Highly Accurate Whole Transcriptome Inference and Synthesis through Generative Adversarial Networks

Suneng Fu

Guangzhou Laboratory, Guangdong, China

Abstract

Transcriptome profile is the most extensive and standardized among all biological data, but its lack of inherent structure impedes the application of deep learning tools. To overcome this barrier, the global structure and neighborhood relationship of protein-coding genes is resolved through uniform manifold approximation and projection (UMAP)[1] of high quality gene expression data. The resultant transcriptome is a bird-shaped manifold, with tissue- and disease-specific genes congregating as body parts. Cancer and control tissue transcriptome images are visually distinct in the proliferation zone, the immunity zone, and a tissue-specific zone. Convolutional neural networks (CNNs) trained with lung squamous cell carcinoma (LUSC) and control tissue transcriptome images readily classify LUSC transcriptome with over 95% accuracy. More importantly, generative adversarial networks (GAN)[2] trained with the PanCancer Atlas transcriptome image set infer whole transcriptomes from top 200 genes with a Pearson correlation coefficient of 0.9535. Surveillance of GAN-generated, pseudo-LUSC transcriptomes identifies mitochondria electron transport complex I and immunity as tier-I and tier-II lung cancer classifiers with a two-fold median survival time (MST) differentiation. Further interpolation of individual LUSC patients' GAN latent space reveals a patient-specific, proteasome-driven tumorigenesis mechanism, providing opportunities for personalized treatment. Therefore, GANs have the potential to significantly accelerate in silico biological discovery and precision medicine through whole transcriptome inference and synthesis.

INTRODUCTION

Transcriptional regulation enables the specification of cell fate, physiological response, and disease pathogenesis. Since the advent of microarray and RNA sequencing technologies, transcriptome profiles have become the most comprehensive among all omics data[3-5]. However, machine learning tools with transcriptomic data remain limited and has yet to transform the field of biology as it has for computer vision and natural language processing[6].

Machine learning is most successful when dealing with structured data, like images and languages. The recent success in protein structure modeling with AlphaFold and Rosetta are also examples of machine learning with structured, ordered data[7,8]. Therefore, significant efforts have been made to capture the power of computer vision-oriented tools to transform transcriptome learning by converting gene expression lists to images. For example, Lyu and Haque first used the chromosome location of genes as coordinates to convert the one-dimensional transcriptome profile to two-dimensional images[9], which was replicated in more recent efforts by Chen and colleagues[10]. Sharma and colleagues used t-distributed stochastic neighbor embedding (t-SNE) in DeepInsight[11]. Bazgir and colleagues used the Bayesian metric multidimensional scaling approach in REFIND[12]. Although all of them were moderately successful in sample classification, the images were not biologically intuitive, nor did they demonstrate new learning capabilities.

Here I successfully embedded the mouse and human transcriptome as a phoenix-like object, thus named phoenix transformation, and it demonstrated unprecedented deep learning capacities for the transcriptome latent space.

RESULTS

Transcriptome structurization through manifold embedding

I envision transcriptomes as structural entities with the placement of genes in the transcriptome object analogous to the positioning of atoms in a protein. Further, I conceive gene-gene relationships as *x*-*y* coordinates and gene expression levels as *z*-coordinates in the three-dimensional transcriptome space. As genes that function together co-express temporally and spatially, coexpression-based mapping shall allow genes with similar expression patterns to congregate in a manifold and form structural features in the same way as atoms in an amino acid residue. The uniform manifold approximation and projection (UMAP)[1] algorithm excels in preserving both global and local structures and is selected for computing the neighborhood relationship (i.e., the *x*-*y* coordinates) of genes based on coexpression data.

A set of criteria were applied to filter gene expression datasets: 1) each dataset shall have a large number of samples and biological diversity, 2) biological diversity shall be represented in a balanced manner, 3) all datasets shall be centrally collected and processed and the total number of datasets used shall be as few as possible to minimize non-biological variance. Two sets of mouse RNA bulk-sequencing data fit this criteria: the ENCSR574CRQ dataset that examines mouse early embryogenesis from the ENCODE3 project[13] and the GSE132040 dataset studying mouse aging from the Tabula Muris Senis consortium[14]. They were supplemented with a small dataset (DRA000484) focusing on zygotes and fertilization from the DBTMEE project[15].

Based on these three datasets, an Euclidian distance-based neighborhood relationship of protein-coding genes in the mouse genome is projected into a two-dimensional space (Supplementary Fig. 1). Together, a total of 20,424 protein-coding genes in the mouse genome are mapped onto 18, 545 unique coordinate pairs (Supplementary Table 1). The coordinates for 16, 752 protein-coding genes in human genomes are transferred from their mouse homologs due to the lack of quality transcriptome data for embedding (Supplementary Table 2).

A few more decisions are made for the final rendition of the transcriptome structure. First, as atoms have volumes in a protein space, genes in the transcriptome space are represented as points of ~30 pixels rather than a single pixel. Such a treatment also reduces information sparsity in the space and makes it amenable to the dense neural network nature inherent to computer vision-based deep learning. Second, gene expression levels are scaled to 14x $\log_2$-transformed fragments per kilobase exon per million reads (FPKM), clipped between [1, 255], rendered in a five-color rainbow gradient, and projected onto the UMAP1/2 plane. Third, the transcriptome information is rendered in ascending orders of expression, forming a top view of the transcriptome structure. In the end, all transcriptome objects have the same geometrics on the *x*-*y* plane, and they only differ on the *z*-axis, colorization. The detailed information for computing and rendering the transcriptome structure is provided in Methods and Supplementary Fig. 1a.

Fig. 1a shows the mouse transcriptome resolved as a bird-shaped structure; tissue-specific and functional-related genes congregate to form body parts of the embedding manifold. The "head" (zone A) consists of sperm-specific genes adjoin with the "neck" (zone B) that enriches for genes that are selectively expressed in reproductive cells/tissues, including sperms, oocytes, and the testis (Fig. 1a, b; Supplementary Fig. 1b). Underneath the reproductive zone is the body plan zone (zone C) that enriches for genes expressed during early embryogenesis (Fig. 1a; Supplementary Table 6). The body-plan zone is then connected to the skin (E/F), neuronal tissue (G), muscle (H), and lung/mitochondria/metabolism-related genes (I) on the left, and proliferation (J), immunity (K), digestive tissue (L/O)-related genes on the right (Fig. 1a, b; Supplementary Fig. 1b; Supplementary Table 3). Underneath the body-plan zone lies a large group of poorly-resolved house-keeping genes involved in transcription, translation, degradation, cell adhesion, and migration in the abdominal and foot areas (D, M, N, P, Q; Fig. 1a, Supplementary Table 3).

The transcriptome image cannot only be used for distinguishing tissue source but also differentiate circadian time of tissue collection[16]. Fig. 1c shows the neuronal zone gene expression in the SCN transcriptome image reaches its trough at ZT14, coinciding with SCN inactivity starting around 10 pm[17]. In contrast, neuronal zone gene expression signal is specifically elevated in the ZT0 lung transcriptome image (Supplementary Fig. 1c), coincides with awakening and increased physical activity during dawn.

Transcriptome image classification through convolutional neural networks

The proliferation zone provides a potential visual landmark for cancer transcriptomes. For example, the lung squamous cell carcinoma (LUSC) transcriptome images consistently have elevated signals in the proliferation zone than non-tumor controls from the same patient, although the level of upregulation varies (Fig. 2a-c). Besides the proliferation zone, the LUSC transcriptome images also show a variable degree of downregulation in the immunity zone (Fig. 2b) and a lung-specific zone gene expression (Fig. 2c). These distinct features allow over 95% classification accuracy of LUSC or normal lung transcriptome images through convolutional neural networks (Fig. 2d; Supplementary Fig. 3). Even training with partial transcriptome image consisting the proliferation zone (J), the immunity zone (K), and the lung/mitochondria/metabolism zone (I) reached a similar level of accuracy in LUSC *vs.* normal classification (Fig. 2e, f).

Besides cancer diagnosis, the visible and variable reductions in the immunity zone and the tissue-specific zone provide potential stratification markers for cancer. Broadly, almost all cancer transcriptome profiles from the PanCaner Atlas dataset (abbreviated as PanCanAtlas onward)[18] may be classified into immunity-hot, where the immunity gene is expressed at similar or higher levels than the proliferation genes, and immunity-cold, where immunity genes are expressed at significantly lower levels than proliferation genes (Supplementary Fig. 2). More specifically, the

expression the surfactant protein A2 in the lung-specific zone is an important prognosis marker for lung cancer patient survival. Fig. 2g shows that surfactant protein A2 (SFTPA2) downregulation is correlated with reduced median survival time (MST) after diagnosis for lung adenocarcinoma[19] patients (LUAD, 45 months in $SFTPA2^{low}$ *vs.* 105 month in $SFTPA2^{high}$, $p = 0.0016$). To the contrary, low SFTPA2 expression is a protective marker for LUSC[20] patients: 70 months for $SFTPA2^{low}$ *vs.* 36 month in $SFTPA2^{high}$, $p = 0.0026$.

Whole transcriptome inference through transcriptome image autocompletion

Unlike convolutional neural networks, generative adversarial network (GAN) has the ability to learn rules of play from ground up[2]. If successfully executed on the transcriptome images, it may allow us to learn rules of transcriptional regulation and mechanisms of development and disease *in silico*. A potentially verifiable test for GAN's learning capability on the transcriptome image is image completion and super resolution, where ground truth exists. Therefore, I asked whether GANs can infer gene expression for the whole transcriptome with minimal information input through a combination of image completion GAN (pix2pix)[21] and super-resolution GAN (SRGAN)[22].

The PanCanAtlas dataset with both cancer and normal tissue transcriptomes was chosen for GAN training. For data balance, the dataset was divided into 93 classes based on tissue source and disease conditions, and each class (excluding the LUSC class) was sampled 20 times. Each of the sampled transcriptome profiles was then used to synthesize three sets of images: a 1024x1024 image and a 3072x3072 image for the full transcriptome, and another 1024x1024 image for only the top 200 highly-expressed genes (Supplementary Fig. 4a). I did not choose a fixed set of 200 genes because, in real-life RNAseq situations of low sample availability, the quantification of the

most abundant transcripts is the easiest to obtain. Detailed information about the training is provided in Supplementary Fig. 4a and the methods section.

After training (Fig. 3a; Supplementary Fig. 4), the combined pix2pix-SRGAN model is able to infer highly realistic transcriptome images with top200 gene inputs (Fig. 3b). Statistically, the inferred transcriptome and the real transcriptome show a Pearson correlation coefficient (R) of 0.9535 (Fig. 3c) and a mean absolute error (MAE) of 10.492 (~0.743 $\log_2$FPKM, Fig. 3d) for the training set, and R of 0.9335 and MAE of 12.786 for the LUAD testing set. Importantly, the pix2pix/SRGAN model is able to infer transcriptome profiles in disease conditions (e.g. LUSC) not included in the training process. Fig. 3c and d show that the pix2pix/SRGAN model infer LUSC and LUAD transcriptome profiles from top200 hints with similar levels of precision: MAE 12.785 for LUAD *vs.* 13.286 for LUSC, R of 0.9335 for LUAD and 0.9323 for LUSC. The model's ability to infer LUSC transcriptome is not solely because of its similarity to LUAD. Fig. 3e shows that the gene expression difference between the real and inferred LUSC and LUAD transcriptomes is linearly correlated (R^2 = 0.5967, p < 0.001), although the fold of change is moderately compressed (slope = 0.5890, Fig. 3e).

The RNA-seq based pix2pix/SRGAN model may be expanded further for training with a limited number of microarray-based samples to achieve an equivalent level of accuracy (Supplementary Fig. 3e, f). The transcriptome images inferred from microarray-based top200 hints can be classified by pre-trained convolutional models with 100% accuracy as being cancer or normal (Fig. 3f). If only the IJK regions of the image is considered, the accuracy is 95% (Fig. 3f).

De novo **transcriptome synthesis and interpolation with StyleGAN generator**

Unlike pix2pix and SRGAN, the StyleGAN can synthesize highly realistic images *de novo*[23], which provides a near cost-free opportunity to study gene regulation and disease mechanism. To

test the generative capacity of GANs in transcriptome synthesis, transcriptome profiles from all 93 classes (100 samples per class) of the PanCanAtlas dataset is used to train a conditional StyleGAN2-ADA network (Fig. 4a, Supplementary Fig. 5a). After 2,400 kimg training, the Frechet Inception Distance (FID) score, a measurement of distance between real and fake images, of the StyleGAN model reached 5.85, comparable to training with real images reported by the original StyleGAN2-ADA paper[23].

