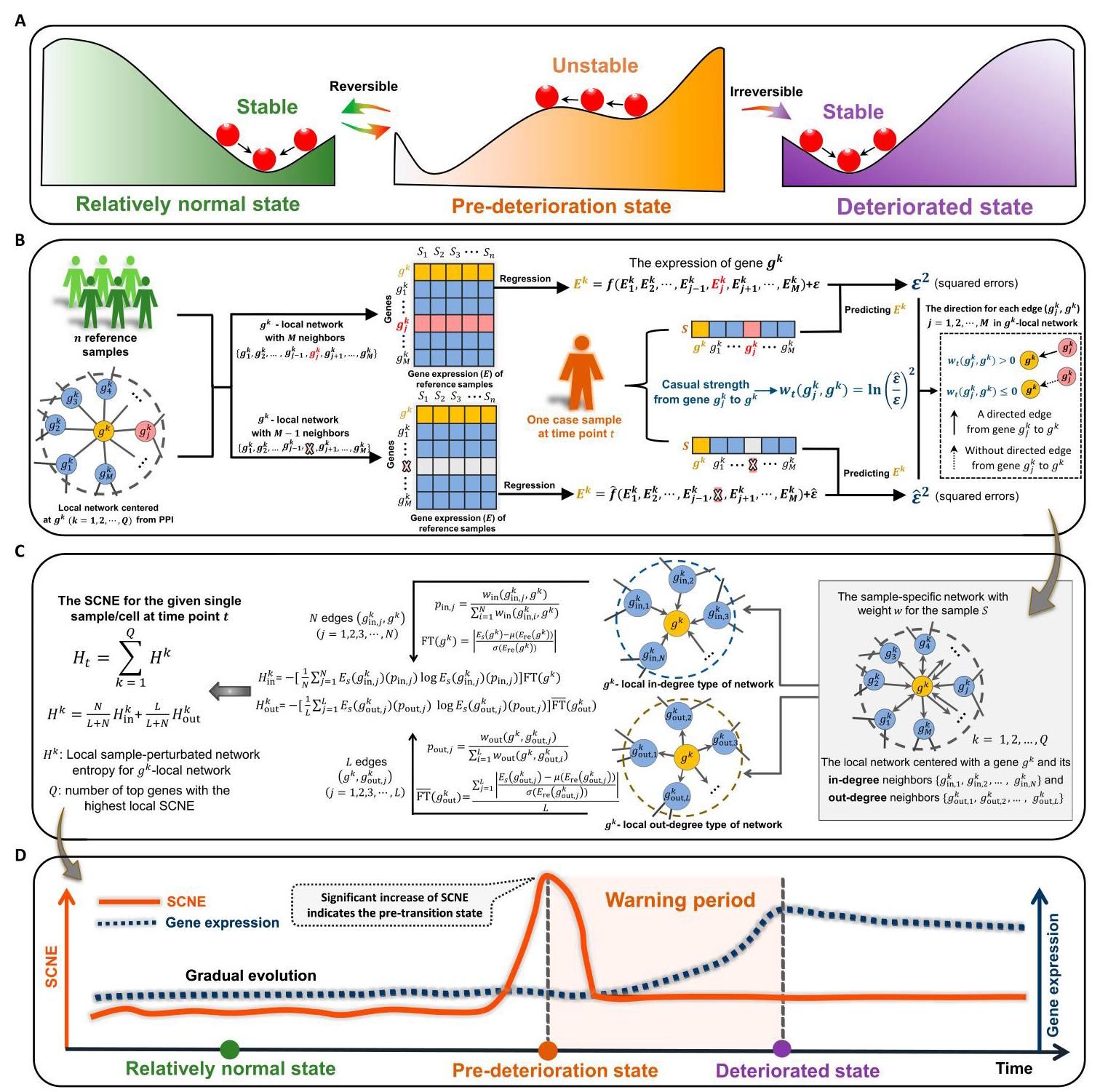


图 1.利用 SCNE 揭示退化前状态的示意图。(A) 疾病演变可分为三种不同的状态：稳定的相对正常状态、恶化前状态和另一种稳定的恶化状态。退化前状态的特点是复原力减弱和易感性增加，通过适当的干预措施通常可以恢复到相对正常的状态。(B) 在针对一组给定参考样本的验证预测指导下，通过利用植根于因果推理的统计概念，构建特定样本因果关系网络。(C) 为每个局部因果关系网络计算局部 SCNE，并对特定样本触发的分子间因果关系与参考样本的动态变化进行定量测量。(D) SCNE 指数用于量化复杂疾病的危急程度，它的显著增加是接近恶化前状态的指标。

随着高通量测序技术的快速发展，许多推断因果网络的方法被引入，如基于交叉图的框架 [16]、GRNBoost2 [17]、DNRS [18] 和 NME [19]。然而，这些方法包含了基于特定时间多样本数据的因果关系推断，这限制了它们在真正的个体化临床医学中的应用。

在这项研究中，我们提出了一种创新的定量方法，即样本特异性因果关系网络熵（SCNE），它可以推断出每个样本/细胞的样本特异性/细胞特异性因果关系网络，并通过量化分子间因果关系从相对正常状态到临界状态的动态变化，有效地作为恶化前状态的指标。具体来说，在针对一组来自相对正常状态的参考样本进行验证预测的指导下，利用植根于因果推理的统计概念（图 1B）[20]，实现对特定样本因果关系网络的推断。随后，计算每个局部因果关系网络的局部 SCNE，测量特定样本/细胞触发的分子间因果关系与参考样本/细胞间因果关系的动态变化（图 1C）。复杂疾病的临界点可通过 SCNE 量化，其明显增加可作为临界点或恶化前状态的指标（图 1D）。为了展示 SCNE 的稳健性和有效性，我们对受到不同程度噪声影响的模拟数据进行了数值模拟验证。随着噪声强度的增加，与其他已有的单样本方法相比 [11-15]，我们提出的方法在捕捉即将到来的临界点方面表现出一贯的稳定性和鲁棒性。同样，我们的 SCNE 方法在实际数据中也表现出了更好的性能，包括 TCGA 数据库中的肾透明细胞癌（KIRC）、胃腺癌（STAD）和肺腺癌（LUAD）。此外，通过将 SCNE 方法应用于结肠直肠癌上皮细胞恶化（EPCD）单细胞数据和流感感染数据，我们成功地发现了复杂疾病的关键信号，这表明预测的前恶化状态在严重疾病恶化开始之前就已显现。上述发现与临床和实验观察结果一致。此外，还进行了功能分析，以评估相应的 SCNE 信号生物标志物的有效性。简而言之，我们从单样本数据的角度介绍了一种新的计算方法，即 SCNE，它在批量和单细胞表达数据的分析中都表现出了很高的有效性和鲁棒性，为临床应用中的个体疾病诊断和精准医疗提供了一个全新的视角。

