scientific reports

Check for updates

OPEN

A deep-learning model for characterizing tumor heterogeneity using patient-derived organoids

Kosuke Takagi[1], Motoki Takagi[2,3]✉, Gen Hiyama[2] & Kazuhito Goda[4]

Genotypic and phenotypic diversity, which generates heterogeneity during disease evolution, is common in cancer. The identification of features specific to each patient and tumor is central to the development of precision medicine and preclinical studies for cancer treatment. However, the complexity of the disease due to inter- and intratumor heterogeneity increases the difficulty of effective analysis. Here, we introduce a sequential deep learning model, preprocessing to organize the complexity due to heterogeneity, which contrasts with general approaches that apply a single model directly. We characterized morphological heterogeneity using microscopy images of patient-derived organoids (PDOs) and identified gene subsets relevant to distinguishing differences among original tumors. PDOs, which reflect the features of their origins, can be reproduced in large quantities and varieties, contributing to increasing the variation by enhancing their common characteristics, in contrast to those from different origins. This resulted in increased efficiency in the extraction of organoid morphological features sharing the same origin. Linking these tumor-specific morphological features to PDO gene expression data enables the extraction of genes strongly correlated with intertumor differences. The relevance of the selected genes was assessed, and the results suggest potential applications in preclinical studies and personalized clinical care.

Heterogeneity is a complex attribute of cancer tumors[1,2]. Multiple factors, including genome instability and microenvironmental differences, produce cellular diversity, which results in complex dynamics such as those observed in proliferation and metastasis processes[3–5]. This leads to heterogeneity at the molecular, cellular, and tissue levels.

A high degree of variety in morphology, which arises from cellular heterogeneity consisting of distinct subpopulations, is a characteristic of tumors. These types of heterogeneity can be observed across patients and within a single tumor and are known as inter- and intratumor heterogeneity, respectively[6–11]. A better characterization of individual variability based on morphological features and genomic profiling is needed for drug design and clinical implementation in therapeutics, diagnostics, and personalized medicine for cancerr[12–23]. However, there is still a lack of understanding of the relationship between phenotypic changes and the genomic mechanisms underlying cancer due to the complexity of cellular dynamics.

In an effort to understand the complexity of cancer diseases, organoids serve as promising models for mimicking tumor behavior with three-dimensional multicellular structures in vitro[24–28]. Among the various types of organoids, patient-derived organoids (PDOs) are those established from patient tumor tissues. PDOs are produced by culturing one cell or multiple cells obtained from patients' tissues in vitro, which are acquired from, for example, surgical or biopsy samples of patient tumor tissues. In this study, we used Fukushima patient-derived tumor organoids (F-PDOs), which are a series of PDOs from various types of tumor tissues established by the Fukushima Translational Research Center[29–33]. This organoid model was developed under a concept different from those using isolated single cells from patient tumors. F-PDOs were derived from minced tissues acquired from the primary solid tumor tissues of patients. Because minced pieces of tissues maintain the local structures of multiple cells, we can expect better reflection of the features of the source tumor. In fact, histological observation and the gene expression analysis[30] revealed that they retain the function and architecture of their source tissues even after long-term culture. By maintaining intertumor differences across patients, F-PDOs have the potential to model diverse characteristics of tumors by reproducing the responses to various factors, such as environmental changes or chemical stimulation[29–33]. Thus, we used PDOs that allow direct measurements of such dynamics under in vitro conditions by contrasting inter- and intratumor differences.

[1]Research and Development, Advanced Core Technology Japan Unit 2, Evident Corp. Hachioji, 192-0033 Tokyo, Japan. [2]Translational Research Center, Fukushima Medical University, 960-1295 Fukushima, Japan. [3]JeiserBio Inc, 220-0004 Yokohama, Japan. [4]Research and Development, Advanced Biological Engineering Japan, Evident Corp., 192-0033 Hachioji, Japan. ✉email: motokitakagi02@gmail.com

In addition to the basic biological profiles, such as gene expression and drug responses obtained through PDO experiments, microscopy images might provide substantial information on organoids and offer insight into their complex dynamics[28,34–37]. These image datasets were applied to identify genes with distinct patterns of gene expression and characterize tumors[38,39] using machine learning methods. For these approaches, PDO cultures contribute to augmenting the training datasets by reproducing clonal populations. In deep neural networks, in general, the size and variety of training datasets have an impact on the accuracy of learning. Using microscopy images of F-PDOs, we introduced a machine learning-based approach that can reduce the complexity and extract representations of input image datasets accurately[40–46]. Morphological features common to organoids of the same origin were identified to distinguish them from other organoids. We subsequently determined the features characterizing the intertumor heterogeneity. Integrating data on biological profiles, including the morphological data obtained as described above, might facilitate the prediction of clinical responses and the understanding of the mechanisms underlying complex processes in tumors[46–53]. In this report, the extracted image features were applied to select genes that might be associated with morphological changes due to intertumor heterogeneity. To clarify the relevance of the selected genes, further analyses of gene pathways were performed and correlations with the drug response were calculated.

Results

Overview of the analysis model

In this study, we used F-PDO datasets for gene expression, drug response, and microscopy images of lung cancer organoids (Fig. 1a). As illustrated in Fig. 1b, the input datasets were sequentially processed by two different deep neural networks (DNNs), the convolutional network for images and subsequent regression models applied to predict gene expression. The first DNN generated the image representation, which was obtained by reducing the dimensions of the input image datasets into vectors with smaller dimensions. Here, the DNN was trained to distinguish the original label, which was the label of the organoid line. Then, the representations were designed to contain the information enhancing the difference between the organoids across patients. In the subsequent process, gene sets were extracted by using this information. Based on the prediction accuracy of the gene expression from the image representations, the genes were selected to reflect the morphological changes in the organoids.

