Supplementary Information: Single-sample network module biomarkers (sNMB) reveals the pre-deterioration stage of disease progression

Contents

A. The detailed theory of dynamical system for numerical simulation

A regulatory network with 8 genes (see Fig. 3A in the main text) is applied to conduct a numerical simulation for detecting the pre-disease stage based on sDNM method. Such molecular regulatory networks are usually employed to study various biological processes, such as transcription, translation, and diffusion (Chen et al., 2002; Chen et al., 2009). The following 8 differential equations represent a gene regulation of a network with 8 genes. In this network, the gene regulation represented in the Michaelis-Menten form is linearly proportional to the concentrations of the corresponding genes.

$$
\begin{cases}
\dfrac{dz_1(t)}{dt} = \dfrac{(4-3|p|)z_2(t)}{10(1+z_2(t))} - \dfrac{(4+3|p|)}{10} z_1(t) + \zeta_1(t) \\
\dfrac{dz_2(t)}{dt} = \dfrac{(4-3|p|)z_1(t)}{10(1+z_1(t))} - \dfrac{(4+3|p|)z_2(t)}{10(1+z_2(t))} + \zeta_2(t) \\
\dfrac{dz_3(t)}{dt} = \dfrac{(6|p|-10)}{10} + \dfrac{(5-3|p|)}{10(1+z_1(t))} + \dfrac{(5-3|p|)}{10(1+z_2(t))} - z_3(t) + \zeta_3(t) \\
\dfrac{dz_4(t)}{dt} = \dfrac{(6|p|-12)}{10} + \dfrac{(6-3|p|)z_1(t)}{10(1+z_1(t))} + \dfrac{(6-3|p|)z_2(t)}{10(1+z_2(t))} - \dfrac{6}{5} z_4(t) + \zeta_4(t) \\
\dfrac{dz_5(t)}{dt} = \dfrac{(6|p|-14)}{10} + \dfrac{(7-3|p|)z_1(t)}{10(1+z_1(t))} + \dfrac{(7-3|p|)z_2(t)}{10(1+z_2(t))} - \dfrac{7}{5} z_5(t) + \zeta_5(t) \\
\dfrac{dz_6(t)}{dt} = -\dfrac{3}{5} + \dfrac{1}{10(1+z_1(t))} + \dfrac{1}{10(1+z_2(t))} + \dfrac{1}{5(1+z_5(t))} + \dfrac{1}{5(1+z_7(t))} \\
+ \dfrac{z_8(t)}{5(1+z_8(t))} - \dfrac{8}{5} z_6(t) + \zeta_6(t) \\
\dfrac{dz_7(t)}{dt} = \dfrac{z_8(t)}{10(1+z_8(t))} - \dfrac{19}{10} z_7(t) + \zeta_7(t) \\
\dfrac{dz_8(t)}{dt} = \dfrac{z_7(t)}{10(1+z_7(t))} - \dfrac{19}{10} z_8(t) + \zeta_8(t)
\end{cases}
\qquad (S1)
$$

where p represents a scalar control parameter and $\varsigma_i(t)$ (i = 1, 2, ..., 8) are Gaussian noises with zero means and covariances $\mathrm{k}_{ij} = Cov(\varsigma_i, \varsigma_j)$. $z_i(t)$ (i = 1, 2, ..., 8) represents the concentrations of mRNA-i. In Eq.(S1), the degradation rates of mRNAs is $\mathrm{R}=(\dfrac{(4+3|p|)}{10}, \dfrac{(4+3|p|)}{10}, 1, \dfrac{6}{5}, \dfrac{7}{5}, \dfrac{8}{5}, \dfrac{19}{10}, \dfrac{19}{10})$. $\overline{Z} = (\overline{z_1}, \overline{z_2}, \overline{z_3}, \ldots, \overline{z_8}) = (0, 0, 0, \ldots, 0)$ is the stable equilibrium point of the differential equations Eq.(S1) . Based on the Euler scheme (Kloeden et al., 1999), the differential equations Eq.(S1) is transformed into the difference equations $Z(k+1) = f(Z(k), S)$ with a small time interval Δt . The result is as follows:

