Identifying the critical state of cancers by single-sample Markov flow entropy

Juntan Liu [1], **Yuan Tao** [1], **Ruoqi Lan** [1], **Jiayuan Zhong** [Corresp., 1, 2], **Rui Liu** [1], **Pei Chen** [Corresp. 1, 3]

[1] School of mathematics, South China University of Technology, Guangzhou, Guangdong Province, China

[2] School of Mathematics and Big Data, Foshan University, Foshan, China

[3] Pazhou Lab, Guangzhou 510330, China, Guangzhou, Guangdong Province, China

Corresponding Authors: Jiayuan Zhong, Pei Chen
Email address: 201720127132@mail.scut.edu.cn, chenpei@scut.edu.cn

Abstract

Background: The progression of complex diseases sometimes undergoes a drastic critical transition, at which the biological system abruptly shifts from a relatively healthy state (before-transition stage) to a disease state (after-transition stage). Searching for such a critical transition or critical state is crucial to provide timely and effective scientific treatment to patients. However, in most conditions where only a small sample size of clinical data is available, resulting in failure when detecting the critical states of complex diseases, particularly only single-sample data.

Methods: In this study, different from traditional methods that require multiple samples at each time, a model-free computational method, single-sample Markov flow entropy (sMFE), provides a solution to the identification problem of critical states/pre-disease states of complex diseases, solely based on a single-sample . Our proposed method was employed to characterize the dynamic changes of complex diseases from the perspective of network entropy.

Results: The proposed approach was verified by unmistakably identifying the critical state just before the occurrence of disease deterioration for four tumor datasets from The Cancer Genome Atlas (TCGA) database. In addition, two new prognostic biomarkers, optimistic sMFE (O-sMFE) and pessimistic sMFE (P-sMFE) biomarkers, were identified by our method and enable the prognosis evaluation of tumors.

Conclusions: The proposed method has shown its capability to accurately detect pre-disease states of four cancers and provide two novel prognostic biomarkers, O-sMFE and P-sMFE biomarkers, to facilitate the personalized prognosis of patients. This is a remarkable achievement that could have a major impact on the diagnosis and treatment of complex diseases.

1 **Identifying the critical state of cancers by single-sample Markov flow**

2 **entropy**

3 **Juntan Liu[1], Yuan Tao[1], Ruoqi Lan[1], Jiayuan Zhong[1,3,*], Rui Liu[1], Pei Chen[1,2,*].**

4 *[1]School of Mathematics, South China University of Technology, Guangzhou 510640, China*

5 *[2]Pazhou Lab, Guangzhou 510330, China*

6 *[3]School of Mathematics and Big Data, Foshan University, Foshan 528000, China*

7 * **Corresponding authors**.

8 E-mail:

9 Pei Chen, chenpei@scut.edu.cn; Jiayuan Zhong, 201720127132@mail.scut.edu.cn.

Abstract

11 **Background:** The progression of complex diseases sometimes undergoes a drastic critical transition, at
12 which the biological system abruptly shifts from a relatively healthy state (before-transition stage) to a
13 disease state (after-transition stage). Searching for such a critical transition or critical state is crucial to
14 provide timely and effective scientific treatment to patients. However, in most conditions where only a small
15 sample size of clinical data is available, resulting in failure when detecting the critical states of complex
16 diseases, particularly only single-sample data.
17 **Methods:** In this study, different from traditional methods that require multiple samples at each time, a
18 model-free computational method, single-sample Markov flow entropy (sMFE), provides a solution to the
19 identification problem of critical states/pre-disease states of complex diseases, solely based on a single-
20 sample. Our proposed method was employed to characterize the dynamic changes of complex diseases
21 from the perspective of network entropy.
22 **Results:** The proposed approach was verified by unmistakably identifying the critical state just before the
23 occurrence of disease deterioration for four tumor datasets from The Cancer Genome Atlas (TCGA)
24 database. In addition, two new prognostic biomarkers, optimistic sMFE (O-sMFE) and pessimistic sMFE
25 (P-sMFE) biomarkers, were identified by our method and enable the prognosis evaluation of tumors.
26 **Conclusions:** The proposed method has shown its capability to accurately detect pre-disease states of
27 four cancers and provide two novel prognostic biomarkers, O-sMFE and P-sMFE biomarkers, to facilitate
28 the personalized prognosis of patients. This is a remarkable achievement that could have a major impact
29 on the diagnosis and treatment of complex diseases.
30

31 **Keywords:** critical state; single-sample Markov flow entropy (sMFE); dynamic network biomarker (DNB);
32 prognostic biomarker.

33

34

35

1. Introduction

The time evolution of a complex biological system is sometimes viewed as a time-dependent nonlinear dynamic system (Gorban et al. 2021; Liu et al. 2020), with the abrupt transition seen as the phase shift at a bifurcation point (Scheffer et al. 2009). Therefore, from the viewpoint of dynamic systems, we can roughly model the progression of complex diseases (such as cancer, COVID-19, diabetes, etc.) as three distinct stages or states (Fig. 1A): (1) a before-transition state, a stable state with low fluctuation; (2) a critical/ pre-disease state, an unstable state before the start of qualitative changes, high fluctuation (according to DNB theory(Chen et al. 2012b), the variance of gene expression increases dramatically), sensitive to perturbation; (3) an after-transition/disease state, another stable state with low fluctuation after the qualitative transition. The pre-disease state is unstable and can be reverted to the before-transition state with appropriate intervention, while the after-transition/disease state is a stable state with high resilience and is nearly irreversible(Chen et al. 2012a; Liu et al. 2014). Therefore, it is essential to identify the early warning signs of the critical transition during disease progression and recognize the pre-disease state, so that timely medical intervention can be administered to prevent or delay the progression of the disease.

Recently, the dynamic network biomarker (DNB) has been proposed to qualitatively describe the critical state of a biological system (Chen et al. 2019). Specifically, DNB theory uses differential equation theory to derive that systems exhibit three statistical properties when they reach the bifurcation point, allowing us to measure this dynamic difference at the molecular network level, not just gene expression. The proposal of DNB theory provides a basic framework for identifying the pre-disease states of complex diseases(Chen et al. 2016). For instance, successfully detect the pre-disease state of metabolic syndromes (Koizumi et al. 2019; Liu et al. 2013b), and respiratory viral infections(Gao et al. 2022), identify immune checkpoint blockades (Lesterhuis et al. 2017), and assess cell fate commitment(Peng et al. 2022; Zhong et al. 2022). However, most developed methods are based on multiple samples, so multiple samples are required at each time, which limits their application in most real-world cases where only a single sample is available.

Specifically, inspired by the previous study (Guo et al. 2021a; Guo et al. 2021b), this study proposes a model-free computational approach, single-sample Markov flow entropy (sMFE), which utilizes the inferred direct interaction network and the DNB concept to characterize the perturbation of a single sample, thus detecting the early-warning signal of critical transition during a complex disease process. Precisely, given reference samples from the normal cohort, sMFE is calculated for each single sample against the directed network., constructed based on rewiring the protein–protein interaction (PPI) network by gene networks inference using projection and lagged regression (GNIPLR) method (Zhang et al. 2021) (Fig. 1B). Two novel insights were provided in this computational method. On the one hand, by eliminating indirect interactions between genes, the established directed network more realistically reflects the molecular interactions in the biological network, and thus dynamic changes during complex disease progression can be more accurately captured. On the other hand, at the single-sample level, sMFE provides a reliable method for quantifying the pre-disease state or critical point of complex disease (Fig. 1C).

To test the effectiveness of sMFE, four different tumor datasets from TCGA, including esophageal

73   carcinoma (ESCA), colon adenocarcinoma (COAD), kidney renal clear cell carcinoma (KIRC), lung

74   adenocarcinoma (LUAD), were used to validate our algorithm. In all datasets, the critical states identified

75   by sMFE preceded severe disease deterioration, i.e., a significant alteration in sMFE score indicated an

76   early warning of the critical transition to disease deterioration. By searching the literature, we found that the

77   following clinical characteristics during the deterioration of diseases (e.g., cancer) occur, and distant

78   metastasis occurs in the later stage of cancer development, which is difficult to recover by external

79   treatment and can be regarded as the basis for disease deterioration. Specifically, the critical state for

80   ESCA was identified in stage IIIA disease before distant metastasis, and stage IIB was identified as the

81   critical state for COAD before lymph node metastasis; stage II was identified critical state for KIRC before

82   the tumor migrates to form distant metastasis, and that of LUAD was identified in stage IIIB disease before

83   the tumor cells invaded distant tissues or organs. In addition, we further analyze the signaling gene and

84   propose two types of biomarkers, i.e., optimistic sMFE (O-sMFE) and pessimistic sMFE (P-sMFE)

85   biomarkers (see Supplementary Figure S1 for the specific procedure), combined with survival analysis.