Representation of images

First, the images were analyzed with a neural network, and the representations of the images were obtained. As explained in the Introduction and shown in the images (Fig. 2a), F-PDOs in culture had a variety of morphologies. To identify the characteristics common to organoids that share the same origin, a convolutional neural network

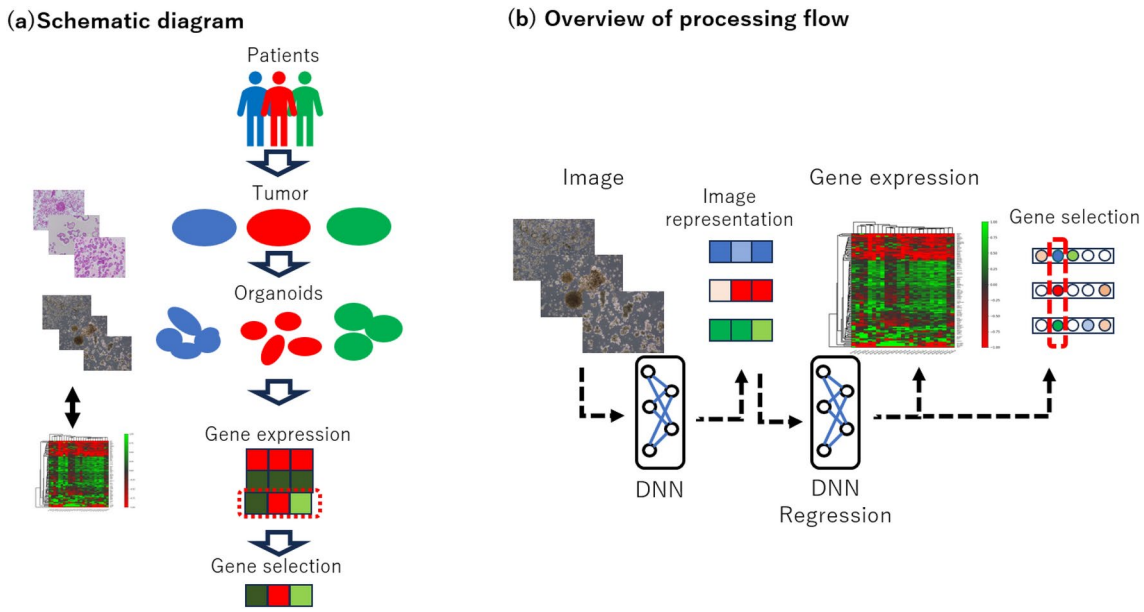

Fig. 1. Overview of model. (**a**) Schematic diagram of the study. We analyzed datasets of Fukushima-patient derived organoids (F-PDOs) that contained microscopy images, gene expression data, and drug response profiles. The data were acquired from in vitro cultured tumor tissue samples from different patients. By integrating these datasets, subsets of genes were selected through correlation analysis of images and gene expression data. (**b**) Overview of the processing flow. Two different deep neural networks (DNNs) were applied sequentially to datasets containing microscopy images and gene expression data from F-PDOs. The first one was trained with the image to distinguish between the original tissues. Next, we predicted the gene expression from the extracted representations of the images. Based on the results of the second model, the genes that might have reflected the difference between the original tissues were selected.

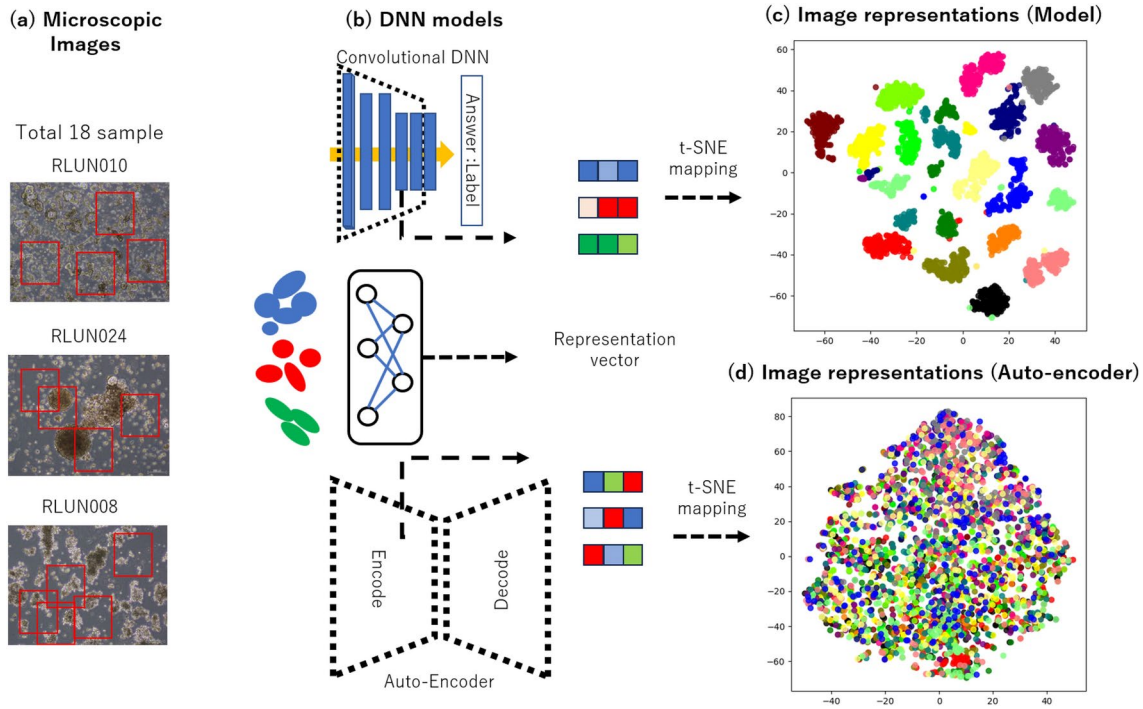

Fig. 2. Image representations. (**a**) Microscopy images. Microscopy images obtained with a brightfield 10x objective with a color CCD are displayed. Labels such as "RLUN010" indicate the origin of the PDO according to, patient and tumor. The input image was preprocessed by randomly cropping image patches, as indicated by the red rectangles on the images. (**b**) DNN models. The upper half of the diagram shows a convolutional neural network (CNN) of our model. This CNN was trained to answer the label of each image. It encoded the input image, and the image representation was obtained by vector outputs from the middle layer. It was compared with the autoencoder model illustrated in the lower column, in which the model was trained to reconstruct the input image instead of answering the label. (**c**) Image representations of the model. The representations of the images were obtained as 10-dimensional vectors from the intermediate layer and were projected on 2-dimensional space by using the t-distributed stochastic neighbor embedding (t-SNE) algorithm. For this mapping, the image labels were distinguished by different colors. (**d**) Image representations of the autoencoder network. The results in (**c**) were compared with those of the autoencoder network model. The t-SNE plot was made in the same manner as in the case of (**c**).