$$
\begin{cases}
z_1(k+1) = z_1(k) + \left[\dfrac{(4-3|p|)z_2(t)}{10(1+z_2(t))} - \dfrac{(4+3|p|)}{10} z_1(t) + \zeta_1(t) \right] \Delta t \\
z_2(k+1) = z_2(k) + \left[\dfrac{(4-3|p|)z_1(t)}{10(1+z_1(t))} - \dfrac{(4+3|p|)z_2(t)}{10(1+z_2(t))} + \zeta_2(t) \right] \Delta t \\
z_3(k+1) = z_3(k) + \left[\dfrac{(6|p|-10)}{10} + \dfrac{(5-3|p|)}{10(1+z_1(t))} + \dfrac{(5-3|p|)}{10(1+z_2(t))} - z_3(t) + \zeta_3(t) \right] \Delta t \\
z_4(k+1) = z_4(k) + \left[\dfrac{(6|p|-12)}{10} + \dfrac{(6-3|p|)z_1(t)}{10(1+z_1(t))} + \dfrac{(6-3|p|)z_2(t)}{10(1+z_2(t))} - \dfrac{6}{5} z_4(t) + \zeta_4(t) \right] \Delta t \\
z_5(k+1) = z_5(k) + \left[\dfrac{(6|p|-14)}{10} + \dfrac{(7-3|p|)z_1(t)}{10(1+z_1(t))} + \dfrac{(7-3|p|)z_2(t)}{10(1+z_2(t))} - \dfrac{7}{5} z_5(t) + \zeta_5(t) \right] \Delta t \\
z_6(k+1) = z_6(k) + \left[-\dfrac{3}{5} + \dfrac{1}{10(1+z_1(t))} + \dfrac{1}{10(1+z_2(t))} + \dfrac{1}{5(1+z_5(t))} + \dfrac{1}{5(1+z_7(t))} \right. \\
\left. + \dfrac{z_8(t)}{5(1+z_8(t))} - \dfrac{8}{5} z_6(t) + \zeta_6(t) \right] \Delta t \\
z_7(k+1) = z_7(k) + \left[\dfrac{z_8(t)}{10(1+z_8(t))} - \dfrac{19}{10} z_7(t) + \zeta_7(t) \right] \Delta t \\
z_8(k+1) = z_8(k) + \left[\dfrac{z_7(t)}{10(1+z_7(t))} - \dfrac{19}{10} z_8(t) + \zeta_8(t) \right] \Delta t
\end{cases} \qquad (S2)
$$

Where $Z(k)$ is the vector $Z(t)$ at the time instant $k\Delta t$. The Jacobian matrix of Eq.(S2) is defined as $J = \left. \dfrac{\partial f(Z(k);S)}{\partial Z} \right|_{Z=\overline{Z}}$, with

$$
J = e^{\Delta t \bullet A} \qquad (S3)
$$

By taking $\Delta t = 1$, we can obtain eight distinct eigenvalues from Eq.(S3), in which the largest eigenvalue satisfies $0.66^{|p|} \to 1$ when $p \to 0$. Therefore, if $p \in (0,1]$, the equilibrium point $\overline{Z}$ is stable. The special point $p = 0$ is a bifurcation point, at which the system undergoes a critical transition. Based on the theoretical model Eq.(S2), the numerical simulation dataset of the 8-gene expressions was collected under varying parameter p ranging from -0.45 to 0.15.

Besides, by using the simulated dataset generated from Eq. (S1), we have analyzed the critical signals when the number of reference samples varies. It can be seen from Figure S1 that the sNMB index accurately indicates the tipping point under different reference sample sizes. Therefore, the number of reference samples within a range barely affects the evolution tendency of the signal curve (e.g., abrupt increase when approaching the tipping point).

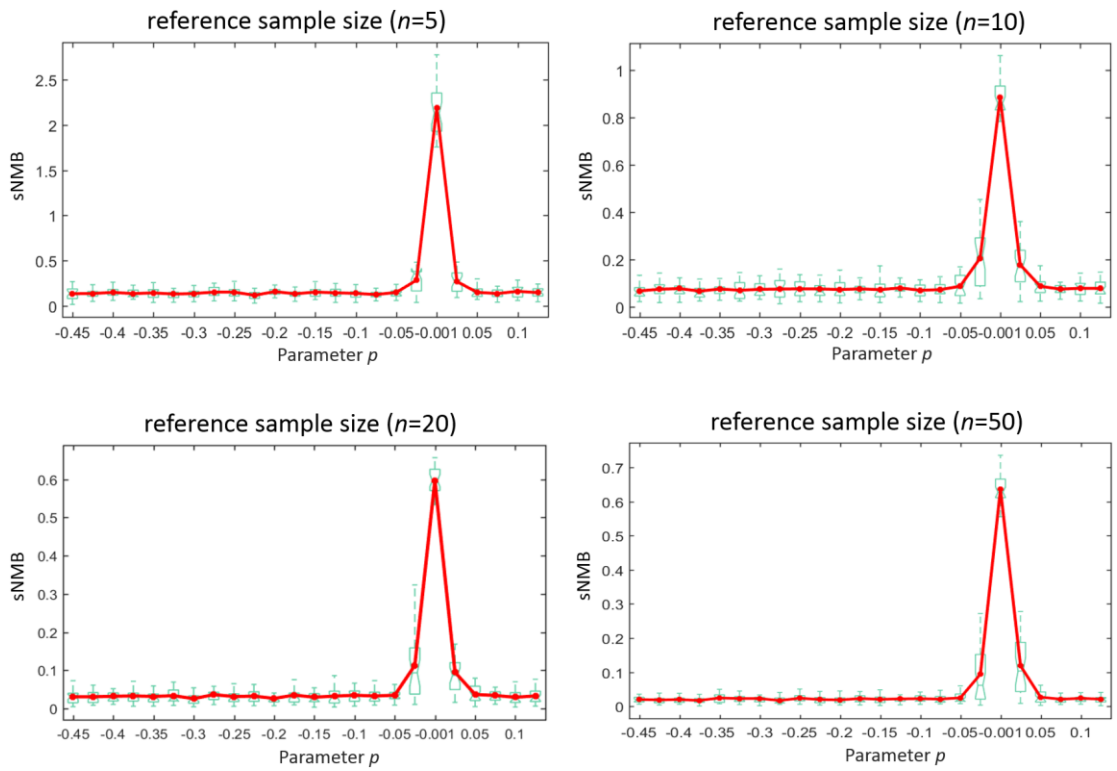


Figure S1: The critical signals under different reference sample sizes.