Two sets of experiments were conducted to probe what was learned by the conditional StyleGAN model. First, the StyleGAN generator synthesized two classes of transcriptomes: class 54 for liver and class 56 for lung. All 72 transcriptome images (36 for each class) were super-resolutioned to 3072x3072 pixels to render gene expression on numeric levels and clustered (Supplementary Fig. 5b, Supplementary Table 4). Supplementary Fig. 5c shows that the synthetic liver and lung transcriptomes were well separated from each other. Therefore, StyleGAN learned the boundaries between the liver and lung transcriptome latent space.

Second, the StyleGAN generator synthesized 62 fake LUSC (class 58) transcriptome images (Supplementary Fig. 6a; Supplementary Table 5). Hierarchical cluster analysis of all 62 synthetic LUSC transcriptomes showed a two-step bifurcation into three major subsets (Fig. 4b; Supplementary Fig. 6b). Gene set enrichment analysis for differentially regulated genes (DEG) between clusters I and II shows over-representation of cell adherence, mitochondria electron transport complex I, and EGFR signaling functions (Supplementary Fig. 6b; Supplementary Table 14), and DEGs between cluster IIa and IIb are enriched for membrane proteins and immune response functions (Fig. 5c; Supplementary Table 5). For comparison, DEGs at both tier-I and tier-II branch points in real LUSC transcriptomes are enriched for immunity genes but not mitochondria electron transport complexes (Supplementary Table 5).

To evaluate the validity of StyleGAN's *de novo* discovery, LUSC patients were classified based on mitochondria complex I and immunity gene expression levels, and their prognosis value for survival was examined (Fig. 4c). The results show that neither mitochondria complex I nor immunity alone has prognosis value for patient survival (Supplementary Fig. 6c, d), but patients with combined high mitochondria and low immunity gene expression ($C^{high}I^{low}$) have an MST of 64 months *vs.* 39 months for patients with low mitochondria complex I and high immunity ($C^{low}I^{high}$) gene expression (Fig. 4d). Similarly, combining high mitochondria or low immunity gene expression with low surfactant protein expression also extends MST to 80 months ($C^{high}S^{low}$, Fig. 4e) and 89 months ($I^{low}S^{low}$, Fig. 4f), respectively. Therefore, StyleGAN training with PanCanAtlas transcriptome image generates novel and clinically-relevant insights.

The StyleGAN latent space provides an additional opportunity to understand disease history and trajectory through latent space interpolation and feature vector transfer. To explore this functionality, the PanCanAtlas dataset was used to train an unconditional StyleGAN model augmented with the microarray-based LUSC dataset (Fig. 4a; Supplementary Fig. 6a). The trained model was then used to invert four LUSC patients normal and cancer transcriptomes (Supplementary Fig. 6b-d). Then the w+ latent space between each of the normal and cancer transcriptome was interpolated and projected into a manifold (Fig. 4g, h; Supplementary Table 6). The results show that all four patients' control transcriptomes are very similar to each other on UMAP2, but they take two distinct approaches to reach their respective cancer states (Fig. 4h). For example, the cell cycle transcription factor E2F7 increases when the normal lung tissue transcriptome of patients 122 and 126 progress toward the tumor state (122T and 126T) along the UMAP2 axis but does not change for patients 130 and 144 (Fig. 4i). Network analyses show that 40 of the top100 genes induced along with the descending UMAP2 gradient are centered around

ubiquitin (UBX, Fig. 4j). Therefore, proteasome upregulation is a candidate tumorigenesis driver for patients 122 and 126, and the proteasome inhibitor Bortezomib may work most effectively for them.

DISCUSSION

In summary, this study deciphers the underlying structure of the mouse transcriptome, and it enables an imagery transformation of transcriptome profiles suitable for computer vision-based deep learning. The transformation process is termed Phoenix transformation for the resultant bird-shaped manifold.

The proof-of-principle capabilities demonstrated herein are either comparable or superior to the state-of-the-art techniques in transcriptome classification and gene expression inference[24]. For example, the combined pix2pix/SRGAN model exceeds the state-of-art gene expression inference algorithms in transcriptome coverage (15,000 *vs.* 9,500), need of informational input (highest 200 *vs.* fixed 1000 genes), and robustness (trained with 10% *vs.* 80% of total datasets)[24-27]. Although the apparent MAE of ~13, equivalent of ~0.9 $\log_2$FPKM, underperforms the state-of-art algorithms based on 1000 hallmark genes, which has an MAE of ~0.45 for RNAseq data, the difference is largely accounted for by the scaling difference: expression is scaled between [0, 18] $\log_2$FPKM herein and between [4, 15] in other methods[24-27]. Additionally, by inferring from 200 highest expressing genes, this approach also has broad utility for transcriptome inference for low-availability samples, a major advantage over traditional methods[25,28]. For example, it may be even possible to infer the whole transcriptome based on single cell sequencing data.

Despite its enormous capability, the spatial resolution of the UMAP-embedded transcriptome structure may be improved further by incorporating more high power, low noise transcriptome data. In this regard, integrating single cell RNA sequencing data collections may be worthy of

consideration. Each transcriptome may also be converted into not one but two images: one rendered in the order of ascending gene expression and the other in descending order. Additionally, the network architecture and hyperparameters may be fine-tuned for transcriptome images. With these improvements, the field of genomics may be fully open to computer vision and broadly accelerate *in silico* gene function inference, development and disease mechanism synthesis, drug target discovery, and precision medicine.

Methods

Tools

The following tools are used for Phoenix transformation and subsequent experiments: R Studio, PyCharm, Prism 9, pix2pix, SRGAN, StyleGAN2-ADA, DAVID web server, STRING web server.

Data

UMAP is calculated from GSE132040, ENCSR574CRQ, and DRA000484. The downloaded RNAseq datasets are combined in Supplementary Table 1. The baboon dataset used for visualization of transcriptome change in circadian is GSE98965 and provided in Supplementary Table 6. The LUSC dataset (GSE18842, 19804, 27262) used for visualization and training is downloaded from CuMiDa and provided in Supplementary Table 9. The PanCaner Atlas (PanCanAtlas) dataset used for generative adversarial network training is downloaded via GDC Data Transfer Tool (DTT) and provided in Supplementary Table 12.

Projection of mouse protein-coding genes onto a two-dimensional space

Protein-coding genes from RNAseq datasets GSE132040, ENCSR574CRQ, and DRA000484 are combined, rounded to integer values, and saved in Supplementary Table 1. This dataset is imported into R and used for gene projection with plot_umap function from the bio_plotr package. The number of neighborhood is set at 15, the distance matrix is Euclidian, and the dimension is set at 2. The resulting coordinates are rotated 30 clockwise before re-set the origin at (-17.5, -17.5). The new coordinates are then multiplied by 16X to fit in a 512x512 square. The mouse coordinates are transferred to the human and baboon protein-coding genes through homologue mapping. The coordinates for mouse, human, and baboons are provided in Supplementary Tables 4, 5, 7.

Transcriptome data transformation

For mouse and human RNAseq transcriptome datasets, the log2-transformed FPKM values are multiplied by 14, added 1, rounded, and clipped between [1, 255]. For the LUSC microarray datasets, the transformation is performed through (n-3) * 1.7 * 14 + 1, rounded and clipped between [1, 255], where n is the probe signal from microarray. For the baboon dataset, the transformation is min(round((log(n, 2)+3)*14), 255), where n is the FPKM values of individual genes. The transformed data is provided in Supplementary Tables 2, 7, 10, 13.

Tabular transcriptome data to image conversion

The transcriptome data is first sorted in ascending orders of expression before being rendered into images by ggplot2 in R. The x- and y-axis ranges are set at (0, 512), point size 0.1, color scheme 5-color gradient. For images of 512x512 pixels, the size is set 1.79 inches; for 1024x1024 pixels, size set at 3.58 inches; for 3072x3072 pixels, size set at 10.74 inches. For the production of top200

gene images, the transcriptome data is sorted in ascending orders. The last 200 rows of genes plus references of 0 and 255 set at coordinates of (20, 20), and (15, 15) are printed by ggplot2 and saved though ImageMagick. Partial transcriptome images (IJK images) for the lung-specific zone, the proliferation zone, and the immune zone are printed similarly. The PanCanAtlas RNAseq dataset is used to generate two image sets, a 9300 image set where each of the 93 sample classes is sampled 100 times, and a 46500 image set where each class is sampled 500 times.

The ggplot2-printed images have four channels: RGBA; they are converted to three-channel sRGB formats by ImageMagick truecolor option. The 768x768 images used for training are resized from 1024x1024 images.

Network training

The Keras convolutional neural network used for image classification consists six Conv2D layers (16-512), six MaxPooling2D layers, and two Dense layers (512, 1). It is trained with the microarray-based LUSC dataset, batch_size of 20, learning_rate of 0.0001. Training and validation data split was 5:5.

The pix2pix discriminator consists six Conv2D layers (64, 128, 256, 512, 512, 1), and the generator has seven blocks each for the encoder (64, 128, 256, 512x4) and the decoder (512x4, 256, 128, 64). It was trained with the PanCanAtlas dataset, batch_size of 16, lr of 0.0002. For data balancing, the RNAseq-based PanCanAtlas dataset excluding LUSC transcriptomes is sampled 20 times per class and 1840 images representing 1634 unique samples out of 19120 total samples. Some classes have less than 20 samples, so the number of unique samples is always less than the number of images used for training. For cross-platform training, the 1840 RNAseq-based images are

combined with the microarray-based LUSC dataset that is sampled 7 times per sample for a total of 1624 images from 232 samples.

The SRGAN discriminator consists eight Conv2D layers (64x2, 128x2, 256x2, 512x2), and the generator has one Sequential block, six ResidualBlocks, and an UpsampleBlock. For inference of RNAseq-based transcriptomes, the SRGAN is trained with paired 768x768 transcriptome images inferred from the pix2pix model and real 3072x3072 transcriptome images from the PanCanAtlas dataset. For inference of microarray-based transcriptome, the SRGAN model is trained with real, paired 768x768 and 3072x3072 images from the PanCanAtlas dataset. The training includes 20 images per class, 1840 images in total, 7:3 train validation split, crop_size 178, batch_size 4.

The StyleGAN2-ADA network consists seven blocks (512-8) in the discriminator and eight blocks (4-512) in the generator network, and both has 65536 features. Both the conditional and unconditional models are trained with 46500 images from the PanCanAtlas dataset. After training, pseudo LUSC transcriptome image is synthesized with the generator from the conditional model, and the interpolating transcriptome images are synthesized after projecting normal and LUSC transcriptomes into the w+ space of the unconditional model.

Accuracy of training

The Keras convolutional neural nets is evaluated by a testing dataset comprised 20 transcriptome images (10 tumor, 10 paired normal controls) not included in the 532-set microarray-based LUSC expression data.

The performance of the pix2pix/SRGAN models in inferring whole transcriptome gene expression from top200 hints are evaluated in three steps. First, it is evaluated with the ground truth expression

data from the 1634 RNAseq-based PanCanAtlas transcriptomes and 532 microarray-based LUSC included in the training set based on mean absolute error and Pearson correlation coefficient. Second, the model is evaluated with the ground truth of two test datasets, including 555 RNAseq-based LUAD transcriptome profiles and 546 LUSC transcriptome profiles. The 1634 PanCanAtlas training set does not include any samples from LUSC patients but does include 20 LUAD transcriptome profiles excluded from the testing dataset. Third, the model is evaluated by the convolutional neural network trained for classifying the inferred normal *vs.* LUSC tumor transcriptomes from 20 microarray transcriptome samples not included in the 532-sample microarray-based LUSC training set.

Biological relevance of the transformation

Functional enrichment of genes in each of the compartment of the UMAP-transformed mouse protein-coding genome is performed at the DAVID web server. Network analysis and functional enrichment of genes in the network is performed on the STRING web server.

Hierarchical cluster analysis of the overall subtypes of lung squamous cell carcinoma (LUSC) real and synthetic transcriptomes is performed in R using the hclust package with default settings using the transformed data scaled between [1, 255]. The differentially regulated genes between clusters are determined by two-tailed *t*-test, and the top1000 genes with the lowest *p*-values are evaluated by DAVID for functional enrichment.