(CNN) was trained to determine the label of each image. The CNN is a basic deep learning method commonly used for analyzing the spatial patterns of images. The representation of the image, or the latent representation, was subsequently obtained by extraction from the middle layer of the network model, as illustrated in Fig. 2b.

The images were visualized using t-distributed stochastic neighbor embedding (t-SNE), in which these 10-dimensional latent values were projected into a lower-dimensional space (Fig. 2c). Each plot corresponded to a single image patch and was colored differently according to the label value, displaying the accuracy of the training with a scatter plot, on which a set of patches of the same label were distributed in proximity, apart from those of the different labels. As we expected, the results indicated that the representations exhibited different patterns according to the differences in the organoid lines.

In this study, images were encoded into representations that distinguished between inter- and intratumor heterogeneity. Fewer overlapping distributions across different colored labels indicated that the differences among the original tumors, which corresponded to the intertumor heterogeneity, were distinguished by being encoded differently. Conversely, the differences within the same labeled group, or intratumor heterogeneity, exhibited small fluctuations. This encoding enhanced the intertumor differences by compressing the intratumor differences.

These results were compared to those of another neural network model with an autoencoder (AE) by using the same datasets (Fig. 2b,d). The AE model is widely used to extract image features via unsupervised learning[41,42], where the input datasets are trained to reconstruct images that are the same as the input images. Compared with our model (Fig. 2c), the AE model generated representations that were distributed with mixed patterns regarding the label, as shown in the t-SNE plot (Fig. 2d). Even though the AE model captured the image features of each patch, the information of the original tissues was lost compared with that of the model representation (Fig. 2c), in which the difference between the original tumors, or the intertumor heterogeneity, was enhanced and apparent. Thus, in the following analysis, we used the results obtained by our model shown in Fig. 2c.

Image-based gene mapping and gene selection

Using the generated representations of images that enabled us to distinguish intertumor heterogeneity, we selected the gene subset responsible for the morphological changes specific to each original tumor. For this purpose, we analyzed the correlations between image representations and gene expression. As illustrated in Fig. 3a, another regression model of the DNN was trained to predict the gene expression of each organoid line using the image representations as the input data. In this training, the regression model was learned to decrease the difference between the predicted and actual values of gene expression datasets. For the representation of each image patch, the model predicted the gene expression levels of 14,400 genes. The scatter plot on the right-hand side of Fig. 3a indicates the accuracy of the prediction for all genes, for which the correlation coefficient was calculated as 0.78.

The accuracy of this image-based gene mapping depends on the specific features of each gene. For 10 repeated trainings, the coefficient of variation, or the ratio of the standard deviation to the mean, was averaged at 0.080 for the accuracy of each single gene. The learning result of this mapping was stable with small deviations, and the prediction accuracy of each gene reflected the specific relationships to the image representation.

The differences in accuracy could be considered to arise from the sensitivity to fluctuations, which corresponded to the intratumor heterogeneity in this representation. As illustrated in Fig. 3b, large responses to fluctuations in the image representation caused a loss of accuracy. Conversely, the stability against the representation differences within the same labeled images produced accurate predictions. Then, the gene subset closely related to the morphological changes associated with intertumor heterogeneity could be selected using the accuracy of image-to-gene mapping.