B. The verification of identified critical stage by the Kaplan-Meier (log-rank) survival analysis

B1. A schematic illustration for the verification of the identified critical stage

To validate the identification of the critical stage, the prognosis results respectively for before-transition and after-transition samples were exhibited and compared through Kaplan-Meier (log-rank) survival analysis. For example, we carried out the following steps to validate a critical transition of tumor disease at stage IIIB. Specifically, the survival time of samples from the before-transition stage is significantly longer than that of after-transition samples. However, there was no statistical difference in survival time between before and after any other stages. An illustrative process for validating the identified critical stage is shown as follows.

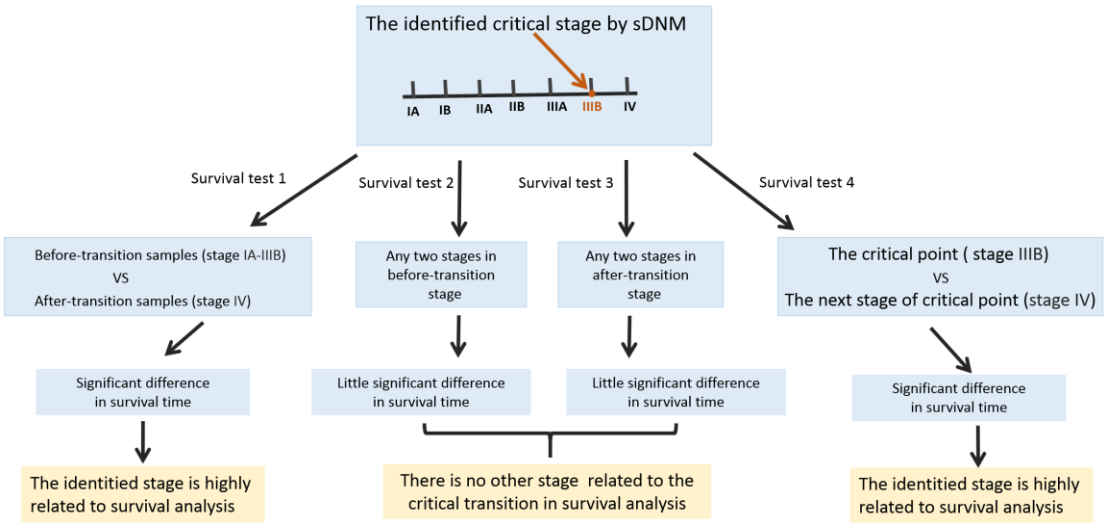


Figure S2: A schematic illustration for the verification of the identified critical stage.

B2. Validating the identified critical stage for three tumor disease

In this study, the proposed method is applied to three tumor datasets, including stomach adenocarcinoma (STAD), Esophageal carcinoma (ESCA), and Rectum adenocarcinoma (READ) from the cancer genome atlas (TCGA). Each tumor dataset was composed of tumor and tumor-adjacent samples. We grouped the tumor samples into different cancer stages according to corresponding clinical staging information of TCGA. Specifically, The tumor samples were grouped into seven stages (stage IA, IB, IC, IIA, IIB, IIIA, IIIB, and IV) for STAD, six stages (stage I, IB, IC, IIA, IIB, IIIA, IIIB, and IV) for ESCA, and four stages (stage I, II, III, and IV) for the READ. Overall, there are three criteria for the dataset selection: 1, they are all fine stages datasets, which indicate the state of cancer; 2, both tumor and tumor-adjacent samples are available in these datasets; 3, clinical information is available in all these datasets so as to carry out the survival analysis. The detailed number of samples within each stage is shown as the following table S1.

Table S1 The detailed number of samples within each stage of tumor disease dataset.

		STAD	ESCA	READ
Stage I	**Stage IA**	9	15	34
	Stage IB	18		
Stage II	**Stage IIA**	23	31	53
	Stage IIB	29	28	

Stage III	**Stage IIIA**	27	11	53
	Stage IIIB	20	14	
Stage IV	**Stage IV**	15	8	26
TA samples	**TA samples**	33	10	11

Based on before-transition and after-transition samples, we carried out the Kaplan-Meier (log-rank) survival analysis to validate the critical stage identified by sDNM method. The prognostic results were seen from the following Figure S3, which consistently indicated that the survival time of before-transition samples is significantly longer than that of after-transition samples. Besides, there is no significant difference among the survival time of before-transition samples. These results illustrated that the early-warning signals of a critical transition of survival time can be accurately detected by the sDNM method, that is, the identified critical stage can be validated by the sDNM method.