86   Statistically, O-sMFE biomarkers were associated with good prognosis, while P-sMFE biomarkers were

87   linked to poor prognosis. Furthermore, our results were validated by functional and pathway enrichment

88   analyses.

89   2. Materials & Methods

90   2.1 Theoretical background

91   The DNB theory (Chen et al. 2012b; Liu et al. 2013a) is proposed to visualize the evolution of complex

92   diseases into three periods: a normal (before-transition) state, a pre-disease (critical) state, and a disease

93   (after-transition) state. The normal state is stable and resilient to perturbations, while the pre-disease state

94   is unstable and reversible with appropriate intervention. The disease state is also stable and resilient, but

95   is almost irreversible (Litt et al. 2001; Liu et al. 2020; McSharry et al. 2003; Venegas et al. 2005). Moreover,

96   when a complex system is near the critical point, a dominant group of DNB biomolecules can be identified,

97   based on observed data, that satisfy the following three conditions (Chen et al. 2012a):

98   ●   The Pearson Correlation Coefficient (PCC_{in}) between any pair of members in the DNB group rapidly

99   increases;

100   ●   The Pearson Correlation Coefficient (PCC_{out}) between one member of the DNB group and any other

101   non-DNB member rapidly decreases;

102   ●   The standard deviation (SD_{in}) or coefficient of variation for any member in the DNB group drastically

103   increases.

104   Our DNB theory suggests that phase transitions are characterized by a set of highly fluctuating and highly

105   correlated features/variables, which implies an imminent transition to the disease state. Therefore, we can

106   quantify the tipping points based on these three conditions, provide early-warning signals of disease

107 deterioration, and then determine the biomolecular dominant group to make up DNB members. These three

108 conditions are the basis of DNB theory and have many applications in disease progression and biological

109 processes to predict its critical state(Liu et al. 2013a; Liu et al. 2020).

110 Markov Chain Entropy (MCE)(Shi et al. 2018) was constructed to represent the potential of samples

111 containing scRNA-Seq or bulk RNA-Seq information. This method requires the normalized RNA Seq profile

112 and a defined interaction network between the genes.

113 Specifically, given an interaction network G with n_v nodes and n_E edges, and the transition probability matrix

114 $P = \left\{ p_{ij} \right\}$, where $i, j = 1, \ldots, n_E$. Where $p_{ij} = 0$ if there is no edge between two different nodes i and j.

115 Then the Markov Chain Entropy based on network G is defined as

$$\text{MCE} = -\sum_{(i,j)\in \bar{E}} \pi_i p_{ij} \log\left(\pi_i p_{ij} \right) \tag{1}$$

117 Where π_i was node i's expression normalized to [0,1]. In this paper, MCE was split into each node to form

118 Markov flow entropy (MFE) (Fig. 1B). For centered gene g^k, the MFE is defined as

$$\text{MFE}(g^k) = -\sum_{j=1}^{M} \pi_i p_{ij} \log\left(\pi_i p_{ij} \right) \tag{2}$$

120 Where M denotes the number of the 1st-order neighbors $(g_1^k, \ldots, g_M^k)$ of g^k, p_{ij} denotes the transition

121 probability from g^k to g_j^k, and $\pi_i p_{ij}$ means the probability transmitted from g^k to g_j^k along the directed edge.

122 **2.2 Algorithm to identify the tipping point based on sMFE**

123 Reference samples (healthy or relatively healthy cells; tumor-adjacent samples here) and only one disease

124 sample were required in the sMFE method to identify the tipping point with the following algorithm (Fig. 1B).

125 **[Step 1]** Constructing a global directed network N^G by mapping the genes to a protein–protein interaction

126 (PPI) network from STRING (http://string-db.org) (Szklarczyk et al. 2015), with a confidence level of 0.800.

127 Discard isolated nodes that do not have any links with other nodes. The direct interaction network was

128 constructed based on rewiring the protein–protein interaction (PPI) network by GNIPLR (Gene networks

129 inference using projection and lagged regression) (Zhang et al. 2021)method. Tumor-adjacent samples

130 were used to infer the gene regulatory network.

131 **[Step 2]** Mapping data into the global network N^G constructed in the previous step to form expression

132 matrix. Mapping datasets of four cancers downloaded from the TCGA database to the global network N^G,

133   then the gene expression matrix formed for further analysis.

134   **[Step 3]** Calculating the local Markov flow entropy (MFE) for each gene. Specifically, we extract each gene

135   g^k's local network N^k (k=1, 2, …, L) from the global network N^G (Fig. 1B), the number of the local network

136   is *L*, and $\left\{ g_1^k,\ldots,g_M^k \right\}$ are the 1st-order neighbors of g^k. Then, based on n reference samples, the formula

137   for calculating local Markov flow entropy $MFE_k^n(t)$ is

$$MFE_k^n(t) = -\sum_{i=1}^{M} \overline{x_k^n p_{ki}^n} \log\left(\overline{x_k^n p_{ki}^n} \right), \qquad (3)$$

139   with

$$p_{ki}^n(t) = \frac{\left| PCC^n\left(g_i^k(t), g^k(t) \right) \right|}{\sum_{j=0}^{M} \left| PCC^n\left(g_j^k(t), g^k(t) \right) \right|}, \qquad (4)$$

141   where the constant *M* represents the count of neighbors in the local network N^k and $PCC^n\left(g_i^k(t), g^k(t) \right)$

142   represents the Pearson correlation coefficient (PCC) between the center gene g^k and its 1-st neighbor g_i^k

143   based on *n* reference samples at time point *t*. Specifically, $g_i^k = g^k$ if $j = 0$. $\overline{x_k^n}$ is the mean of the canter

144   gene g^k's normalized expression based on *n* reference samples at time point *t*.

145   **[Step 4]** Calculating the differential MFE called $sMFE_k(t)$. Under *n*+1 mixed samples, where a single

146   sample from an individual was added to *n* reference samples at time point *t*, the same with the prior step,

147    $MFE_k^{n+1}(t)$ is calculated, i.e.,

$$MFE_k^{n+1}(t) = -\sum_{i=1}^{M} \overline{x_k^{n+1}} p_{ki}^{n+1} \log\left(\overline{x_k^{n+1}} p_{ki}^{n+1} \right) \qquad (5)$$

149   Then, the $sMFE_k(t)$ is calculated as

$$sMFE_k(t) = \frac{1}{\log M} \left| SD^n\left(g^k(t) \right) * MFE_k^{nt}(t) - SD^{n+1}\left(g^k(t) \right) + MFE_k^{n+1}(t) \right|, \qquad (6)$$

151   where $SD^n(g^k(t))$ and $SD^{n+1}(g^k(t))$ represent the standard deviation of the gene expression for the center

152   gene g^k based on *n* reference samples and *n*+1 mixed samples at time point *t*, respectively.

153   **[Step 5]** The fifth step is to calculate the global single-sample Markov flow entropy $sMFE(t)$, i.e.,

$$sMFE(t) = \frac{1}{Q} \sum_{k=1}^{Q} sMFE_k(t), \qquad (7)$$

155   where constant Q represents the count of top 5% genes with highest local *sMFE*. Furthermore, t-test was

156   applied to test if the identified critical state significantly different with the other states. The alternative state

157   was viewed as critical state if $P<0.05$ (see supplementary information). From the formula of sMFE, we can

158   see that when the interaction between the center node and the 1-st neighbor node converges (that is, the

159   transition probability is the same), the sMFE value reaches the maximum value. Conversely, when the

160   interaction between the center node and one 1-st neighbor node is extremely strong (the transition

161   probability equals 1), the sMFE value reaches the minimum. From DNB theory, it can know that in the

162   sample at the critical point period, the interaction between DNB molecules and their neighboring nodes will

163   increase extremely so that it can form a significant difference from the reference sample, and sMFE can

164   measure this difference, and when the sample reaches the critical point, the significance increases so that

165   it can warn the disease evolution process. According to DNB theory (Chen et al. 2012b; Liu et al. 2013a),

166   when the system approaches the critical point, the sMFE score can effectively characterize the fluctuation

167   of the network, and thus can be regarded as an early warning signal prior to critical deterioration.