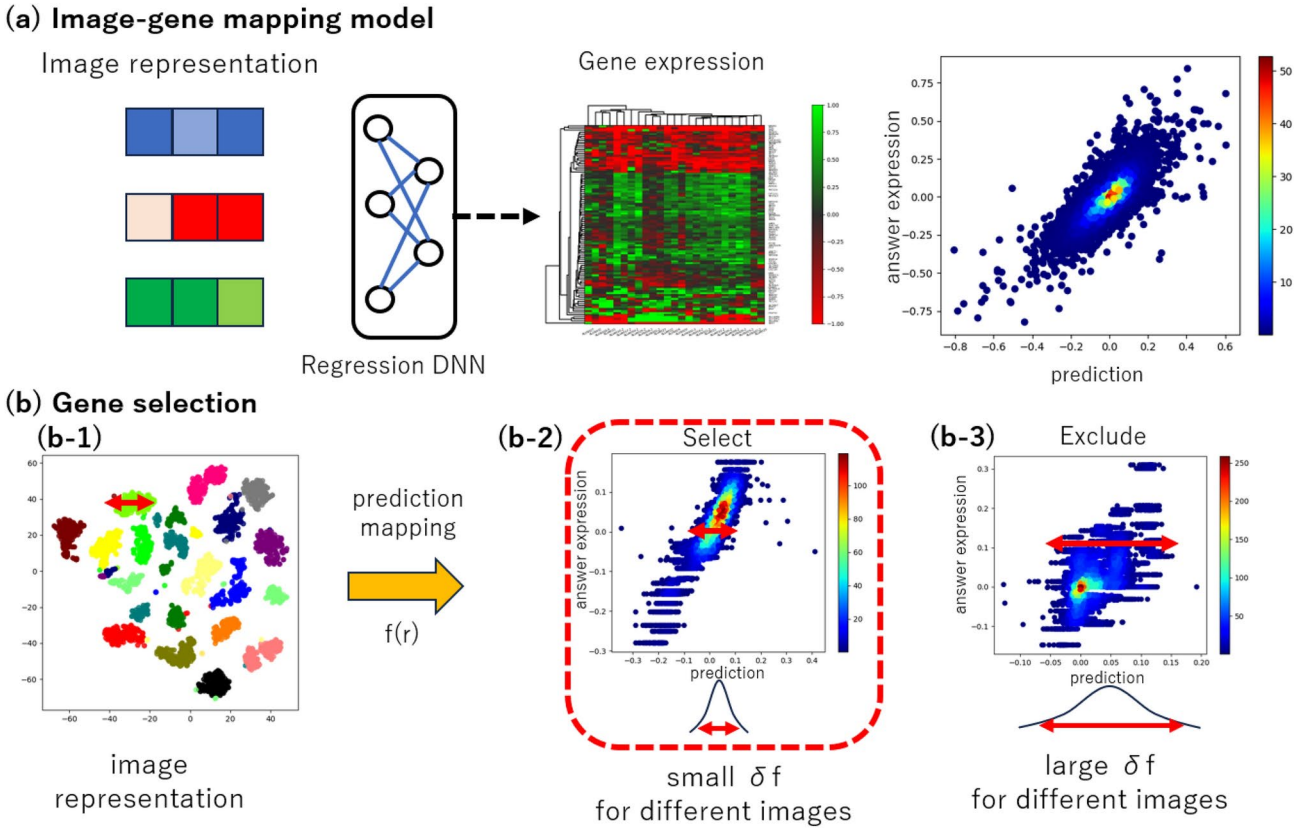

Fig. 3. Image-gene mapping and gene selection (**a**) Image-gene mapping model. The gene expression profile was predicted by another DNN model from the image representation vectors calculated by the DNN model, as shown in the previous figure. On the right-hand side of this column, a scatter plot is shown, with the answers on the horizontal axis and the predictions on the vertical axis. In this plot, the color indicates the density of the distribution. (**b**) Gene selection scheme. The gene selection flow is shown in this panel, where gene subsets with higher accuracies were selected. The image-gene mapping shown in (**a**) produced different accuracy results for each gene. The subpanels (b-2) and (b-3) show the plots for the top 10 most accurate genes and the worst 10 genes, respectively. On these panels, the predicted values on the vertical axis were plotted against the answers, which were the real expression values on the horizontal axis. For those with higher accuracy, the predictions were stable and differed only slightly from the answers (b-2). This indicated that the prediction for higher values was robust for perturbations in the image representation. Conversely, the lower groups predicted wider deviations from the answer (b-3). As shown by the red arrows overlaid on these panels, perturbations in the image representation, shown as a red arrow in (b-1), resulted in narrow and wide deviations, respectively, in (b-2) and (b-3).

Correlations with drug responses and pathway analysis

To assess the biological relevance of the gene selection obtained as described above, we analyzed the correlations of the gene selection with drug responses, which is one of the phenotypic profiles for PDOs. First, an overview of the drug response profiles is given in a plot (Fig. 4a), where the average effects and the standard deviations across the different organoid groups are plotted. In this dataset, the effects of drugs on inhibiting tumor growth were measured according to the viability of the treated organoids, and the area under the curve (AUC) for different doses was calculated (Materials and Methods). Larger and smaller values indicated ineffectiveness and strong effects, respectively. In this plot (Fig. 4a), lower values for the AUC on the left-hand side indicated strong effects with rapid decreases in viability under the treatments, and higher values distributed on the right-hand side indicated an ineffectiveness in preserving viability. On both sides, lower values of standard deviations indicated

