MFmap: A semi-supervised generative model matching cell lines to tumours and cancer subtypes

Xiaoxiao Zhang^{1,2}, Maik Kschischo^{1*}

1 Department of Mathematics and Technology, RheinAhrCampus, University of Applied Sciences Koblenz, 53424 Remagen, Germany

2 Department of Informatics, Technical University of Munich, 81675 Munich, Germany

* kschischo@rheinahrcampus.de

Abstract

Translating *in vitro* results from experiments with cancer cell lines to clinical applications requires the selection of appropriate cell line models. Here we present MFmap (model fidelity map), a machine learning model to simultaneously predict the cancer subtype of a cell line and its similarity to an individual tumour sample. The MFmap is a semi-supervised generative model, which compresses high dimensional gene expression, copy number variation and mutation data into cancer subtype informed low dimensional latent representations. The accuracy (test set F_1 score $> 90\%$) of the MFmap subtype prediction is validated in ten different cancer datasets. We use breast cancer and glioblastoma cohorts as examples to show how subtype specific drug sensitivity can be translated to individual tumour samples. The low dimensional latent representations extracted by MFmap explain known and novel subtype specific features and enable the analysis of cell-state transformations between different subtypes. From a methodological perspective, we report that MFmap is a semi-supervised method which simultaneously achieves good generative and predictive performance and thus opens opportunities in other areas of computational biology.

Author summary

Cancer researchers perform experiments with cell lines to better understand the biology of cancer and to develop new anti-cancer treatments. A prerequisite to translate promising results from these *in vitro* experiments to clinical applications is to use the most appropriate cell line for a given tumour or cancer subtype. We present MFmap (model fidelity map), a deep learning technique to integrate cancer genomic data from patients with cell line data. The MFmap neural network compresses complex genomic features from thousands of genes into a small set of features called latent representations. This makes cell line and tumour data comparable and allows cancer researchers to select the best cell line which closely resembles a specific type of tumours or even an individual tumour. By classifying cancer cell lines into subtypes, MFmap offers a new possibility to predict the effect of therapeutic compounds in a particular tumour subtype. For the example of an aggressive brain tumour we demonstrate that MFmap can be used to study cell-state transformations during the disease course. In addition, MFmap is a promising machine learning method with potential applications in many other areas of biology and medicine.

Introduction

Tumour-derived cell lines are important model systems for developing new anti-cancer treatments and for understanding cancer biology [1–3]. They are comparably cost efficient, easy to handle under laboratory conditions and do not inflict ethical issues arising in research involving human or animal subjects. Yet, promising cell line experiments are rarely translated to clinical applications. In some cases, there are remarkable differences between cell lines and the primary tumours they were derived from [2–4]. This is also the reason why the assignment of clinically informative tumour subtypes to cell line models [3–5] is not a straightforward task.

To narrow the gap between preclinical findings and tumour treatment, it is necessary to select appropriate cell line models for a given tumour sample or a given cancer subtype. Several attempts to evaluate similarities and differences between cell lines and bulk tumours have focused on associations between corresponding data modalities including mutation, copy number, gene expression and methylation [6–12]. An important data resource comes from collaborative projects like NCI-60 [13] and the Cancer Cell Line Encyclopaedia (CCLE) [5, 14], who have generated large-scale pharmacogenomics data from patient-derived cell lines across organs. Other efforts like Sanger Genomics of Drug Sensitivity in Cancer (GDSC) [15], Connectivity Map (CMAP) [16], the Cancer Therapeutics Response Portal (CTRP v1 and CTRP v2) [17, 18] further expanded the datasets. On the other hand, The Cancer Genome Atlas (TCGA) [19] and the International Cancer Genome Consortium (ICGC) [20] systematically characterised molecular profiles of thousands of tumours. These complementary data resources are valuable for understanding the complexity of cancer biology and connecting *in vitro* pharmacogenomic profiles to patient molecular characteristics, potentially informing anti-cancer treatment strategies.

Integrative analyses considering multiple data types of both cell lines and bulk tumours are still challenging and new analysis concepts tailored towards specific questions are an ongoing research topic. For instance, Collector [21] preselects the most frequent genomic alterations and defines cancer subtypes based on a sequence of these alterations. Although such a preselection of genomic alterations integrates prior knowledge about cancer mutational patterns, it neglects complementary information contained in other data types. Furthermore, Collector relies on a binary matrix of genomic alterations. This matrix is very sparse, since samples harbouring the same alterations are very rare. Therefore, the statistical power to detect appropriate cell lines for tumours might be limited.

A recent study [22] highlighted that independent classifiers based on different data types to predict cell line identity often yield inconsistent results. For example, predictions based on the mutation spectrum and oncogenic mutations can be contradictory, although both features are derived from mutation data. Complementary information from different data sources is integrated by the MAGNETIC-framework [23] into gene modules. Gene set enrichment analysis (GSEA) is then used to interpret these modules as pathways. MAGNETIC is indeed a powerful technique for integrating multiple molecular datasets and prior knowledge, but it does not conclude to what extent a cell line is suitable as a tumour model. The maui framework assigns cancer subtype labels to cell lines by extracting relevant features from multiple data types using a variational autoencoder (VAE) [24]. However, most of the maui embedded features are weakly associated with subtype labels and are therefore difficult to interpret.

Here, we propose MFmap, a new semi-supervised VAE architecture and objective function which combines good classification accuracy with good generative performance. We exploit these properties to derive subtype informed low dimensional representations for both cell lines and bulk tumours from high dimensional multi-omics data including gene expression, mutation and copy number variation. The latent representations can

then be used to assess the similarity between a cell line and a tumour. We provide cell line by tumour dissimilarity matrices for CCLE and TCGA for the ten different cancer types listed in Table 1. In addition, MFmap predicts cancer subtype labels for cell lines. We demonstrate, how these predicted cancer subtypes can be used to transfer information from cell-line-based drug sensitivity screens to patient cohorts. We also show, that the latent representations learnt by MFmap are biologically interpretable. Finally, we illustrate how the generative nature of the MFmap model can be exploited for studying subtype transformations during cancer progression. At http://h2926513.stratoserver.net:3838/MFmap_shiny/ we provide a resource enabling researchers to select the most relevant cell line for a cancer patient.

Table 1. The sample size of TCGA and CCLE data used for training and testing MFmap.

TCGA code	study name	number of subtypes	TCGA sample size	CCLE sample size
BRCA	Breast invasive carcinoma	4	484	51
COADREAD	Colon adenocarcinoma	4	414	54
ESCA	Esophageal carcinoma	2	169	27
HNSC	Head and neck squamous cell carcinoma	4	278	29
LUAD	Lung adenocarcinoma	3	227	70
LUSC	Lung squamous cell carcinoma	4	178	22
PAAD	Pancreatic adenocarcinoma	2	149	40
SKCM	Skin cutaneous melanoma	3	260	49
UCEC	Uterine corpus endometrial carcinoma	3	234	28
GBMLGG	Glioblastoma multiforme and lower grade glioma	7	621	55

Methods

Matching cell lines and tumours as a semi-supervised learning problem

MFmap is a semi-supervised deep neural network which integrates gene expression, copy number variation (CNV) and somatic mutation data with subtype classification. Each tumour sample t consists of a pair of $(\mathbf{x}_t, y_t)$, where $\mathbf{x}_t \in \mathbb{R}^D$ denotes the high dimensional molecular features and $y_t \in \{1, \dots, h\}$ is the cancer subtype label. For a cell line c , the cancer subtype is unknown and only the molecular features $\mathbf{x}_c$ are available. The index c or t will be suppressed, whenever we refer to a single observation. The MFmap neural network is trained in a semi-supervised manner using both cell line data $\mathcal{D}_{cl}^{train} = \{\mathbf{x}_c\}_{c=1}^{C_{train}}$ and tumour data $\mathcal{D}_{tu}^{train} = \{(\mathbf{x}_t, y_t)\}_{t=1}^{T_{train}}$. Here, we used cell line data from CCLE and tumour data from TCGA.

One aim of MFmap is to use semi-supervised classification to infer the cancer subtype y_c of a cell line c . A second aim is to assess the similarity between a cell line and a tumour. Instead of comparing the high dimensional molecular features $\mathbf{x}_t$ and $\mathbf{x}_c$ directly, we first encode them into low dimensional latent representations $\mathbf{z}$ (see next section for details). Then, the similarity of a tumour sample t and a cell line c is

measured as the cosine coefficient between the corresponding latent representation vectors $\mathbf{z}_t$ and $\mathbf{z}_c$. We will also show that these latent representations $\mathbf{z}$ carry interpretable biological information.

The molecular data $\mathbf{x} = (\mathbf{x}_{DNA}, \mathbf{x}_{RNA})$ consist of gene expression profiles $\mathbf{x}_{RNA}$ and network smoothed mutation and CNV profiles $\mathbf{x}_{DNA}$. We will refer to these two parts as RNA and DNA view, respectively. The DNA view is obtained from the original binary mutation and CNV matrices (Fig 1(A)), which indicate the occurrence of a mutation or CNV event targeting a gene in a given tumour sample or cell line. These very sparse matrices are first projected onto an annotated cancer network [25]. By using a network diffusion algorithm [26], a mutation or CNV signal hitting a single gene is propagated to neighbouring nodes in the network, thereby enriching the mutation or CNV data by cancer network information. All molecular features were translated and scaled to the interval between zero and one.

Fig 1. Overview of MFmap. (A) In a preprocessing step, mutation and CNV profiles are transformed to network smoothed DNA profiles. The original mutation and CNV data are represented as a binary matrix indicating the presence/absence of a DNA alteration in a given tumour sample or cell line. This sparse matrix is projected onto a cancer reference network (CRN) [25] and a network diffusion algorithm propagates this information to network neighbours, resulting in a dense DNA mutation or CNV matrix (DNA features). (B) The smoothed DNA features (DNA view) combined with gene expression data (RNA view) form the input of MFmap. The neural network architecture of MFmap has three components: encoder, decoder and classifier, encoded by different colours. The encoder maps sample features to a distribution $q(\mathbf{z}|\mathbf{x})$ for the latent representation $\mathbf{z}$ with mean value $\boldsymbol{\mu}(\mathbf{x})$ and covariance $\boldsymbol{\sigma}^2(\mathbf{x})$. The classifier outputs a molecular subtype probability $p(y|\mathbf{z})$ and the decoder models a density $p(\mathbf{x}|\mathbf{z})$ for the reconstruction of the DNA and RNA views. During semi-supervised training, the molecular subtypes of tumour samples are used. (C) For visualisation, the latent representations of bulk tumour samples are used to generate a reference map. Cell lines are then projected to the reference map. The colour coding of individual samples or cell lines (dots) indicates the tumour subtype or the predicted subtype, respectively. The density of the tumour samples is indicated by background contour lines coloured according to the subtypes.