168   **2.3 Data processing and functional analysis**

169   The Cancer Genome Atlas (TCGA) database ((GDC (cancer.gov)) was used to download four unrelated

170   clinical tumor datasets: esophageal carcinoma (ESCA), colon adenocarcinoma (COAD), (kidney renal clear

171   cell carcinoma (KIRC), and lung adenocarcinoma (LUAD). RNA-seq data from tumor and tumor-adjacent

172   samples, as well as clinical information, were included in these datasets. Then we classified the tumor

173   samples into different stages based on clinical (stage) information downloaded from TCGA, samples lacking

174   stage information were discarded, and Supplementary Table S1 provided the full clinical staging

175   information.

176   The molecular global template network was constructed using the steps shown in Fig. 1A.

177   (i)   The first step in constructing the molecular global template network was to download the

178   protein–protein interaction networks for Homo sapiens from STRING (http://string-db.org). This

179   data was then used to generate a global template network, which was composed of nodes

180   representing proteins and edges representing interactions between proteins. All interaction

181   used for discussion was picked with high confidence (higher than 0.8). Nodes that are not

182   connected to any other nodes are removed from the PPI network.

183   (ii)   Second, the direct interaction network was constructed based on rewiring the protein–protein

184   interaction (PPI) network by GNIPLR (Gene networks inference using projection and lagged

185   regression) (Zhang et al. 2021)method. Tumor-adjacent samples were used to infer the gene

186   regulatory network (GRN; directed network) for subsequent analysis (detailed information for

187   how to choose adjustable parameters in GNIPLR were added in Supplementary information).

188   (iii) In the end, the directed network, created in prior step, was adapted to map the gene expression

189   of each RNA-seq dataset to form a molecular network for further analysis.

190   The Kyoto Encyclopedia of Genes and Genomes (KEGG) database Mapper tool (KEGG Mapper Color)

191 and the Database for Annotation, Visualization and Integrated Discovery (DAVID) Functional Annotation

192 Tool (DAVID: Functional Annotation Tools (ncifcrf.gov) were used to conduct the enrichment analysis.

193 These statistical analysis carried in the paper was done using R Software v4.0.3 (R: The R Project for

194 Statistical Computing (r-project.org)), and the R Survival package was used to make Cox survival analysis.

195 The visualization results and network analysis are implemented using Cytoscape

196 software(www.cytoscape.org).

3. Results

3.1 Identifying the critical transition points of four cancers with sMFE

199 The sMFE algorithm was used to identify the critical transition points of four cancers (ESCA, COAD, KIRC,

200 and LUAD) from TCGA datasets. These datasets contain much clinical information, in which doctors divide

201 cancer patients into different stages according to the size of the tumor diameter, such as KIRC samples

202 were divided into four stages (I, II, III, and IV), while LUAD samples were divided into seven stages (IA, IB,

203 IIA, IIB, IIIA, IIIB, and IV). The tumor-adjacent (TA) samples were deemed relatively normal and used as a

204 reference for each cancer when calculating sMFE according to the proposed algorithm. The average sMFE

205 score curve for each cancer stage was calculated to reflect the dynamic change and was shown in Fig. 2A-

206 D. Based on DNB theory, the critical state during disease progression can be reflected by the tipping point

207 of the sMFE score curve (Fig. 1A). Kaplan-Meier (log-rank) survival analysis was performed on the

208 prognosis of before-transition and after-transition samples (Fig. 2E-H) to further validate the identified

209 critical state.

210 As seen in Fig. 2A (yellow curve), the average sMFE score of ESCA increased significantly (P=0.0077)

211 at stage IIIA, suggesting an upcoming critical transition after stage IIIA.; that is, the tumor migrates and

212 invades to form distant metastasis at stage IIIB and ultimately cause distant metastasis (stage IV) (Dong et

213 al. 2022). While at the six-time points, there was little significant difference among the average gene

214 expression of differentially expressed genes (DEGs) (the gray curve in Fig. 2A). Of note, samples taken

215 before and after the transition point (Stage IIIA) had significantly different survival curves (P<0.0001; Fig.

216 2E), and patients taken after the transition point (Stage IIIA) had shorter survival times than those before

217 the transition.

218 When applied to COAD, a significant change (P=0.0045; Fig. 2B) in sMFE was detected around stage

219 IIB, suggesting lymph node metastasis and tumor invasion of other adjacent organs in stage III (Hari et al.

220 2013). As presented in gray in Fig. 2B, dynamic changes (P=0.16) in the mean gene expression of DEGs

221 were detected in stage IIIB and therefore failed to provide timely early-warning signals of the critical

222 transition. Moreover, it is worth noting that survival analysis showed that the survival time of samples taken

223 before the transition point (Stage IIB) was significantly different than that taken after the transition point

224 (Stage IIB) (P<0.0001; Fig. 2F). Furthermore, patients from after the identified critical state (Stage IIB) had

225 shorter survival times than those from before the transition.

226   For KIRC dataset, a drastic increase (*P*=0.014) in the sMFE score was observed in stage II (transition
227   point; Fig. 2C), indicating an upcoming critical transition into the disease state (stage III/IV), where the tumor
228   migrates to form distant metastasis at stage III (Gu & Zhao 2019). In terms of mean gene expression,
229   dynamic changes (*P*=0.25; the gray curve in Fig. 2C) were detected in stage III and therefore failed to
230   provide timely early-warning signals of the critical transition. Furthermore, survival analysis demonstrated
231   that survival curves had a significant difference (*P*<0.0001) between stage I~II and stage III~IV KIRC
232   samples according to clinical information (Fig. 2G). In addition, the survival time of patients after stage II
233   was significantly shorter than before stage II.

234   The red curve presented the LUAD data in Fig. 2D, the drastic increase of sMFE occurred at stage IIIB
235   (*P*=0.0027), indicating an upcoming critical transition after stage IIIB; that is, stage IV was characterized by
236   a distant metastasis process, in which the tumor cells invaded distant tissues or organs (Chiang &
237   Massagué 2008). However, the gray curve in Fig. 2D shows little significant increase in the mean
238   expression of DEGs occurring among the six-time points. Of note, the survival curves of samples taken
239   before and after the transition point (stage IIIB) tend to show a significantly different (*P*=0.0024; Fig. 2H).
240   In addition, patients with stage after stage IIIB had shorter survival times than those before. As shown,
241   sMFE is an effective approach to identify the critical state of cancers.

242   **3.2 The dynamic evolution of gene regulatory networks**

243   To validate the critical states identified by sMFE, we further examined them from the network perspective.
244   First, for each sample, the top 5% of genes with the highest sMFE scores were considered the signaling
245   genes. Dynamic network biomarkers (DNBs) consist of common signaling genes in samples in the identified
246   critical state. In DNB theory, these DNB molecules exhibit statistical properties when the system reaches a
247   critical point. Biologically, they may be involved in important biological reaction processes during disease
248   progression. These DNB molecules were mapped to the STRING database to form a connected PPI
249   network, and their dynamic changes in the process of disease evolution were studied from the network
250   level.

251   The landscape of the local sMFE score for ESCA data was depicted in Fig. 3A. It was worth noting that
252   the local sMFE scores of the signaling genes showed an abruptly increased collaborative manner around
253   stage IIIA. Furthermore, in the network structure, a noticeable change occurred at stage IIIA (Fig. 3E),
254   indicating an impending critical transition (Dong et al. 2022), which spontaneously come together with the
255   experimental results. When applied to the COAD dataset, the landscape of the local sMFE score was shown
256   in Fig. 3B. It was discovered that the peak of local sMFE scores for signaling genes appeared at stage IIB.
257   Furthermore, there was a significant change in the network structure at stage IIB (Fig. 3F), indicating lymph
258   node metastasis and tumor invasion of other adjacent organs in stage III (Hari et al. 2013). As shown in
259   Fig. 3C, for KIRC, showing that the peak local sMFE of signaling genes appeared at stage II. Moreover, in
260   the PPI network, a drastic change occurred in stage II (Fig. 3G), where the tumor gradually migrates to form
261   distant metastasis at stage III (Gu & Zhao 2019). Results for LUAD In Fig.3D, the local sMFE score

262 landscape peaked in stage IIIB and there was a significant change in the network structure of stage IIIB
263 (Fig. 3H), i.e., stage IV was characterized by a distant metastasis process, in which the tumor cells invaded
264 distant tissues or organs (Chiang & Massagué 2008). In addition, Supplementary Figure S2 provided results
265 of the dynamic network changes across all stages.