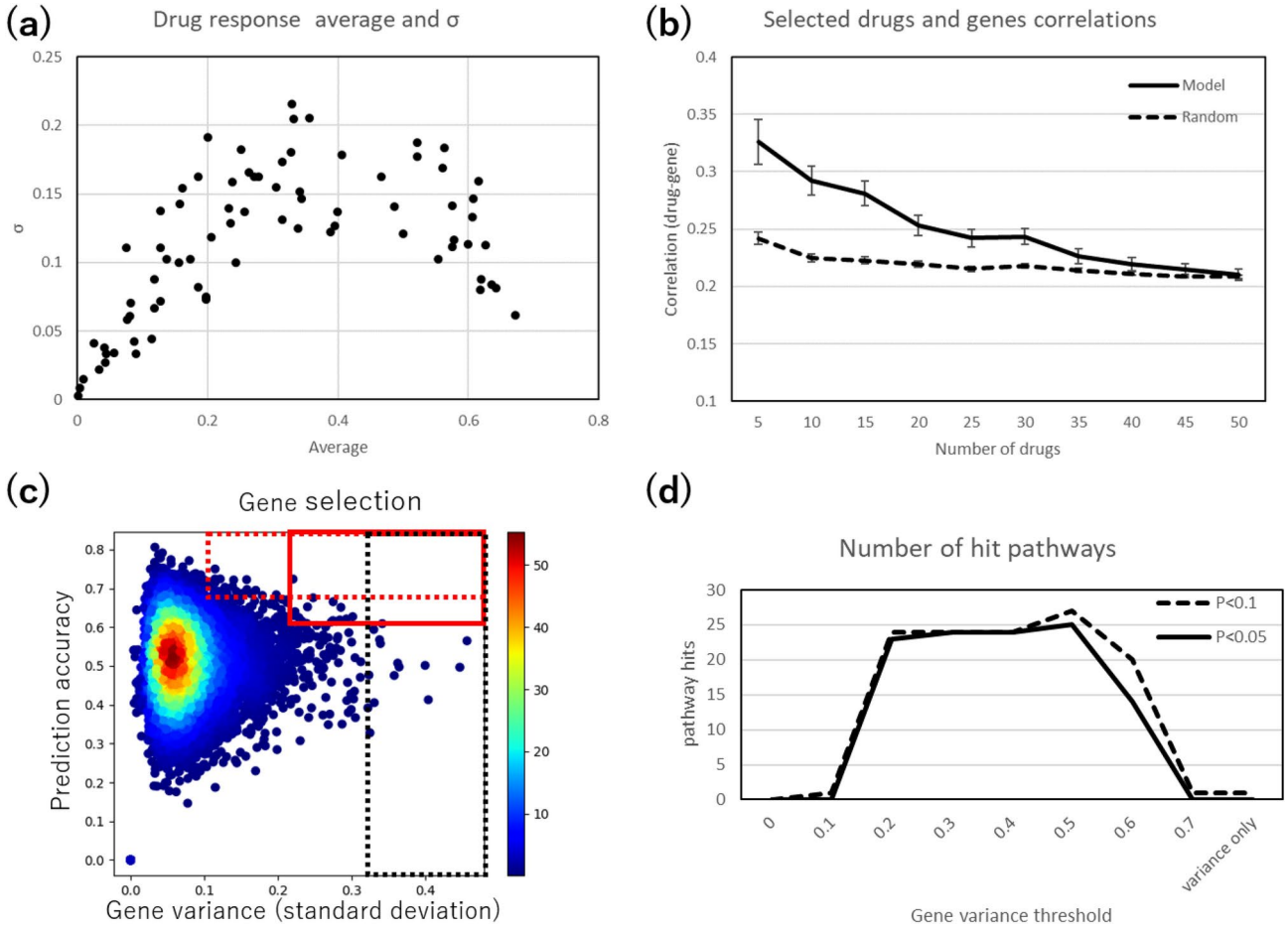

Fig. 4. Drug response and pathways (**a**) Averages and standard deviations of drug responses. As an overview of the drug response profiles of 76 anticancer agents for 18 PDO samples, the averages of drug effects to inhibit tumor growth were plotted on the vertical axis against their standard deviations across different PDOs. (**b**) Correlations between selected drugs and genes. After excluding reagents that were uniformly effective and ineffective on both sides of the plot in (**a**), drugs with relatively high standard deviations were selected, and their correlations with the gene expression levels were evaluated on the vertical axis. In this graph, 20 genes were selected in the process explained in the previous figure (Fig. 3) according to the prediction accuracy of image-gene mapping. The different numbers of selected drugs (from 5 to 50 drugs with 5 steps on the horizontal axis) had relatively high values compared with the results with randomly selected genes shown with dashed lines. (**c**) Gene selection with variance and accuracy. The variance of each gene was calculated across 18 samples for the gene expression level. The prediction accuracy for each gene was subsequently plotted against the standard deviation, which is the square root of variance, where 10,000 randomly selected genes were selected. The rectangular regions represent the selected regions with different thresholds. The red solid rectangle corresponds to a case with a moderate variance threshold. The results are compared to those of the red dotted rectangle, which has a lower threshold, and the black dotted rectangle, which has a higher threshold. (**d**) Number of hit pathways. For the selected genes, enriched pathways were identified using the DAVID database. The numbers of hit pathways were plotted against the variance thresholds on the horizontal axis. As shown in Fig. 4a, 20 genes with high accuracy scores were identified, and pathways with p values<0.1 and<0.05 were selected for analysis. At the right end, we compared the results of the case where the top 20 genes with large variances were taken regardless of the accuracy scores.

that these effects were observed uniformly across PDOs, whereas relatively high standard deviations in the middle range suggested that the effects of these drugs varied according to tumor characteristics.

To clarify the impact of gene selection on the characterization of intertumor heterogeneity, the correlations between gene expression and drug responses across different organoid lines were estimated, as shown in Fig. 4b. Here, we selected drugs according to the standard deviation values because large standard deviations in drug response suggest heterogeneous effects across patients. For the selected pairs of drugs and genes, the estimated averages of the correlations were greater than those for the randomly selected genes. Moreover, the estimated averages increased as the number of drugs decreased, where the difference in drug effectiveness between tumors was significant with a small number of selected drugs. These findings suggest the contribution of gene selection to characterizing differences in drug responses across patients.

The same selection scheme using standard deviations across patients was applied to refine the efficiency of the gene selection. Small changes in the expression levels across different organoid lines indicated that genes with small variances were inactive or common. Conversely, large variances among organoids might indicate that these genes vary largely by reflecting the features of each organoid and contrasting differences more significantly. Then, as illustrated in Fig. 4c, genes were selected based on the combination of gene expression variance and prediction accuracy.

For further validation of the biological relevance of the gene selection, enrichments in Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways, were identified using the DAVID online tool for gene ontology. Pathway enrichment analysis is a method that extracts relevant and significant gene subsets statistically from large sets of genes, based on the knowledge of interpretable pathways such as signaling, metabolic, and disease-related pathways. According to the selection criteria explained above, for the different thresholds for the variance, the genes with higher accuracies were selected. The top 20 genes were extracted, and the number of hit pathways was plotted against the variance threshold (Fig. 4d). In this plot, rapid increases were observed in the lower and higher range limits, where only a few pathways were found. Moreover, in the intermediate range of the threshold, the number of hit pathways was stably high. The significance of this result was assessed in comparison with the case at the right end of this plot, where the result was selected based on only the variance, and no pathway was hit with the top 20 genes with larger variances. This result shows that gene selection based on variance and prediction accuracy improved the statistical significance in determining biological pathways. The selected genes and pathways at the peak of the plot, 0.5 in Fig. 4d, are listed in Tables 1 and 2.

Discussion

Image-based neural network analysis reduced the complexity of the datasets and contributed to the extraction of image features of organoids in culture. Owing to intratumor heterogeneity, organoids in culture exhibit wide variations in morphology, such as size, shape and other morphological measurements, even when they are derived from the same tissue. This diversity might have augmented the ability of organoids to understand the complex dynamics of tumors. To extract features that were common to the original tissue, we introduced a neural

Gene symbol	Gene name	Accuracy	Standard-deviation
HLA-DRA	Major histocompatibility complex, class II, DR alpha	0.776	0.335
HLA-DPB1	Major histocompatibility complex, class II, DP beta 1	0.723	0.251
CD74	CD74 molecule	0.680	0.386
C3	Complement C3	0.676	0.248
IFI6	Interferon alpha inducible protein 6	0.670	0.257
HLA-DRB3	Major histocompatibility complex, class II, DR beta 3	0.665	0.275
ANXA3	Annexin A3	0.652	0.240
IGFBP3	Insulin like growth factor binding protein 3	0.651	0.286
S100P	S100 calcium binding protein P	0.630	0.299
ITGAM	Integrin subunit alpha M	0.627	0.302
APOD	Apolipoprotein D	0.627	0.338
HLA-DMA	Major histocompatibility complex, class II, DM alpha	0.625	0.229
TRIM29	Tripartite motif containing 29	0.621	0.293
VGLL1	Vestigial like family member 1	0.619	0.284
HPGD	15-hydroxyprostaglandin dehydrogenase	0.617	0.275
PLAAT3	Phospholipase A and acyltransferase 3	0.616	0.232
ARHGDIB	Rho GDP dissociation inhibitor beta	0.614	0.280
LIPH	Lipase H	0.613	0.273
VAMP8	Vesicle associated membrane protein 8	0.610	0.239
UCHL1	Ubiquitin C-terminal hydrolase L1	0.609	0.342

Table 1. Selected genes (threshold at 0.5). The selected genes for which the variance threshold was 0.5 are listed according to the accuracy scores. For 20 genes selected according to the scheme shown in Fig. 4c, the gene names and the corresponding gene symbols are listed with the prediction accuracy scores and the standard deviations across different organoids.