# 成果

# 验证 SCNE 方法的有效性和稳健性

为了评估所提出的 SCNE 方法的有效性和适应性，我们设计了一个由 18 个节点组成的调控网络（图 S1），该网络由随机微分方程系统控制，如式（1）所示。S1 用于生成模拟数据，以确定临界阶段，因为系统接近分叉点。这种调控网络模型通常以迈克尔-门顿（Michaelis-Menten）形式表示，常用于描述各种生物现象中的基因调控网络 。通过在 -0.50 至 0.15 的范围内改变参数 ，可以从调控网络中生成模拟数据。有关动态系统的更多详情，请参阅补充材料 A 部分。

如图 2A 所示，在特定参数 （分叉点）附近，观察到 SCNE 分数突然急剧上升，表明临界点即将到来。为了更直观地展示相对正常状态和恶化前状态之间的不同动态，图 2B 展示了不同节点的局部 SCNE 的景观演变。显而易见，当系统远离临界点时，所有节点的局部 SCNE 分数都持续较低，但当系统接近临界点时，特定节点（即 DNB 成员）的局部 SCNE 分数会急剧上升。此外，图 2C 显示了调控网络的动态演变，在临界点附近，由 DNB 成员组成的子网络的配置发生了明显变化，表明网络状态即将发生转变。此外，为了展示所提方法的复原能力，我们使用受到不同程度噪声干扰的样本，对 SCNE 和其他现有的单样本方法 [11-13] 进行了比较分析（图 2D）。随着噪声强度的增加，SCNE 方法在识别生物过程临界点方面表现出更强的鲁棒性和有效性，这体现在它始终能够在噪声较强的情况下产生灵敏度更高、表观分数更高的临界信号。此外，我们还利用六节点调控网络生成模拟数据，以探讨 DNB 理论的关键指数与 SCNE 指数之间的关系（图）。S3 和 S4）。上述数值模拟说明，我们提出的 SCNE 方法有能力仅从特定样本中提取高维信息。

# 识别个人流感感染的恶化前状态

在与流感感染有关的时间历程数据集中，一组 17 人被分为两个不同的类别：9 名有症状的受试者出现流感感染的临床症状，8 名无症状的受试者没有临床症状。如图 3A 所示，这些个体的基因表达数据是在 16 个不同的采样时间点采集的，时间跨度从 -24 到 。在每个人的情况中，前四个时间点的基因表达谱被视为参考组，代表其相对健康的状态。个性化的 SCNE 分数（在公式中定义为 ）。9）的算法计算出 17 个个体中的每个个体。SCNE 分数的增加可作为疾病发生的早期预警信号，特别是表明临床症状出现的阶段。在 9 名有症状的受试者中，SCNE 分数在流感症状出现前明显上升，而 8 名无症状的受试者的 SCNE 分数则相对稳定（图 3B）。因此，在九名有症状的受试者中可以观察到流感症状的早期预警信号，而在八名无症状的受试者中则看不到此类信号。图 3C 显示了九名症状受试者的个性化 SCNE 分数，揭示了每个症状受试者在出现临床症状之前的退化前状态。因此，我们的 SCNE 方法令人信服地证明了它能够有效、准确地识别单个流感病毒感染的恶化前状态。

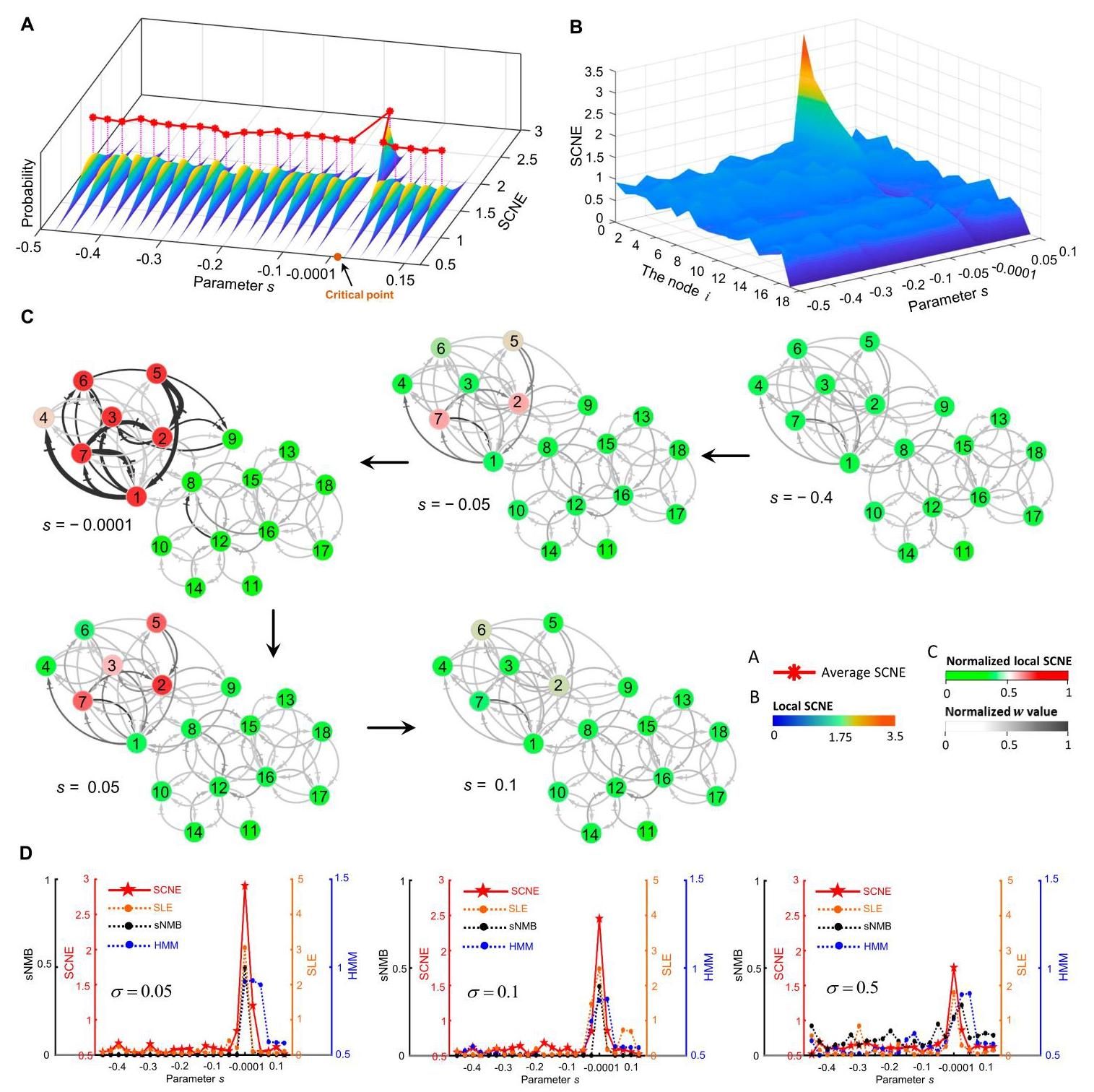


图 2.通过数值模拟验证 SCNE 方法的性能。(A) 很明显，SCNE 分数在临界点附近突然增加。(B) 描绘了不同节点的局部 SCNE 地形演变。值得注意的是，当系统接近分叉点时，DNB 成员的局部 SCNE 出现了明显的上升。(C）调控网络的动态演化揭示了临界点附近由 DNB 成员组成的子网络结构的显著变化。(D) 比较了 SCNE 和以前的单样本方法的性能。在低噪声水平下（ =0.05 或 0.1），以检测临界转换为目标的 SCNE 方法与单样本方法没有明显区别。然而，当噪声水平较高时（ =0.5），SCNE 方法表现出更高的鲁棒性和有效性。