266 **3.3 Prognostic biomarkers of tumors in the sMFE method**

267 In order to identify biomarkers that were effective in predicting prognoses. Indeed, signaling genes could
268 be categorized into two types of molecules for prognostic prediction as common biomarkers for all samples:
269 those for samples with poor prognosis, termed pessimistic sMFE (P-sMFE) biomarkers; and those with
270 good prognosis, termed optimistic sMFE (O-sMFE) biomarkers. Furthermore, those two types of biomarkers
271 are not only the signaling genes, pessimistic/optimistic biomarkers, but also the non-differentially expressed
272 genes usually screened out by most researchers in the first step. More details about the identification
273 method are provided in Supplementary Figure S1.

274   The meaning of these two biomarkers can also be understood in this way, samples of optimistic sMFE
275 biomarkers that appeared in their signaling gene had better-predicted prognoses than those without; in
276 other words, these samples were expected to survive longer than others. Conversely, samples with
277 pessimistic sMFE biomarkers in their signaling genes had lower predicted prognoses than those without;
278 that is, these samples were expected to survive for a shorter time than other samples. A total of 4 genes
279 were identified as optimistic sMFE biomarkers for ESCA (Fig. 4A and Supplementary Table S2), while 6
280 genes were identified as pessimistic sMFE biomarkers (Fig. 4A and Supplementary Table S2). Specifically,
281 survival times of samples with O-sMFE biomarkers *SCYL3* and *TRPS1* (*P*=0.016 and 0.036 respectively)
282 were longer than those without, as shown in Fig. 4A. Conversely, samples with P-sMFE biomarkers *EPCAM*
283 and *POLK* (*P*=0.0057 and 0.0064 respectively) had shorter survival times than those without.

284   In the analysis for COAD, *AGR2* (*P*=0.044; Fig. 4B) and *PCM1* (*P*=0.046; Fig. 4B) were identified as
285 O-SMFE biomarkers, and it's seen the survival time of these identified samples tend to be longer than that
286 of other samples according to survival curve analysis (Fig. 4B). Additionally, *ATP1A3* (*P*<0.0001; Fig. 4B)
287 and *PARL* (*P*=0.00093; Fig. 4B) were identified as P-SMFE biomarkers, survival analysis in Fig. 4B showed
288 the survival time of these identified samples tend to be shorter than that of other samples. When applied to
289 KIRC dataset, samples with O-sMFE biomarkers *MAML2* and *RNF5* (*P*=0.0058 and 0.0034 respectively)
290 had better prognoses, with longer survival times, compared to other samples. Conversely, samples with P-
291 sMFE biomarkers *BST1* and *VPS39* (*P*=0.0085 and 0.00011 respectively) had worse prognoses, with
292 shorter Median Survival Times of almost five years. As shown in Fig. 4D, for LUAD, *ACOX2* (*P*=0.021; Fig.
293 4D) and *KHDRBS3* (*P*=0.027; Fig. 4D) were identified as O-SMFE biomarkers and *FLNC* (*P*=0.0021; Fig.
294 4D) and *FZD1* (*P*=0.0091; Fig. 4D) were identified as P-SMFE biomarkers.

295   We propose a method to uncover novel biomarkers, drug targets and key regulators, which are usually
296 overlooked by traditional studies due to their non-differential expression. In addition, all identified
297 optimistic/pessimistic sMFE biomarkers for the four cancers were supplemented in Supplementary Table

298 S2.

3.4 Functional analysis of common signaling genes

300 Common signaling genes for the four cancers were extracted for functional analysis (Fig. 5A). As shown in
301 Fig. 5B, not only were there occurred rich intersections among signaling genes from different cancers, but
302 their biological functions were also closely related. In addition, we performed the KEGG pathway
303 enrichment analysis of common signaling genes, where their samples located in tipping points, to explore
304 the biological processes in which they participated. For these four tumor datasets, as shown in Fig. 5C-F,
305 these signaling genes were significantly associated with cancer-associated pathways. For example Cell
306 cycle(Hartwell & Kastan 1994), PI3K-Akt signaling pathway(Martini et al. 2014), Spliceosome(Yang et al.
307 2022), Wnt signaling pathway(Zhou et al. 2022). In addition, Figure 5G and H illustrated that common
308 signaling genes were mainly enriched in the above-mentioned pathway, which suggests that these common
309 signaling genes may serve an important contributor to cancer development and progression. For instance,
310 *Ras-related C3 botulinum toxin substrate 2 (RAC2)* promotes abnormal proliferation of quiescent cells by
311 enhanced *JUNB* expression via the *MAL-SRF* pathway(Pei et al. 2018), *PIK3CA*-mutation was reported
312 related to metastatic breast cancer(Higgins et al. 2012; Mosele et al. 2020), *Protein tyrosine kinase 2*
313 *(PTK2)* was a novel therapeutic target to overcome acquired *EGFR-TKI* resistance in non-small cell lung
314 cancer and as a Driver of *Radioresistance* in *HPV-negative Head and Neck CancerPTK2/FAK and*
315 *Radioresistance*(Skinner et al. 2016; Tong et al. 2019), *RhoA* is associated with invasion and poor
316 prognosis in colorectal cancer(Jeong et al. 2016).

4. Discussion

318 It is crucial to hunt for the critical state of complex biological systems. Suppose an early warning signal of
319 the critical state can be provided before the disease deterioration. In that case, appropriate time can be
320 provided to prevent or at least prepare for a catastrophic deterioration. However, small sample sizes are a
321 challenge in biological studies and clinical practice, especially in cancer, and often lead to model errors and
322 biases in the analysis. To overcome these problems when identifying transition points or critical states that
323 precede disease states, the single-sample Markov flow entropy (sMFE) method was proposed in this study.
324 Its effectiveness and reliability have been tested across multiple diseases.

325 In this study, we developed a computational method, single-sample Markov flow entropy (sMFE), to
326 explore the dynamic changes in combined effects on molecular associations and thus characterize the
327 perturbation of a single sample close to the tipping point. In addition, functional enrichment analysis
328 revealed that the common signaling genes for four cancer datasets are involved in significant biological
329 processes or pathways, for instance, Cell cycle(Hartwell & Kastan 1994), PI3K-Akt signaling
330 pathway(Martini et al. 2014), Spliceosome(Yang et al. 2022), Wnt signaling pathway(Zhou et al. 2022) (Fig.
331 5). Further analysis revealed two distinct types of biomarkers derived from the signaling genes (Liu et al.
332 2019), i.e., pessimistic sMFE (P-sMFE) and optimistic sMFE (O-sMfE) biomarkers. Furthermore, those two
333 types of biomarkers not only the signaling genes, pessimistic/optimistic biomarkers, but also the non-

334 differentially expressed genes and are therefore usually screened out by most researchers in the first step,
335 so they can provide novel insight for further study into the molecular mechanisms of how tumor onset and
336 disease deterioration work.

337 In a word, the sMFE approach has several important advantages. First, this method is a model-free
338 data-driven and single-sample-based algorithm, which is conducive to the development of personalized
339 medicine. Second, our method was based on the direct interaction network, constructed from an a priori
340 knowledge-based PPI network by GNIPLR, which is different from the past studies that mainly focused on
341 undirected networks. Third, the sMFE method provided two distinct types of biomarkers applicable for
342 identifying novel biomarkers unlike any previous one, drug targets in drug development and prognostic
343 indicators for prognostic analysis.

5. Conclusions

345 The sMFE method provides a solution to the identification problem of critical states/pre-disease states of
346 complex diseases solely based on a single sample, making it suitable for real-world clinical data. In addition
347 to effectively identifying the critical states or transition points of the four cancers, our method provided two
348 new prognostic biomarkers, optimistic sMFE (O-sMFE) and pessimistic sMFE (P-sMFE) biomarkers.
349 Therefore, this method has great practical application potential in the fields of personalized medicine,
350 identification of molecular mechanisms of disease progression, and preventive medicine.
351
352

List of abbreviations

354 sMFE: single-sample Markov flow entropy; DNB: the dynamical network biomarker; LUAD: lung
355 adenocarcinoma; COAD: colon adenocarcinoma; KIRC: kidney Renal Clear Cell Carcinoma; ESCA:
356 esophageal carcinoma; TCGA: the cancer genome atlas; SD: standard deviation; PCC: Pearson correlation
357 coefficient; signaling genes: the top 5% of genes with the highest sMFE score; O-sMFE biomarker:
358 optimistic sMFE biomarker; P-sMFE biomarker: pessimistic sMFE biomarker. DEGs: differentially
359 expressed genes.
360

Declarations

Ethics approval and consent to participate

363 Not applicable.
364

Consent for publication

366 Not applicable.
367

Availability of data and materials

Kidney renal clear cell carcinoma (KIRC), lung adenocarcinoma (LUAD), esophageal carcinoma (ESCA), colon adenocarcinoma (COAD), are available from the cancer genome atlas (TCGA) database (http://cancergenome.nih.gov). The source code of algorithm is accessed in supplementary material sMFE.rar.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by the National Natural Science Foundation of China (Nos. 62172164, 12026608, and 11971176), Guangdong Basic and Applied Basic Research Foundation (Grant Nos. 2019B151502062, 2021A1515012317).