Pearson correlation coefficient analysis between UMAP2 and genes in the interpolating transcriptome is performed in Microsoft Excel. Only genes with expression levels of higher than 42 (equivalent to 3 log2-transformed FPKM) across all samples are included for Pearson

correlation coefficient analysis. The value of 42 is arbitrarily set to reduce the impact of large fold-of-change from low-expressing genes/samples. Pearson correlation coefficient analysis for evaluating the pix2pix/SRGAN model performance includes all genes in the transcriptome not gated by expression levels.

Median survival time (MST) analysis is performed in Prism 9. Genes with maximal variation across samples in each category are selected for analysis. The mitochondria electron transport chain complex I genes have lower variations across samples, and the average of the top six variable genes (NDUFA2, A6, B7, B11, S7, and V1) are taken.

Data availability

All data needed to reproduce the results will be available from GitHub (https://github.com/SamTransformer/Phoenix_Transformation).

Code availability

All codes needed to reproduce the results will be available from GitHub (https://github.com/SamTransformer/Phoenix_Transformation).

Acknowledgement

I thank Angelo Pisco from Chan Zukerberg Biohub for technical assistance with the Tabula Muris Senis dataset. I thank Tero Karras, Derrick Schultz, Jun-Yan Zhu, and Lornatang for sharing their StyleGAN-ada, pix2pix, and SRGAN PyTorch repositories in GitHub, and Diego Porres, Janne Hellstein, and Julian Pinzaru for technical assistances with the implementation. I thank Qiangfeng Zhang from Tsinghua and Cong Yu from Google for careful reading of this manuscript and my

past students for their hard work and inspirations. This work is supported by Guangzhou National Laboratory.

Competing interests

S.F has filed a provisional patent application (PCT/CN2022/085963) relating to transcriptome image transformation and machine learning.

References

1 McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv pre-print server*, doi:None arxiv:1802.03426v1 (2020).

2 Goodfellow, I. J. *et al.* in *NIPS.*

3 Lashkari, D. A. *et al.* Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proceedings of the National Academy of Sciences* **94**, 13057-13062, doi:10.1073/pnas.94.24.13057 (1997).

4 Wilhelm, B. T. *et al.* Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature* **453**, 1239-1243, doi:10.1038/nature07002 (2008).

5 Nagalakshmi, U. *et al.* The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science* **320**, 1344-1349, doi:10.1126/science.1158441 (2008).

6 Eraslan, G., Avsec, Ž., Gagneur, J. & Theis, F. J. Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics* **20**, 389-403, doi:10.1038/s41576-019-0122-6 (2019).

7 Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583-589, doi:10.1038/s41586-021-03819-2 (2021).

8 Anishchenko, I. *et al.* Protein tertiary structure prediction and refinement using deep learning and Rosetta in CASP14. *Proteins* **89**, 1722-1733, doi:10.1002/prot.26194 (2021).

9 Lyu, B. & Haque, A. *Deep Learning Based Tumor Type Classification Using Gene Expression Data* (Cold Spring Harbor Laboratory, 2018).

10 Chen, X., Chen, D. G., Zhao, Z., Balko, J. M. & Chen, J. Artificial image objects for classification of breast cancer biomarkers with transcriptome sequencing data and convolutional neural network algorithms. *Breast Cancer Research* **23**, doi:10.1186/s13058-021-01474-z (2021).

11 Sharma, A., Vans, E., Shigemizu, D., Boroevich, K. A. & Tsunoda, T. DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture. *Scientific Reports* **9**, doi:10.1038/s41598-019-47765-6 (2019).

12 Bazgir, O. *et al.* Representation of features as images with neighborhood dependencies for compatibility with convolutional neural networks. *Nature Communications* **11**, doi:10.1038/s41467-020-18197-y (2020).
13 He, P. *et al.* The changing mouse embryo transcriptome at whole tissue and single-cell resolution. *Nature* **583**, 760-767, doi:10.1038/s41586-020-2536-x (2020).
14 Almanzar, N. *et al.* A single-cell transcriptomic atlas characterizes ageing tissues in the mouse. *Nature* **583**, 590-595, doi:10.1038/s41586-020-2496-1 (2020).
15 Park, S.-J., Shirahige, K., Ohsugi, M. & Nakai, K. DBTMEE: a database of transcriptome in mouse early embryos. *Nucleic Acids Research* **43**, D771-D776, doi:10.1093/nar/gku1001 (2015).
16 Mure, L. S. *et al.* Diurnal transcriptome atlas of a primate across major neural and peripheral tissues. *Science* **359**, doi:10.1126/science.aao0318 (2018).
17 Deboer, T., Vansteensel, M. J., Détári, L. & Meijer, J. H. Sleep states alter activity of suprachiasmatic nucleus neurons. *Nature Neuroscience* **6**, 1086-1090, doi:10.1038/nn1122 (2003).
18 Ding, L. *et al.* Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics. *Cell* **173**, 305-320.e310, doi:10.1016/j.cell.2018.03.033 (2018).
19 Cancer Genome Atlas Research, N. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543-550, doi:10.1038/nature13385 (2014).
20 Network, C. G. A. R. Comprehensive genomic characterization of squamous cell lung cancers. *Nature* **489**, 519-525, doi:10.1038/nature11404 (2012).
21 Isola, P., Zhu, J.-Y., Zhou, T. & Alexei. Image-to-Image Translation with Conditional Adversarial Networks. *arXiv pre-print server*, doi:None
arxiv:1611.07004 (2018).
22 Ledig, C. *et al.* Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network. *arXiv pre-print server*, doi:None
arxiv:1609.04802 (2017).
23 Karras, T. *et al.* Training Generative Adversarial Networks with Limited Data. *arXiv pre-print server*, doi:None
arxiv:2006.06676v2 (2020).
24 Li, W., Yin, Y., Quan, X. & Zhang, H. Gene Expression Value Prediction Based on XGBoost Algorithm. *Front Genet* **10**, 1077, doi:10.3389/fgene.2019.01077 (2019).
25 Subramanian, A. *et al.* A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **171**, 1437-1452 e1417, doi:10.1016/j.cell.2017.10.049 (2017).
26 Chen, Y., Li, Y., Narayan, R., Subramanian, A. & Xie, X. Gene expression inference with deep learning. *Bioinformatics* **32**, 1832-1839, doi:10.1093/bioinformatics/btw074 (2016).
27 Dizaji, K. G., Chen, W. & Huang, H. Deep Large-Scale Multitask Learning Network for Gene Expression Inference. *J Comput Biol* **28**, 485-500, doi:10.1089/cmb.2020.0438 (2021).
28 Cleary, B., Cong, L., Cheung, A., Lander, E. S. & Regev, A. Efficient Generation of Transcriptomic Profiles by Random Composite Measurements. *Cell* **171**, 1424-1436.e1418, doi:10.1016/j.cell.2017.10.023 (2017).

435

436 **Fig. Legends:**

437 **Fig. 1 Uniform manifold approximation and projection (UMAP) of the transcriptome**

438 **a**, Compartmentalization of the transcriptome manifold. Each compartment is color-coded based
439 on their enrichment (or lack of) for tissue-specific genes or gene ontologies.

440 **b**, Representative mouse tissue transcriptome images. The fragments per kilobase of exon per
441 million reads (FPKM) expression of each gene was log2 transformed then multiplied by 14,
442 clipped between [1, 255]. Color scales from 0 (red) to 255 (purple).

443 **c**, Representative baboon circadian transcriptome from the suprachiasmatic nucleus (SCN). Inserts
444 show zoom-in views of a section of neuron-specific gene expression oscillation across circadian
445 time at 3072x3072 pixel resolution. Color scale is the same as **b**.

446 **Fig. 2 Lung squamous cell carcinoma (LUSC) transcriptome characteristics and**
447 **classification**

448 **a**, Transcriptome images of normal (N) and tumor (T) tissues from patient 109 of dataset
449 GSE19804. Zoom-in views show upregulation of genes in the proliferation zone (upper right) and
450 downregulation of lung-specific genes (lower left) in the stage 3b tumor transcriptome (109T, 3b)
451 *vs.* normal controls (109N).

b, Transcriptome images of normal (N) and tumor (T) tissues from patient 102 of the same dataset as **a**. Patient 102 was also diagnosed with a stage 3b LUSC but showed less downregulation in lung-specific gene expression than in **a**.

c, Transcriptome images of normal (N) and tumor (T) tissues from patient 40 of the same dataset as **a**. Zoom-in view (lower left of the red dashed line) shows downregulation of immune genes.

d, Convolutional neural network-based classification of normal and LUSC transcriptome images. Network was trained for 100 epochs, and the loss and accuracy of the training and validation image set were plotted.

e, Partial transcriptome image of normal and tumor tissues from patient 109. Only genes in the lung-specific zone (I), the proliferation zone (J), and the immune zone (K) were rendered in the image.

f, Convolutional neural network-based classification of normal and LUSC transcriptome images with zones I, J, and K alone.

g, Median survival time (MST) after diagnosis of lung adenocarcinoma (LUAD) patients with different expression levels of surfactant protein A2 (SFTPA2). *p*-value was calculated for Log-rank (Mantel-Cox) test.

h, MST of LUSC patients with different expression levels of SFTPA2.

Fig. 3 Transcriptome inference through generative adversarial network (GAN) learning

a, Two-step transcriptome auto completion through pix2pix and super resolution GAN (SRGAN). The pix2pix performs training from transcriptome images containing only the top200 highest-expressing genes to full-transcriptome image at 768x768 pixel resolution. SRGAN performs training from 768x768 pixels to 3072x3072 pixels to resolve overlapping expression signals caused by close approximation of coordinates.

b, Representative images synthesized by the pix2pix and SRGAN after training. "Inferred" denotes images generated by GANs based on the top200 hints, and "Real" denotes the ground truth image.

c-d, Distribution of mean absolute error (MAE, **c**) and Pearson correlation (**d**) for 16,752 genes inferred from top200 genes across three datasets *vs*. their ground truth expression levels. N = 1634 for the PanCanAtlas training set (Train), 555 for LUAD, and 546 for LUSC. The training set did not include any samples from LUSC, and the 20 LUAD samples included in the training set was excluded from the LUAD testing data.

e, Correlation between real and inferred LUAD and LUSC transcriptome. The average 14x $\log_2$FPKM difference between LUAD and LUSC was calculated for each gene and plotted on *x* (real)- and *y* (inferred)-axes. N = 555 for LUAD, 546 for LUSC.

f, Classification of full (ALL) and partial (IJK) transcriptome images inferred from top200 genes. A set of 20 transcriptomes (10 normal and 10 LUSC patients) were generated from top200 genes and classified by the models trained in Fig. 2d and 2f. The experiment was repeated three times.

Fig. 4 *De novo* transcriptome synthesis and interpolation

a, Illustration of conditional and unconditional StyleGAN-ADA training with PanCanAtlas dataset, *de novo* transcriptome synthesis, and latent space interpolation.

b, Transcript level for representative genes in cell adhesion, mitochondria electron transport complex I, and immunity in LUSC transcriptome subtypes generated through conditional StyleGAN-ADA training.

c, Transcript level distribution for representative genes in Complex I, SFTPA2, and CD48 in their respective high- and low-expressing groups in clinical LUSC transcriptome samples. The expression level of Complex I was averaged among NDUFA2, A6, B7, B11, S7, and V1.

d-f, Probability of survival after diagnosis for patients expressing different level of mitochondria complex I (C), surfactant protein A2 (S), and immunity (CD48).

g, Representative images of the proliferation and immunity zone interpolating from normal to tumor transcriptome in the w+ space. Individual patient's normal and tumor transcriptome was inverted into unconditional StyleGAN w+ space, linear interpolated in 9 steps, and the interpolating transcriptomes from step 0 to step 9 were synthesized from the generative model. Arrow points to genes upregulated in the proliferation zone during the interpolating process while arrowheads denotes genes downregulated in the immunity zone.

h, Manifold projection of the interpolating transcriptomes shows distinct tumorigenesis trajectory for patients 122/126 *vs.* 130/144 on UMAP2.

i, Inverse correlation between cell cycle gene E2F7 expression and UMAP2 in patients 122/126 but not in patients 130/144.

j, STRING network analysis for top100 genes inversely correlated with UMAP2. STRING identifies 41 of the top100 genes reside in the same network as ubiquitin (UBX).