# 识别肿瘤疾病恶化前的状态

为了评估 SCNE 方法在识别肿瘤疾病恶化前状态方面的有效性，我们对从癌症基因组图谱（TCGA）数据库中获得的三个肿瘤数据集（KIRC、STAD 和 LUAD）采用了这种方法。对于每个肿瘤样本，我们都计算了个体特异性 SCNE 分数（如公式中的定义）。9）的算法。每个阶段的 SCNE 平均值被用来定量测量退化前的状态。我们提出的方法分析表明，KIRC 的退化前状态被确定为 II 期，STAD 为 IIIA 期，LUAD 为 IIIA 期（图 4A 至 C）。具体而言，如图 4A 所示，在 KIRC 数据集中，SCNE 分数在第一阶段和第二阶段之间显著增加 ，这意味着在第二阶段之后出现了临界恶化事件。事实上，III 期的特点是肾脏周围的脂质水平迅速升高，随后肿瘤侵入肾静脉[23]。在 STAD 数据集中，SCNE 评分在 IIIA 期（图 4B）出现了显著的过渡 ，此后肿瘤向邻近组织浸润或向远处器官扩散，最终导致远处转移[24]。同样，在 LUAD 数据集中，SCNE 评分在 IIIA 期急剧上升 ，表明病情即将严重恶化，这与肿瘤细胞在 IIIB 期至 IV 期表现出向远离原发部位的组织或器官浸润的能力这一观察结果相吻合[25]。然而，从图 4A 至 C 中深蓝色代表的曲线中可以观察到，差异表达基因的基因表达量并没有显示出临界转换的信号。此外，与其他现有的六种单样本方法 [11-15]（表和图 S6）相比，我们提出的方法在揭示疾病进展过程中的预恶化状态方面表现出更好的性能。此外，我们还对我们提出的 SCNE 与另一种方法进行了比较分析，后者利用的是背景蛋白质-蛋白质相互作用（PPI）网络，没有引入区别。结果表明，我们提出的 SCNE 方法提供的信号明显强于其他方法（图 S7）。

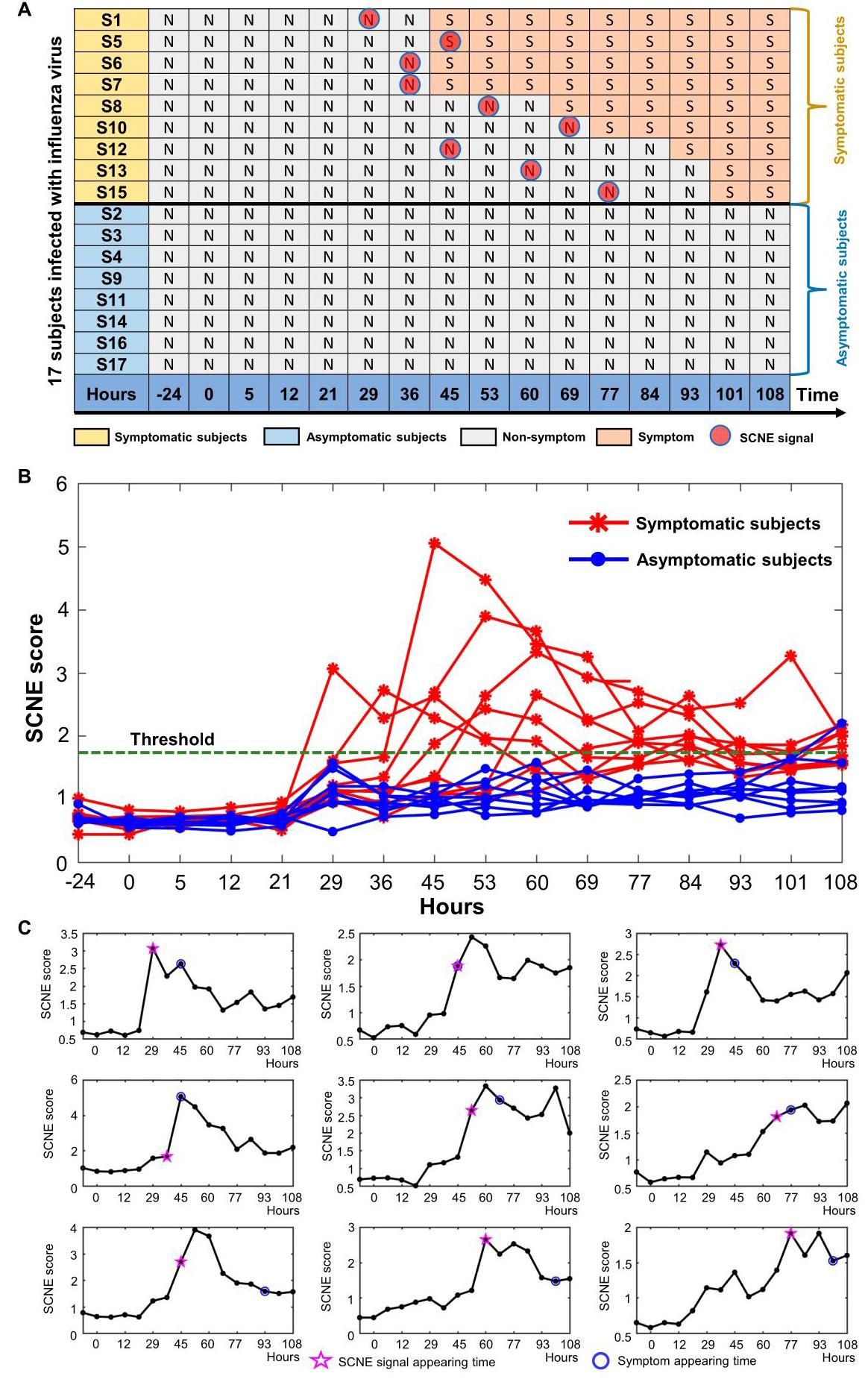


图 3.确定流感感染者的状态。(A) 总表显示了流感感染数据集中 17 人的综合信息。(B) 17 个个体的 SCNE 分数曲线。红色曲线表示 9 名有症状受试者的 SCNE 分数，蓝色曲线表示 8 名无症状受试者的 SCNE 分数。(C) 九名症状患者的个性化 SCNE 评分曲线。蓝色圆圈表示临床观察到最初流感症状的时间，橙色星星表示 SCNE 分析确定的恶化前状态。