Authors' contributions

RL, PC, and JYZ conceived the research. JTL, YT, and RQL performed the real data analysis. JTL, and JYZ performed the functional analysis. All authors wrote the paper. All authors read and approved the final manuscript.

Acknowledgements

Not applicable.

Authors' information

[1]School of Mathematics, South China University of Technology, Guangzhou 510640, China

[2]Pazhou Lab, Guangzhou 510330, China

[3]School of Mathematics and Big Data, Foshan University, Foshan 528000, China

Additional file

Additional file. **Supplementary information of sMFE**.

References

Chen L, Liu R, Liu Z-P, Li M, and Aihara K. 2012a. Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. *Scientific Reports* 2:1-8.

Chen LN, Liu R, Liu ZP, Li MY, and Aihara K. 2012b. Detecting early-warning signals for sudden deterioration of complex diseases by dynamical network biomarkers. *Scientific Reports* 2:8. 10.1038/srep00342

Chen P, Chen E, Chen L, Zhou XJ, and Liu R. 2019. Detecting early‐warning signals of influenza outbreak based on dynamic network marker. *Journal of cellular and molecular medicine* 23:395-404.

406 Chen P, Liu R, Li Y, and Chen L. 2016. Detecting critical state before phase transition of complex biological systems
407   by hidden Markov model. *Bioinformatics* 32:2143-2150.

408 Chiang AC, and Massagué J. 2008. Molecular basis of metastasis. *New England Journal of Medicine* 359:2814-2823.

409 Dong X, Chen X, Lu D, Diao D, Liu X, Mai S, Feng S, and Xiong G. 2022. LncRNA miR205HG hinders HNRNPA0
410   translation: anti‑oncogenic effects in esophageal carcinoma. *Molecular oncology* 16:795-812.

411 Gao R, Yan J, Li P, and Chen L. 2022. Detecting the critical states during disease development based on temporal
412   network flow entropy. *Briefings in Bioinformatics*.

413 Gorban AN, Tyukina TA, Pokidysheva LI, and Smirnova EV. 2021. Dynamic and thermodynamic models of adaptation.
414   *Phys Life Rev* 37:17-64. 10.1016/j.plrev.2021.03.001

415 Gu T, and Zhao X. 2019. Integrating multi-platform genomic datasets for kidney renal clear cell carcinoma subtyping
416   using stacked denoising autoencoders. *Scientific Reports* 9:1-11.

417 Guo W-F, Yu X, Shi Q-Q, Liang J, Zhang S-W, and Zeng T. 2021a. Performance assessment of sample-specific network
418   control methods for bulk and single-cell biological data analysis. *PLoS Computational Biology* 17:e1008962.

419 Guo WF, Zhang SW, Feng YH, Liang J, Zeng T, and Chen LN. 2021b. Network controllability-based algorithm to target
420   personalized driver genes for discovering combinatorial drugs of individual patients. *Nucleic acids research*
421   49:21. 10.1093/nar/gkaa1272

422 Hari DM, Leung AM, Lee J-H, Sim M-S, Vuong B, Chiu CG, and Bilchik AJ. 2013. AJCC Cancer Staging Manual 7th edition
423   criteria for colon cancer: do the complex modifications improve prognostic assessment? *Journal of the*
424   *American College of Surgeons* 217:181-190.

425 Hartwell LH, and Kastan MB. 1994. Cell cycle control and cancer. *Science* 266:1821-1828.

426 Higgins MJ, Jelovac D, Barnathan E, Blair B, Slater S, Powers P, Zorzi J, Jeter SC, Oliver GR, and Fetting J. 2012.
427   Detection of Tumor PIK3CA Status in Metastatic Breast Cancer Using Peripheral BloodPIK3CA Mutations in
428   Plasma of Patients with Metastatic Breast Cancer. *Clinical cancer research* 18:3462-3469.

429 Jeong D, Park S, Kim H, Kim C-J, Ahn TS, Bae SB, Kim HJ, Kim TH, Im J, and Lee MS. 2016. RhoA is associated with
430   invasion and poor prognosis in colorectal cancer. *International Journal of Oncology* 48:714-722.

431 Koizumi K, Oku M, Hayashi S, Inujima A, Shibahara N, Chen L, Igarashi Y, Tobe K, Saito S, and Kadowaki M. 2019.
432   Identifying pre-disease signals before metabolic syndrome in mice by dynamical network biomarkers.
433   *Scientific Reports* 9:1-11.

434 Lesterhuis WJ, Bosco A, Millward MJ, Small M, Nowak AK, and Lake RA. 2017. Dynamic versus static biomarkers in
435   cancer immune checkpoint blockade: unravelling complexity. *Nature Reviews Drug Discovery* 16:264-272.

436 Litt B, Esteller R, Echauz J, D'Alessandro M, Shor R, Henry T, Pennell P, Epstein C, Bakay R, Dichter M, and
437   Vachtsevanos G. 2001. Epileptic seizures may begin hours in advance of clinical onset: A report of five
438   patients. *Neuron* 30:51-64. 10.1016/s0896-6273(01)00262-8

439 Liu R, Aihara K, and Chen L. 2013a. Dynamical network biomarkers for identifying critical transitions and their driving
440   networks of biologic processes. *Quantitative Biology* 1:105-114.

441 Liu R, Chen P, and Chen L. 2020. Single-sample landscape entropy reveals the imminent phase transition during
442   disease progression. *Bioinformatics* 36:1522-1532.

443 Liu R, Wang X, Aihara K, and Chen L. 2014. Early diagnosis of complex diseases by molecular biomarkers, network
444   biomarkers, and dynamical network biomarkers. *Medicinal research reviews* 34:455-478.

445 Liu X, Chang X, Leng S, Tang H, Aihara K, and Chen L. 2019. Detection for disease tipping points by landscape dynamic
446   network biomarkers. *Natl Sci Rev* 6:775-785. 10.1093/nsr/nwy162