Specification of MFmap as a semi-supervised generative model

The MFmap neural network (Fig 1(B)) is a new variant of a semi-supervised VAE [27]. The observable data are considered to be drawn from the probability distributions $p(\mathbf{x}, y)$ for tumour samples and $p(\mathbf{x})$ for cell lines. These distributions are modelled as marginals over the latent variable $\mathbf{z} = (z_1, \dots, z_d)^T \in \mathbb{R}^h$, such that

$$p(\mathbf{x}, y) = \int p(\mathbf{x}, y, \mathbf{z}) d\mathbf{z}, \quad p(\mathbf{x}) = \sum_{y=1}^h p(\mathbf{x}, y). \quad (1)$$

To facilitate biological interpretation of the latent representations, we set the dimension d of the latent space equal to the number of cancer subtypes h . In other applications of the MFmap model, one could also consider d as a tuneable hyper-parameter.

For the generative model, we assume $\mathbf{x}$ and y to be conditionally independent given the latent variable $\mathbf{z}$. Accordingly, the joint distribution can be factorised as

$$p(\mathbf{x}, y, \mathbf{z}) = p(\mathbf{x}|\mathbf{z}) p(y|\mathbf{z}) p(\mathbf{z}). \quad (2)$$

These distributions are specified as

$$p(\mathbf{z}) = \mathcal{N}(\mathbf{z}|0, \mathbf{I}) \quad (3a)$$

$$p(y|\mathbf{z}) = \text{Cat}(y|\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{z})) \quad (3b)$$

$$p(\mathbf{x}|\mathbf{z}) = \mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}). \quad (3c)$$

Here, $p(\mathbf{z})$ is the prior distribution for the latent representation vector. We denote the Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ by $\mathcal{N}(\cdot|\boldsymbol{\mu}, \Sigma)$. The parameter $\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{z})$ of the categorical distribution $p(y|\mathbf{z})$ depends on the latent representation $\mathbf{z}$. For the decoder $p(\mathbf{x}|\mathbf{z})$ one can choose a suitable distribution $\mathbf{f}_{\boldsymbol{\theta}}$ with parameters depending on the latent representations $\mathbf{z}$ [27]. The functions $\mathbf{z} \mapsto \boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{z})$ and $\mathbf{z} \mapsto \mathbf{f}_{\boldsymbol{\theta}}(\cdot|\mathbf{z})$ are represented as neural networks. The parameters of these decoder networks are jointly denoted as $\boldsymbol{\theta}$.

For the mfMAP model we initially used a Gaussian distribution $\mathbf{f}_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})$ to model the outputs. However, we found that rescaling the molecular features $\mathbf{x}$ to the interval $[0, 1]$ and using a Bernoulli distribution for $\mathbf{f}_{\boldsymbol{\theta}}$ improved the semi-supervised classification accuracy (see Results section). Then, each single output of the decoder neural network $\mathbf{z} \mapsto \mathbf{f}_{\boldsymbol{\theta}}(\cdot|\mathbf{z})$ can be interpreted as the probability, that the corresponding molecular feature is active or not. For instance, for the i -th component $(\mathbf{x}_{RNA})_i$ of the RNA-view, the corresponding output can be regarded as the probability that the i -th gene is expressed.

Posterior inference, i.e. the evaluation of $p(y, \mathbf{z}|\mathbf{x})$ using Bayes theorem, is often intractable, because the marginal likelihood $p(\mathbf{x})$ in Eq (1) requires integrating over $\mathbf{z}$. Therefore, a variational distribution $q(y, \mathbf{z}|\mathbf{x})$ is introduced to approximate the true posterior [24, 27]. We assume that the variational distribution reflects the conditional independence $\mathbf{x} \perp y|\mathbf{z}$ of the generative model in Eq (2). This implies

$$q(\mathbf{x}, y|\mathbf{z}) = q(\mathbf{x}|\mathbf{z}) q(y|\mathbf{z}). \quad (4)$$

For consistency we assume that $q(y|\mathbf{z})$ in Eq (4) is identical to $p(y|\mathbf{z})$ in Eq (3b) and is represented by the same neural network mapping $\mathbf{z}$ to the categorical parameter $\boldsymbol{\pi}_{\boldsymbol{\theta}}(\mathbf{z})$. For the variational distribution $q(\mathbf{z}|\mathbf{x})$ we choose a Gaussian

$$q_{\phi}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma})) \quad \text{with} \quad (\boldsymbol{\mu}(\mathbf{x}), \log \boldsymbol{\sigma}(\mathbf{x})) = \mathbf{g}_{\phi}(\mathbf{x}) \quad (5)$$

with parameters $\boldsymbol{\mu}(\mathbf{x})$ and $\boldsymbol{\sigma}(\mathbf{x})$. The parameters are represented by the encoder neural network $\mathbf{g}_{\phi}$, which is itself parametrised by ϕ . The overall architecture of MFmap (Fig 1(B)) is thus formed by three neural networks, the encoder Eq (5), the classifier Eq (3b) and the decoder Eq (3c).

Training of MFmap using a semi-supervised loss function

Variational inference involves maximising an evidence lower bound (ELBO) to the log-likelihood of the observational data [24, 27]. For a single cell line sample $\mathbf{x}_c \in \mathcal{D}_{cl}$ one can derive a lower bound to the log-likelihood

$$\log p(\mathbf{x}_c) = \log \left(\sum_y \int p(\mathbf{x}_c, y, \mathbf{z}) d\mathbf{z} \right) \geq \mathfrak{L}(\mathbf{x}_c), \quad (6)$$

which is identical to the ELBO of the basic VAE [24] for unsupervised learning

$$\mathfrak{L}(\mathbf{x}) = E_{q(\mathbf{z}|\mathbf{x})} [\log p(\mathbf{x}|\mathbf{z})] - D_{KL}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})), \quad (7)$$

consisting of a reconstruction loss term and a Kullback-Leibler (KL) divergence term. For a single labelled tumour sample $(\mathbf{x}_t, y_t) \in \mathcal{D}_{tu}$ we have for the log-likelihood

$$\log p(\mathbf{x}_t, y_t) = \log \left(\int p(\mathbf{x}_t, y_t, \mathbf{z}) d\mathbf{z} \right) \geq \mathcal{L}_{tu}(\mathbf{x}_t, y_t), \quad (8)$$

where the ELBO for labelled examples reads

$$\mathcal{L}_{tu}(\mathbf{x}, y) = \mathcal{L}(\mathbf{x}) + E_{q(\mathbf{z}|\mathbf{x})} [\log p(y|\mathbf{z})]. \quad (9)$$

To derive this ELBO (see Supporting Information), we exploited the conditional independence assumption $\mathbf{x} \perp y|\mathbf{z}$ for both the generative model (Eq (2)) and the inference model (Eq (4)). The additional term in Eq (9) in comparison to Eq (7) can be interpreted as a classification loss. Given a tumour sample $(\mathbf{x}_t, y_t)$, the probability for the cancer subtype label $p(y_t|\mathbf{z})$ is a function of $\mathbf{z}$, which is inferred from $q(\mathbf{z}|\mathbf{x}_t)$. This distribution is in turn determined by the molecular feature vector $\mathbf{x}_t$.

We found empirically that the semi-supervised classification accuracy during training was relatively poor when using these exact negative ELBOs as loss functions. This is in line with previous findings that achieving both good semi-supervised classification accuracy and good generative performance is often difficult in VAEs [28] or other generative models [29]. Motivated by the work from [30], we added the negative entropy $\mathcal{H}[p(y|\mathbf{z})]$ of the distribution $p(y|\mathbf{z})$ to the unsupervised ELBO $\mathcal{L}$ in Eq (7) and to the supervised ELBO $\mathcal{L}_{tu}$ in Eq (9). In summary, the MFmap loss functions for the unlabelled cell line and the labelled tumour data are respectively given by

$$\begin{aligned} \mathcal{U}(\mathbf{x}) &= -\mathcal{L}(\mathbf{x}) + \mathcal{H}[p(y|\mathbf{z})] \\ &= -E_{q(\mathbf{z}|\mathbf{x})} [p(\mathbf{x}|\mathbf{z})] + D_{KL}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) + \mathcal{H}[p(y|\mathbf{z})] \end{aligned} \quad (10a)$$

$$\begin{aligned} \mathcal{S}(\mathbf{x}, y) &= -\mathcal{L}_{tu}(\mathbf{x}) + \mathcal{H}[p(y|\mathbf{z})] \\ &= -E_{q(\mathbf{z}|\mathbf{x})} [p(\mathbf{x}|\mathbf{z})] + D_{KL}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) + \mathcal{H}[p(y|\mathbf{z})] - E_{q(\mathbf{z}|\mathbf{x})} [\log p(y|\mathbf{z})]. \end{aligned} \quad (10b)$$

This entropy regularisation encourages the classification boundaries to be located in low sample density regions [30] in the latent space, which improves the generalisation performance of the model. As shown below (see Results section), the semi-supervised classification accuracy was very convincing, when using this entropy regularisation.

During training, mini-batches $b = 1, \dots, B$ from the cell line $\mathcal{D}_{cl}^{(b)} \subset \mathcal{D}_{cl}^{Train}$ and tumour data $\mathcal{D}_{tu}^{(b)} \subset \mathcal{D}_{tu}^{Train}$ are used to minimise

$$\sum_{\mathbf{x}_c \in \mathcal{D}_{cl}^{(b)}} \mathcal{U}(\mathbf{x}_c) + \sum_{(\mathbf{x}_t, y_t) \in \mathcal{D}_{tu}^{(b)}} \mathcal{S}(\mathbf{x}_t, y_t) \quad (11)$$

over different epochs. To check whether all terms in the MFmap loss function in Eq (10) can be jointly optimised, we recorded the values of each term in each training epoch and calculated their pair-wise correlations. The reconstruction loss $-E_{q(\mathbf{z}|\mathbf{x})} [p(\mathbf{x}|\mathbf{z})]$, the KL-divergence $D_{KL}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))$, the entropy $\mathcal{H}[p(y|\mathbf{z})]$ and the classification loss $-E_{q(\mathbf{z}|\mathbf{x})} [\log p(y|\mathbf{z})]$ are highly correlated (Fig 2), what suggests that they are optimised simultaneously.