Supplemental Fig. Legends:

Fig. S1 UMAP-based transformation of the transcriptome from list structure to 2-D images

a, Illustration of the phoenix transformation. RNAseq results from ENCODE3 (ENCSR574CRQ), Tabula Muris Senis (GSE132040), and DBTMEE (DRA000484) are combined to calculate the k-neighborhood relationship for protein-coding genes in the mouse genome. The UMAP1/2 coordinates are rotated 30 degrees, reset the origin at (-17.5, -17.5), multiplied by 16, and rounded to integers as *x*- and *y*-coordinates. FPKM gene expression values are log2 transformed, multiplied by 14, added 1, and clipped between [1, 255].

b, Representative baboon tissue transcriptome images at ZT0. Color scale is set from 0 (red) to 255 (purple). MUG: muscle gastrocnemius, OES: esophagus, SPL: spleen, TEST: testis.

c, Representative lung (LUN) circadian transcriptome. Genes in the neuronal zone is specifically induced at ZT0. Color scale is the same as **b**.

Fig. S2 Immune-cold and hot cancers

Representative immune-cold and hot transcriptome images from the PanCanAtlas dataset. Yellow arrows denote immune zone gene expression are visibly lower than neighboring proliferation zone in immune-cold cancers; blue arrows denote equivalent or higher level gene expression in the immune zone than in the proliferation zone in immune-hot cancer transcriptomes.

Fig. S3 Keras convolutional neural network training for lung squamous cell carcinoma (LUSC) transcriptome classification

A total of 532 microarray-based transcriptome data 1:1 split between LUSC and paired, normal transcriptomes were combined from three studies, divided into training and validation sets (7:3 split), and trained for 100 epochs.

Fig. S4 Generative adversarial network training for image auto completion (pix2pix) and super resolution (SRGAN)

a, Illustration of the training process.

b, Loss of the generative model (G_Loss) during pix2pix training process.

c, G_ G_Loss during SRGAN training for 4X super resolution from 768x768 pixels to 3072x3072 pixels (SRGAN768_3072). The 768x768 pixel pictures are either resized from 1024x1024 pixel pictures through ImageMagick or synthesized by the pix2pix generator.

d, Peak signal to noise ratio (PSNR) for SRGAN768_3072 during the training process.

e-f, MAE (**e**) and Pearson (**f**) correlation of the pix2pix/SRGAN model trained across PanCanAtlas RNAseq and LUSC microarray transcriptome data. N = 532 for the LUSC training data (Train) and 20 for the testing data (Test).

f, Peak signal to noise ratio (PSNR) for SRGAN768_3072 during training.

Fig. S5 Conditional StyleGAN training and pseudo-LUSC transcriptome synthesis

a, Fréchet inception distance (FID) of conditional StyleGAN training. Each of the 93 classes in the PanCanAtlas dataset were randomly sampled 500 times to generate a dataset of 46,500 images, trained for 2500 kimg. The conditional model with FID of 5.87 at 2400 kimg was used for follow up analyses.

b, Illustration of the workflow downstream of a trained conditional StyleGAN-ADA model.

c, Pseudo-LUSC transcriptome landscape. A total of 62 fake LUSC transcriptomes (s0-61) were generated from the conditional StyleGAN-ADA model and clustered. Gene ontologies (GOs) of the differentially expressed genes (DEG) between cluster A *vs.* cluster B/C as well as those between cluster B *vs.* cluster C were analyzed through DAVID (Database for Annotation, Visualization and Integrated Discovery, https://david.ncifcrf.gov/).

d-e, MST of LUSC patients expressing high or low levels of mitochondria electron transport chain complex I (**d**) and the immunoglobin-like receptor CD48 (**e**).

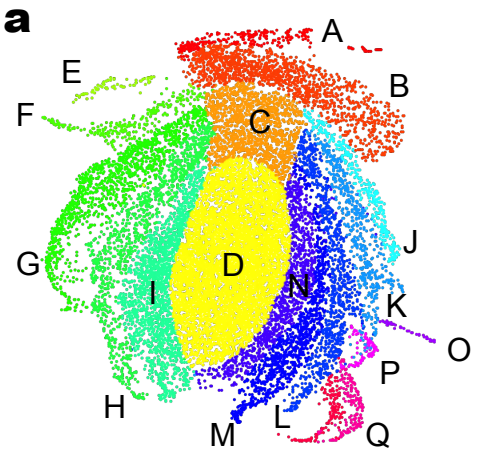
Fig. S6 Unconditional StyleGAN training and latent space interpolation

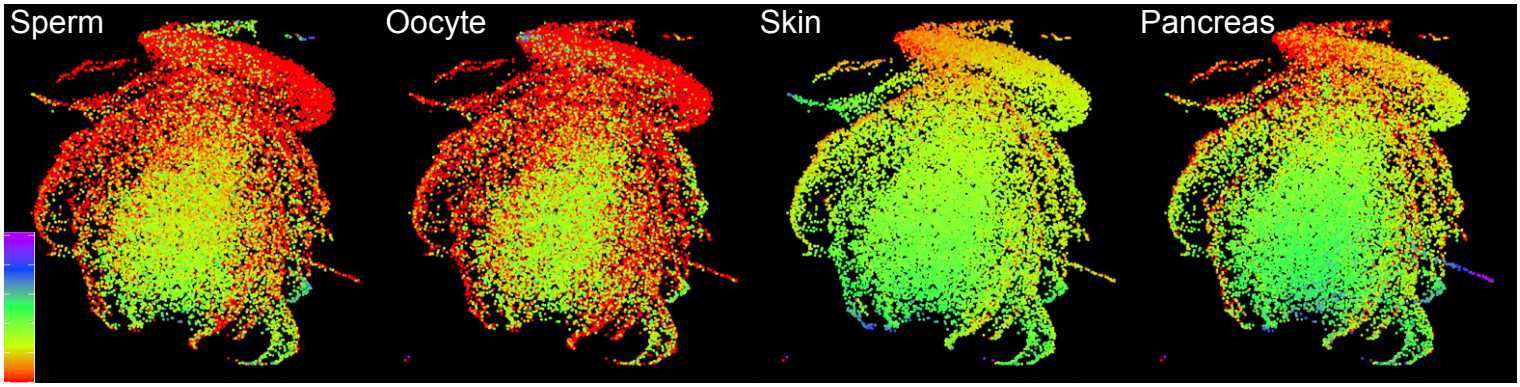
a, FID score of unconditional StyleGAN-ADA training. The unconditional model was first trained with the PanCanAtlas dataset for 2,700 kimg to reach a FID score of 5.92. The model was then retrained for 1,300 more kimg with the PanCanAtlas dataset supplemented by the microarray-based LUSC dataset to reach FID of 6.12. The transcriptomes of the LUSC dataset were over sampled 10,000 times to provide balance between RNAseq data and microarray data in the final training set.

b, Illustration of the workflow downstream of an unconditional StyleGAN-ADA model training.

c, Cluster analysis of the normal and cancer transcriptomes from the GSE19804 dataset. Patient 122 and 126 were selected for their distance away from the rest of samples.

d, Transcriptome images of 122/126/130/144 normal and tumor tissues. Patient 122T and 126T showed higher level of induction in the proliferation zone (right, big white box) and more complete suppression of the lung-specific zone (left, small white box) than 130T and 144T.





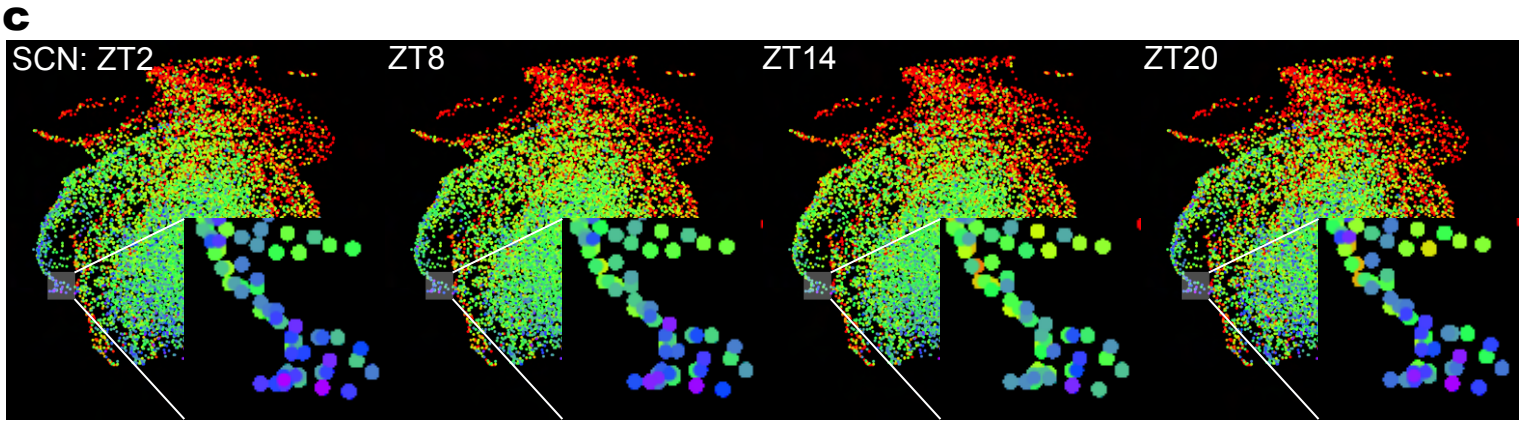


Fig. 1 Uniform manifold approximation and projection (UMAP) of the transcriptome
a, Compartmentalization of the transcriptome manifold. Each compartment is color-coded based on their enrichment (or lack of) for tissue-specific genes or gene ontologies (GO).
b, Representative mouse tissue transcriptome images. The fragments per kilobase of exon per million reads (FPKM) expression of each gene was log2 transformed then multiplied by 14, clipped between {1, 255}. Color scales from 0 (red) to 255 (purple).
c, Representative baboon circadian transcriptome from the suprachiasmatic nucleus (SCN). Inserts show zoom-in views of a section of neuron-specific gene expression oscillation across circadian time at 3072x3072 pixel resolution. ZT, zeitgeber time; SCN, suprachiasmic nucleus; color scale same as **b**.