447   Liu X, Liu R, Zhao X-M, and Chen L. 2013b. Detecting early-warning signals of type 1 diabetes and its leading
448   biomolecular networks by dynamical network biomarkers. *BMC medical genomics* 6:1-10.
449   Martini M, De Santis MC, Braccini L, Gulluni F, and Hirsch E. 2014. PI3K/AKT signaling pathway and cancer: an
450   updated review. *Annals of medicine* 46:372-383.
451   McSharry PE, Smith LA, and Tarassenko L. 2003. Prediction of epileptic seizures: are nonlinear methods relevant?
452   *Nature medicine* 9:241-242.
453   Mosele F, Stefanovska B, Lusque A, Dien AT, Garberis I, Droin N, Le Tourneau C, Sablin M-P, Lacroix L, and Enrico D.
454   2020. Outcome and molecular landscape of patients with PIK3CA-mutated metastatic breast cancer. *Annals*
455   *of Oncology* 31:377-386.
456   Pei H, Guo Z, Wang Z, Dai Y, Zheng L, Zhu L, Zhang J, Hu W, Nie J, and Mao W. 2018. RAC2 promotes abnormal
457   proliferation of quiescent cells by enhanced JUNB expression via the MAL-SRF pathway. *Cell cycle* 17:1115-
458   1123.
459   Peng H, Zhong J, Chen P, and Liu R. 2022. Identifying the critical states of complex diseases by the dynamic change
460   of multivariate distribution. *Brief Bioinform* 23. 10.1093/bib/bbac177
461   Scheffer M, Bascompte J, Brock WA, Brovkin V, Carpenter SR, Dakos V, Held H, Van Nes EH, Rietkerk M, and Sugihara
462   G. 2009. Early-warning signals for critical transitions. *Nature* 461:53-59.
463   Shi J, Teschendorff AE, Chen W, Chen L, and Li T. 2018. Quantifying Waddington's epigenetic landscape: a comparison
464   of single-cell potency measures. *Brief Bioinform*. 10.1093/bib/bby093
465   Skinner HD, Giri U, Yang L, Woo SH, Story MD, Pickering CR, Byers LA, Williams MD, El-Naggar A, and Wang J. 2016.
466   Proteomic Profiling Identifies PTK2/FAK as a Driver of Radioresistance in HPV-negative Head and Neck
467   CancerPTK2/FAK and Radioresistance. *Clinical cancer research* 22:4643-4650.
468   Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, and
469   Tsafou KP. 2015. STRING v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic*
470   *acids research* 43:D447-D452.
471   Tong X, Tanino R, Sun R, Tsubata Y, Okimoto T, Takechi M, and Isobe T. 2019. Protein tyrosine kinase 2: a novel
472   therapeutic target to overcome acquired EGFR-TKI resistance in non-small cell lung cancer. *Respiratory*
473   *research* 20:1-14.
474   Venegas JG, Winkler T, Musch G, Melo MFV, Layfield D, Tgavalekos N, Fischman AJ, Callahan RJ, Bellani G, and Harris
475   RS. 2005. Self-organized patchiness in asthma as a prelude to catastrophic shifts. *Nature* 434:777-782.
476   10.1038/nature03490
477   Yang H, Beutler B, and Zhang D. 2022. Emerging roles of spliceosome in cancer and immunity. *Protein & Cell* 13:559-
478   579.
479   Zhang Y, Chang X, and Liu X. 2021. Inference of gene regulatory networks using pseudo-time series data.
480   *Bioinformatics*. 10.1093/bioinformatics/btab099
481   Zhong J, Han C, Wang Y, Chen P, and Liu R. 2022. Identifying the critical state of complex biological systems by the
482   directed-network rank score method. *Bioinformatics*. 10.1093/bioinformatics/btac707
483   Zhou Y, Xu J, Luo H, Meng X, Chen M, and Zhu D. 2022. Wnt signaling pathway in cancer immunotherapy. *Cancer*
484   *Letters* 525:84-96.
485

Figure 1

Figure1.The schematic of the sMFE method for detecting the critical state of complex diseases.

(A) The schematic diagram of the evolution of complex diseases. In dynamical systems theory, the evolution of disease can generally be divided into three stages or states, namely a relatively normal (before-transition) state, a pre-disease (critical) state, and a disease (after-transition) state. Among them, the relatively normal and the pre-disease states are reversible, while the disease state is irreversible. In addition, the expression of each molecular network in the before-transition and after-transition states is low-fluctuating and highly volatile in the critical state; **(B)** Single-sample Markov flow entropy (sMFE) algorithm. Given a number of reference samples that can be obtained from normal cohort, and a directed network that used the GNIPLR (Gene networks inference using projection and lagged regression) method [14] to orient the PPI network. Specifically, both the reference samples and the single-sample to be determined were mapped to this directed network, which can be divided into local networks. sMFE is calculated on a local network based on a single sample of any individual. For each local network centered on gene k, the local sMFE score is calculated; **(C)** During the evolution of complex diseases, the sMFE score [Equation (7)]) suddenly increases in the pre-disease state. This mutation in the sMFE indicates a complex biological system's tipping point (or critical state).
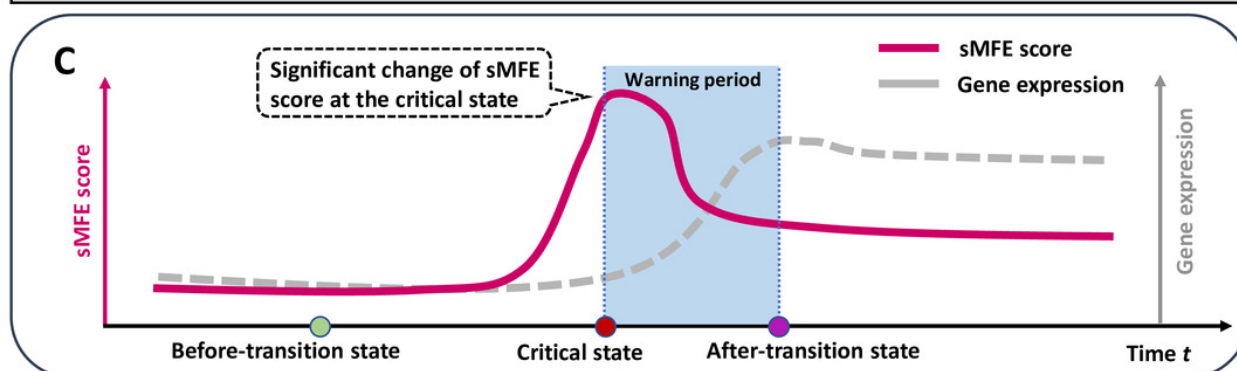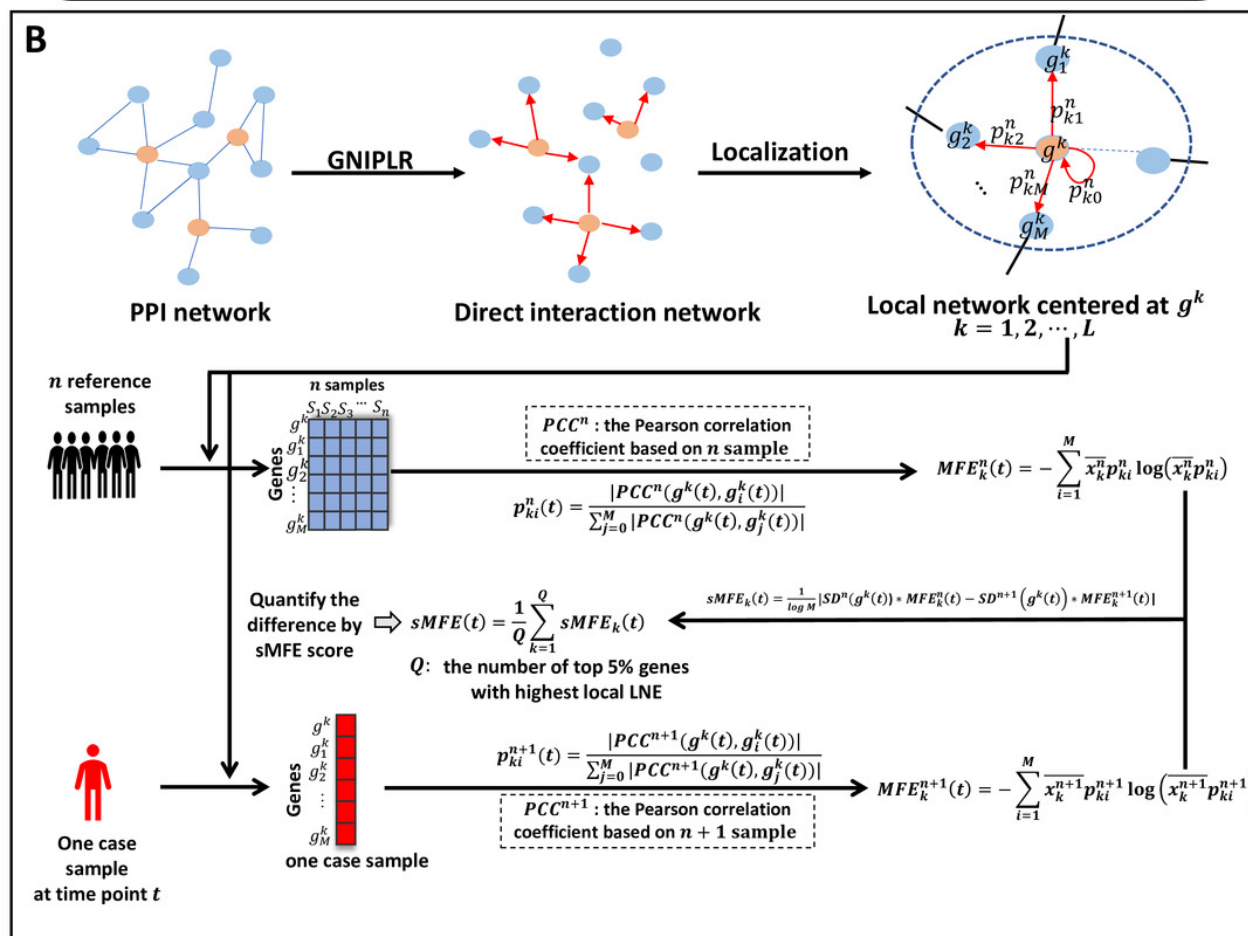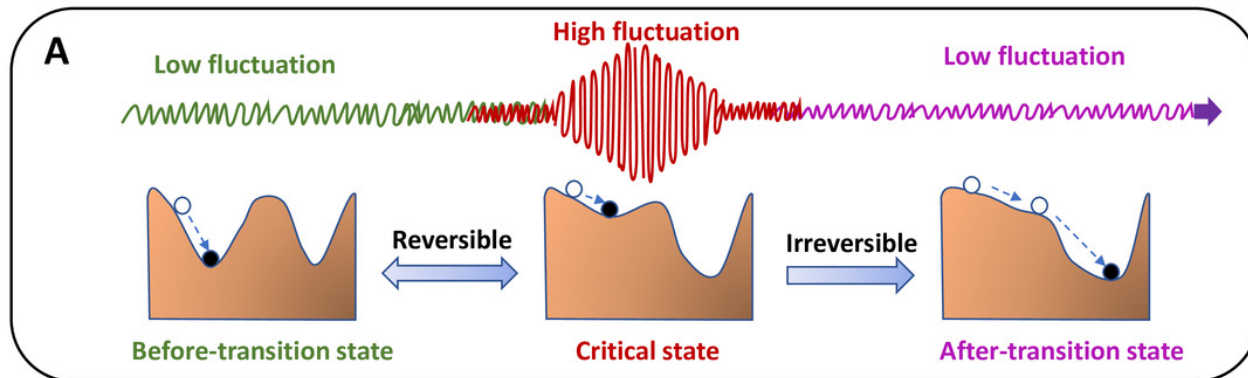
Figure 2