Fig 2. Joint optimisation of the reconstruction loss, the KL divergence, entropy and the classification loss with the MFmap loss function. The plot shows the pairwise correlation of different terms in the MFmap loss function Eq (10) during different training epochs.

Visualisation of individual samples

The MFmap latent representation $\mathbf{z}$ can be used to visualise and organise the associations of individual tumour samples and cell lines (Fig 1(C)). Inspired by the visualisation concept of Onco-GPS (OncoGenic Positioning System) [31], we used the tumour samples with known subtypes to generate a reference map for the cancer subtypes. In this reference map, the components $z_1, \dots, z_h$ of the latent representation are presented as a graph with h corner points in a plane. The location of these corner points is determined by multidimensional scaling and is chosen so as to reflect the distances in the h -dimensional latent space as good as possible (see Supporting Information for details). An individual tumour sample can now be visualised as a point located in the area between the corner points. The location of such a point is given by a superposition of the corner positions weighted by the latent representation magnitudes of individual samples. In addition, the subtypes of the tumour samples are colour coded. The contour lines and the background colour shading represent the sample density in the region.

Once the reference map is established, individual cell lines can be projected to this map, where the colour of each dot encodes the subtype *predicted* by the MFmap classifier. This projection is based on the latent representation values of the cell line samples. Since our aim is to analyse the fidelity of a cell line as an oncological model for a given tumour or a cancer subtype, we name our framework the model fidelity map (MFmap).

Results

Evaluating the MFmap classification and generative performance

A direct evaluation of the MFmap subtype prediction for cell lines is impossible because there are no ground truth labels available. However, the classification accuracy on an unseen test dataset of bulk tumours provides an indirect evaluation of the subtype prediction performance. In Table 2 we used 20% of the tumour samples as independent test set and evaluated the classification performance using four multi-class classification metrics: overall accuracy, weighted precision, weighted recall, and weighted F_1 score. Similar results can be obtained, when 10% of the tumour samples are used for testing (data not shown).

The good classification results for GBMLGG are intriguing, because the G-CIMP-High, G-CIMP-Low and LGm6-GBM subtypes were derived from methylation data [32], which were not used to train MFmap. This indicates that MFmap is able to extract DNA and RNA patterns reflecting features originally derived from different methylation status.

In addition, we tested how well the MFmap autoencoder part reconstructs the molecular features $\mathbf{x}$. To this end, we first sampled a latent representations from the encoder $q(\mathbf{z}|\mathbf{x})$ for a given input $\mathbf{x}$ from the real data. Then, we correlated these original molecular features with the output sampled from the decoder distribution $p(\mathbf{x}|\mathbf{z})$. The histogram of Pearson correlation coefficients in Fig 3 shows a high input-output correlation for most molecular features for three exemplary cancer types: breast invasive carcinoma (BRCA), colorectal adenocarcinoma (COADREAD) and glioblastoma multiforme and lower grade glioma (GBMLGG). Taken together, MFmap can combine very good classification accuracy with good generative performance.

Table 2. MFmap subtype classification performance estimated for unseen tumour samples. Here, 20% of the bulk tumour data were randomly selected as an independent test set.

accuracy	precision	recall	F_1 score	organ
0.97	0.97	0.97	0.97	BRCA
0.96	0.96	0.96	0.96	COADREAD
1.00	1.00	1.00	1.00	ESCA
0.99	0.99	0.99	0.99	GBMLGG
0.91	0.92	0.91	0.91	HNSC
0.96	0.96	0.96	0.96	LUAD
0.94	0.95	0.94	0.94	LUSC
0.97	0.97	0.97	0.97	PAAD
1.00	1.00	1.00	1.00	SKCM
0.96	0.96	0.96	0.96	UCEC

Fig 3. The generative performance of MFmap. The histogram shows sample-wise correlation coefficients between input features (DNA and RNA views) and reconstructed features output by the MFmap decoder.

Selecting the optimal cell line for a given tumour

The heatmaps in Fig 4 represent pairwise cell line by tumour dissimilarity matrices for three exemplary cancer types BRCA, COADREAD and GBMLGG. In addition, the subtypes of bulk tumours annotated from [32–34] and the subtypes of cell lines predicted by the MFmap classifier are displayed. For a better visualisation, cell lines and tumours are clustered based on their pairwise cosine dissimilarity scores. The

Fig 4. Pairwise dissimilarity between CCLE cell lines and TCGA bulk tumours. The colour coding in the heatmaps indicates the pairwise dissimilarity which was obtained from the latent representations of cell lines and tumours for the three exemplary cancer types (A) breast invasive carcinoma (BRCA), (B) colorectal adenocarcinoma (COADREAD) and (C) glioblastoma multiforme and lower grade glioma (GBMLGG). Tumours (columns) and cell lines (rows) were clustered according to the dissimilarity score, which ranges from 0 (very similar) to 2 (very dissimilar). The subtype classification of each cell line was predicted from the classification layer of the MFmap neural network. The tables display the sample size for the different subtypes or predicted subtypes.

similarity of a cell line c to a tumour t is defined as the cosine of the angle between their latent representations $\mathbf{z}_c$ and $\mathbf{z}_t$. Accordingly, the dissimilarity between c and t is defined as $d(c, t) = 1 - \mathbf{z}_c \cdot \mathbf{z}_t / \|\mathbf{z}_c\| \|\mathbf{z}_t\|$. A dissimilarity of $d(c, t) = 0$ indicates perfect alignment between the latent representations of the cell line and the tumour, whereas a dissimilarity $d(c, t) = 1$ indicates orthogonal latent representations. The highest dissimilarity of $d(c, t) = 2$ would be achieved for antipodal latent vectors. Based on this dissimilarity matrix, researchers can select the best cell lines for a given tumour or a given tumour subtype. And, vice versa, the relevance of promising experimental results observed *in vitro* can be checked by selecting a subset of tumours most likely resembling the cell line characteristics. The pairwise dissimilarity matrices between TCGA bulk tumours and CCLE cell lines and cell line subtype predictions for all tumour types listed in Table 1 are provided on our website

(http://h2926513.stratoserver.net:3838/MFmap_shiny/).

These results also indicate, for which subtypes suitable cell line models exist and for

which subtypes cell lines should be prioritised for future *in vitro* model development [21]. Each BRCA subtype is represented by at least three cell lines (Fig 4(A)) and the heatmap shows that these cell lines are very similar to the corresponding tumours of the same subtype. However, only three cell lines represents the HER2-enriched subtype. The four subtypes of COADREAD tumours are also well represented by at least six highly similar cell lines in CCLE (Fig 4(B)).

For GBMLGG, the Mesenchymal-like tumour subtype is represented by 31 cell lines with high similarity scores. Many TCGA tumour samples have the molecular subtype Codel and G-CIMP-high, but they are only represented by seven and nine cell lines, respectively. Only two cell lines were classified as Classic-like and a single cell line has the predicted subtype LGm6-GBM. The PA-like tumour subtype is not represented by any cell line.

Predicting drug sensitivity in cancer patient sub-cohorts using MFmap and *in vitro* drug screens

Predicting patient therapeutic response is one important goal of subtype stratification. To explore the translational potential of the subtypes predicted by MFmap we estimated the association between predicted subtypes and drug sensitivity of all compounds available in the CTRP dataset [18]. For each cancer type listed in Table 1 and each compound, we compared the drug sensitivity among different cell line subtypes predicted by the MFmap classifier. Drug sensitivity is quantified in CTRP by the area under the dose response curve (AUC). We used an ANOVA to test for differences in the mean AUC among the predicted subtypes. At a false discovery rate (FDR) cutoff of 25%, we found 18, six and 16 compounds in BRCA, GBMLGG and UCEC to show significant subtype specificity, respectively. For the other seven cancer types in Table 1, there are no significant AUC differences across the different subtypes. Note that the sample size per subtype is very small, which might explain why statistically significant results can only be obtained for three cancer types.

For BRCA, the compound with the strongest association between subtype and drug sensitivity is Lapatinib (ANOVA p-value=2.95e-05). Lapatinib is a tyrosine kinase inhibitor used in combination therapy for HER2-positive breast cancer [35]. Our results suggest that cell lines of molecular subtype HER2-enriched are more sensitive to Lapatinib treatment (Fig 5(A)) in comparison to other three subtypes. Although there are only three cell lines representing the HER2-enriched subtype, this finding is in line with the known inhibitive mechanism of Lapatinib on the HER2/neu and epidermal growth factor receptor (EGFR) pathways. This result highlights the potential of MFmap as a tool for translating *in vitro* drug screening results to patient sub-cohorts. Our analysis also suggests that larger sample sizes and a better coverage of underrepresented subtypes are essential to increase the statistical power for detecting subtype specificity from cell line drug screens.

Fig 5. Cancer subtype specific drug sensitivity of CCLE cell lines. The subtypes of breast invasive carcinoma (BRCA) cell lines respond differentially to the compounds Lapatinib and Oligomycin A. Treatment response to the compounds KHS101 and Bortezomib in of glioblastoma multiforme and lower grade glioma (GBMLGG) cell lines is subtype specific. The drug sensitivity is summarised by the area under the dose response curve (AUC) and p-values refer to an ANOVA of the AUC differences among different subtypes.

Another drug with significant variations of the AUC values across the different BRCA subtypes is Oligomycin A (ANOVA p-value =1.39e-4), a compound targeting oxidative phosphorylation via an inhibition of the ATP synthase. The potential of

Oligomycin A as a therapeutic compound to prevent metastatic spread in breast cancer has recently been highlighted [36]. The results in Fig 5 (B) suggest that treatment with Oligomycin A might be most efficient for the HER2-enriched and Luminal A or Luminal B subtypes.

The drug sensitivities of KHS101 and Bortezomib are significantly associated with GBMLGG subtypes (KHS101: ANOVA p-value=2.3e-04; Bortezomib: ANOVA p-value=2.3e-04). The synthetic small molecule KHS101 was shown to promote tumour cell death in diverse glioblastoma multiforme cell line models [37]. Our analysis suggests that the G-CIMP-low subtype is more sensitive to KSH101 treatment (Fig 5(C)) compared to the other six GBMLGG subtypes. G-CIMP-low is an IDH mutant glioma subtype with poor clinical outcome in recurrent glioma [32].