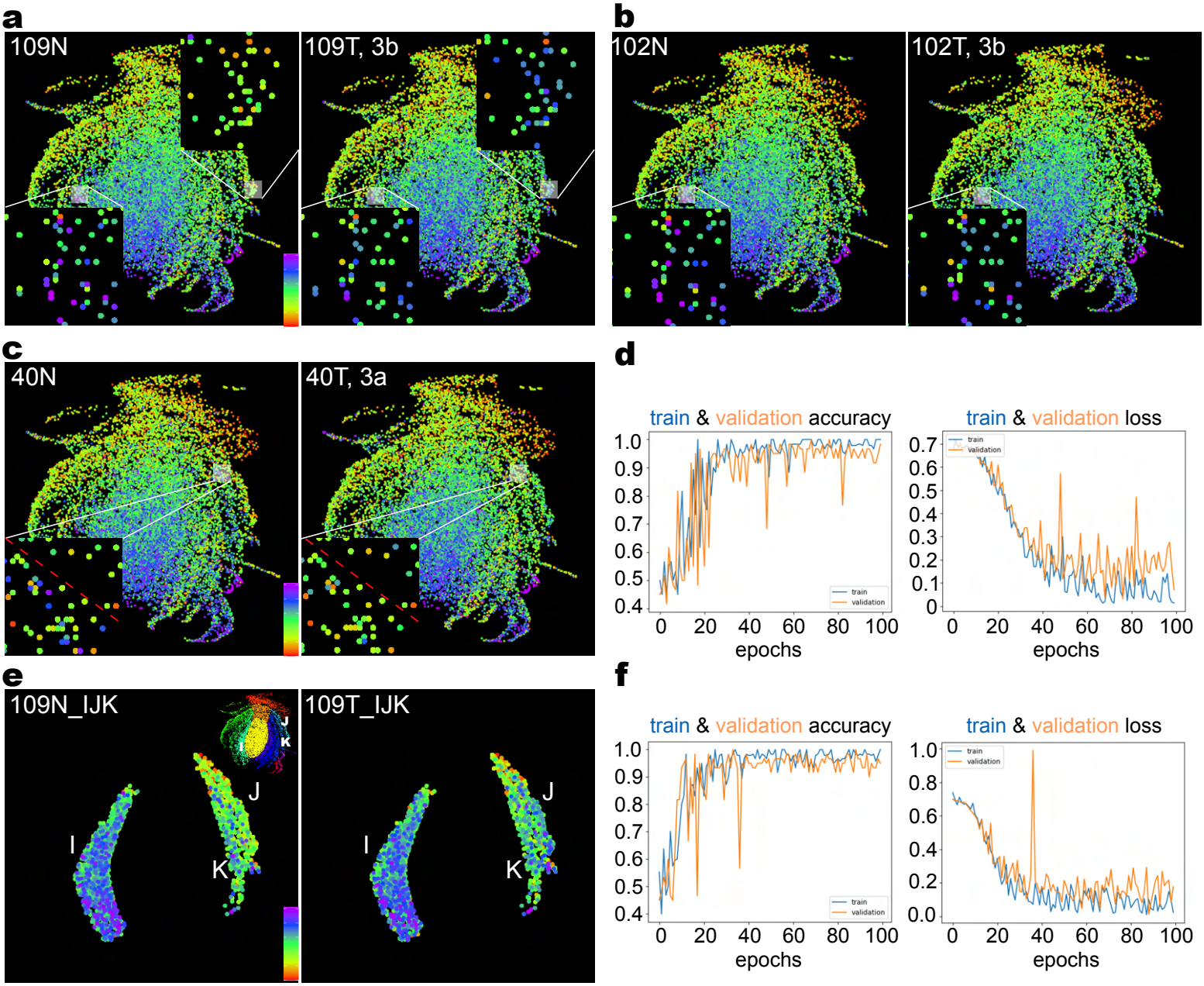


Fig. 2 Lung squamous cell carcinoma (LUSC) transcriptome characteristics and classification
a-c, Transcriptome images of normal (N) and tumor (T) tissues from patient #109 (**a**, stage 3b), #102 (**b**, stage 3b), and #40 (**c**, stage 3a) from dataset GSE19804. Zoom-in views in (**a**) show upregulation of genes in the proliferation zone (upper right) and downregulation of lung-specific genes (lower left), in (**b**) show less downregulation in lung-specific gene expression than in a, in (**c**) shows downregulation of immune genes.
d, Convolutional neural network-based classification of normal and LUSC transcriptome images. Network was trained for 100 epochs, and the loss and accuracy of the training and validation image set were plotted.
e, Partial transcriptome image of normal and tumor tissues from patient #109. Only genes in the lung-specific zone (I), the proliferation zone (J), and the immune zone (K) were rendered in the image.
f, Convolutional neural network-based classification of normal and LUSC transcriptome images with zones I, J, and K alone.
g, Median survival time (MST) after diagnosis of lung adenocarcinoma (LUAD) patients with different expression levels of surfactant protein A2 (SFTPA2). *p*-value was calculated for Log-rank (Mantel-Cox) test.
h, MST of LUSC patients with different expression levels of SFTPA2.

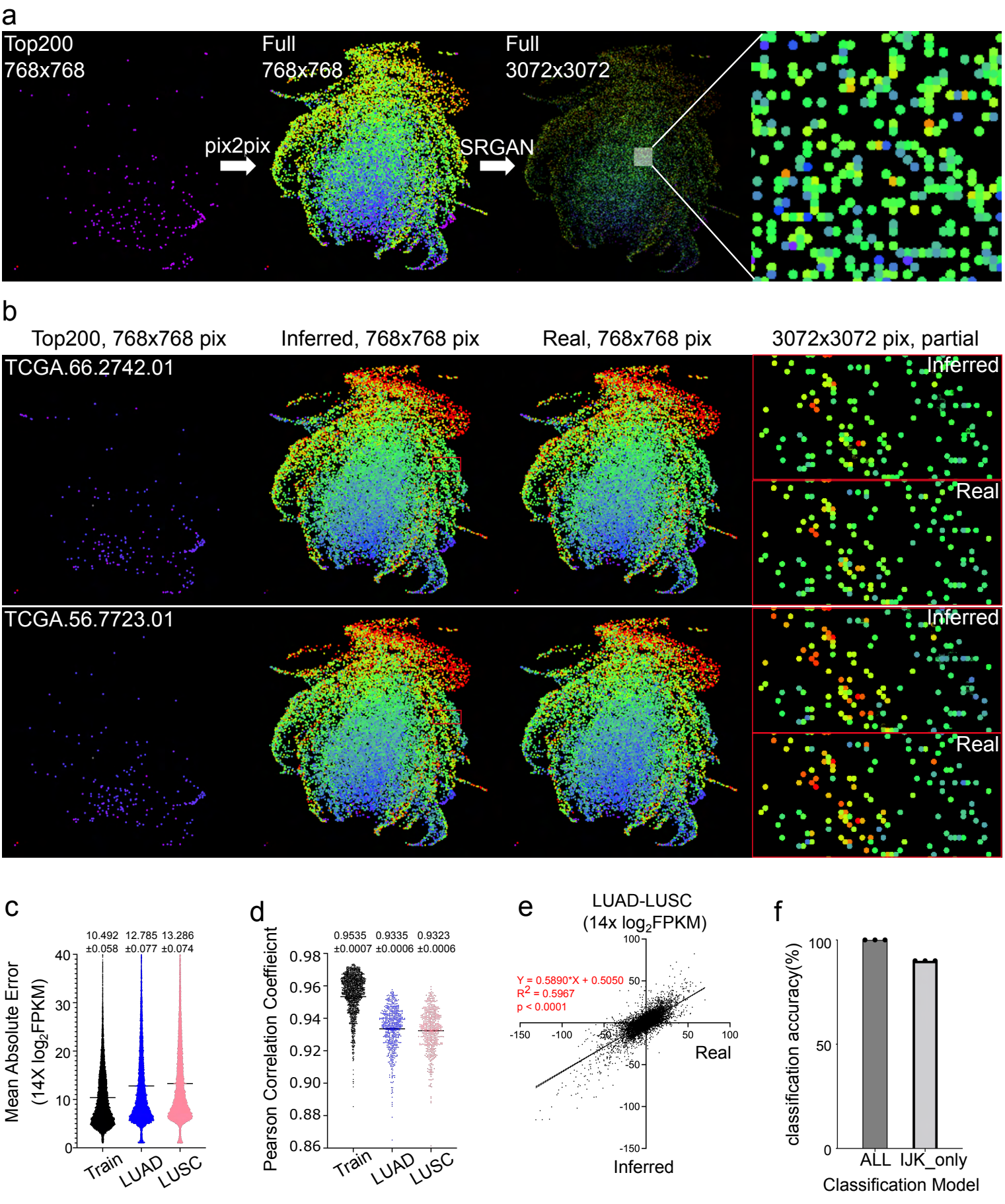


Fig. 3 Transcriptome inference through generative adversarial network (GAN) learning

a, Two-step transcriptome auto completion through pix2pix and super resolution GAN (SRGAN). The pix2pix performs training from transcriptome images containing only the top200 highest-expressing genes to full-transcriptome image at 768x768 pixel resolution. SRGAN performs training from 768x768 pixels to 3072x3072 pixels to resolve overlapping expression signals caused by close approximation of coordinates.

b, Representative images synthesized by the pix2pix and SRGAN after training. “Inferred” denotes images generated by GANs based on the top200 hints, and “Real” denotes the ground truth image.

c-d, Distribution of mean absolute error (MAE, **c**) and Pearson correlation coefficient (**d**) for 16,752 genes inferred from top200 genes across three datasets vs. their ground truth expression levels. N = 1634 for the PanCanAtlas training set (Train), 555 for LUAD, and 546 for LUSC. The training set did not include any samples from LUSC, and the 20 LUAD samples included in the training set was excluded from the LUAD testing data.

e, Correlation between real and inferred LUAD and LUSC transcriptome. The average 14x log2FPKM difference between LUAD and LUSC was calculated for each gene and plotted on x (real)- and y (inferred)-axes. N = 555 for LUAD, 546 for LUSC.

f, Classification of full (ALL) and partial (IJK) transcriptome images inferred from top200 genes. A set of 20 transcriptomes (10 normal and 10 LUSC patients) were generated from top200 genes and classified by the models trained in Fig. 2d and 2f. The experiment was repeated three times.