Figure2. Detection of the critical states for the four cancers.

Dynamic changes between sMFE and the mean DEGs gene expression across four cancer datasets: (A) ESCA; (B) COAD; (C) KIRC; (D) LUAD. **Survival analysis for samples adopt from before and after the critical state for four cancers:** (E) ESCA; (F) COAD; (G) KIRC; (H) LUAD.
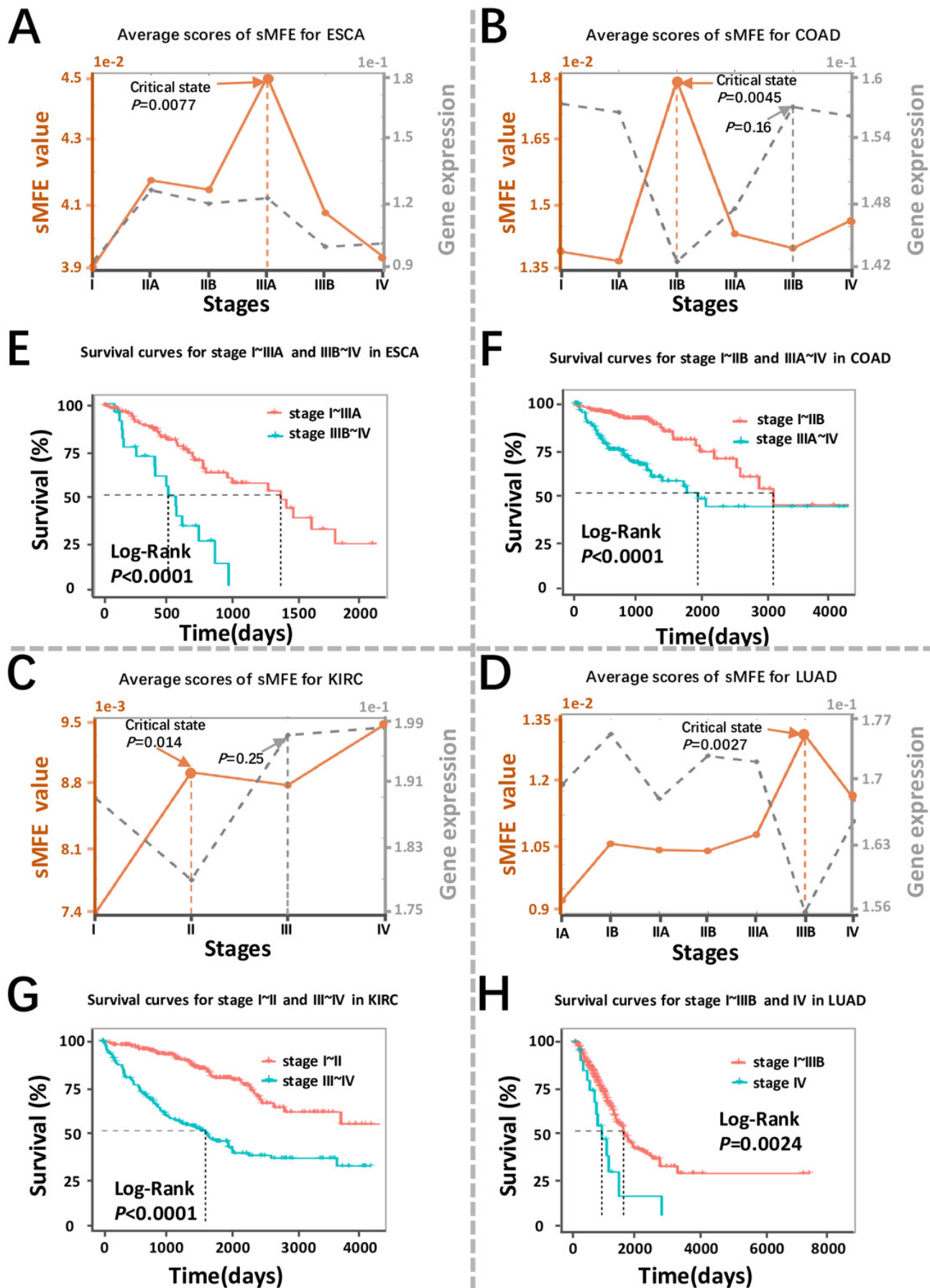
Figure 3

Figure3. The dynamic evolution of the regulatory networks for signaling genes in ESCA, COAD, KIRC, and LUAD.

The landscape of local sMFE values illustrates the dynamic evolution of network entropy in a global view for ESCA (A), COAD (B), KIRC (C), and LUAD (D), respectively. Key gene regulatory networks were reconstructed using sMFE for the signaling genes (top 5% genes with the largest local sMFE value). The network structure was derived by mapping s ignaling genes to the STRING PPI network. We discarded all the isolated nodes without any links to other nodes. The c olor of each node represents the mean local sMFE value , and the color of each edge represents the absolute value of the Pearson correlation coefficient . (A) In ESCA, the signaling gene group/module evolved, and there was a significant change in the network structure at stage IIIA. (B) Similarly, the network sMFE score significantly changed at stage IIB in COAD. (C) The network sMFE score significantly changed at stage II in KIRC. (D) The network sMFE score significantly changed at stage IIIB in LUAD.

A sMFE landscape of ESCA   **B** sMFE landscape of COAD   **C** sMFE landscape of KIRC   **D** sMFE landscape of LUAD

★ sMFE signal

E Network change of ESCA   **F** Network change of COAD   **G** Network change of KIRC   **H** Network change of LUAD

IIB   IIA   I   IIIA

IIIA(Critical state)   IIB(Critical state)   II(Critical state)   IIIB(Critical state)