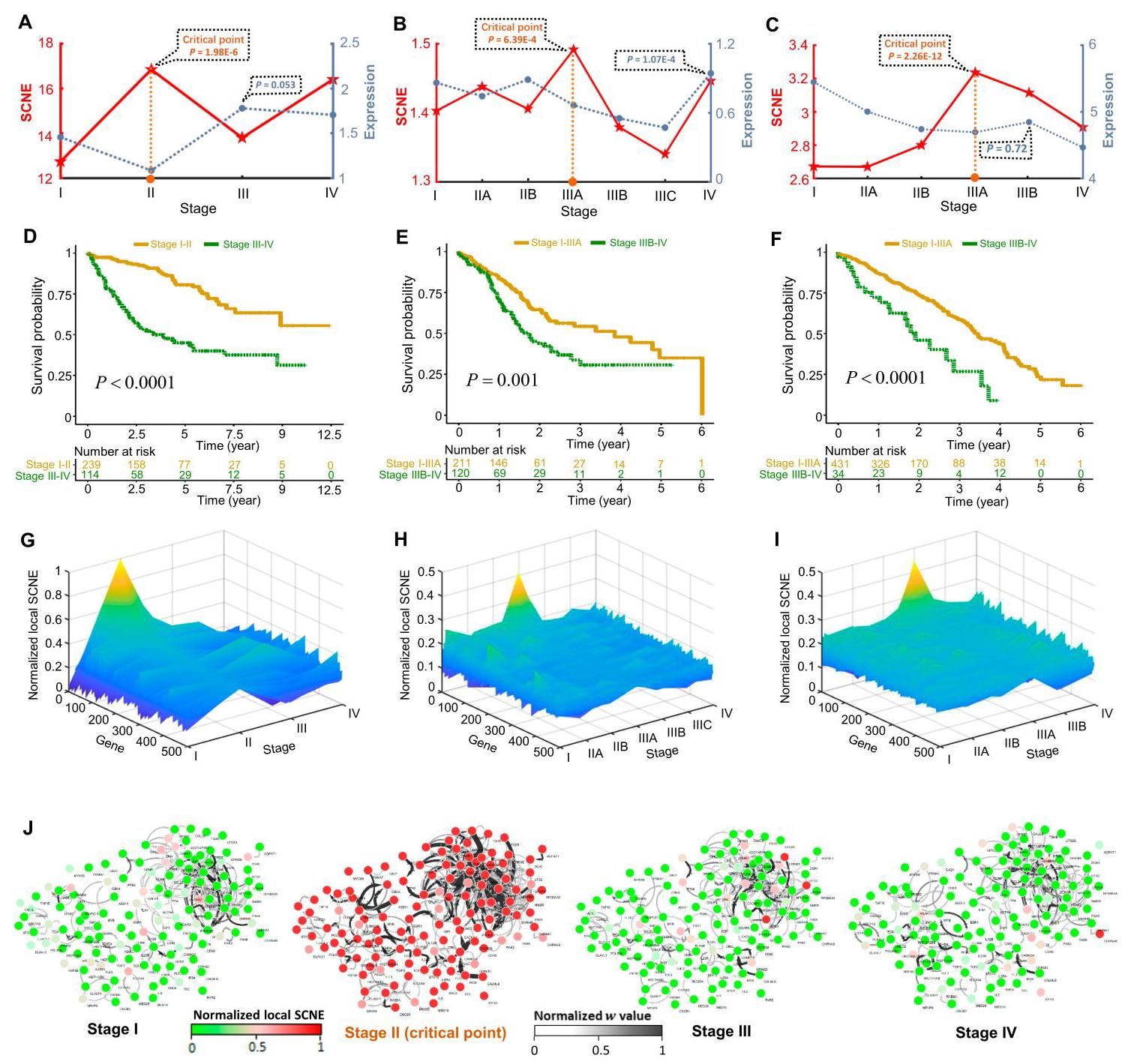


图 4.确定肿瘤疾病的恶化前状态。在三种肿瘤类型中分析了 SCNE 的动态性能和基因表达：(A) KIRC、(B) STAD 和 (C) LUAD。比较了三种肿瘤类型在恶化前和恶化后的存活时间：(D) KIRC、(E) STAD 和 (F) LUAD。我们为三个数据集生成了信号基因和非信号基因的局部 SCNE 评分图谱：(G) KIRC、(H) STAD 和 (I) LUAD。(J) KIRC 数据集展示了信号基因网络的动态演化。

为了验证已确定的恶化前状态，我们根据卡普兰-梅耶（log-rank）生存分析法，分别对临界转换前后的样本进行了预后分析。如图 4D 至 F 所示，恶化前状态前后的存活率曲线表现出明显的差异，KIRC、STAD 和 LUAD , 和 , 分别观察到显著的 值）。这些结果突出表明，与达到恶化前阶段后诊断的患者相比，在确定的恶化前状态之前诊断的患者预后明显改善。有关肿瘤存活率分析的更多详细信息，请参阅图 5。S8 和 S9。因此，SCNE 评分能够识别与存活时间相关的恶化前状态的预警信号。此外，为了全面概述局部 SCNE 的变化，图 4G 至图 I 显示了信号基因和非信号基因的局部 SCNE 图谱，其中一组信号基因在退化前状态或临界点时显示出局部 SCNE 的突然激增。此外，如图 4J 所示，我们直观地展示了 KIRC 数据集中信号基因（SCNE 得分最高的前 5%基因）构建的网络的动态演化。在临界点上，网络结构发生了明显的变化，标志着向疾病恶化的临界过渡。

# 识别结直肠癌 EPCD 的预恶化状态

为了深入了解结直肠癌发生的分子机制，研究人员利用良性细胞、TUBA1B+H2AFZ+ HMGB2+ HIST1H4C+ 细胞和恶性细胞这三个不同的亚群建立了 EPCD 的轨迹（图 5A）。通过分析基因表达模式，EPCD 的进展被分为六个不同时期或群组（图 5B）。具体细节见我们之前的研究[26]。图 5C 显示，第 4 组的 SCNE 评分出现了明显的变化 ，这表明了向 EPCD 过渡的关键信号，并揭示了退化前上皮细胞亚群。此外，与现有的其他六种单样本方法相比，我们提出的 SCNE 在检测 EPCD 期间的前恶化状态方面表现出更优越的性能（表）。在已确定的恶化前状态中，代表局部 SCNE 值升高最多的前 5%的基因子集被选为信号基因。此外，还利用信号基因及其相邻差异表达基因构建的调控网络来探索肿瘤进展的网络水平分子调控机制。如图 5D 所示，在退化前状态之后，网络内的基因表达模式发生了明显的变化，基因表达水平出现了明显的变化，从高到低或相反。图 显示，这些相邻的差异表达基因（DEGs）在癌症相关信号通路中表现出显著的富集，包括磷脂酰肌醇 3 激酶（PI3K）-Akt 信号通路[27]、细胞衰老[28]和 FoxO 信号通路[29]。