Bortezomib targets the ubiquitin-proteasome pathway and is used for the treatment of multiple myeloma, but has also been discussed as treatment for glioma [38]. Our results in Fig 5 (D) show that the Codel and G-CIMP-high subtypes have larger AUCs. The results for LGm6-GBM and Classic-like are not conclusive because there are not enough cell lines representing these subtypes.

Biological characterisation of latent representations learnt by MFmap

The pattern of MFmap learnt latent representations z can be used as a signature for cancer subtypes. For example, in BRCA, the basal-like subtype is characterised by a pattern of low values of components z_1 and z_4 and high values of z_2 and z_3 (Fig 6(A)). HER2-enriched tumours are characterised by high values of z_1 and z_3 and z_4 . Luminal A and B subtypes can be distinguished by z_4 . Similarly, cancer subtypes in COADREAD and GBMLGG are highly associated with their latent representations learnt by MFmap (Fig 6(B,C)).

Fig 6. Association of MFmap latent representations and cancer subtypes.

The dimension of the latent representation h is set to the number of cancer subtypes. The boxplots display latent representations of different subtypes of TCGA samples in the three exemplary cancer types (A) breast invasive carcinoma (BRCA), (B) colorectal adenocarcinoma (COADREAD) and (C) glioblastoma multiforme and lower grade glioma (GBMLGG). Cancer subtypes are colour encoded and sorted by their median latent representations.

To further investigate the biological meaning of the latent representations we analysed the association between z and pathway activities in TCGA reference datasets. We used single sample gene set enrichment analysis (ssGSEA) [39] to assess sample-wise pathway activities. The pathway signatures were compiled from several sources including 10 curated oncogenic signalling pathways [40], 19 curated specific DNA damage repair (DDR) pathways [41], 14 expert-curated specific DDR processes and DDR associated processes [42]. This collection was combined with MsigDB (v7.0) [43] chemical and genetic perturbations (CGP) and canonical pathways (CP) collections (MsigDB C2 collection) and MsigDB (v7.0) hallmark gene sets (MsigDB H collection). The degree of associations was quantified by the information coefficient and the Pearson correlation coefficient and the statistical significance was assessed by permutation tests. To tackle class imbalance in the different subtypes, we applied SMOTE upsampling [44].

We used COADREAD as a proof of concept, because it has four well characterised molecular subtypes CMS1-CMS4 [34]. The CMS1 subtype is characterised by micro-satellite instability (MSI), whereas CMS4 tumours are micro-satellite stable. The CMS4 subtype is also distinguished from CMS1 by epithelial mesenchymal

transformation (EMT) characteristics, accompanied by prominent stromal invasion and angiogenesis. These mutually exclusive characteristics are clearly reflected in the magnitudes of the latent representation components. The top gene sets associated with component z_2 are "WATANABE COLON CANCER MSI VS MSS UP" and "KOINUMA COLON CANCER MSI UP", whereas z_4 is associated with the activity of gene sets annotated as "HALLMARK ANGIOGENESIS" and "HALLMARK EPITHELIAL MESENCHYMAL TRANSITION". Clearly, high values of z_2 are a characteristics of the CMS1 subtype, whereas high values of z_4 are a distinctive feature of CMS4 tumours. This example illustrates that a meaningful way to guide biological interpretation of the latent representations is to associate them to single sample pathway activity.

The same method was applied to annotate latent representations of GBMLGG (Fig 7(A)), which has seven subtypes [32]. The Mesenchymal-like and PA-like are stratified by gene expression profiles and the G-CIMP-high, G-CIMP-low and LGM6-GBM are methylation based. The Codel subtype describes IDH-mutant samples harbouring a co-deletion of chromosome arm 1p and 19q. Many pathways associated with latent representation z_1 are related to the neurotransmitter release cycle, which is also a characteristics of the Verhaak proneuronal subtype [45]. Pathways correlated to latent representation z_2 are related to the mesenchymal cell type, hypoxia and angiogenesis, which characterises the Verhaak mesenchymal subtype. The activity of the Fanconi Anemia (FA) DNA repair pathway is highly correlated with latent representation z_3 . DNA damage response deficiency and amplified oncogenic MYC signalling characterises tumours with large values of latent representation z_4 . Latent representation z_5 is related to the neurotransmitter release cycle and dysfunctional metabolism; latent representation z_6 to mitotic checkpoint deficiency. Many pathways associated with latent representation z_7 are involved in mismatch repair deficiency, replication stress and cell cycle dysregulation and also related to the classical subtype in the earlier classification of Verhaak [45].

Fig 7. Characterising the MFmap learnt latent representations in glioblastoma multiforme and lower grade glioma (GBMLGG). (A) The top heatmap shows the latent representation z of TCGA tumour samples (columns). The tumour samples are ordered based on a hierarchical clustering of z and their subtypes are colour encoded. The heatmap at the bottom displays sample-wise pathway activities that are significantly associated with the latent representations $z_1, \dots, z_7$. Pathway activities were computed using the ssGSEA algorithm [39]. For better visualisation, we upsampled the input data of MFmap and ssGSEA to get a balanced sample size in each subtype. (B) The MFmap reference map is formed by projecting the latent representations z of bulk tumours into two dimensions using multidimensional scaling. It consists of seven dominant components represented by black nodes. The length of their connections is given by the Euclidean distance of the dominant components in the latent space. The annotation of the seven dominant nodes is based on the correlation between z and pathway activity scores (see A). The background colour encodes sample subtypes, and the background contour encodes sample density. Individual bulk tumours are displayed as dots on the MFmap reference map. (C) Cell line samples are projected to the MFmap reference map. In both (B) and (C), the subtype of bulk tumours and predicted subtype of cell lines are colour coded. Subtype specific sample size for bulk tumours and cell lines is reported in the legend table.

Individual samples and their relationships can be displayed in the MFmap reference map (Fig 7(B)), a visualisation tool adapted from OncoGPS [31]. Here, the seven corners of the map correspond to the respective latent representations $z_1, \dots, z_7$ in GBMLGG. The corner locations are determined by multidimensional scaling on the

latent representations of bulk tumours. Individual bulk tumour samples are displayed as dots in the regions between the corner points with locations determined by a weighted vector sum of the seven corner locations (see Supporting Information for details). The subtype of each tumour sample is indicated by colours. The density of the tumour samples of a given subtype is depicted by the contour lines and the corresponding colour shading. Fig 7(B)) shows that samples of the same subtype clustered together and the inter-cluster distance is large. Projecting cell lines to the MFmap reference map (Fig 7(C)) helps to visualise the relationship between their predicted subtypes and their latent representations.

Modelling cellular state transformations using latent space arithmetics

Cancerous neoplasms undergo various biochemical changes during cancer evolution and in response to selective pressure. One example is the transition from a proneural to a mesenchymal phenotype in glioblastoma, which is characterised by acquired therapeutic resistance and more aggressive potential [46]. In the DNA methylation based subtype classification of [32], the G-CIMP-high methylation phenotype tends to have the proneural molecular subtype [45] (see Fig 7(B)). Given that the latent representations learnt by MFmap clearly distinguish these different subtypes, we asked, whether the generative nature of the semi-supervised VAE can also be exploited to study such cancer subtype transformations.

To this end, we used the latent representations of the G-CIMP-high tumours and the Mesenchymal-like tumours (see Fig 7 (B)) and computed the centroid vectors $\bar{z}_{\text{G-CIMP-high}}$ and $\bar{z}_{\text{Mesenchymal-like}}$ for the corresponding tumour samples. The difference $\delta = \bar{z}_{\text{Mesenchymal-like}} - \bar{z}_{\text{G-CIMP-high}}$ was used as a latent perturbation vector. By adding δ to the latent representation of each G-CIMP-high tumour (Fig 8 (A)) we obtained the latent representation of *in silico* samples (Fig 8 (B)), which are located in the "Mesenchymal-like region" of the reference map. We used these latent representation vectors of the *in silico* samples as input to the decoder of the MFmap network. We then checked, whether key molecular features of real Mesenchymal-like samples are reflected by these generated samples. Based on the available biological knowledge, we focussed on the most prominent onco-markers of the G-CIMP-high subtype: mutation status of the alpha thalassemia/mental retardation syndrome X-linked (ATRX), isocitrate dehydrogenase (IDH) and TP53 genes. The original G-CIMP-high tumours show a high propensity towards mutations in these genes, indicated by relatively higher network smoothed mutation scores (Fig 8(C)), although not all samples are necessarily harbouring these mutations. In contrast, the predicted mutation scores for the perturbed *in silico* samples in Fig 8(B) are much lower, indicating a lower propensity to IDH1, ATRX or TP53 mutations. This is in agreement with the observed tendency of Mesenchymal-like tumours for these mutations [46]. This example not only highlights the good generative performance of MFmap but also hints at potential applications on integrative analysis of cancer evolution dynamics.

Discussion

Limited success in translating *in vitro* therapeutic markers to clinical applications highlights that not all cell lines are good models for a given cancer subtype. Selecting the most appropriate cell line for a given tumour or a set of tumours is crucial for understanding cancer biology and developing new anti-cancer treatments. Here, we provide a computational framework and a resource for cancer researchers to select the best cell lines for a TCGA tumour or a cancer subtype from ten different cancer types

Fig 8. *In-silico* perturbation analysis of cellular state changes during disease transformation from the G-CIMP-high to the Mesenchymal-like subtype in glioblastoma multiforme and lower grade glioma (GBMLGG).

(A) The G-CIMP-high tumours from TCGA are projected to the MFmap reference map. (B) By perturbing the latent representation vectors of these G-CIMP-high tumours we generate artificial tumour samples located in the Mesenchymal-like region of the MFmap reference map (compare Fig 7(B)). (C) Boxplots of the sample mutation status (network smoothed mutation scores) of marker genes IDH1, ATRX1 and TP53 before and after perturbation.

(http://h2926513.stratoserver.net:3838/MFmap_shiny/). The quantitative similarity score enables researchers to judge, whether a given tumour or a subtype of tumours is well represented by a cell line.

The assignment of cancer subtype labels to cell lines enables cell biologists to optimise experimental planning and to focus their research on clinically relevant model systems. We found that our semi-supervised MFmap model can classify tumours with a very high accuracy. Further analysis of drug sensitivity profiles supports that the subtype prediction for cell lines is biologically meaningful. Our analysis shows that HER2-enriched cell lines are most sensitive to Lapatinib, in agreement with prior knowledge about drug efficiency of this compound. As an example for the translation of *in vitro* pharmacogenomic data, we predict that the G-CIMP-low subtype is more sensitive to the new synthetic compound KHS101 compared to other GBMLGG subtypes.