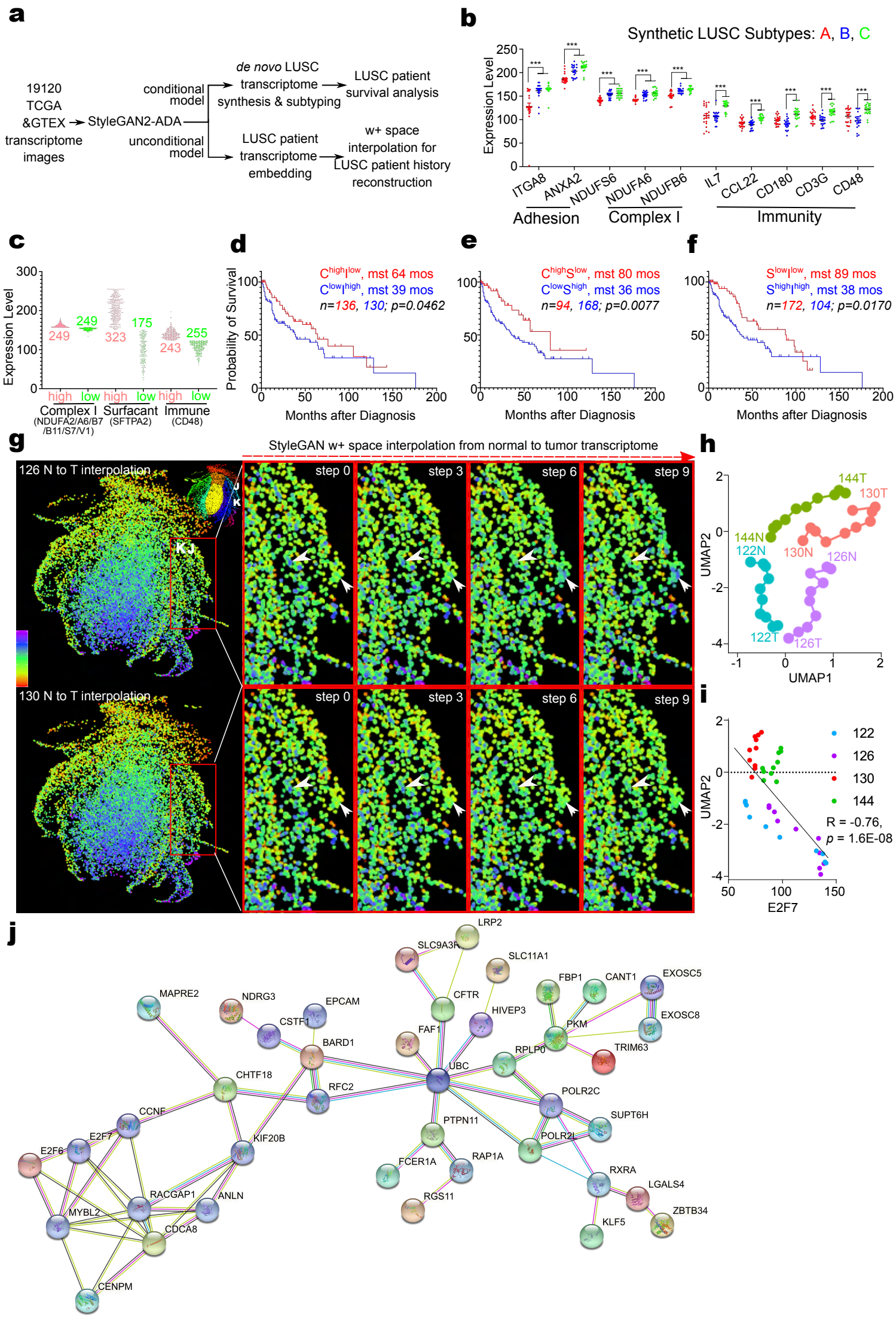

a
19120 TCGA &GTEX transcriptome images
StyleGAN2-ADA
conditional model
unconditional model
de novo LUSC transcriptome synthesis & subtyping
LUSC patient survival analysis
LUSC patient transcriptome embedding
w+ space interpolation for LUSC patient history reconstruction
b
Synthetic LUSC Subtypes: A, B, C
Expression Level
ITGA8
ANXA2
NDUFS6
NDUFA6
NDUFB6
IL7
CCL22
CD180
CD3G
CD48
Adhesion
Complex I
Immunity
c
Expression Level
249
249
323
175
243
255
high low high low high low
Complex I (NDUFA2/A6/B7 /B11/S7/V1)
Surfacant (SFTPA2)
Immune (CD48)
d
Probability of Survival
C^high I^low, mst 64 mos
C^low I^high, mst 39 mos
n=136, 130; p=0.0462
Months after Diagnosis
e
C^high S^low, mst 80 mos
C^low S^high, mst 36 mos
n=94, 168; p=0.0077
Months after Diagnosis
f
S^low I^low, mst 89 mos
S^high I^high, mst 38 mos
n=172, 104; p=0.0170
Months after Diagnosis
g
StyleGAN w+ space interpolation from normal to tumor transcriptome
126 N to T interpolation
130 N to T interpolation
step 0
step 3
step 6
step 9
K
J
h
144T
130T
144N
122N
130N
126N
122T
126T
UMAP2
UMAP1
i
122
126
130
144
R = -0.76,
p = 1.6E-08
UMAP2
E2F7
j
LRP2
SLC9A3R
SLC11A1
FBP1
CANT1
EXOSC5
EXOSC8
MAPRE2
NDRG3
EPCAM
CSTF1
BARD1
CFTR
HIVEP3
FAF1
PKM
RPLP0
TRIM63
UBC
CHTF18
RFC2
POLR2C
SUPT6H
CCNF
E2F7
E2F6
KIF20B
PTPN11
POLR2L
FCER1A
RAP1A
RXRA
LGALS4
RACGAP1
ANLN
RGS11
ZBTB34
MYBL2
CDCA8
KLF5
CENPM

Fig. 4 *De novo* transcriptome synthesis and interpolation

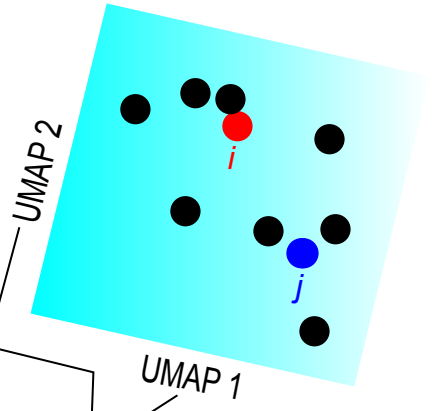
a, Illustration of conditional and unconditional StyleGAN-ADA training with PanCanAtlas dataset, *de novo* transcriptome synthesis, and latent space interpolation.
b, Transcript level for representative genes in cell adhesion, mitochondria electron transport complex I, and immunity in LUSC transcriptome subtypes generated through conditional StyleGAN-ADA training.
c, Transcript level distribution for representative genes in Complex I, SFTPA2, and CD48 in their respective high- and low-expressing groups in clinical LUSC transcriptome samples. The expression level of Complex I was averaged among NDUFA2, A6, B7, B11, S7, and V1.
d-f, Probability of survival after diagnosis for patients expressing different level of mitochondria complex I (C), surfactant protein A2 (S), and immunity (CD48).
g, Representative images of the proliferation and immunity zone interpolating from normal to tumor transcriptome in the w+ space. Individual patient's normal and tumor transcriptome was inverted into unconditional StyleGAN w+ space, linear interpolated in 9 steps, and the interpolating transcriptomes from step 0 to step 9 were synthesized from the generative model. Arrow points to genes upregulated in the proliferation zone during the interpolating process while arrowheads denotes genes downregulated in the immunity zone.
h, Manifold projection of the interpolating transcriptomes shows distinct tumorigenesis trajectory for patients 122/126 vs. 130/144 on UMAP2.
i, Inverse correlation between cell cycle gene E2F7 expression and UMAP2 in patients 122/126 but not in patients 130/144.
j, STRING network analysis for top100 genes inversely correlated with UMAP2. STRING identifies 41 of the top100 genes reside in the same network as UBX.

a

selecting datasets

	Sample 1	Sample 2	...	Sample n
Gene 1	12	3	...	6764
Gene 2	1	5782	...	1
Gene 3	35765	7	...	0
...	...	...	...	...
...	...	...	...	...
Gene i	2789	37	...	4281
...	...	...	...	...
...	...	...	...	...
Gene j	88	278	...	972
...	...	...	...	...
...	...	...	...	...

$$q_{ij}^{UMAP} = \left(1 + a\left\|z_i - z_j\right\|^{2b}\right)^{-1}$$

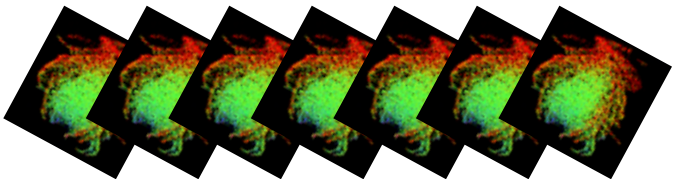


scaled between {0, 512}

log_2FPKM * 14 + 1, clipped {1, 255}

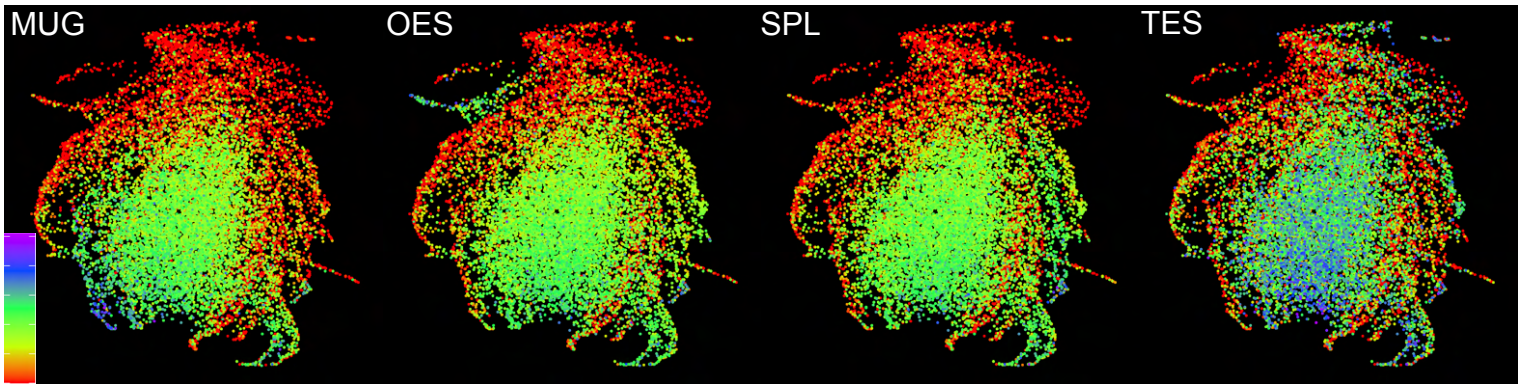
	Cor.X	Cor.Y	Sample 1	...	Sample n
Gene 1	3	12	3	...	255
Gene 2	500	398	6	...	1
Gene 3		255	7	...	1
...		...	...	...	...
...		...	...	...	...
Gene i	256	223	12	...	79
...		...	...	...	...
...		...	...	...	...
Gene j	345	123	9	...	53
...		...	...	...	...
Max_Ref	15	15	255	255	255
Min_Ref	20	20	0	0	0

Sample 1 Sample 2 ...Sample k..... Sample n

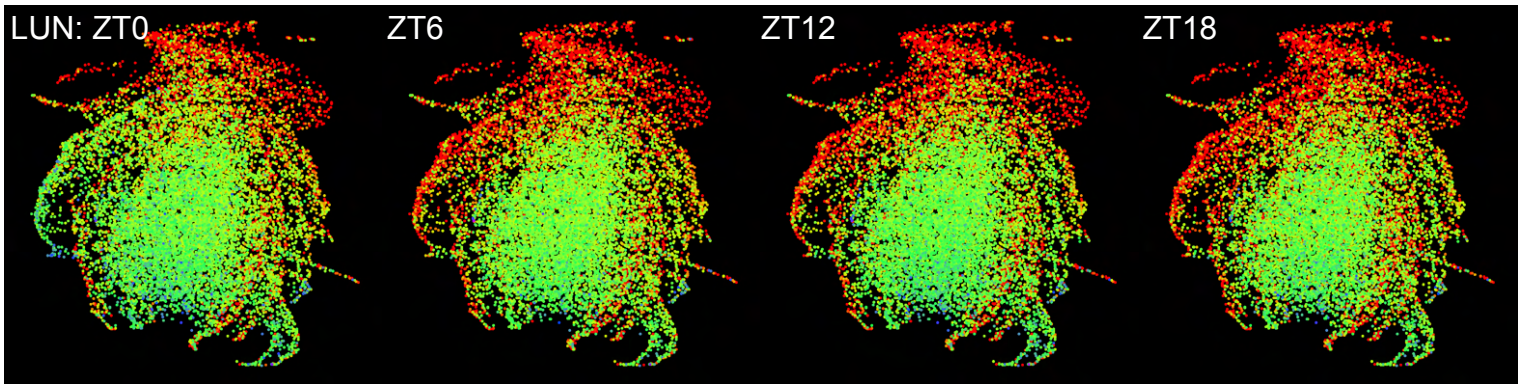


1: Each gene is represented by a point of 30 pixels;
2: A five-color rainbow scheme is used to render gene expression levels;
3: Genes are rendered in ggplot2 in ascending expression levels to emphasize high-expression genes;
4: For each sample, images are printed at 512x512, 1024x1024, and 3072x3072 sizes.