图 5F 显示，对核心基因及其一阶邻近基因进行的功能分析揭示了 FoxO 信号通路中与细胞周期调控相关的潜在机制。FoxO 信号通路是一种负责管理细胞增殖和凋亡的平衡驱动机制，与多种肿瘤类型的发展、侵袭和转移密切相关 。如图 5F 所示，在上皮细胞退化过程中，观察到信号基因 的表达明显增加，而一阶相邻基因 和 P27 的表达明显减少，这表明 对 起着负向调控作用。随后，下游分子（包括与增殖相关的基因，如 和 ）的表达水平也明显升高，这意味着在 EPCD 进展过程中，细胞周期可能会被打乱，并可能出现异常增殖。因此，活化的 SGK 主要通过磷酸化引发 FOXO 转录因子的构象变化，导致其转录活性减弱。这一错综复杂的过程会导致细胞周期抑制因子 P27 的表达减少，进而促进细胞异常增殖，增强癌细胞的存活和繁殖能力 [32]。有研究表明，P27 在良性病变和正常组织中的表达水平远远高于恶性组织[33]，这与我们的研究结果一致。总之，SGK 基因对一阶相邻 的表达产生抑制作用，最终导致细胞周期抑制因子 P27 的下调。这种分子级联最终导致细胞异常增殖，有可能促使前体上皮细胞从衰退状态发展为末期恶性上皮细胞。如图 5G 所示，在转化生长因子- (TGF- ) 信号通路的背景下，MMP7+ FABP1+ TFF1+ CKB+ 上皮细胞亚群（群组 6）和巨噬细胞亚群向 TUBA1B+ 上皮细胞亚群（群组 4）传递了强有力的增殖信号，这可能表明上皮细胞的恶化即将加剧。此外，在肿瘤微环境（TME）中，巨噬细胞在TGF- 信号通路中表现出最高的受体-配体通信概率（图S1o），这意味着它们对EPCD的传播有潜在贡献。此外，TGF- 还可以通过调节 与其他蛋白质之间的相互作用来调节 FoxO 信号通路的活性，从而影响细胞的增殖和存活 [34]。TGF- 可与 TGFBR2 和 TGFBR1 形成稳定的异质四聚复合物，在调节上皮细胞的增殖动态方面发挥着重要作用 [35]。在我们的研究中， 作为信号基因，而 TGFBR2 和 TGFBR1 作为一阶邻接基因。它们之间的相互影响以及对其他信号基因的潜在反馈最终可能有助于调节细胞内的通讯和细胞增殖的动态。

表.不同单样本检测方法的性能比较

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 数据集 | KIRC | STAD | LUAD | EPCD |
| SCNE | 第二阶段（P = 1.98E-6） | IIIA 阶段（P = 6.39E-4） | IIIA 阶段（P = 2.26E-12） | 第 4 组（P = 1.43E-10） |
| 系统性红斑狼疮 | 无 | IIIA 期（P = 0.039） | IIIB 期（P = 5.62E - 6） | 第 6 组（P = 2.18E-5） |
| SNMB | 无 | IIIA 期（P = 0.038） | IIIB 阶段 | 第 4 组 |
| SHMM | 第二阶段（P = 1.04E-5） | IIIA 期（P = 0.046） | IV 期（P = 3.39E-5） | 无 |
| TNFE | IV 期（P = 1.35E-66） | IIIC 期（P = 0.016） | IIIB 期（P = 0.0012） | 无 |
| PDNB | 无 | IIIA 期（P = 0.034） | IV 期（P = 0.021） | 第 5 组 (P = 0.0056) |
| LDNB | 第二阶段（P = 0.016） | 第四阶段 | IIIB 期（P = 3.42E-04） | 第 6 组 |

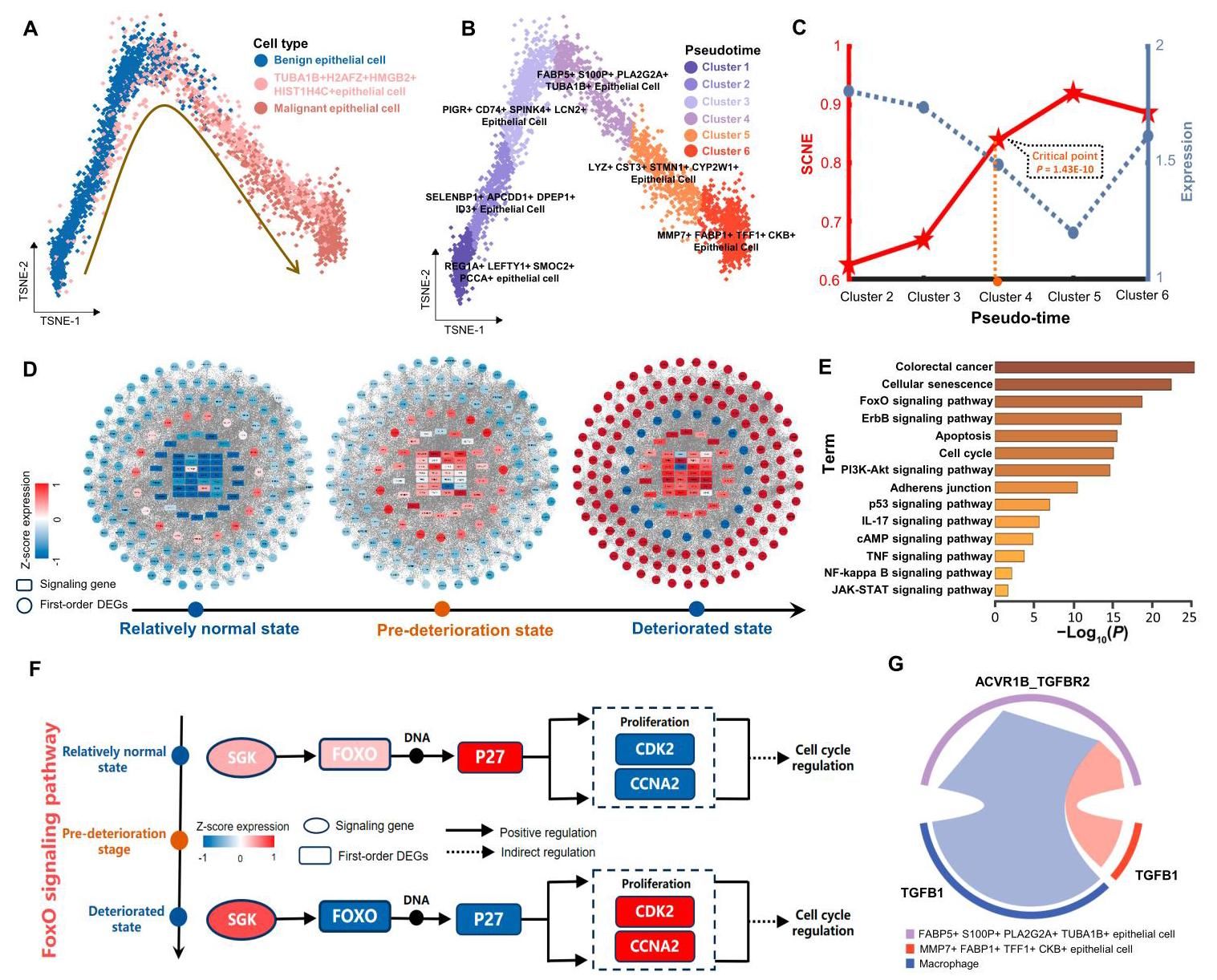


图 5.确定结直肠癌 EPCD 的预恶化状态。(A) 根据三个不同的亚群构建了上皮细胞的退化轨迹。(B) EPCD 的进展分为六个不同时期或组群。(C) 分析了 SCNE 和基因表达在 EPCD 发展过程中的动态表现。(D) 研究了信号基因及其邻近 DEGs 在 EPCD 过程中形成的调控网络的动态演变。(E) 对一阶差异表达邻近因子进行了 KEGG 通路富集分析。(F) 对基因及其差异表达的邻近基因进行的功能分析发现了 FoxO 信号通路中与细胞周期调控相关的信号机制。(G) 不同细胞亚群之间通过 TGF- 信号通路进行细胞交流。