KEGG pathway	Gene count	P-value
Leishmaniasis	6	3.71E-08
Tuberculosis	7	6.25E-08
Staphylococcus aureus infection	6	1.13E-07
Phagosome	6	1.13E-06
Antigen processing and presentation	5	2.87E-06
Hematopoietic cell lineage	5	7.47E-06
Asthma	4	8.99E-06
Allograft rejection	4	1.68E-05
Graft-versus-host disease	4	2.28E-05
Type I diabetes mellitus	4	2.45E-05
Systemic lupus erythematosus	5	2.71E-05
Intestinal immune network for IgA production	4	3.63E-05
Autoimmune thyroid disease	4	4.61E-05
Cell adhesion molecules	5	4.75E-05
Inflammatory bowel disease	4	8.51E-05
Viral myocarditis	4	9.32E-05
Th1 and Th2 cell differentiation	4	2.39E-04
Rheumatoid arthritis	4	2.47E-04
Th17 cell differentiation	4	3.84E-04
Herpes simplex virus 1 infection	6	4.00E-04
Toxoplasmosis	4	4.17E-04
Influenza A	4	0.001467
Epstein-Barr virus infection	4	0.002367
Human T-cell leukemia virus 1 infection	4	0.003097
Transcriptional misregulation in cancer	3	0.028254
Legionellosis	2	0.075077
Arachidonic acid metabolism	2	0.081522

Table 2. Pathways. The table lists the pathways identified for the set of 20 selected genes listed in Table 1(a) by using the DAVID gene ontology online tool. The number of genes involved in each pathway and the p values are also shown.

network model that was designed to answer the labels assigned to the origins (Fig. 2). Within various features expressed in the images, the training of this model might have enhanced the factors common to the same label and sorted those that have effects in distinguishing different labels. This model was applied to visualize the intertumor differences in Fig. 2c, where images of the same PDO had similar features with distributions around their centers apart from the other images with different labels. Additionally, the same distribution showed that intratumor heterogeneity could be regarded as a perturbation. It was subsequently surmised that inter- and intratumor morphological variations could be distinguished.

The extracted image features, which are common to organoids sharing the same original tissue, were further applied in the prediction of gene expression. In image-gene mapping based on these features, it could be expected that the gene subsets responsible for the intertumor morphological changes would demonstrate high accuracy because these genes would vary according to the extracted features representing the intertumor differences. However, in the same prediction, gene factors related strongly to the intratumor differences could be excluded as elements with lower accuracy because sensitivity to the perturbation caused deviations from the expected answers and decreased the accuracy of the prediction (Fig. 3b). Additionally, because the expression of genes that are not correlated with morphological changes can be determined independently of the information of the image features, they cannot be predicted by the image information. Selection based on the prediction accuracy led to the extraction of only the gene subsets related to the features common to each original tissue, reflecting the intertumor differences.

Validation of this gene selection was performed by mapping to the other phenotypic profiles and analyzing the correlations with the drug responses. Fig. 4b shows that these selected genes had stronger correlations with drug responses than did the randomly selected genes. These findings suggest that selection functioned to leave only genes related to intertumor differences. Thus, our gene selection method could be applied to extract gene subsets that have an impact on resistance to drug stimulation in tumors.

To clarify the biological meaning of the selected genes, the extracted genes and associated pathways are shown in Tables 1 and 2. Multiple genes known as cancer-related genes are involved in the disease process of cancer through pathways such as "phagosome," "cell adhesion," and "transcriptional misregulation in cancer." For example, genes such as IGFBP3 and ITGAM are known to regulate intracellular and extracellular processes and play roles in the progression, tumor growth, invasion, and metastasis of cancer[18–23]. From the gene list, we can assume that variations in the immune system across individuals represent one of the origins of the

differences between organoid lines. For example, extracted genes such as HLA-DRA, HLA-DPB1, and CD74 are known to be related to immunological processes through pathways such as the antigen presentation pathway. Variations in the expression levels of these genes affect the immune microenvironment of tumors by mediating immune evasion and have clinical significance, especially in immunotherapy[22,23]. This evidence could explain how extracted genes work to characterize tumor differences.

The results indicate that the efficiency of pathway identification and the correlations of selected gene expression levels with drug responses were improved by means of gene selection. Moreover, these results suggest the potential to bridge the gaps between phenotypes and genotypes by identifying the factors responsible for each subtype. These findings might lead to accurate predictions in clinical practice and the identification of potential gene targets for drug development.

One of the features of our model is the sequential structure composed of the different DNNs applied for the images and the gene expression separately (Fig. 1b). Preprocessing such as the type introduced in our model (Fig. 2) was effective for datasets with large variety (Fig. 2d) to reduce their complexity (Fig. 2c). Additionally, this preprocessing allows the extraction of relevant factors as shown in Fig. 3, improving the prediction in contrast with general approaches which apply a single DNN model directly.

The flexibility of the extension is another advantage of our model. By substituting the preprocessing model with other models, we would be able to apply this model to other issues. For example, an extension for the intratumor heterogeneity, instead of intertumor heterogeneity, would be accomplished by relabeling the answers for the preprocessing in Fig. 2b. Because intratumor heterogeneity is one of the major factors affecting the drug resistance of each tumor, this extension would be necessary for further accurate analyses of drug responses in more results than those we have given (Fig. 4b). However, this extension for intratumor analysis would require more detailed information about the differences between organoids derived from the same unique tumor source and those between the local spatial regions on each organoid. Because advanced methods such as image analysis, including segmentation or attention maps, which correlate the local regions and the extracted image features[28,37], should be introduced, such extension would be a future issue.

These results promote further preclinical and clinical studies using PDOs. PDOs enable a variety of tumor experiments by allowing chemical stimulation and environmental changes in vitro. For example, as an application of our approach, relabeling PDO groups according to the experimental conditions instead of the patient number used in this study would help to extract morphological representations characterizing each group and associated gene subset. These methods would allow various experimental designs for detailed studies of mechanisms and would enable us to identify molecular targets associated with specific reactions and changes in tumors. These results suggest the usefulness of imaging data and the potential application of this model in the development of biomarkers and drugs.

Materials and methods

F-PDO data

Fukushima patient-derived tumor organoids (F-PDOs) are organoids that were established from primary solid tumor tissues obtained from surgical or biopsy samples from patients[29–33]. The minced tissues from tumors from individual patients were cultured as suspension cultures without enzyme treatment. In this study, we used F-PDO datasets for gene expression data, drug response profiles, and microscopy images of lung organoids[29–33]. F-PDOs and these data were supplied by the Fukushima Translational Research Center and the images were acquired from Evident Corp. They were measured and collected as follows.

