Appendix

Fangliangzi Meng[1,2], Hongrun Zhang[3], Ruodan Yan[4], Guohui Chuai[1,2✉], Chao Li[4,5✉], and Qi Liu[1,2✉]

[1] Key Laboratory of Spine and Spinal Cord Injury Repair and Regeneration (Tongji University), Ministry of Education, Orthopaedic Department of Tongji Hospital, Frontier Science Center for Stem Cell Research, Bioinformatics Department, School of Life Sciences and Technology, Tongji University, Shanghai, China
[2] National Key Laboratory of Autonomous Intelligent Unmanned Systems, Frontiers Science Center for Intelligent Autonomous Systems, Ministry of Education, Shanghai Research Institute for Intelligent Autonomous Systems, China
[3] Cancer Research UK Cambridge Institute, UK
[4] Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK
[5] School of Science and Engine, School of Medicien, University of Dundee, UK
qiliu@tongji.edu.cn

1 Dataset division

In Fig. S1, we divide 23 TCGA datasets into training set(85%) and test set(15%) through stratified sampling based on tumor types. The training set undergoes further splitting into five folds for cross-validation. Each split is used to pretrain an SNN model, train the main branch, and perform early stopping. The process includes finding parameters within the training splits and final evaluation on the test set. The evaluation is done using the reserved test set (15%) after the cross-validation process is complete.

2 Visualization of embeddings

2.1 t-SNE plots at different stages

In Fig. S2, from left to right, the t-SNE plots respectively show the embeddings in three stages: original embeddings after feature extraction, embeddings after the Siamese network but without DANN, and embeddings after DANN. In the left plot, part of the original embeddings from certain tissues tend to cluster together. After Siamese, it is not surprising that samples from the same origin are more closely clustered together, because genes contain abundant information related to tissue specificity. Following DANN, WSI embeddings are dispersed evenly in the representation space, indicating that tissue information unrelated to TME subtypes has been relatively eliminated and TME-related features are kept. In detail, comparative t-SNE plots for each dataset, following training with ABMIL and PathoTME, are presented in Fig. S3.

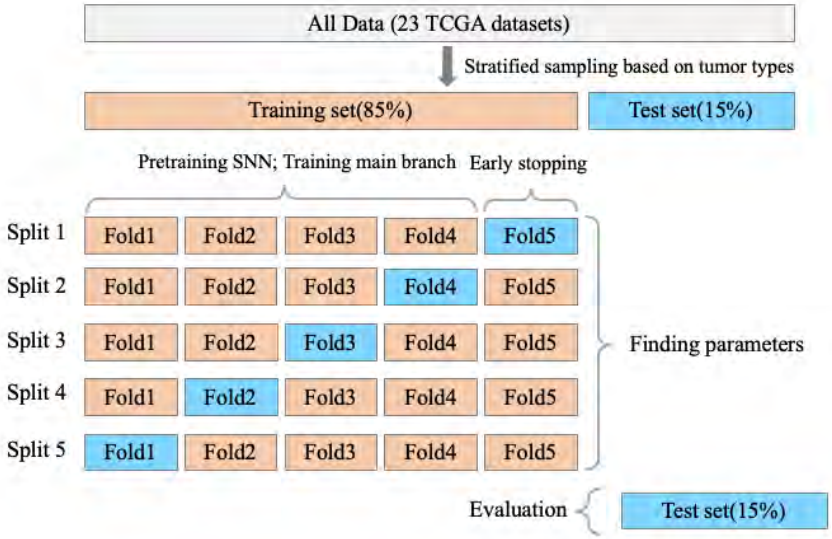

Fig. S1. Dataset division on pancancer TCGA dataset.

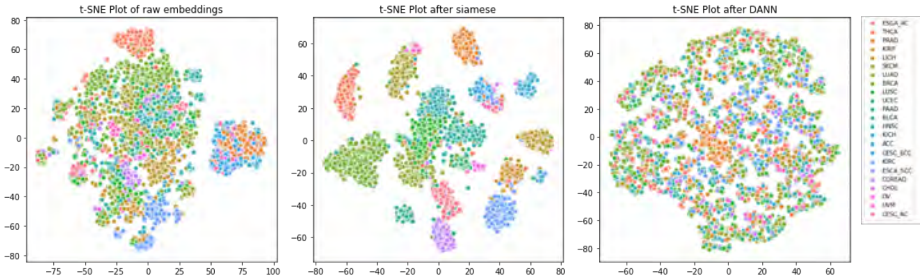

Fig. S2. Distribution of WSI embeddings in representation space at different stages.

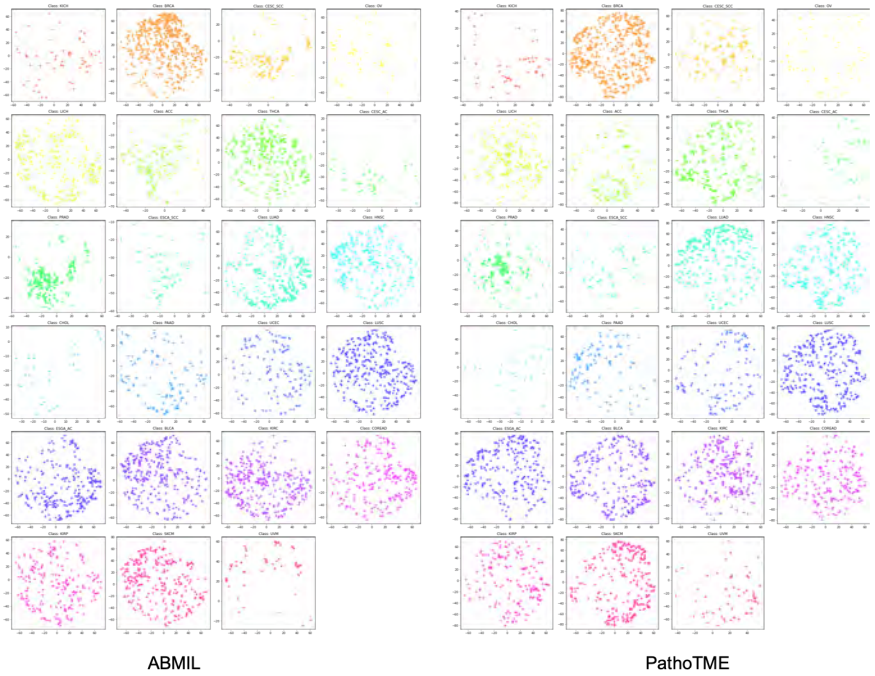

Fig. S3. t-SNE plots on each TCGA dataset.

2.2 t-SNE plots of our method on single dataset

In Fig. S3, we plot representation space of PathoTME and ABMIL with tissue labels on each TCGA dataset. It illustrates that after training with PathoTME, the spatial clustering based on tissue type is reduced, compared with ABMIL.