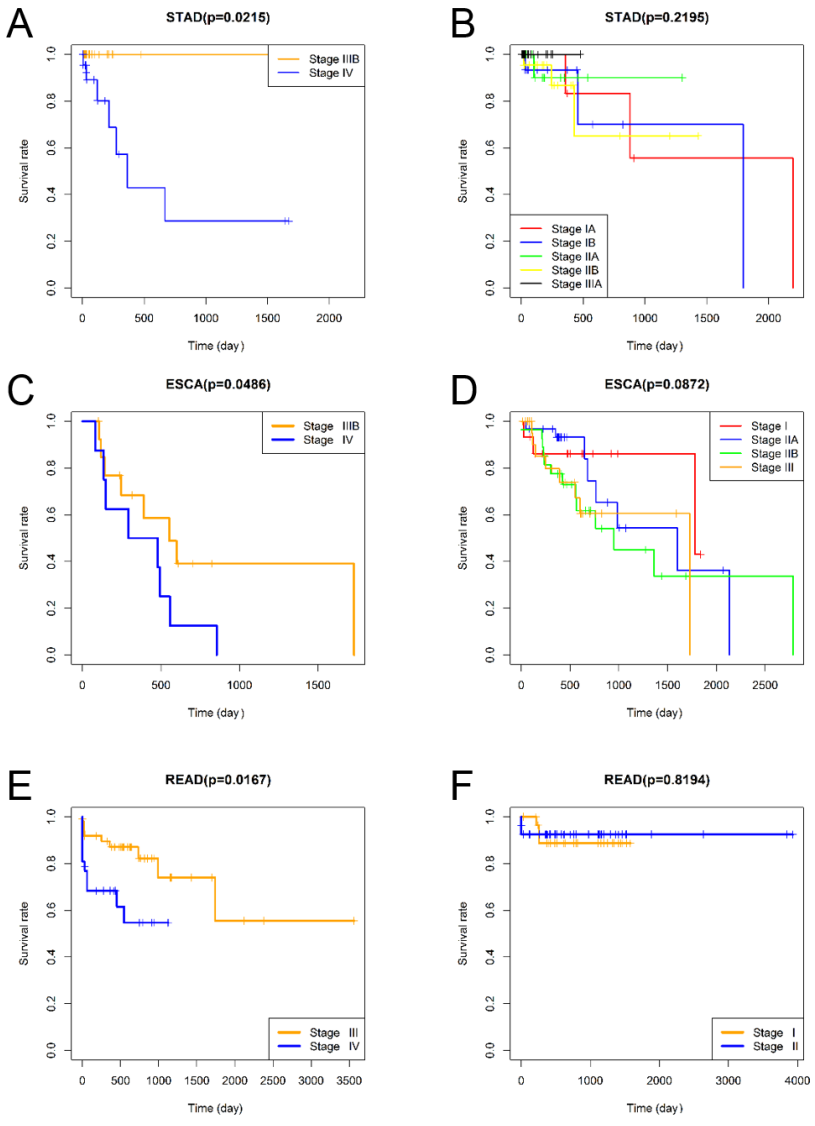


Figure S3: **A.** The critical point (IIIB) of STAD vs the next stage of critical point (IV) of STAD in survival analysis. **B.** Any two stages in before-transition period (IA-IIIA) of STAD in survival analysis. **C.** The critical point (IIIB) of ESCA vs the next stage of critical point (IV) of ESCA in survival analysis. **D.** Any two stages in before-transition period (I-III) of ESCA in survival analysis. **E.** The critical point (III) of READ vs the next stage of critical point (IV) of READ in survival analysis. **F.** Two stages in before-transition period (I-II) of READ in survival analysis.

In addition, at identified pre-deterioration stage (the critical point), the top 5% genes with the highest local sNMB scores were picked out as the signaling genes for further functional analyses. As shown in Figure S4, we performed the KEGG enrichment analysis of the specific signaling genes for three tumors (STAD, READ, and ESCA).

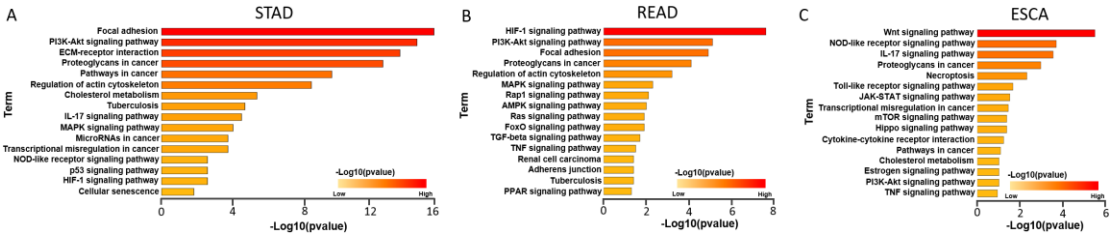


Figure S4: KEGG enrichment analysis of the specific signaling genes for (A) STAD, (B) READ, and (C) ESCA.

B3. The critical signals under different settings of the adjustable parameter *T*.