Our finding that only BRCA, GBMLGG and UCEC show significant subtype specific drug sensitivity variation merits further investigation. One important reason is the small number of cell lines representing some cancer subtypes, which prevents us from finding statistically significant variations of drug sensitivity across the different subtypes. This highlights the need to prioritise cell line development for underrepresented disease variants [21]. However, it can not be ruled out that for some cancers the known subtype classifications are not predictive of drug sensitivity. This suggests that clinically relevant subtype stratification should take into account drug sensitivity.

By embedding the original gene expression space, somatic mutation space and copy number space of bulk tumours and cell lines into a lower dimensional latent space, MFmap extracts latent features that are strongly associated with cancer subtypes. For COADREAD and GLMBGG, we have illustrated that the abstract latent representations can be annotated biologically using their associations with pathway activities. This makes the latent representations interpretable and allows to study the molecular and clinical heterogeneity of this disease. In principle, MFmap can be complemented by other modalities such as methylation or proteomics data. However, for our purpose we found that gene expression and DNA features in combination with the prior knowledge about tumour subtypes contains sufficient information.

Our proof of principle analysis of the transformation between two different tumour subtypes presents a new approach for studying tumour evolutionary processes in a more integrative way [47]. The small sample size of some multi-region sequencing or single-cell sequencing studies limits the ability to infer robust evolutionary patterns. By projecting these data to the MFmap reference map obtained from training on large sets of bulk tumour data one could deduce useful phenotypic information for individual patients. We believe that this can leverage information gathered in large cancer genomic studies like TCGA to guide personalised clinical decision making.

The MFmap is based on a new semi-supervised neural network architecture combining a basic VAE with an additional classifier. Such semi-supervised learning tasks are very common in the biomedical research field, because it is often easier to

acquire a large number of measurements than to obtain the corresponding labels. Based on the good predictive and generative performance of MFmap together with the evidence provided here, that MFmap can learn biologically and clinically meaningful information, we are convinced that the MFmap model can be adapted to other semi-supervised tasks in oncology and beyond.

Supporting Information

All statistical and machine learning analysis was done with R 3.6.3 and Python 3.6, the source code is available at <https://github.com/mcmzxx/mfMap.py>.

TCGA bulk tumour subtype annotations

The TCGA bulk tumour subtype annotations were collected from literatures listed in Table 3

Table 3. Subtype labels for 10 TCGA bulk tumours annotated from literatures.

TCGA code	study name	reference
BRCA	Breast invasive carcinoma	[33]
COADREAD	Colon adenocarcinoma	[34]
ESCA	Esophageal carcinoma	[48]
HNSC	Head and neck squamous cell carcinoma	[49]
LUAD	Lung adenocarcinoma	[50]
LUSC	Lung squamous cell carcinoma	[51]
PAAD	Pancreatic adenocarcinoma	[52]
SKCM	Skin cutaneous melanoma	[53]
UCEC	Uterine corpus endometrial carcinoma	[54]
GBMLGG	Glioblastoma multiforme and lower grade glioma	[32]

Bulk tumor molecular data processing

Copy number, gene expression and mutation data from bulk tumours for 10 cancer types listed in Table 1 were obtained from Firehose Broad GDAC portal (<http://gdac.broadinstitute.org/>). Copy number $\log_2$ ratio segment data genome were input into GISTIC2.0 (version 2.0.23) [55] to obtain thresholded CNV values. These data were dichotomised and stored in a sample by gene binary matrix, where 0 indicates a diploid copy number level and 1 indicates one of the following CNV levels: homozygous deletion, heterozygous deletion, low-level amplification, high-level amplification.

RSEM normalised gene expression data were preprocessed as follows: genes with missing values in more than 20% of the samples were excluded. For the remaining genes, missing values were imputed using `impute.knn` function from `impute` package [56]. The resulting gene expression values were transformed by adding 1.0 to each value and taking the logarithm (pseudo-log-transformation).

Mutation data were downloaded as MAF files and synonymous mutations were excluded. The single-sample MAF files were further aggregated into a binary matrix M , with entries $(M)_{i,j} = 1$ indicating that there is a nonsynonymous mutation at j in patient i .

Cell line molecular data processing

Cell line data for the same cancer types in Table 1 were downloaded from the CCLE portal (<https://portals.broadinstitute.org/ccle>) and processed in the same way as bulk tumours, except that ensemble ids were mapped to gene symbols using human genome Genecode V19.

Molecular data normalisation

All molecular data input into MFmap neural network are normalised to the 0 – 1 range.

Bulk tumor clinical data

Multiple subtype classification schemes for TCGA BRCA, COADREAD and GBMLGG exist and we decided to follow the most frequently used schemes. COADREAD has four consensus molecular subtypes CMS1-4 obtained from gene expression data [57]. GBMLGG [32] has seven subtypes (IDHmut-codel or Codel, G-CIMP-low, G-CIMP-high, Classic-like, Mesenchymal-like, LGm6-GBM, PA-like) which were derived from multi-modal data sets including mutation, methylation and gene expression patterns. BRCA subtypes (Basal, Her2, Luminal A, Luminal B, and Normal) are based on PAM50 (Prediction Analysis of Microarray using 50 gene set) [58] features derived from gene expression data and obtained from [33]. In our study, only Basal, Her2, Luminal A, Luminal B samples were kept. The subtype annotations of all studied cancer types listed in Table 1 can be found in Table 3.

Cell line drug sensitivity data

CCLE cell line sample annotation data were obtained from CCLE portal and drug sensitivity data were downloaded from the Cancer Therapeutics Response Portal (CTRP [18], www.broadinstitute.org/ctrp).

Correcting for batch effects between bulk tumour and cell line gene expression

Batch effects between bulk tumours and cell lines were corrected by using the function *Combat* from the R package *SVA* [59]. The label bulk tumour or cell line was used as the covariate.

Handling class imbalance

Cancer subtype datasets especially GBMLGG are unbalanced, which is a major reason of overfitting for machine learning models. To overcome this issue, we applied Synthetic Minority Over-sampling Technique (SMOTE) [44] implemented in *SmoteClassif* function of *UBL* [60] R package to oversample the minority subtypes, creating an equal balance with majority subtypes.

Propagating copy number and mutation profiles on the protein-protein interaction network

Mutation profiles were mapped to human cancer network *A* curated by pyNBS [25] as protein-protein interaction network (PPI) source, which aggregates different interaction types. The propagating function in pyNBS with optimal signal diffusion distance parameter setting $\alpha = 0.7$ was applied to perform network propagation, described by the

iterative process $x_{t+1} = \alpha x_t A + (1 - \alpha)x_0$ with a closed solution $x_\infty = (1 - \alpha A)^{-1}x_0$. Here x_0 is the sample-by-gene matrix, A is the adjacency matrix of the PPI, and α is the parameter controlling the diffusion distance. The same approach was used for dichotomised copy number profiles. The smoothed copy number profiles and mutation profiles (sample by gene matrices) were combined into a single DNA-view matrix.

Pathway activity scores

Gene expression data for MsigDB [43] gene sets were input to ssGSEA [39], which outputs sample-wise pathway activity scores.

Biological annotation of latent representations

For a given component z_k of the latent representation, we selected all pathways whose activities are significantly associated with z_k (association was measured using the information coefficient, FDR threshold of 5%). Some pathways are associated with more than one latent representation. To resolve these ambiguities, the Pearson correlation coefficients were estimated as well. The selection of pathways was further refined by using the fold change of z_k intensities between subtypes with the highest two latent representations and the lowest two latent representations. Here fold change and significance were estimated by a linear model implemented in the limma package [61], only those pathways with FDR adjusted p-value less than 0.05 were further kept for human review.

Generating MFmap visualisations

Let $Z = (z_{ij})$ be the $n \times h$ -matrix of latent representations. The element z_{ij} is the value of the latent representations $j \in \{1, \dots, h\}$ for patient $i \in \{1, \dots, n\}$. The MFmap visualisation proceeds in three steps:

Step 1: Generate coordinates C_{1j}, C_{2j} of the MFmap prominent component nodes j by projecting the columns of matrix Z to 2-D space. The projection is chosen in order to preserve the distances between samples (Sammon projection). Delaunay triangulation on the 2-D coordinates is used to connect neighbouring nodes.

Step 2: Project samples onto the MFmap layout. The coordinates (S_{1i}, S_{2i}) of sample $i \in \{1, \dots, n\}$ are given as

$$S_{li} = \frac{1}{\sum_{j=1}^h z_{ij}^\alpha} \sum_{j=1}^h C_{lj} z_{ij}^\alpha, \quad l = 1, 2 \quad (12)$$

where α is a tuneable hyper-parameter controlling the distance between nodes.

Step 3: Generate MFmap contour lines and background colours based on the estimated sample density of each subtype. The density is obtained from a Gaussian kernel density estimate on the coordinate lattice corresponding to the projected MFmap latent representations. The probability estimate for the subtype with the highest probability is then used for MFmap visualisation.

In-silico perturbation analysis

Let $z_i^{(s)}$ denote the latent space representation of a tumour sample or cell line i with subtype s . An *in silico* latent sample $\tilde{z}_i = z_i^{(s)} + \delta$ is obtained by adding the latent space perturbation δ . Then, the artificial data point $\tilde{x}_i$ is sampled from the decoder

$\tilde{x}_i \sim p(\mathbf{x}|\tilde{z}_i)$. For Fig 8 we used the $s = \text{G-CIMP-high}$ as the subtype of the original data. The perturbation $\delta = \bar{z}^{(s)} - \bar{z}^{(s')}$ with $s' = \text{Mesenchymal-like}$ was obtained as the difference between the mean latent space vectors of both subtypes.