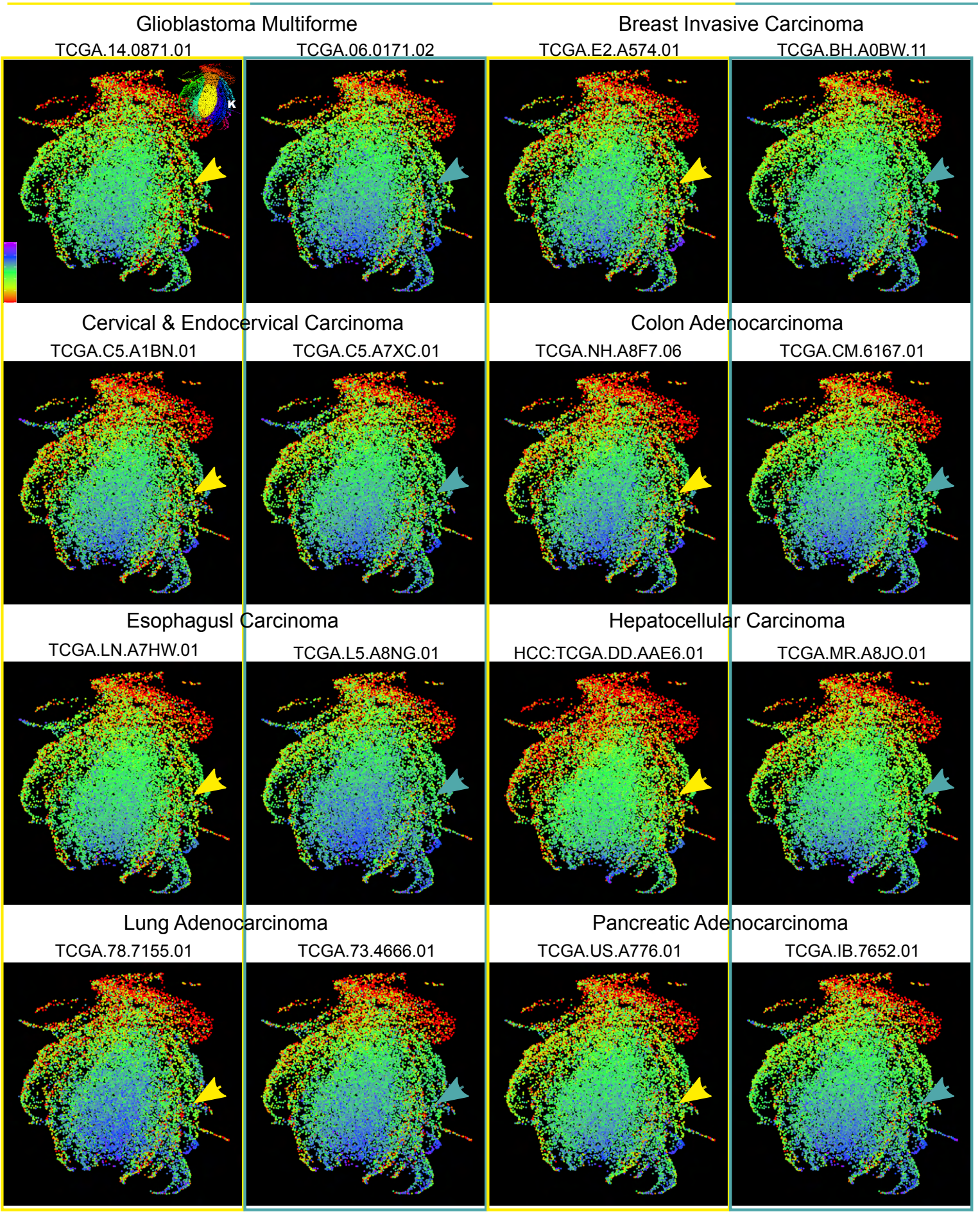
b



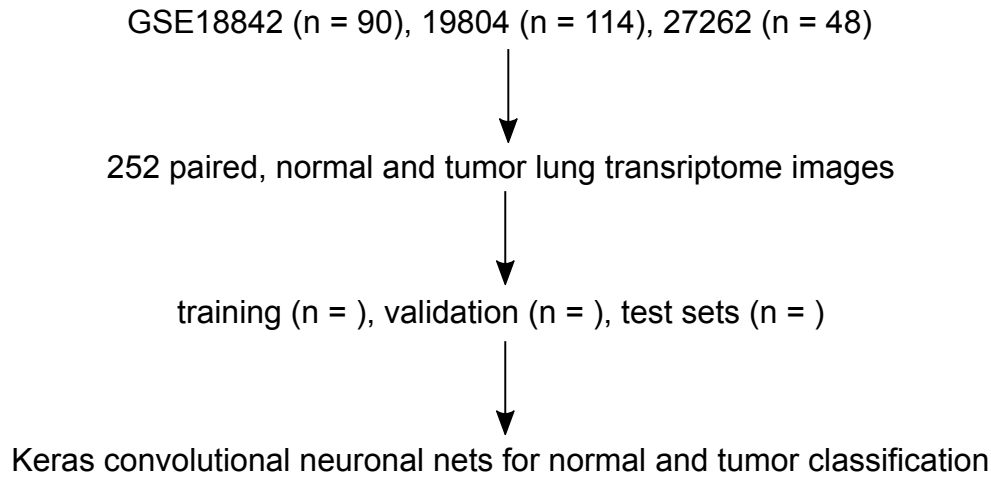
c



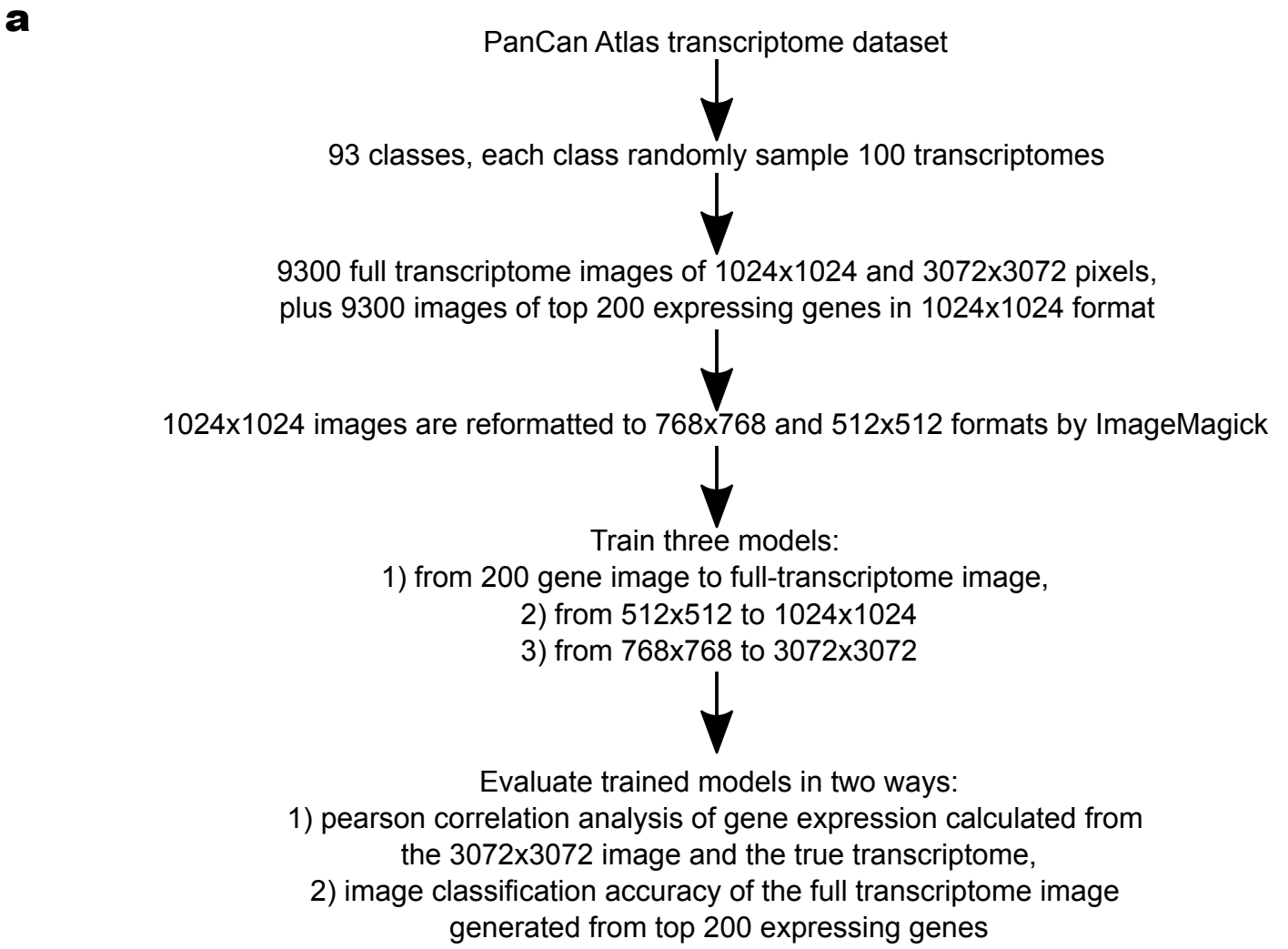
Supplementary Fig. 1 UMAP-based transformation of the transcriptome from list structure to 2-D images
a, Illustration of the phoenix transformation. RNAseq results from ENCODE3 (ENCSR574CRQ), Tabula Muris Senis (GSE132040), and DBTMEE (DRA000484) are combined to calculate the k-neighborhood relationship for protein-coding genes in the mouse genome. The UMAP1/2 coordinates are rotated 30 degrees, reset the origin at (-17.5, -17.5), multiplied by 16, and rounded to integers as x- and y-coordinates. FPKM gene expression values are log2 transformed, multiplied by 14, added 1, and clipped between {1, 255}.
b, Representative baboon tissue transcriptome images at ZT0. Color scale is set from 0 (red) to 255 (purple). MUG: muscle gastrocnemius, OES: esophagus, SPL: spleen, TEST: testis.
c, Representative lung (LUN) circadian transcriptome. Genes in the neuronal zone is specifically induced at ZT0. Color scale is the same as **b**.

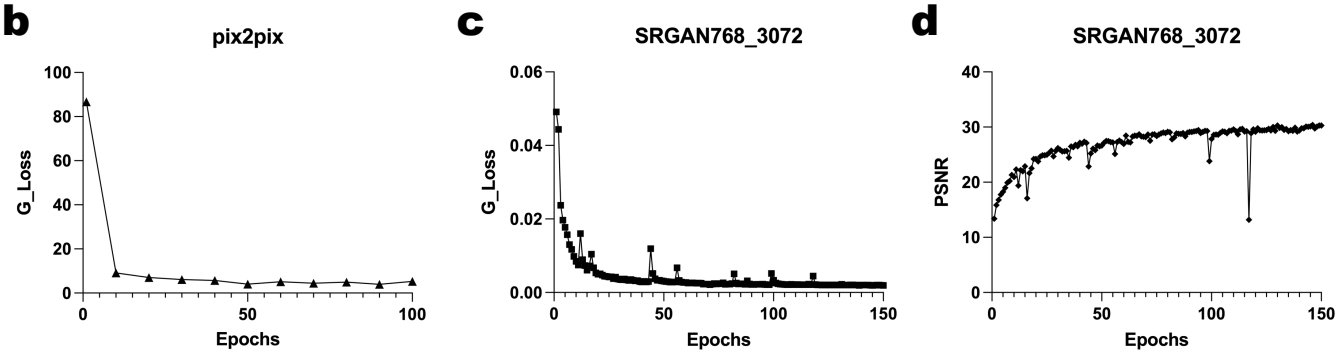


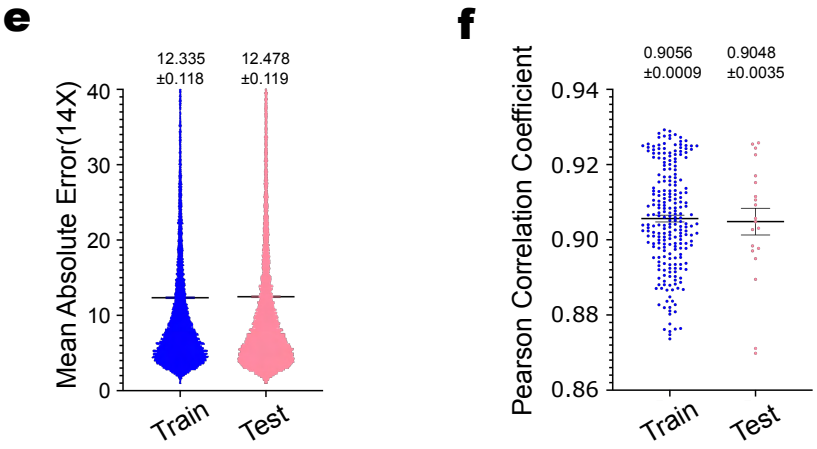
Supplementary Fig. 2 Immune-cold and hot cancers
Representative immune-cold and hot transcriptome images from the PanCanAtlas dataset. Yellow arrows denote immune zone gene expression are visibly lower than neighboring proliferation zone in immune-cold cancers; blue arrows denote equivalent or higher level gene expression in the immune zone than in the proliferation zone in immune-hot cancer transcriptomes.