# 讨论

在大多数复杂疾病中，检测突然恶化的关键信号至关重要。然而，现有的临界状态检测方法在应用于样本量有限的大容量 RNA 序列数据或单细胞数据时，受到数据固有的大量噪声的限制 [36]。在这项研究中，我们提出了一种在特定样本水平上的稳健计算方法，称为 SCNE，它能够为每个个体构建特定样本的因果关系网络，并有效识别与疾病恶化相关的临界点或恶化前状态。通过对模拟数据集和五个真实数据集实施 SCNE 方法，我们有效地确定了即将发生严重恶化的关键信号。这些数据集的成功预测验证了 SCNE 分数在仅根据特定样本量化临界点方面的有效性。此外，还通过数值模拟验证了 SCNE 在不同噪声强度下的鲁棒性，并利用肿瘤数据集与现有的单样本方法进行比较，证明了 SCNE 在检测疾病相关临界点方面的良好性能。

我们的 SCNE 方法的优点可简述如下。首先，通过引入 SCNE，我们的方法有效地减轻了 omics 数据中存在的大量噪声的影响，从而提高了数据的鲁棒性和可靠性。其次，与其他现有的六种单样本方法相比，所提出的方法可以推断出特定样本的因果关系网络，并提供其 SCNE 分数来量化复杂疾病的临界度，在揭示疾病进展过程中的恶化前状态方面具有更好的性能。第三，SCNE 方法不仅可作为向恶化状态关键转变的指标，还能精确定位与重要生物过程有关的相应信号生物标志物。最后，作为一种无模型计算方法，SCNE 适用于各种 omics 数据类型，包括体细胞和单细胞数据集。然而，SCNE 方法的局限性包括它依赖于 PPI 网络作为背景网络，而且必须有一个由相对健康样本组成的参照组（详见补充材料 B 部分）。此外，如何准确解释复杂的多阶段疾病过程中每个临界点的生物医学意义仍是一个尚未解决的问题，这也将是我们今后研究工作的重点。

# 材料与方法

# 理论背景

从动态系统的角度来看，疾病演变一般可理解为一个随时间演变的非线性动态过程，在这一过程中，突然恶化的状态被视为在分岔点发生的状态质变[37]。因此，这种开发过程通常分为三种状态（图 1A）：(a) 在临界过渡之前的相对正常的稳定状态；(b) 对干扰的敏感性增强的恶化前状态，标志着疾病严重恶化的临界过渡；(c) 以疾病发作或恶化为标志的另一种稳定的恶化状态。一般来说，相对正常的状态和恶化前的状态与恶化后的状态没有明显区别。因此，传统的统计方法（如差异基因表达分析）在区分退化前状态时可能会面临挑战（图 1D）。

最近，出现了 DNB 理论概念[1,4]，该概念的提出是为了根据多样本数据定量确定复杂系统发展过程中的临界状态或临界点。特别是当复杂系统接近临界点时，DNB 分子主要满足以下两个统计条件 [38]：DNB 分子的标准偏差急剧增大，DNB 分子之间的相关性迅速增加。事实上，DNB 的特性表明，分子表达的波动及其因果调控强度的改变可以作为系统临界转换的信号[39]。因此，我们分析了研究中构建的因果网络，以展示表达波动和因果关系强度的动态变化，这预示着复杂疾病即将恶化。在我们的研究中，当系统接近临界状态时，存在一组主导变量，它们被定义为 DNB 分子，根据观测数据满足以下两个标准：(a) 主导变量的表达偏差或波动急剧增加，(b) 主导变量之间的因果强度迅速增加（详见补充材料 A 部分）。

从特定样本因果关系网络的角度来看，我们所提出的 SCNE 方法旨在精确定位从相对正常状态向恶化状态转变的关键信号。基于因果推断的统计概念[20]，SCNE 可通过验证预测分析推断分子间的因果效应，并重建特定样本的因果关系网络。具体来说，我们提出的方法以 PPI 网络为背景，整合个体化基因表达数据，从而推断出特定样本的因果关系网络。值得注意的是，在构建样本特异性因果关系网络时，考虑了基因之间的动态调控关系。我们进行了分析论证，以探索网络拓扑结构和动态是如何变化到恶化前状态的，并确定了在推动系统走向恶化过程中发挥关键作用的关键动态分子（如图 S5 和表 S1、S2 所示）。总之，通过量化所构建因果网络的动态变化，SCNE 方法能够作为恶化前状态的有力指标。该算法的源代码可在 https://github.com/zhongjiayuan/SCNE\_project 免费获取。

# 使用 SCNE 揭示恶化前状态的算法

通过收集描述相对健康阶段的参考样本，采用 SCNE 计算方法，利用病例样本揭示临界点或恶化前状态，并在随后的章节中对其过程进行全面描述。

[Step 1] 构建时间点 的案例样本特定因果关系网络 。利用 PPI 网络和给定的参考样本，可以通过因果关系强度指数 ，为案例样本构建特定于样本的因果关系网络，其定义如公式所示。1.具体来说，如果 的值超过零，则表明存在一条从基因 到 的有向边 ；反之，则不存在这样的有向边。因此，根据因果关系强度指数 来确定每条边 的方向，我们就可以为时间点的案例样本构建一个特定于样本的因果关系网络 。

其中 和 表示从公式中得到的测试误差。当应用于案例样本时，分别为 2 和 3。

其中，符号 代表以 th gene 为中心的局部网络 中 th gene 的表达量，向量 定义为 ，其中 代表 的回归系数。同样， 表示 的回归系数。具体来说，对于以 PPI 网络中的一个基因 为中心的局部网络 ，其一阶相邻基因为 ，我们假设所有一阶相邻基因 都是中心节点 的原因，即任何相邻基因 的表达式变化都可能影响 的表达式变化。一组相对健康的参考样本作为训练样本，用于确定回归模型 和 ，而每个时间点的单个病例样本 则被指定为测试样本。通过将测试样本分别输入 和 ，得到输出 和 ，从而推断出特定样本的因果关系网络。更详细的说明见补充材料 部分。

[步骤 2] 从特定样本因果关系网络中提取每个局部因果关系网络 。本地因果关系网络由两类网络组成：本地内度网络和本地外度网络。具体来说， - 本地因果关系网络 以基因 为中心，它的 一阶内度邻居 对应 内度边， 一阶外度邻居 对应 外度边。中心基因 与其一阶内度邻接基因（记为 ）以及其一阶外度邻接基因（记为 ）之间的边权重由因果关系强度指数 决定。

[步骤 3] 计算每个局部因果网络的局部 SCNE 分数。具体而言，在 -局部因果关系网络 （由 一阶内度邻居和 一阶外度邻居组成）的背景下，其局部 SCNE 分数通过以下公式计算（公式4).