The gene expression dataset contained profiles estimated from the expression levels of 14,400 genes for 25 different organoid lines[29–33]. Each of the organoid lines was derived from a different patient. The total amount of RNA corresponding to each gene was measured, and the measured signals were converted into primary expression ratios of samples and the human common reference RNA. The data were normalized and converted to log2 values.

Drug sensitivity was measured by the area under the curve (AUC) of growth-inhibition assays[29–33]. For the cells treated with 76 anticancer agents at different doses, cell viability was assessed using an ATP assay, and the (AUC) was calculated. The data were taken from 19 different F-PDOs.

Microscopy images of the F-PDOs in culture were acquired with a bright-field 10x objective with a color CCD. To distinguish the origins of each organoid, they were assigned labels such as RLUN10 and RLUN24. For each labeled sample, 20 images with 1,360 $\times$ 1,024 pixels were collected for analysis. In total, datasets of PDOs for 18 organoid lines with 14,400 genes and 76 drugs were included, except for the missing profiles for each label.

Model for the extraction of image representations

The details of the process for the images, which are illustrated in Figs. 1b and 2b, are as follows. For the learning datasets, image patches with 64 $\times$ 64 pixels were generated from the original images, where locations were selected at random and 100 patches were taken for each single image. The CNN consisted of a convolution layer, following layers, a flattened layer and two fully connected layers, and the outputs were ultimately processed using the Softmax function. The convolution layer had 32 $\times$ 32 $\times$ 3 dimensions, and the intermediate layers had 128 and 10 dimensions. This model was subsequently optimized to minimize the loss function, the sparse categorical cross entropy. The training was repeated for 100 epochs with a batch size of 100. This and other subsequent neural network models were implemented by using TensorFlow and Keras. t-SNE, as shown in Fig. 2, and clustering calculations, as shown in the heatmaps in Figs. 1 and 3, were performed using scikit-learn, a Python library.

Autoencoder model in comparison to the model

For the autoencoder model in Fig. 2d, the same input images were encoded with a set of layers consisting of 32 $\times$ 32 $\times$ 3 convolutional layers, a dropout layer (dropout rate=0.1), and a 2 $\times$ 2 max pooling layer. This layer set was applied 3 times and processed with the flattened layer, the layer with 1,024 nodes, and the other layer with 10 nodes. These parameters were inversely decoded. The loss function was optimized with the mean squared error. It was then trained with a batch size of 100 for 100 epochs. The middle layer with 10 nodes generated the latent representations shown in this panel, where the t-SNE images were drawn the same as those in the model case.

Regression model for gene expression prediction

The neural network illustrated in Fig. 3a consisted of five layers, which were fully connected linearly without activation functions. They had dimensions of 10, 18, 54, and 162 and an output dimension of 14400. The loss function was taken with the mean squared error. The training batch size was 50, which was randomly selected from the input gene set, and the training process was repeated for 15 epochs.

Pathway analysis

For Fig. 4d, genes were selected under the criteria of prediction accuracy and variance. For each selection, the threshold variances were varied from 0 to 0.7 σ_{max} with σ_{max}, which was the maximum value of the standard deviation. Pathway analysis was subsequently performed using the Database for Annotation, Visualization and Integrated Discovery (DAVID, http://david.abcc.ncifcrf.gov/) an online software program that can identify Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways for given gene sets.

Other calculations

Other calculations of Pearson's correlation coefficient for Fig. 3 and Fig. 3b were performed using scikit-learn.

Data availability

The programs for our deep-learning model with sample datasets is available on GitHub ((https://github.com/ktakagievd/Model4FPDOv1). The other related datasets and the program used during the present study are available from the corresponding author upon reasonable request. The F-PDOs are commercially available by contacting Fukushima Medical University (https://www.fmu.ac.jp/home/trc/en/contract-research-provision/f-pdo/).

Received: 14 May 2024; Accepted: 20 September 2024
Published online: 01 October 2024

References

1. Hanahan, D. & Weinberg, R. Hallmarks of cancer: The next generation. *Cell* **144**(5), 646–674 (2011).
2. Fouad, Y. & Aanei, C. Revisiting the hallmarks of cancer. *Am. J. Cancer Res.* **7**(5), 1016–1036 (2017).
3. Burrell, R., McGranahan, N., Bartek, J. & Swanton, C. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature* **501**, 338–345 (2013).
4. Flavahan, W., Gaskell, E. & Bernstein, B. Epigenetic plasticity and the hallmarks of cancer. *Science* **357**, 6348 (2017).
5. Yuan, Y. Spatial heterogeneity in the tumor microenvironment. *Cold Spring Harb. Perspect. Med.* **6**, a026583 (2016).
6. Marusyk, A. & Polyak, K. Tumor heterogeneity: Causes and consequences. *Biochim. Biophys. Acta* **1805**(1), 105 (2010).
7. Gerlinger, M., Rowan, A., Horswell, S., Larkin, J. & Endesfelder, D. Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N. Engl. J. Med.* **366**(10), 883–892 (2012).
8. Marusyk, A., Almendro, V. & Polyak, K. Intra-tumour heterogeneity: A looking glass for cancer?. *Nat. Rev. Cancer* **12**, 323–334 (2012).
9. Swanton, C. Intratumour heterogeneity: Evolution through space and time. *Cancer Res.* **72**(19), 4875–4882 (2012).
10. McGranahan, N. & Swanton, C. Biological and therapeutic impact of intratumor heterogeneity in cancer evolution. *Cancer Cell* **27**, 15–26 (2015).
11. Dagogo-Jack, I. & Shaw, A. Tumour heterogeneity and resistance to cancer therapies. *Nat. Rev. Clin. Oncol.* **15**, 81–94 (2018).
12. Collins, F. & Varmus, H. A new initiative on precision medicine. *N. Engl. J. Med.* **372**, 793–795 (2015).
13. Greaves, M. & Maley, C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012).
14. Ritchie, M., Holzinger, E., Li, R., Pendergrass, S. & Kim, D. Methods of integrating data to uncover genotype-phenotype interactions. *Nat. Rev. Genet.* **16**, 85–97 (2015).
15. Aronson, S. & Rehm, H. Building the foundation for genomics in precision medicine. *Nature* **526**, 336–342 (2015).
16. Young, A., Benonisdottir, S., Przeworski, M. & Kong, A. Deconstructing the sources of genotype-phenotype associations in humans. *Science* **365**(6460), 1396–1400 (2019).
17. Kamat, M., Blackshaw, J., Young, R., Surendran, P. & Burgess, S. Phenoscanner v2: An expanded tool for searching human genotype-phenotype associations. *Bioinformatics* **35**(22), 4851–4853 (2019).
18. Yu, H. & Rohan, T. Role of the insulin-like growth factor family in cancer development and progression. *J. Natl Cancer Inst.* **92**, 1472–1489 (2000).
19. Hamidi, H. & Ivaska, J. Every step of the way: Integrins in cancer progression and metastasis. *Nat. Rev. Cancer* **18**, 533–548 (2018).
20. Schmid, M., Khan, S., Kaneda, P., Pathria, M. M. & Shepard, R. Integrin cd11b activation drives anti-tumor innate immunity. *Nat. Commun.* **9**, 5379 (2018).
21. Song, Q., Hawkins, G., Wudel, L., Chou, P.-C. & Forbes, E. Dissecting intratumoral myeloid cell plasticity by single cell RNA?. *Seq. Cancer Med.* **8**, 3072–3085 (2019).
22. Morad, G., Helmink, B., Sharma, P. & Wargo, J. Hallmarks of response, resistance, and toxicity to immune checkpoint blockade. *Cell* **184**, 5309–5337 (2021).
23. Schaafsma, E., Fugle, C., Wang, X. & Cheng, C. Pan-cancer association of hla gene expression with cancer prognosis and immunotherapy efficacy. *Br. J. Cancer* **125**, 422–32 (2021).
24. Clevers, H. Modeling development and disease with organoids. *Cell* **7**(16), 1586–1597 (2016).
25. Fatehullah, A., Tan, S. & Barker, N. Organoids as an in vitro model of human development and disease. *Nat. Cell Biol.* **18**, 246–254 (2016).
26. Drost, J. & Clevers, H. Organoids in cancer research. *Nat. Rev. Cancer* **18**, 407–418 (2018).