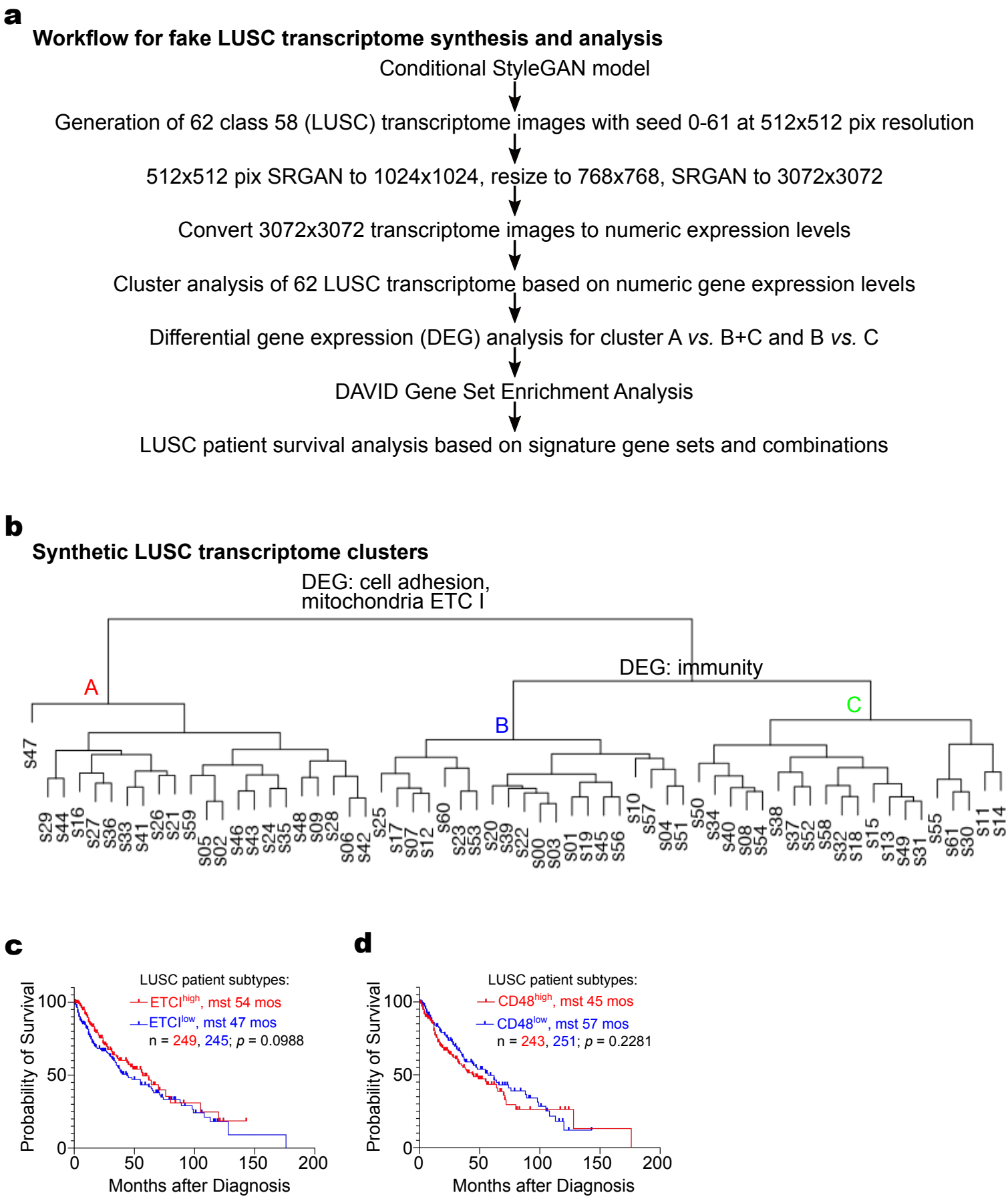
Supplementary Fig. 3 Keras convolutional neural network training for lung squamous cell carcinoma (LUSC) transcriptome classification
A total of 532 microarray-based transcriptome data 1:1 split between LUSC and paired, normal transcriptomes were combined from three studies, divided into training and validation sets (7:3 split), and trained for 100 epochs.



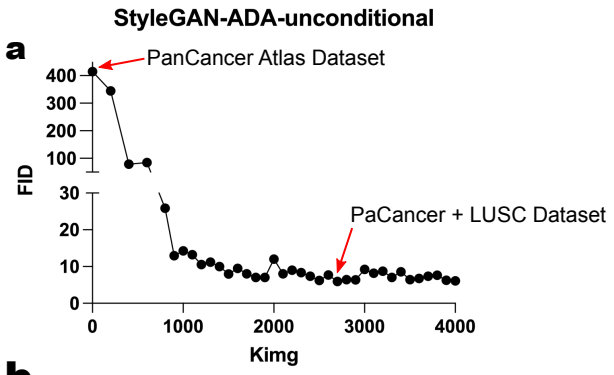




Supplementary Fig. 4 Generative adversarial network training for image auto completion (pix2pix) and super resolution (SRGAN)
a, Illustration of the training process.
b, Loss of the generative model (G_Loss) during pix2pix training process.
c, G_Loss during SRGAN training for 4X super resolution from 768x768 pixels to 3072x3072 pixels (SRGAN768_3072). The 768x768 pixel pictures are either resized from 1024x1024 pixel pictures through ImageMagick or synthesized by the pix2pix generator.
d, Peak signal to noise ratio (PSNR) for SRGAN768_3072 during the training process.
e-f, MAE (**e**) and Pearson (**f**) correlation of the pix2pix/SRGAN model trained across PanCanAtlas RNAseq and LUSC microarray transcriptome data. N = 532 for the LUSC training data (Train) and 20 for the testing data (Test).



Supplementary Fig. 5 Conditional StyleGAN training and pseudo-LUSC transcriptome synthesis
a, Fréchet inception distance (FID) of conditional StyleGAN training. Each of the 93 classes in the PanCanAtlas dataset were randomly sampled 500 times to generate a dataset of 46,500 images, trained for 2500 kimg. The conditional model with FID of 5.87 at 2400 kimg was used for follow up analyses.
b, Illustration of the workflow downstream of a trained conditional StyleGAN-ADA model.
c, Pseudo-LUSC transcriptome landscape. A total of 62 fake LUSC transcriptomes (s0-61) were generated from the conditional StyleGAN-ADA model and clustered. Gene ontologies (GOs) of the differentially expressed genes (DEG) between cluster A vs. cluster B/C as well as those between cluster B vs. cluster C were analyzed through DAVID (Database for Annotation, Visualization and Integrated Discovery, https://david.ncifcrf.gov/).
d-e, MST of LUSC patients expressing high or low levels of mitochondria electron transport chain complex I (**d**) and the immunoglobin-like receptor CD48 (**e**).



b

Workflow downstream of unconditional StyleGAN model training

Unconditional StyleGAN model

↓

Projection of four pairs of patient-specific normal and cancer tissue transcriptome into w+ space

↓

Interpolation between paired normal and cancer tissue transcriptome in w+ space in 9 steps

↓

SRGAN interpolating image to 3072x3072 images, convert to numeric gene expression values

↓

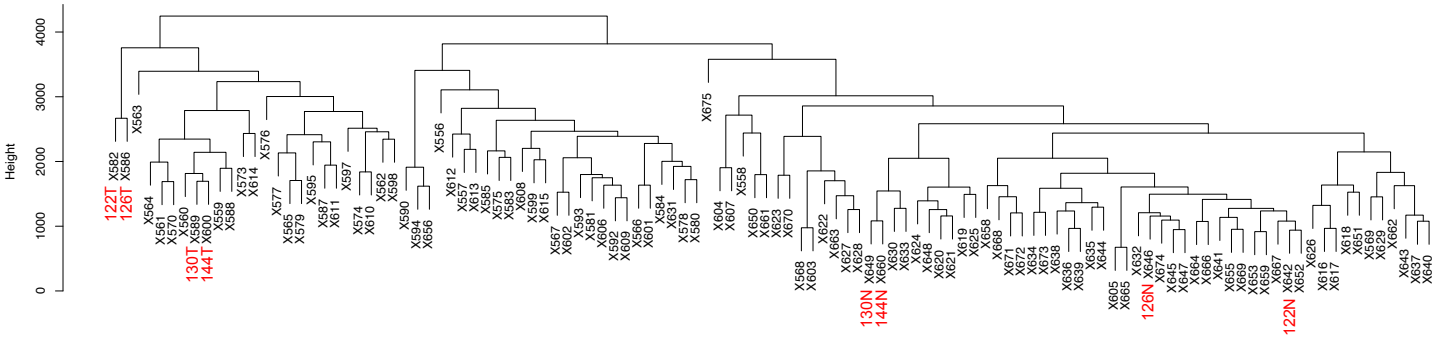
UMAP embedding of 40 transcriptomes

↓

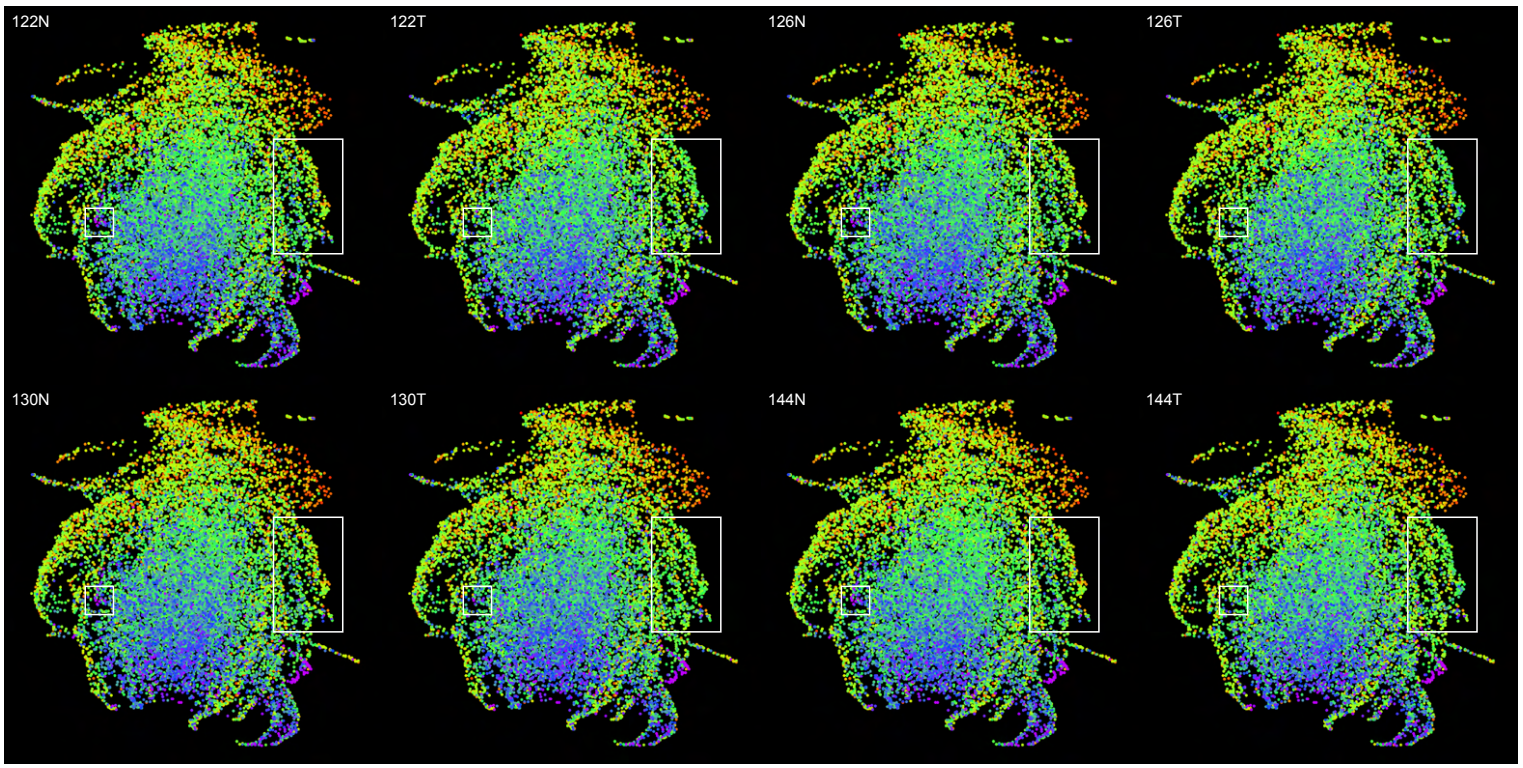
Pearson correlation and STRING analysis of gene expression along UMAP2 as cancer drivers

c

Cluster analysis of GSE19804 dataset



d



Supplementary Fig. 6 Unconditional StyleGAN training and latent space interpolation
a, FID score of unconditional StyleGAN-ADA training. The unconditional model was first trained with the PanCanAtlas dataset for 2,700 kimg to reach a FID score of 5.92. The model was then retrained for 1,300 more kimg with the PanCanAtlas dataset supplemented by the microarray-based LUSC dataset to reach FID of 6.12. The transcriptomes of the LUSC dataset were over sampled 10,000 times to provide balance between RNAseq data and microarray data in the final training set.
b, Illustration of the workflow downstream of an unconditional StyleGAN-ADA model training.
c, Cluster analysis of the normal and cancer transcriptomes from the GSE19804 dataset. Patient 122 and 126 were selected for their distance away from the rest of samples.
d, Transcriptome images of 122/126/130/144 normal and tumor tissues. Patient 122T and 126T showed higher level of induction in the proliferation zone (right, big white box) and more complete suppression of the lung-specific zone (left, small white box) than 130T and 144T.

ZIP File

Click here to access/download

ZIP File

Supplementary_Tables_Archive.zip