Design of MFmap neural network

The MFmap encoder network

The encoder receives a gene expression profile $\mathbf{x}_{rna}$ and a propagated DNA alteration profile $\mathbf{x}_{dna}$ as input. Two hidden layers first encode the two input layers into two 1024-dimensional latent vectors. The second hidden layer then concatenates the two 1024-dimensional latent vectors and then is further encoded into 512-dimensional vector by the third hidden layer. The third hidden layer is fully connected to two output layers representing mean μ and log-transformed variation $\log \sigma^2$ of $q_\phi(\mathbf{z}|\mathbf{x})$. The dimension of the latent representation $\mathbf{z}$ is set as the number of subtypes of cancer. The encoder is given by the following equations:

$$\mathbf{x}_{erna}^1 = \text{ReLU}(W_{erna}^0 \cdot \mathbf{x}_{rna} + \mathbf{b}_{erna}^1) \quad (13)$$

$$\mathbf{x}_{edna}^1 = \text{ReLU}(W_{edna}^0 \cdot \mathbf{x}_{dna} + \mathbf{b}_{edna}^1) \quad (14)$$

$$\mathbf{x}_e^2 = \text{ReLU}(W_e^1 \cdot \mathbf{x}_{erna}^1 \oplus \mathbf{x}_{edna}^1 + \mathbf{b}_e^2) \quad (15)$$

$$\mathbf{x}_e^3 = \text{ReLU}(W_e^2 \cdot \mathbf{x}_e^2 + \mathbf{b}_e^3) \quad (16)$$

$$\mu = \text{ReLU}(W_e^\mu \cdot \mathbf{x}_e^3 + \mathbf{b}^\mu) \quad (17)$$

$$\log(\sigma^2) = \text{ReLU}(W_e^\sigma \cdot \mathbf{x}_e^3 + \mathbf{b}^\sigma) \quad (18)$$

where $W_{erna}^0 \in \mathbb{R}^{k_{rna} \times h_{rna}^1}$, $\mathbf{b}_{erna}^1 \in \mathbb{R}^{1 \times h_{rna}^1}$, $W_{edna}^0 \in \mathbb{R}^{k_{dna} \times h_{dna}^1}$, $\mathbf{b}_{edna}^1 \in \mathbb{R}^{1 \times h_{dna}^1}$, $W_e^1 \in \mathbb{R}^{h_{rna}^1 + h_{dna}^1 \times h^2}$, $\mathbf{b}_e^2 \in \mathbb{R}^{1 \times h^2}$, $W_e^2 \in \mathbb{R}^{h^2 \times h^3}$, $\mathbf{b}_e^3 \in \mathbb{R}^{1 \times h^3}$, $W_e^\mu \in \mathbb{R}^{h^3 \times m}$, $\mathbf{b}^\mu \in \mathbb{R}^{1 \times m}$, $W_e^\sigma \in \mathbb{R}^{h^3 \times m}$, $\mathbf{b}^\sigma \in \mathbb{R}^{1 \times m}$ are trainable parameters of the encoder network. Rectified linear unit (ReLU) activation is defined as $\text{ReLU}(x) = \max(x, 0)$. $\oplus$ denotes concatenating two matrices.

The MFmap decoder network

The MFmap decoder layer structure is symmetric to the encoder with latent representations $\mathbf{z}$ as inputs and reconstructed data as outputs $\mathbf{x}'_{rna}$ and $\mathbf{x}'_{dna}$:

$$\mathbf{x}_d^1 = \text{ReLU}(W_d^0 \cdot \mathbf{z} + \mathbf{b}_d^1) \quad (19)$$

$$\mathbf{x}_d^2 = \text{ReLU}(W_d^1 \cdot \mathbf{x}_d^1 + \mathbf{b}_d^2) \quad (20)$$

$$\mathbf{x}_{drna}^3 = \text{ReLU}(W_{drna}^2 \cdot \top^{h_{rna}^1}(\mathbf{x}_d^2) + \mathbf{b}_{drna}^3) \quad (21)$$

$$\mathbf{x}_{ddna}^3 = \text{ReLU}(W_{ddna}^2 \cdot \perp^{h_{dna}^1}(\mathbf{x}_d^2) + \mathbf{b}_{ddna}^3) \quad (22)$$

$$\mathbf{x}'_{rna} = \sigma(W_{rna}^o \cdot \mathbf{x}_{drna}^3 + \mathbf{b}_{rna}^o) \quad (23)$$

$$\mathbf{x}'_{dna} = \sigma(W_{dna}^o \cdot \mathbf{x}_{ddna}^3 + \mathbf{b}_{dna}^o) \quad (24)$$

Here, $W_d^0 \in \mathbb{R}^{m \times h^3}$, $\mathbf{b}_d^1 \in \mathbb{R}^{1 \times h^3}$, $W_d^1 \in \mathbb{R}^{h^3 \times h^2}$, $\mathbf{b}_d^2 \in \mathbb{R}^{1 \times h^2}$, $W_{drna}^2 \in \mathbb{R}^{h^2 \times h_{rna}^1}$, $\mathbf{b}_{drna}^3 \in \mathbb{R}^{1 \times h_{rna}^1}$, $W_{ddna}^2 \in \mathbb{R}^{h^2 \times h_{dna}^1}$, $\mathbf{b}_{ddna}^3 \in \mathbb{R}^{1 \times h_{dna}^1}$, $W_{rna}^o \in \mathbb{R}^{h_{rna}^1 \times k_{rna}}$, $\mathbf{b}_{rna}^o \in \mathbb{R}^{1 \times h_{rna}^1}$, $W_{dna}^o \in \mathbb{R}^{h_{dna}^1 \times k_{dna}}$, $\mathbf{b}_{dna}^o \in \mathbb{R}^{1 \times h_{dna}^1}$ are trainable parameters of the decoder network. The sigmoid activation is defined as $\sigma(x) = \frac{1}{1+e^{-x}}$. The subsetting operation of the top and bottom j elements of x is denoted by $\top^j(x)$ and $\perp^j(x)$, respectively.

The MFmap classifier network

The classifier serves as a regulariser controlling the capability to learn a latent representation that is cancer subtype relevant. Since cancer subtypes are clinically and biologically meaningful, a higher classification accuracy will encourage the neural network to extract features essential to patients stratification and are more interpretable. The classifier is a neural network with three fully connected layers and takes the expectation $\boldsymbol{\mu}$ (see Eq (17)) of the distribution $p(\mathbf{z}|\mathbf{x})$ with $\mathbf{x} = (\mathbf{x}_{rna}, \mathbf{x}_{dna})$ as input and outputs a probability for each of the h subtypes:

$$\mathbf{x}_c^1 = \text{ReLU}(W_c^0 \cdot \boldsymbol{\mu} + \mathbf{b}_c^1) \quad (25)$$

$$\mathbf{x}_c^2 = \text{ReLU}(W_c^1 \cdot \mathbf{x}_c^1 + \mathbf{b}_c^2) \quad (26)$$

$$\mathbf{x}_o = \text{softmax}(W_c^o \cdot \mathbf{x}_c^2 + \mathbf{b}_c^o), \quad (27)$$

where $W_c^0 \in \mathbb{R}^{m \times h_c^1}$, $\mathbf{b}_c^1 \in \mathbb{R}^{1 \times h_c^1}$, $W_c^1 \in \mathbb{R}^{h_c^1 \times h_c^2}$, $\mathbf{b}_c^2 \in \mathbb{R}^{1 \times h_c^2}$, $W_c^o \in \mathbb{R}^{h_c^2 \times s}$, $\mathbf{b}_c^o \in \mathbb{R}^{1 \times s}$ are trainable parameters of the classifier. The softmax activation is defined as $\text{softmax}(x, y = c) = \frac{e^{x_c}}{\sum_{j=1}^s e^{x_j}}$, given the inputs and label pair $(x, y = c)$.

Details of the MFmap hidden layers

Each hidden layer is a block containing a fully connected layer, a batch normalisation layer and an activation.

Details of the MFmap loss function

The MFmap loss function in Eq (10) for the specific distributional assumptions can be rewritten as

$$\begin{cases} \mathcal{S}(\mathbf{x}, y) = \sum_{v \in \{rna, dna\}} \mathcal{L}_{\text{recon}}(\mathbf{x}_v, \hat{\mathbf{x}}_v) + \mathcal{D}_{KL}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I})) + \mathcal{L}_{\text{CE}}(y, \hat{y}) + \mathcal{H}(\hat{y}), & (\mathbf{x}, y) \in \mathcal{D}_{tu}, \\ \mathcal{U}(\mathbf{x}) = \sum_{v \in \{rna, dna\}} \mathcal{L}_{\text{recon}}(\mathbf{x}_v, \hat{\mathbf{x}}_v) + \mathcal{D}_{KL}(\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I})) + \mathcal{H}(\hat{y}), & \mathbf{x} \in \mathcal{D}_{cl}. \end{cases} \quad (28)$$

In Eq (28), $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$ are estimated parameters of MFmap encoder. $\hat{y}$ denotes the subtype label probability predicted by the MFmap classifier and $\mathcal{L}_{\text{recon}}$, $\mathcal{D}_{KL}$, $\mathcal{L}_{\text{CE}}$, and $\mathcal{H}$ denote the reconstruction loss, KL divergence, cross entropy loss and entropy, respectively. We next detail each term.

The reconstruction loss $\mathcal{L}_{\text{recon}}$ is quantified by the binary cross entropy loss between input data $\mathbf{x}$ and reconstructed data $\hat{\mathbf{x}}$:

$$\mathcal{L}_{\text{recon}}(\mathbf{x}, \hat{\mathbf{x}}) = - \sum_{i=1}^d \mathbf{x}_i \log(\hat{\mathbf{x}}_i) + (1 - \mathbf{x}_i) \log(1 - \hat{\mathbf{x}}_i). \quad (29)$$

The classification loss for sample i is implemented as a cross entropy loss

$$\mathcal{L}_{\text{CE}}(y^{(i)}, \hat{y}^{(i)}) = - \sum_{k=1}^h y_k^{(i)} \log(\hat{y}_k^{(i)}). \quad (30)$$

The entropy of sample i is implemented as

$$\mathcal{H}(\hat{y}^{(i)}) = - \sum_{k=1}^h \hat{y}_k^{(i)} \log(\hat{y}_k^{(i)}). \quad (31)$$

In Eq (30) and Eq (31) $\hat{y}_k^{(i)} = \text{softmax}_k(\mathbf{c}(\mathbf{x}^{(i)}; \theta))$, where $\mathbf{c}(\cdot) = (c_1(\cdot), \dots, c_h(\cdot))$ is the classification model parametrised by θ , and softmax_k is the softmax function for subtype label k .