Three TCGA datasets (STAD, ESCA, and READ) have been employed to analyze the critical signals when the adjustable parameter *T* ranges from the top 1% to top 10%. It can be seen from Figure S5 that the sNMB index accurately indicates the tipping point with a similar tendency. Therefore, different settings of the adjustable parameter *T* do not affect the evolution tendency of the signal curve.

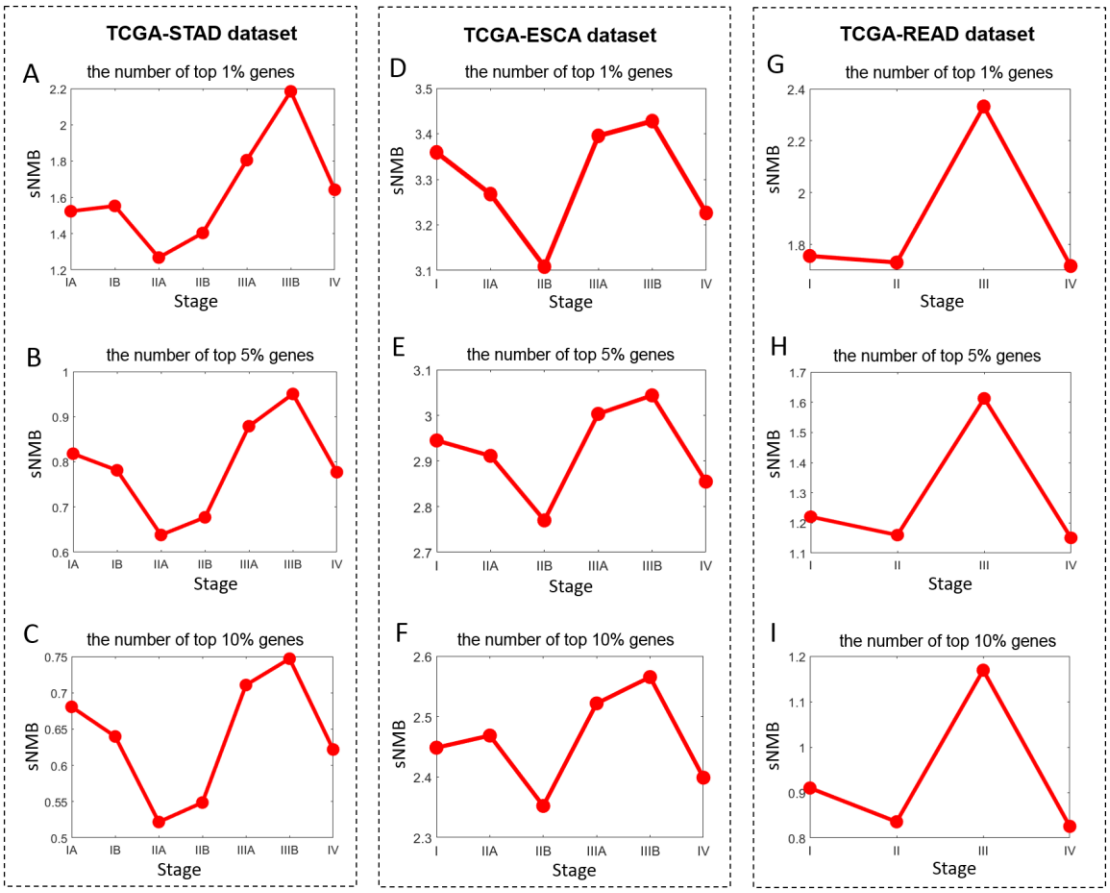


Figure S5: The critical signals of three TCGA datasets under different settings of the adjustable parameter *T*.

C. Compared the sNMB method with other two single-sample methods

We have compared the proposed sNMB method with the previously published single-sample methods, including the single-sample-based hidden Markov model (sHMM) (Liu et al., 2019) and single-sample Kullback-Leibler divergence (sKLD) (Zhong et al., 2020). Specifically, these three methods (sNMB, sHMM, and sKLD) are applied to detect the pre-deterioration stage based on the TCGA-READ dataset, the clinical staging information of which can be seen from Table S2.

For sHMM method, the time-series data need to be divided into the training part ranging from t=1 to t=T-1, and a time point t=T (T>2) is used for detecting the critical point (pre-deterioration stage). Therefore, compare with the proposed sNMB method, the time point of detection from the sHMM method is only available starting from stage II within the time-series data since the time-series data from the initial two-time points, i.e., tumor-adjacent (TA) samples and stage I samples, are used to train the model. As shown in Figure S6A and B, the sudden increase of sNMB score was detected in stage III, but there is no abrupt increase in SSI curve, that is, the proposed sNMB method can detect the critical transition point, while

the sHMM method fails.

In contrast to the proposed sNMB method, the sKLD method requires greater numbers of normal/reference samples to fit a Gaussian distribution for each gene. Thus, the sKLD method may fail to detect the pre-deterioration stage of complex diseases when fewer reference samples are available. Specifically, for the TCGA-READ dataset containing 11 reference samples, any drastic increase is not observed in sKLD curve (Figure S6C), i.e. the sKLD method cannot provide an early-warning signal of critical transition for the READ.