IIIB   IIIA   III   IV

Edge weight |PCC|
0   0.5   1

Scaled sMFE score
0   0.5   1
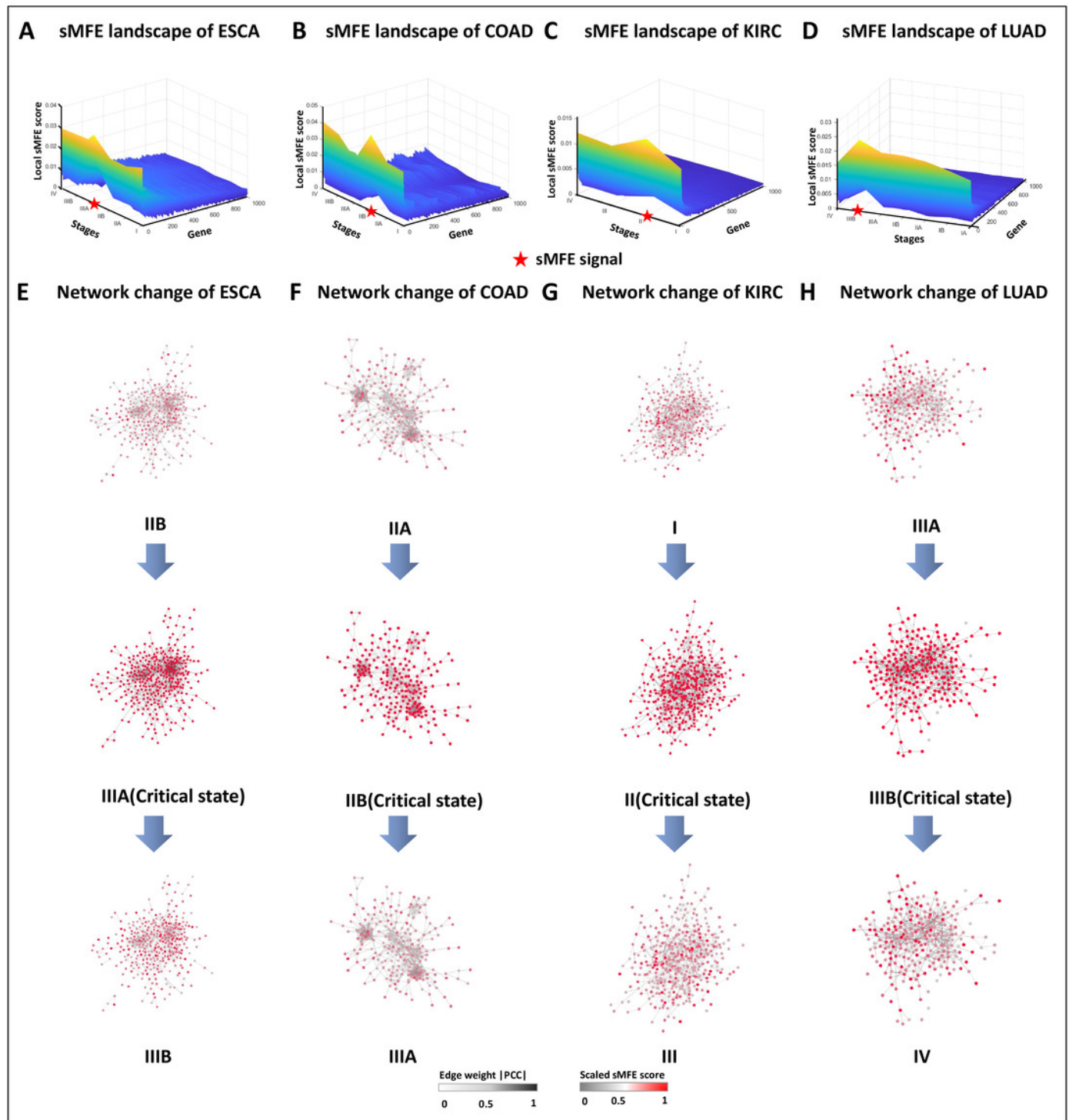
Figure 4

Figure4. Comparison of survival curves between samples with and without O-sMFE and P-sMFE biomarkers for ESCA, COAD, KIRC, and LUAD.

(A) The O-sMFE biomarkers *SCYL3* and *TRPS1* and the P- sMFE biomarkers *EPCAM* and *POLK* in ESCA; **(B)** the O-sMFE biomarkers *AGR2* and *PCM1* and the P-sMFE biomarkers *ATP1A3* and *PARL* in COAD; **(C)** the O-sMFE biomarkers *MAML2* and *RNF5* and the P-sMFE biomarkers *BST1* and *VPS39* in KIRC; **(D)** the O-sMFE biomarkers *ACOX2* and *KHDRBS3* and the P-sMFE biomarkers *FLNC* and *FZD1* in LUAD. Samples with O-sMFE or P-sMFE biomarkers in their signaling genes were deemed samples with biomarkers, while others were deemed samples without biomarkers.
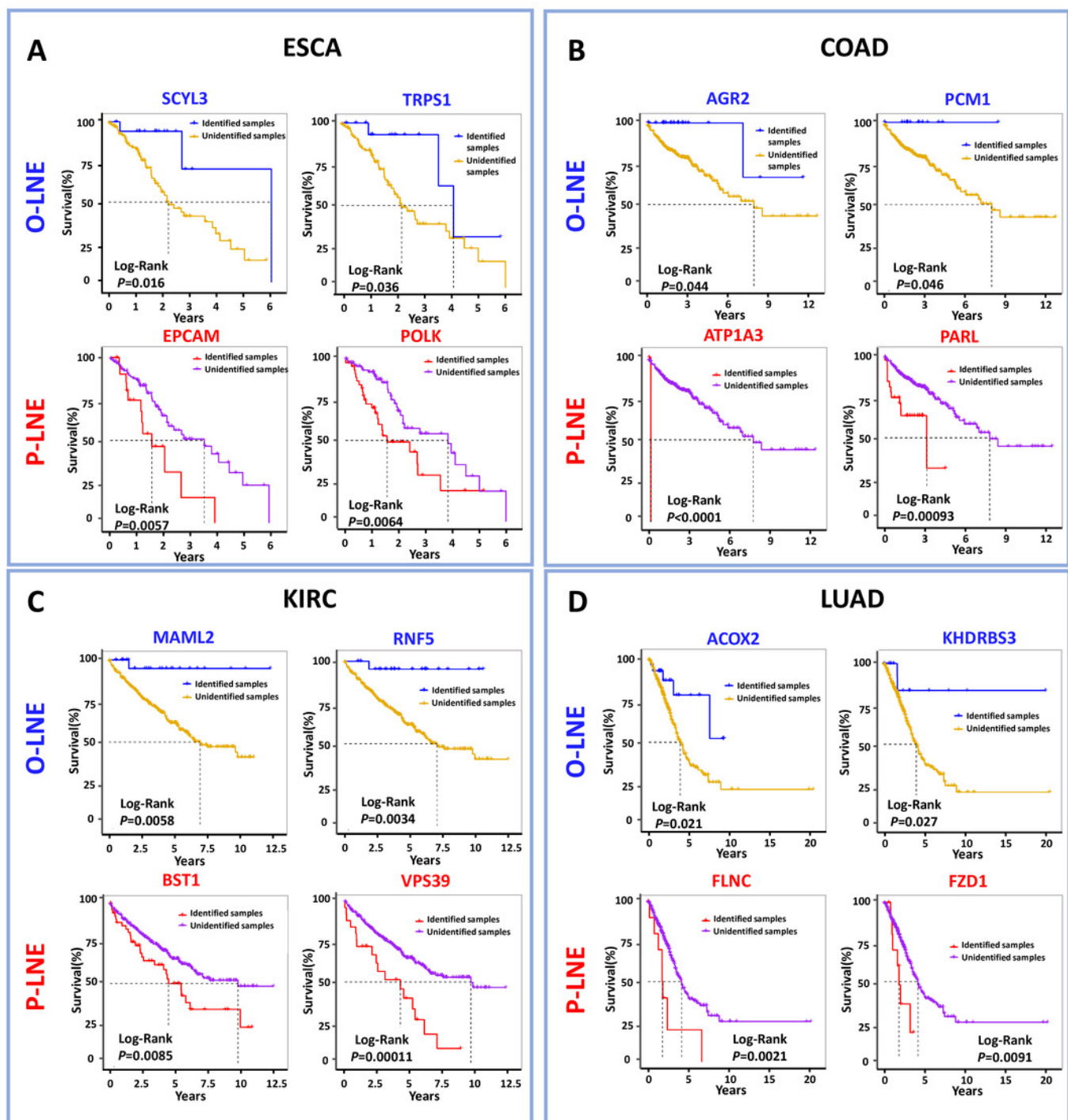
Figure 5

Figure5. Functional analysis of common signaling genes in different cancer datasets.

(A) Common signaling genes among ESCA, LUAD, COAD, and KIRC. (B) There is a considerable overlap of the same signaling genes and their biological functions among the four cancers. The outer ring represents different cancer-signaling genomes, and the inner ring represents identical genes and functions. The interconnections of identical genes are indicated by purple lines and functional links by blue lines. KEGG pathway enrichment analysis of signaling genes in four cancers: ESCA(C), COAD(D), KIRC(E), and LUAD(F). (G) Common signaling genes are involved in multiple cancer-related pathways. (H) Detailed common signaling genes in cancer-related pathways.