In fact, for bulk tumours the MFmap loss function can be viewed as a basic VAE loss plus an entropy regularised classification loss which is equivalent to a modified soft bootstrapping loss proposed by [62] for positive unlabelled learning, extended to multi-class case. It updates the prediction objective based on currently predicted subtype probability. Concretely, the modified soft bootstrapping loss for a pair of feature and subtype label $(\mathbf{x}^{(i)}, y^{(i)})$ is:

$$l_{sb}(\mathbf{x}^{(i)}, y^{(i)}; \theta) = \begin{cases} -\log(\text{softMax}_1(\mathbf{c}(\mathbf{x}^{(i)}; \theta))) - \sum_{k=1}^h \text{softMax}_k(\mathbf{c}(\mathbf{x}^{(i)}; \theta)) \log(\text{softMax}_k(\mathbf{c}(\mathbf{x}^{(i)}; \theta))), & y^{(i)} = 1, \\ \vdots \\ -\log(\text{softMax}_h(\mathbf{c}(\mathbf{x}^{(i)}; \theta))) - \sum_{k=1}^h \text{softMax}_k(\mathbf{c}(\mathbf{x}^{(i)}; \theta)) \log(\text{softMax}_k(\mathbf{c}(\mathbf{x}^{(i)}; \theta))), & y^{(i)} = h. \end{cases} \quad (32)$$

Gradients of the MFmap loss function can be computed with the reparametrisation trick [24], which involves sampling a random vector $\epsilon \sim \mathcal{N}(\epsilon | \epsilon; \mathbf{0}, \mathbf{I})$ and transforming to

$$\mathbf{z} = \boldsymbol{\mu}(\mathbf{x}) + \boldsymbol{\sigma}(\mathbf{x}) \odot \epsilon. \quad (33)$$

This ensures that $\mathbf{z} \sim q(\mathbf{z} | \mathbf{x}) = \mathcal{N}(\mathbf{z} | \boldsymbol{\mu}(\mathbf{x}), \text{diag}(\boldsymbol{\sigma}^2))$.

Details of the MFmap implementation and training process

The MFmap neural network was implemented using PyTorch (version 1.5.0) and trained on two NVIDIA Tesla V100 SXM2 GPUs (each has memory of 32 gigabytes) using the Adam optimiser [63]. The preprocessed data were randomly split into training (90%) and test (10%) sets. Hyperopt [64] was used to search the best model avoiding overfitting by selecting optimal hyper-parameters yielding the minimum total loss divergence between training and validation datasets (ratio training/validation data 9/10). The searching space for hyper-parameter selection is:

- dimensions of first DNA or RNA encoder hidden layer {4096, 2048, 1024}
- dimensions of second DNA or RNA encoder hidden layer {1024, 512, 256}
- dimensions of third DNA or RNA encoder hidden layer {256, 128, 64}
- dimensions of first classifier hidden layer {512, 256, 128, 64}
- dimensions of second classifier hidden layer {128, 64, 32}
- batch size {64, 32}
- learning rate $\{10^{-4}, 10^{-3}, 10^{-2}\}$

The optimised setting found from the minimal validation loss is: learning rate as 10^{-3} ; batch size as 32; first, second and third encoder hidden layer dimension as 1024, 512, 256 respectively; first, second classifier hidden layer dimension as 128, 64 respectively. These hyper-parameters were used for all cancer types in Table 1.

Derivation of the unsupervised and supervised ELBO

We first derive the ELBO for cell line data using Jensen's inequality:

$$\begin{aligned} \log p(\mathbf{x}) &= \log \left(\sum_y \int d\mathbf{z} q(\mathbf{z} | \mathbf{x}) \frac{p(\mathbf{x}, y, \mathbf{z})}{q(\mathbf{z} | \mathbf{x})} \right) \\ &= \log \left(E_{q(\mathbf{z} | \mathbf{x}) p(y | \mathbf{z})} \left[\frac{p(\mathbf{x} | \mathbf{z}) p(\mathbf{z})}{q(\mathbf{z} | \mathbf{x})} \right] \right) \\ &\geq E_{q(\mathbf{z} | \mathbf{x}) p(y | \mathbf{z})} \left[\frac{p(\mathbf{x} | \mathbf{z}) p(\mathbf{z})}{q(\mathbf{z} | \mathbf{x})} \right] \\ &= E_{q(\mathbf{z} | \mathbf{x})} [\log p(\mathbf{x} | \mathbf{z})] - D_{KL}(q(\mathbf{z} | \mathbf{x}) || p(\mathbf{z})). \end{aligned} \quad (34)$$

Here, we have used the conditional independence assumption Eq (3c) and $p(y|z) = q(y|z)$. Similarly, for the labelled examples (bulk tumor samples), we can derive the ELBO for the log-likelihood as

$$\begin{aligned}\log p(\mathbf{x}, y) &= \log \left(\int d\mathbf{z} q(\mathbf{z}|\mathbf{x}) \frac{p(\mathbf{x}, y, \mathbf{z})}{q(\mathbf{z}|\mathbf{x})} \right) \\ &\geq E_{q(\mathbf{z}|\mathbf{x})} \left[\log \left(\frac{p(\mathbf{x}|\mathbf{z})p(y|\mathbf{z})p(\mathbf{z})}{q(\mathbf{z}|\mathbf{x})} \right) \right] \\ &= E_{q(\mathbf{z}|\mathbf{x})} [\log p(\mathbf{x}|\mathbf{z})] - D_{KL}(q(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) + E_{q(\mathbf{z}|\mathbf{x})} [\log p(y|\mathbf{z})] \quad (35)\end{aligned}$$

where we have used the conditional independence of $\mathbf{x}$ and y given $\mathbf{z}$ in both the generative and the inference model.

References

- Sharma SV, Haber DA, Settleman J. Cell line-based platforms to evaluate the therapeutic efficacy of candidate anticancer agents. *Nature Reviews Cancer*. 2010;10(4):241–253. doi:10.1038/nrc2820.
- Kim N, He N, Yoon S. Cell line modeling for systems medicine in cancers (Review). *International Journal of Oncology*. 2014;44(2):371–376. doi:10.3892/ijo.2013.2202.
- Goodspeed A, Heiser LM, Gray JW, Costello JC. Tumor-Derived Cell Lines as Molecular Models of Cancer Pharmacogenomics. *Molecular Cancer Research*. 2016;14(1):3–13. doi:10.1158/1541-7786.MCR-15-0189.
- Kaur G, Dufour JM. Cell lines: Valuable tools or useless artifacts. *Spermatogenesis*. 2012;2(1):1–5. doi:10.4161/spmg.19885.
- Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*. 2012;483(7391):603–607. doi:10.1038/nature11003.
- Mouradov D, Sloggett C, Jorissen RN, Love CG, Li S, Burgess AW, et al. Colorectal Cancer Cell Lines Are Representative Models of the Main Molecular Subtypes of Primary Cancer. *Cancer Research*. 2014;74(12):3238–3247. doi:10.1158/0008-5472.CAN-14-0013.
- Ince TA, Sousa AD, Jones MA, Harrell JC, Agoston ES, Krohn M, et al. Characterization of twenty-five ovarian tumour cell lines that phenocopy primary tumours. *Nature Communications*. 2015;6(1):7419. doi:10.1038/ncomms8419.
- Cheng H, Yang X, Si H, Saleh AD, Xiao W, Coupar J, et al. Genomic and Transcriptomic Characterization Links Cell Lines with Aggressive Head and Neck Cancers. *Cell Reports*. 2018;25(5):1332–1345.e5. doi:10.1016/j.celrep.2018.10.007.
- Domcke S, Sinha R, Levine DA, Sander C, Schultz N. Evaluating cell lines as tumour models by comparison of genomic profiles. *Nature Communications*. 2013;4(1):2126. doi:10.1038/ncomms3126.
- Liu K, Newbury PA, Glicksberg BS, Zeng WZD, Paithankar S, Andrechek ER, et al. Evaluating cell lines as models for metastatic breast cancer through integrative analysis of genomic data. *Nature Communications*. 2019;10(1):2138. doi:10.1038/s41467-019-10148-6.

11. Sinha R, Winer AG, Chevinsky M, Jakubowski C, Chen YB, Dong Y, et al. Analysis of renal cancer cell lines from two major resources enables genomics-guided cell line selection. *Nature Communications*. 2017;8(1):15165. doi:10.1038/ncomms15165. 649-652
12. Yu K, Chen B, Aran D, Charalel J, Yau C, Wolf DM, et al. Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. *Nature Communications*. 2019;10(1):3574. doi:10.1038/s41467-019-11415-2. 653-655
13. Shoemaker RH. The NCI60 human tumour cell line anticancer drug screen. *Nature Reviews Cancer*. 2006;6(10):813–823. doi:10.1038/nrc1951. 656-657
14. Ghandi M, Huang FW, Jané-Valbuena J, Kryukov GV, Lo CC, McDonald ER, et al. Next-generation characterization of the Cancer Cell Line Encyclopedia. *Nature*. 2019;569(7757):503–508. doi:10.1038/s41586-019-1186-3. 658-660
15. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, et al. A Landscape of Pharmacogenomic Interactions in Cancer. *Cell*. 2016;166(3):740–754. doi:10.1016/j.cell.2016.06.017. 661-663
16. Lamb J, Crawford ED, Peck D, Modell JW, Blat IC, Wrobel MJ, et al. The Connectivity Map: Using Gene-Expression Signatures to Connect Small Molecules, Genes, and Disease. *Science*. 2006;313(5795):1929. doi:10.1126/science.1132939. 664-667
17. Basu A, Bodycombe NE, Cheah JH, Price EV, Liu K, Schaefer GI, et al. An Interactive Resource to Identify Cancer Genetic and Lineage Dependencies Targeted by Small Molecules. *Cell*. 2013;154(5):1151–1161. doi:10.1016/j.cell.2013.08.003. 668-671
18. Seashore-Ludlow B, Rees MG, Cheah JH, Cokol M, Price EV, Coletti ME, et al. Harnessing Connectivity in a Large-Scale Small-Molecule Sensitivity Dataset. *Cancer Discovery*. 2015;5(11):1210. doi:10.1158/2159-8290.CD-15-0235. 672-674
19. Chang K, Creighton CJ, Davis C, Donehower L, Drummond J, Wheeler D, et al. The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*. 2013;45(10):1113–1120. doi:10.1038/ng.2764. 675-677
20. Hudson (Chairperson) TJ, Anderson W, Aretz A, Barker AD, Bell C, Bernabé RR, et al. International network of cancer genome projects. *Nature*. 2010;464(7291):993–998. doi:10.1038/nature08987. 678-680
21. Najgebauer H, Yang M, Francies HE, Pacini C, Stronach EA, Garnett MJ, et al. CELLector: Genomics-Guided Selection of Cancer In Vitro Models. *Cell Systems*. 2020;10(5):424–432.e6. doi:10.1016/j.cels.2020.04.007. 681-683
22. Salvadores M, Fuster-Tormo F, Supek F. Matching cell lines with cancer type and subtype of origin via mutational, epigenomic, and transcriptomic patterns. *Science Advances*. 2020;6(27):eaba1862. doi:10.1126/sciadv.aba1862. 684-686
23. Webber JT, Kaushik S, Bandyopadhyay S. Integration of Tumor Genomic Data with Cell Lines Using Multi-dimensional Network Modules Improves Cancer Pharmacogenomics. *Cell Systems*. 2018;7(5):526–536.e6. doi:10.1016/j.cels.2018.10.001. 687-690
24. Kingma DP, Welling M. Auto-Encoding Variational Bayes. arXiv:1312.6114 [Preprint]. 2013. Available from: <https://arxiv.org/pdf/1312.6114.pdf>. 691-692