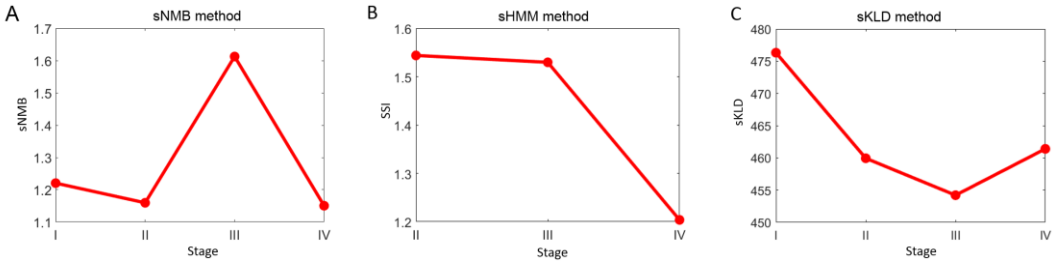


Figure S6: The performance of the three single-sample approaches (i.e., sNMB, sHMM, and sKLD method) in detecting the pre-deterioration stage of READ.

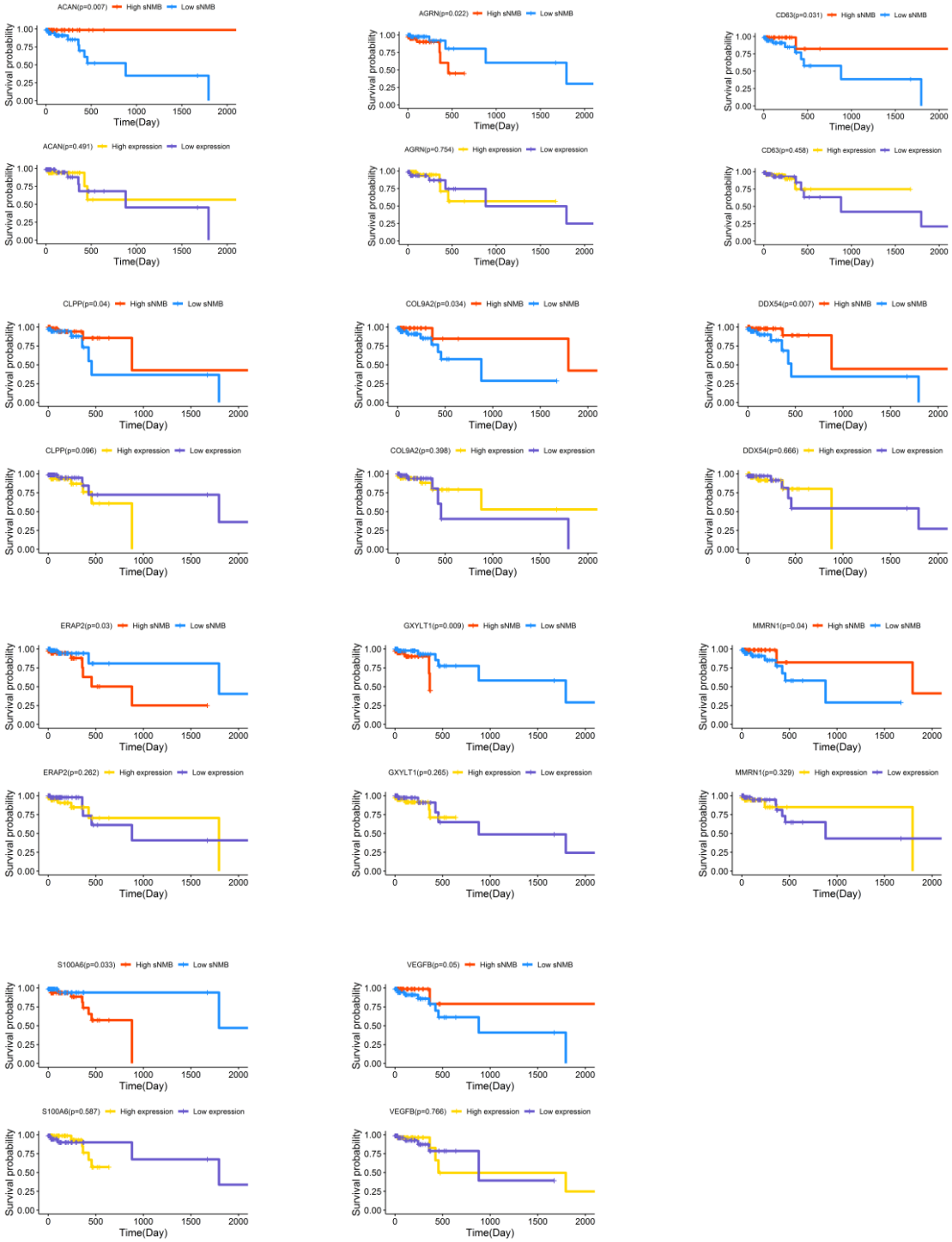
Table S2. the number of tumor samples within each stage in TCGA-READ dataset.