27. Sachs, N. *et al.* Long-term expanding human airway organoids for disease modeling. *EMBO J.* **38**, e100300 (2019).
28. Larsen, B. *et al.* A pan-cancer organoid platform for precision medicine. *Cell Rep.* **36**, 109429 (2021).
29. Higa, A. *et al.* Evaluation system for arrhythmogenic potential of drugs using human-induced pluripotent stem cell-derived cardiomyocytes and gene expression analysis. *J. Toxicol. Sci.* **42**, 755–761 (2017).
30. Tamura, H. *et al.* Evaluation of anticancer agents using patient-derived tumor organoids characteristically similar to source tissues. *Oncol. Rep.* **40**, 635–646 (2018).
31. Takahashi, N. *et al.* An in vitro system for evaluating molecular targeted drugs using lung patient-derived tumor organoids. *Cells* **8**, 4812019 (2019).
32. Takahashi, N. *et al.* Construction of in vitro patient?derived tumor models to evaluate anticancer agents and cancer immunotherapy. *Oncol. Lett.* **21**(5), 1792–1074 (2021).
33. Higa, A. *et al.* High-throughput in vitro assay using patient-derived tumor organoids. *J. Vis. Exp.* **14**, 172 (2021).
34. Rios, A. & Clevers, H. Imaging organoids: A bright future ahead. *Nat. Methods* **15**, 24–26 (2018).
35. Borten, M., Bajikar, S., Sasaki, N., Clevers, H. & Janes, K. Automated brightfield morphometry of 3d organoid populations by organoseg. *Nat. Methods* **15**, 23 (2018).
36. Karolak, A., Poonja, S. & Rejniak, K. Morphophenotypic classification of tumor organoids as an indicator of drug exposure and penetration potential. *PLoS Comput. Biol.* **15**(7), e1007214 (2019).
37. Matthews, J. *et al.* Organoid: A versatile deep learning platform for tracking and analysis of single-organoid dynamics. *PLoS Comput. Biol.* **18**(11), e1010584 (2022).
38. Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **46**, 389–422 (2002).
39. Sawyers, C. The cancer biomarker problem. *Nature* **452**, 548–552 (2008).
40. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**(7553), 436–44 (2015).
41. Hinton, G. & Salakhutdinov, R. Reducing the dimensionality of data with neural networks. *Science* **313**(5786), 504–507 (2006).
42. Hinton, S., Osindero, G. & Teh, Y. A fast learning algorithm for deep belief nets. *Neural Computat.* **18**(7), 1527–1554 (2006).
43. Shorten, C. & Khoshgoftaar, T. A survey on image data augmentation for deep learning. *J. Big Data* **6**, 60 (2019).
44. Chandrashekar, G. & Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **40**(1), 16–28 (2014).
45. Kraus, O., Ba, J. & Frey, B. Classifying and segmenting microscopy images with deep multiple instance learning. *Bioinformatics* **32**(12), i52–i59 (2016).
46. Kassis, T., Hernandez-Gordillo, V., Langer, R. & Griffith, L. Orgaquant: Human intestinal organoid localization and quantification using deep convolutional neural network. *Sci. Rep.* **9**, 12479 (2019).
47. Chaudhary, K., Poirion, O., Lu, L. & Garmire, L. Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clin. Cancer Res.* **24**(6), 1248–59 (2018).
48. Li, Y., Wu, F. & Ngom, A. A review on machine learning principles for multi-view biological data integration. *Brief. Bioinform.* **19**(2), 325–340 (2018).
49. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M. & Blaschke, T. The rise of deep learning in drug discovery. *Drug Discov. Today* **23**(6), 1241–1250 (2018).
50. Esteva, A. *et al.* A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).
51. Calon, A. *et al.* Stromal gene expression defines poor-prognosis subtypes in colorectal cancer. *Nat. Genet.* **47**, 320–329 (2015).
52. Vamathevan, J. *et al.* Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18**(6), 463–477 (2019).
53. Chang, Y. *et al.* Cancer drug response profile scan (cdrscan): A deep learning model that predicts drug effectiveness from cancer genomic signature. *Sci. Rep.* **8**, 8857 (2018).

Author contributions

KT, MT, and KG conceived the research. KT analyzed the results and wrote the manuscript. MT and GH provided the material and data for F-PDO. KG acquired the microscopy images. All the authors reviewed the manuscript and contributed equally to the writing of the final paper.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.