25. Huang JK, Jia T, Carlin DE, Ideker T. pyNBS: a Python implementation for network-based stratification of tumor mutations. *Bioinformatics*. 2018;34(16):2859–2861. doi:10.1093/bioinformatics/bty186. 693
694
695
26. Hofree M, Shen JP, Carter H, Gross A, Ideker T. Network-based stratification of tumor mutations. *Nature Methods*. 2013;10(11):1108–1115. doi:10.1038/nmeth.2651. 696
697
698
27. Kingma DP, Rezende DJ, Mohamed S, Welling M. Semi-Supervised Learning with Deep Generative Models. arXiv:14065298v2 [Preprint]. 2014. Available from: <https://arxiv.org/pdf/1406.5298.pdf>. 699
700
701
28. Feng H, Kong K, Chen M, Zhang T, Zhu M, Chen W. SHOT-VAE: Semi-supervised Deep Generative Models With Label-aware ELBO Approximations. arXiv:201110684 [Preprint]. 2020. Available from: <https://arxiv.org/pdf/2011.10684.pdf>. 702
703
704
705
29. Dai Z, Yang Z, Yang F, Cohen WW, Salakhutdinov R. Good Semi-supervised Learning that Requires a Bad GAN. arXiv:170509783 [Preprint]. 2017. Available from: <https://arxiv.org/pdf/1705.09783.pdf>. 706
707
708
30. Grandvalet Y, Bengio Y. Semi-Supervised Learning by Entropy Minimization. In: *Proceedings of the 17th International Conference on Neural Information Processing Systems. NIPS'04*. Cambridge, MA, USA: MIT Press; 2004. p. 529–536. 709
710
711
712
31. Kim JW, Abudayyeh OO, Yeerna H, Yeang CH, Stewart M, Jenkins RW, et al. Decomposing Oncogenic Transcriptional Signatures to Generate Maps of Divergent Cellular States. *Cell Systems*. 2017;5(2):105–118.e9. doi:10.1016/j.cels.2017.08.002. 713
714
715
716
32. Ceccarelli M, Barthel FP, Malta TM, Sabedot TS, Salama SR, Murray BA, et al. Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell*. 2016;164(3):550–563. doi:https://doi.org/10.1016/j.cell.2015.12.028. 717
718
719
720
33. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*. 2012;486(7403):346–352. doi:10.1038/nature10983. 721
722
723
34. Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, et al. The consensus molecular subtypes of colorectal cancer. *Nature Medicine*. 2015;21(11):1350–1356. doi:10.1038/nm.3967. 724
725
726
35. Higa GM, Abraham J. Lapatinib in the treatment of breast cancer. *Expert Review of Anticancer Therapy*. 2007;7(9):1183–1192. doi:10.1586/14737140.7.9.1183. 727
728
729
36. Davis RT, Blake K, Ma D, Gabra MBI, Hernandez GA, Phung AT, et al. Transcriptional diversity and bioenergetic shift in human breast cancer metastasis revealed by single-cell RNA sequencing. *Nat Cell Biol*. 2020;22(3):310–320. doi:10.1038/s41556-020-0477-0. 730
731
732
733
37. Polson ES, Kuchler VB, Abbosh C, Ross EM, Mathew RK, Beard HA, et al. KHS101 disrupts energy metabolism in human glioblastoma cells and reduces tumor growth in mice. *Science Translational Medicine*. 2018;10(454):eaar2718. doi:10.1126/scitranslmed.aar2718. 734
735
736
737

38. Tang JH, Yang L, Chen JX, Li QR, Zhu LR, Xu QF, et al. Bortezomib inhibits growth and sensitizes glioma to temozolomide (TMZ) via down-regulating the FOXM1–Survivin axis. *Cancer Commun.* 2019;39(1):81. doi:10.1186/s40880-019-0424-2. 738 739 740 741
39. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics.* 2013;14(1):7. doi:10.1186/1471-2105-14-7. 742 743 744
40. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, et al. Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell.* 2018;173(2):321–337.e10. doi:10.1016/j.cell.2018.03.035. 745 746 747
41. Knijnenburg TA, Wang L, Zimmermann MT, Chambwe N, Gao GF, Cherniack AD, et al. Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell Reports.* 2018;23(1):239–254.e6. doi:https://doi.org/10.1016/j.celrep.2018.03.076. 748 749 750 751
42. Pearl LH, Schierz AC, Ward SE, Al-Lazikani B, Pearl FMG. Therapeutic opportunities within the DNA damage response. *Nature Reviews Cancer.* 2015;15(3):166–180. doi:10.1038/nrc3891. 752 753 754
43. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdóttir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. *Bioinformatics.* 2011;27(12):1739–1740. doi:10.1093/bioinformatics/btr260. 755 756 757
44. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research.* 2002;16:321–357. doi:10.1613/jair.953. 758 759 760
45. Verhaak RGW, Hoadley KA, Purdom E, Wang V, Qi Y, Wilkerson MD, et al. Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell.* 2010;17(1):98–110. doi:10.1016/j.ccr.2009.12.020. 761 762 763 764
46. Behnan J, Finocchiaro G, Hanna G. The landscape of the mesenchymal signature in brain tumours. *Brain.* 2019;142(4):847–866. doi:10.1093/brain/awz044. 765 766
47. Williams MJ, Sottoriva A, Graham TA. Measuring Clonal Evolution in Cancer with Genomics. *Annu Rev Genom Hum Genet.* 2019;20(1):309–329. doi:10.1146/annurev-genom-083117-021712. 767 768 769
48. Kim J, Bowlby R, Mungall AJ, Robertson AG, Odze RD, Cherniack AD, et al. Integrated genomic characterization of oesophageal carcinoma. *Nature.* 2017;541(7636):169–175. doi:10.1038/nature20805. 770 771 772
49. Network CGA. Comprehensive genomic characterization of head and neck squamous cell carcinomas. *Nature.* 2015;517(7536):576–582. doi:10.1038/nature14129. 773 774 775
50. Collisson EA, Campbell JD, Brooks AN, Berger AH, Lee W, Chmielecki J, et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature.* 2014;511(7511):543–550. doi:10.1038/nature13385. 776 777 778
51. Hammerman PS, Lawrence MS, Voet D, Jing R, Cibulskis K, Sivachenko A, et al. Comprehensive genomic characterization of squamous cell lung cancers. *Nature.* 2012;489(7417):519–525. doi:10.1038/nature11404. 779 780 781

52. Moffitt RA, Marayati R, Flate EL, Volmar KE, Loeza SGH, Hoadley KA, et al. Virtual microdissection identifies distinct tumor- and stroma-specific subtypes of pancreatic ductal adenocarcinoma. *Nature Genetics*. 2015;47(10):1168–1178. doi:10.1038/ng.3398. 782–785
53. Akbani R, Akdemir KC, Aksoy BA, Albert M, Ally A, Amin SB, et al. Genomic Classification of Cutaneous Melanoma. *Cell*. 2015;161(7):1681–1696. doi:https://doi.org/10.1016/j.cell.2015.05.044. 786–788
54. Levine DA, Getz G, Gabriel SB, Cibulskis K, Lander E, Sivachenko A, et al. Integrated genomic characterization of endometrial carcinoma. *Nature*. 2013;497(7447):67–73. doi:10.1038/nature12113. 789–791
55. Mermel CH, Schumacher SE, Hill B, Meyerson ML, Beroukheim R, Getz G. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*. 2011;12(4):R41. doi:10.1186/gb-2011-12-4-r41. 792–795
56. Janssen KJM, Donders ART, Harrell FE, Vergouwe Y, Chen Q, Grobbee DE, et al. Missing covariate data in medical research: To impute is better than to ignore. *Journal of Clinical Epidemiology*. 2010;63(7):721–727. doi:https://doi.org/10.1016/j.jclinepi.2009.12.008. 796–799
57. Guinney J, Dienstmann R, Wang X, de Reyniès A, Schlicker A, Soneson C, et al. The consensus molecular subtypes of colorectal cancer. *Nature Medicine*. 2015;21(11):1350–1356. doi:10.1038/nm.3967. 800–802
58. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *Journal of Clinical Oncology*. 2009;27(8):1160–1167. doi:10.1200/JCO.2008.18.1370. 803–805
59. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics*. 2012;28(6):882–883. doi:10.1093/bioinformatics/bts034. 806–809
60. Branco P, Ribeiro RP, Torgo L. UBL: an R package for Utility-based Learning. arXiv:160408079 [Preprint]. 2016. Available from: <https://arxiv.org/pdf/1604.08079.pdf>. 810–812
61. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic acids research*. 2015;43(7):e47–e47. doi:10.1093/nar/gkv007. 813–815
62. Reed S, Lee H, Anguelov D, Szegedy C, Erhan D, Rabinovich A. Training Deep Neural Networks on Noisy Labels with Bootstrapping. arXiv:14126596 [Preprint]. 2015. Available from: <https://arxiv.org/pdf/1412.6596.pdf>. 816–818
63. Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv:14126980 [Preprint]. 2014. Available from: <https://arxiv.org/pdf/1412.6980.pdf>. 819–820
64. Bergstra J, Komer B, Eliasmith C, Yamins D, Cox DD. Hyperopt: a Python library for model selection and hyperparameter optimization. 2015;8(1):014008. doi:10.1088/1749-4699/8/1/014008. 821–823

sample-wise correlations between input and reconstructed genome

organ BRCA COADREAD GBMLGG

Figure 3

Figure 4

Figure 5

Figure 6

Figure 7

Figure 8