Stage	TA	I	II	III	IV
Samples	11	34	53	53	26

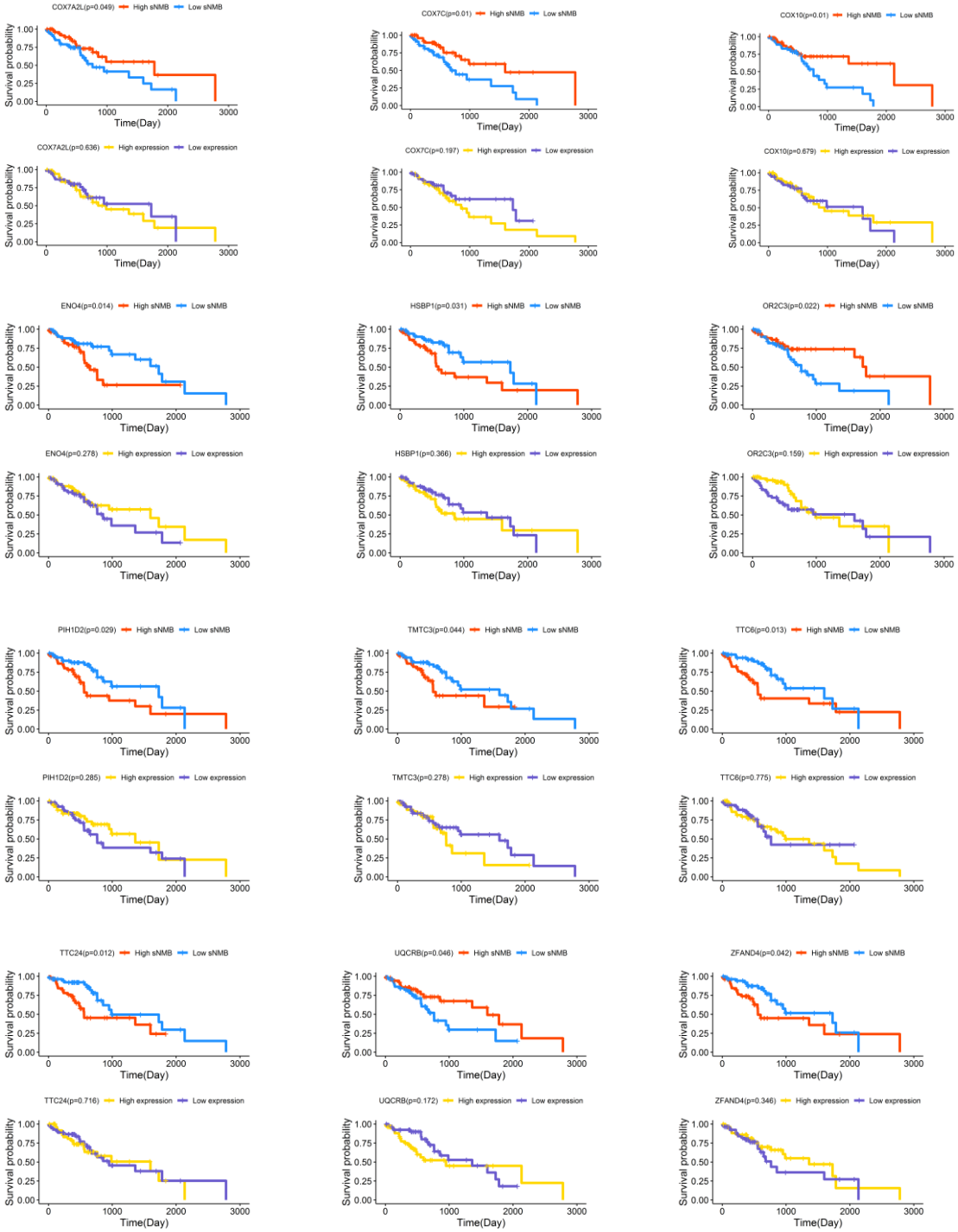
TA refers to the tumor-adjacent samples.

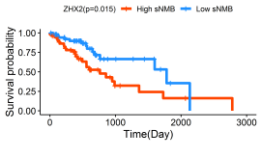
D. The “dark genes” for three tumor datasets