其中 和 的定义如下。

与

(6)

和

(7)

与

其中 代表病例样本中中心基因 的基因表达量，而 和 分别对应参考样本中中心基因 的基因表达量均值和方差。类似地， 表示 的第 个外度邻居 在样本 中的基因表达量。FT 可被视为量化病例样本中基因 相对于参考样本的表达波动/偏差（详见补充材料 F 部分）。

[步骤 4] 计算特定时间点病例样本的 SCNE 分数 。更精确地说，当考虑到局部 SCNE 分数最高的基因子集时，可通过以下公式得出特定样本的 SCNE 分数：

其中，常数 代表一个可配置参数，设定为本地 SCNE 分数最高的前 5%基因的数量。当系统接近临界状态时，由特定变量（DNB 成员）组成的子网络结构会发生明显变化，其特征是 DNB 分子的表达波动（FT）明显增加，它们之间的因果关系强度（ 值）迅速上升（图 S5）。通过在网络层面探索这样一组 DNB 变量的动态信息，就有可能预测定性的状态转换。因此， 指数旨在量化每个单一样本相对于一组给定参考样本所引发的表达波动和因果强度变化，从而提供恶化前状态的预警信号。

[步骤 5] 使用单样本 检验确定恶化前的状态。为了评估 SCNE 分数捕捉临界动态的能力，我们采用了单样本 检验 [40]，以确定相对正常状态和恶化前状态之间是否存在统计学意义上的显著区别。具体来说，下面的单样本 检验统计量 是用来区分一个值 和一个 维向量的平均值 的统计指标。

其中，均值 表示向量 的均值，而 表示其标准差。平均值 和 之间区别的显著性是通过 分布得出的 值来评估的。在本研究中，如果 SCNE 分数 满足以下两个标准，则时间点 将被视为恶化前状态：第一， ，标志着分数呈上升趋势；第二， 与先前信息相比表现出统计意义 。更多详情可参见补充材料 部分。

# 数据处理和功能分析

为了说明 SCNE 方法的功能，我们将其应用于数值模拟和五个真实数据集：来自 TCGA 数据库的 KIRC、STAD 和 LUAD 数据，以及流感感染数据（登录号：GSE30550）和结直肠癌中 EPCD 的单细胞数据（登录号：GSE161277）。肿瘤数据集包括肿瘤和肿瘤邻近样本。利用现有的分期信息将肿瘤样本分为不同的分期，不包括分期信息不完整的样本。与肿瘤相邻的样本被用作参照组，这些样本对应于相对健康的阶段。有关取样条件的更多详情，请参阅补充材料的 部分。对于单细胞结直肠癌数据，采用 Seurat 管道分析单细胞 RNA 序列数据 [41]。为解决不同组织间的生物学差异，使用 软件包 Harmony 实现了批次效应校正[42]。在所有数据集的情况下，我们都会进行一个过滤步骤，剔除缺少相应美国国家生物技术信息中心（NCBI）entrez 基因符号的探针。

使用京都基因组百科全书（https://www.kegg.jp）进行了通路分析。使用 Metascape [43] 和 Cluster-Profiler 软件包 [44] 进行了富集分析。功能结果是通过基因本体论联盟（Gene Ontology Consortium，http://geneontology.org）提供的网络服务工具和 Ingenuity Pathway Analysis 公司提供的客户端软件获取的。网络的可视化是通过 Cytoscape 软件（www.cyto-scape.org）实现的。

# 致谢

资金：这项工作得到了国家自然科学基金的支持（编号.T2341022、12322119、62172164和12271180）、广东省人类数字孪生重点实验室（2022B1212010004）、广东省教育委员会（2023KQNCX073）、广东省自然科学基金（2022A- 1515110759和2023A1515110558）和中央高校基本科研业务费（2023ZYGXZR077）。

作者供稿： R.L.、P.C.和F.L.构思了这项研究。 J.Z.、H.T.、Z.H.和H.C.进行了真实数据分析。所有作者都撰写了论文。所有作者都阅读并批准了最终手稿。

利益冲突：作者声明他们没有利益冲突。

# 数据可用性

本研究采用了五个真实数据集，包括来自 GEO 数据库（http://www.ncbi.nlm.nih.gov/geo/）的流感感染（GSE30550）和结直肠癌中 EPCD 的单细胞数据（GSE161277），以及来自 TCGA 数据库（http://can-cergenome.nih.gov）的 KIRC、STAD 和 LUAD 数据。算法源代码和相关数据可在 https://github.com/zhongjiayuan/ SCNE\_project 上获取。

# 补充材料

A 节至 H 节

图S1 至 S12

表 S1 和 S2

参考文献 [1-12]

# 参考资料

1.Chen L, Liu R, Liu ZP, Li M, Aihara K. 通过动态网络生物标记检测复杂疾病突然恶化的预警信号。科学报告2012;2:Srep00342.

2.Li X、Zhu Q、Zhao C、Qian X、Zhang X、Duan X、Lin W. 利用水库计算检测临界点。研究。2023;6:0174.

3.Liu R, Wang J, Ukai M, Sewon K, Chen P, Suzuki Y, Wang H, Aihara K, Okada-Hatakeyama M, Chen L. 《通过动态网络生物标记物寻找乳腺癌内分泌抵抗过程中的临界点》。J Mol Cell Biol.2019;11(8): 649-664.

4.Liu R, Li M, Liu ZP, Wu J, Chen L, Aihara K. Identifying critical transitions and their leading biomolecular networks in complex diseases.科学报告2012;2:Srep00813.

5.Liu X, Liu R, Zhao XM, Chen L. 利用动态网络生物标记物检测1型糖尿病及其主导生物分子网络的预警信号。 医学基因组学。2013;6 Suppl 2(Suppl 2):S8.

6.Wen Z, Zhang W, Zeng T, Chen L. MCentridFS：从高通量数据中识别多表型模块生物标记物的工具。Mol BioSyst.2014;10(11):2870-2875.

7.Richard A, Boullu L, Herbach U, Bonnafoux A, Morin V, Vallin E, Guillemin A, Papili Gao N, Gunawan R, Cosette J, et al.基于单细胞的分析突出表明，在分化过程中出现不可逆转的承诺之前，细胞与细胞之间的分子变异性激增。PLOS Biol.2016;14(12):Article e1002585.

8.Teschendorff AE, Feinberg AP.统计力学与单细胞生物学的结合Nat Rev Genet.2021;22(7):459-476.

9.Lesterhuis WJ, Bosco A, Millward MJ, Small M, Nowak AK, Lake RA.癌症免疫检查点阻断中的动态与静态生物标志物揭开复杂的面纱。Nat Rev Drug Discov.2017;16(4):264-272.