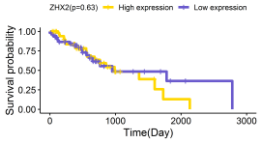
The “dark genes” for STAD

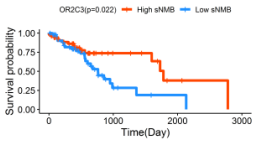


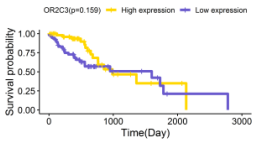
The “dark genes” for ESCA



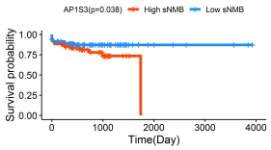


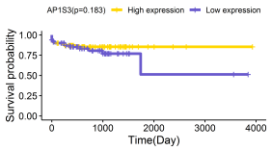


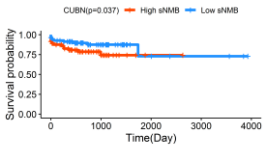


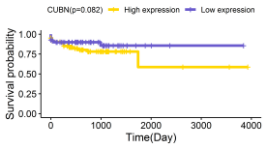


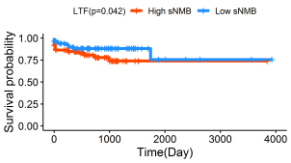
The “dark genes” for READ

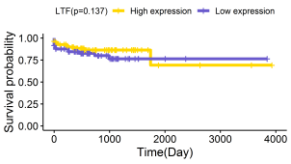


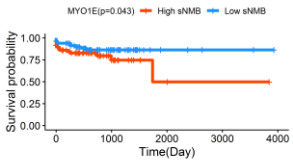


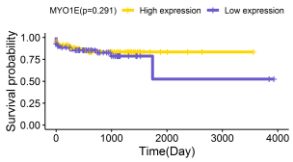


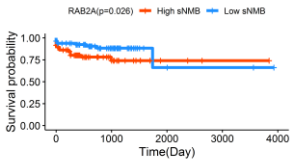


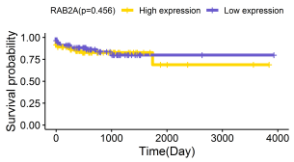


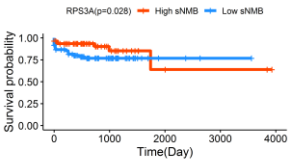


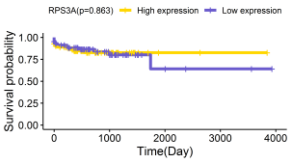


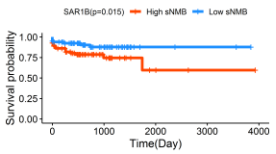


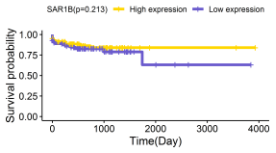


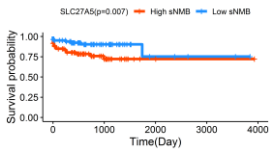


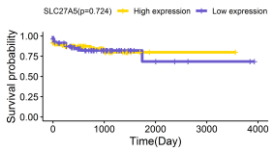


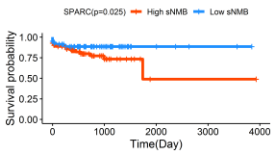


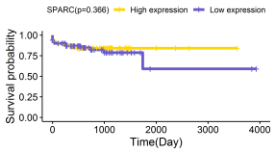


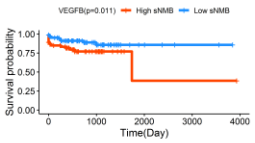


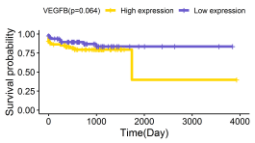


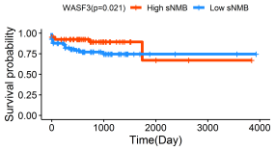


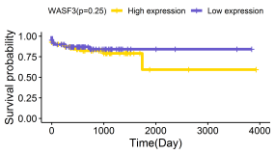


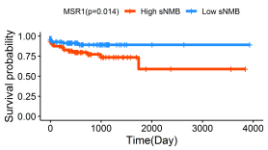


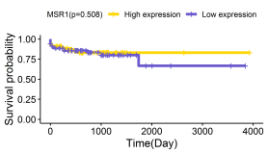


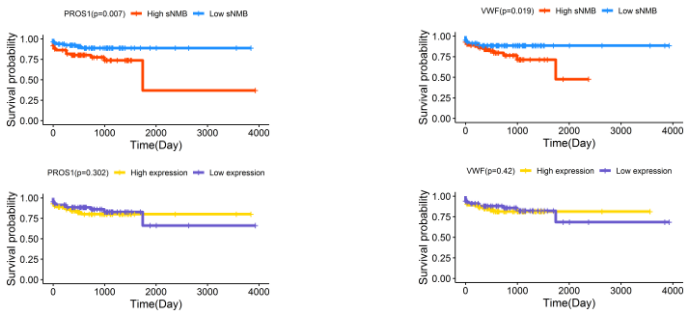












References

Chen, L. & Aihara, K. Stability of genetic regulatory networks with time delay, IEEE Trans. Circuits Syst. I 49, 602–608(2002).

Chen, L., Wang, R. & Zhang, X. Biomolecular Networks: Methods and Applications in Systems Biology, (John Wiley & Sons, Hoboken, New Jersey,2009).

Kloeden, P. & Platen, E. Numerical Solution of Stochastic Differential Equations, (Springer, 1999)

Liu R, Zhong J, Yu X, Li Y, Chen P. Identifying critical state of complex diseases by single-sample-based hidden markov model. Frontiers in genetics. 2019;10:285.

Zhong J, Liu R, Chen P. Identifying critical state of complex diseases by single-sample Kullback-Leibler divergence. BMC genomics. 2020;21(1):87-.