10.Zhong J, Liu H, Chen P. 单样本网络模块生物标志物（sNMB）方法揭示了疾病恶化的前期阶段。J Mol Cell Biol.2022;14(8): 第 mjac052 条。

11.Liu R, Zhong J, Yu X, Li Y, Chen P. 基于单样本的隐马尔可夫模型识别复杂疾病的临界状态。遗传学前沿2019;10:285.

12.Liu R, Chen P, Chen L. 单样本景观熵揭示了疾病进展过程中即将发生的相变。生物信息学2020;36(5):1522-1532.

13.Gao R, Yan J, Li P, Chen L. 基于时序网络流熵检测疾病发展过程中的临界状态。简要生物信息。2022;23(5):Article bbac164.

14.Liang J, Li ZW, Yue CT, Hu Z, Cheng H, Liu ZX, Guo WF.通过多模式优化确定个性化生物标志物，用于癌症患者的疾病预测。简要生物信息。2022;23(5):Article bbac254.

15.Liu X, Chang X, Leng S, Tang H, Aihara K, Chen L. 《通过景观动态网络生物标记检测疾病临界点》。Natl Sci Rev. 2019;6(4):775-785.

16.Ying X, Leng SY, Ma HF, Nie Q, Lai YC, Lin W. Continuity Scaling：准确检测和量化因果关系的严格框架。研究。2022;2022:9870149.

17.Moerman T, Aibar Santos S, Bravo González-Blas C, Simm J, Moreau Y, Aerts J, Aerts S. GRNBoost2 and Arboreto：高效、可扩展的基因调控网络推断。生物信息学2019;35(12):2159-2161.

18.Zhong J, Han C, Wang Y, Chen P, Liu R. 用有向网络秩分法识别复杂生物系统的临界状态。生物信息学2022;38(24):5398-5405.

19.Li L, Xia R, Chen W, Zhao Q, Tao P, Chen L. 通过交叉映射熵推断单细胞因果网络简要生物信息。2023;24(5):Article bbad281.

20.Zhang Y, Li Q, Chang X, Chen L, Lui X.基于交叉验证可预测性的因果网络推断 bioRxiv.2022. https://doi.org/10.1101/2022.12.11.519942.

21.Foo M, Kim J, Bates DG.基因调控网络的建模与控制，以减轻扰动。IEEE/ACM Trans Comput Biol Bioinform.2019;16(2):583-595.

22.Yan J, Li P, Li Y, Gao R, Bi C, Chen L. 基于单样本的网络信息增益进行疾病预测。Fundam Res.2023.

23.Su S, Shahriyari L. RGS5 在肾细胞癌中发挥着重要作用。R Soc Open Sci.2020;7(4):191422.

24.Kwon SJ.评估第 7 届 UICC 胃癌 TNM 分期系统。J 胃癌。2011;11(2):78-85.

25.Steeg PS, Bevilacqua G, Kopper L, Thorgeirsson UP, Talmadge JE, Liotta LA, Sobel ME.与低肿瘤转移潜能相关的新基因证据。J Natl Cancer Inst.1988;80(3):200-204.

26.Huang X, Han C, Zhong J, Hu J, Jin Y, Zhang Q, Luo W, Liu R, Ling F. 动态网络标记物 FOS/ JUN 在退化前上皮细胞中的低表达与结直肠腺瘤向癌的进展相关。J Transl Med.2023;21(1):45.

27.Noorolyai S, Shajari N, Baghbani E, Sadreddini S, Baradaran B. PI3K/AKT 信号通路与癌症的关系。基因2019;698:120-128.

28.Fakhri S、Zachariah Moradi S、DeLiberto LK、Bishayee A. 癌症中的细胞衰老信号传导：抗击人类恶性肿瘤的新型治疗靶点。生物化学药理学2022;199:114989.

29.Zhang Y, Gan B, Liu D, Paik JH.癌症中的 FoxO 家族成员Cancer Biol Ther.2011;12(4):253-259.

30.Arden KC.FOXO 转录因子在哺乳动物细胞中的多种作用表明其在癌症中的多种作用。Exp Gerontol.2006;41(8):709-717.

31.Arden KC.FOXO 动物模型揭示了 FOXO 转录因子的各种不同作用。Oncogene.2008;27(16):2345-2350.

32.Tzivion G、Dobson M、Ramakrishnan G. FoxO 转录因子；受 AKT 和 14-3-3 蛋白调节。Biochim Biophys Acta.2011;1813(11):1938-1945.

33.Pereira SS、Morais T、Costa MM、Monteiro MP、Pignatelli D。分子标记物 p27 在肾上腺皮质肿瘤鉴别诊断中的新作用。Endocr Connect.2013;2(3):137-145.

34.Zhang Y, Alexander PB, Wang XF.控制细胞增殖和存活的 TGF-β 家族信号。Cold Spring Harb Perspect Biol.2017;9(4):a022145.

35.Kang JS, Liu C, Derynck R. TGF- 受体功能的新调节机制。Trends Cell Biol.2009;19(8): 385-394.

36.Li Y, Lu Y, Kang C, Li P, Chen L. 通过多视图图网络从空间解析转录组学中揭示组织异质性和空间暗基因。研究。2023;6:0228.

37.Scheffer M, Carpenter S, Foley JA, Folke C, Walker B. Catastrophic shifts in ecosystems.自然。2001;413(6856): 591-596.

38.Koizumi K, Oku M, Hayashi S, Inujima A, Shibahara N, Chen L, Igarashi Y, Tobe K, Saito S, Kadowaki M, et al.日本传统药物（甘宝方）抑制小鼠代谢综合征前疾病状态（Mibyou）的动态网络生物标志物信号。Evid Based Complement Alternat Med.2020；2020：第 9129134 条。

39.Wen Z, Liu ZP, Liu Z, Zhang Y, Chen L. 一种识别复杂疾病因果网络模块的综合方法在结直肠癌中的应用。J Am Med Inform Assoc.2013;20(4):659-667.

40.Rochon J, Kieser M. 单样本 t 检验的初步拟合优度检验效果详解。Br J Math Stat Psychol.2011;64:410-426.

41.Hao Y, Hao S, Andersen-Nissen E, Mauck WM 3rd, Zheng S, Butler A, Lee MJ, Wilk AJ, Darby C, Zager M, et al.多模态单细胞数据的综合分析。牢房2021;184(13):3573-3587.

42.Korsunsky I、Millard N、Fan J、Slowikowski K、Zhang F、Wei K、Baglaenko Y、Brenner M、Loh PR、Raychaudhuri S.快速、灵敏、准确地整合单细胞数据与和谐。自然方法。2019;16(12):1289-1296.

43.Zhou Y, Zhou B, Pache L, Chang M, Khodabakhshi AH, Tanaseichuk O, Benner C, Chanda SK.Metascape 为分析系统级数据集提供了面向生物学家的资源。Nat Commun.2019;10(1):1523.

44.Yu G、Wang LG、Han Y、He QY：用于比较基因簇间生物主题的 R 软件包。OMICS.2012;16(5):284-287.