CANCER

Biologically relevant integration of transcriptomics profiles from cancer cell lines, patient-derived xenografts, and clinical tumors using deep learning

Slavica Dimitrieva[1]*, Rens Janssens[1], Gang Li[1], Artur Szalata[1], Rajaraman Gopalakrishnan[2], Chintan Parmar[2], Audrey Kauffmann[1]†, Eric Y. Durand[1]†

Cell lines and patient-derived xenografts are essential to cancer research; however, the results derived from such models often lack clinical translatability, as they do not fully recapitulate the complex cancer biology. Identifying preclinical models that sufficiently resemble the biological characteristics of clinical tumors across different cancers is critically important. Here, we developed MOBER, Multi-Origin Batch Effect Remover method, to simultaneously extract biologically meaningful embeddings while removing confounder information. Applying MOBER on 932 cancer cell lines, 434 patient-derived tumor xenografts, and 11,159 clinical tumors, we identified preclinical models with greatest transcriptional fidelity to clinical tumors and models that are transcriptionally unrepresentative of their respective clinical tumors. MOBER allows for transformation of transcriptional profiles of preclinical models to resemble the ones of clinical tumors and, therefore, can be used to improve the clinical translation of insights gained from preclinical models. MOBER is a versatile batch effect removal method applicable to diverse transcriptomic datasets, enabling integration of multiple datasets simultaneously.

INTRODUCTION

Cancer cell lines and patient-derived tumor xenograft (PTX) models continue to play a critical role in preclinical cancer research and drug discovery (*1*–*3*). Thousands of cancer models have been established and propagated in vitro and in vivo in different laboratories, where they have been extensively used in preclinical settings to study the biology of cancer (*4*, *5*), to explore the vulnerabilities of cancer cells (*6*, *7*), to identify new biomarkers (*8*), and to test the efficacy of anticancer compounds (*2*, *9*). Enormous knowledge in cancer biology has been derived from the various experiments conducted on cancer models. Still, many findings from preclinical cancer research are not reproducible in clinical trials (*10*, *11*), and oncology drugs have the highest failure rate compared to compounds used in other disease areas (*12*). One of the major reasons to this lack of translatability is that the cancer models are not perfect, and because of their propagation and differences in growing conditions, they have altered over time, and it is not known how well they represent the biology of the tumors from which they were derived. In addition, many cancer models lack accurate clinical annotations and histopathological classification that are crucial for their utility in cancer research (*13*). For greater clinical translatability, identification of models that sufficiently resemble the biological characteristics and drug responses of patient tumors is of critical importance.

Large collections of molecular data from patient tumors and cancer models have been generated across different cancer types. The Broad-Novartis Cancer Cell Line Encyclopedia (CCLE) (*2*) contains molecular profiles of around 1000 cancer cell lines, which are extensively used as preclinical models for various tumor types in drug discovery studies. In addition, gene expression profiles of >400 PTX models are available via the Novartis Institutes for Biomedical Research Patient-derived Tumor Xenograft Encyclopedia (*9*). Comprehensive molecular characterization of primary and metastatic tumors along with clinical data from >11,000 patients are available from the The Cancer Genome Atlas (TCGA) (*14*), MET500 (*15*), and Count Me In (CMI) (*16*) projects. These efforts provide a powerful opportunity to unravel the systematic differences between cancer cell lines, xenograft models, and patient tumors and to identify the cancer models that sufficiently recapitulate the biology of patient tumors without relying on clinical annotations.

Gene expression profiling accurately reproduces histopathological classification of tumors and is a useful technique for resolving tumor subtypes (*17*–*20*). However, large-scale integration of molecular data from cancer cell lines, xenograft models, and patient tumors is challenging due to the mixture of intrinsic biological signals and technical artifacts. One key challenge is that gene expression measurements from bulk patient biopsy samples are confounded by the presence of human stromal and immune cell populations that are not present in cancer models. In addition, large public datasets can be confounded by hidden technical variables, even when they come from the same source type [e.g., RNA sequencing (RNA-seq) from different patient cohorts]. Existing approaches for removing batch effects do not account for other systematic differences between cancer models or patient tumors or assume that the cell line and tumor datasets have the same subtype composition (*21*, *22*). Previous studies analyzing the differences between cell lines and patient tumors based on transcriptomics profiles have primarily focused on selected cancer types (*23*–*25*). Existing global analysis with the Celligner method (*19*), which leverages a computational approach developed for batch correction of single-cell RNA-seq data, has compared cancer cell lines and patient tumors. However, this method is limited to aligning only two datasets simultaneously, and the Celligner alignment does not consider PTXs.

In recent years, a multitude of studies have used deep learning techniques to transcriptomics data analysis. Particularly, the focus has

[1] Disease Area Oncology, Novartis Institutes for Biomedical Research, CH-4002 Basel, Switzerland. [2] Disease Area Oncology, Novartis Institutes for Biomedical Research, Cambridge, MA, USA.
*Corresponding author. Email: slavica.dimitrieva@novartis.com
†These authors contributed equally to this work.

been on single-cell RNA-seq data, where different autoencoder-based architectures have been proposed and successfully used for data harmonization and mitigating confounding technical effects (*26*–*29*). Inspired by these results, we applied deep learning techniques to explore the fidelity of preclinical models as representatives of patient tumors.

Here, we developed a deep learning–based method, MOBER (Multi-Origin Batch Effect Remover), that performs biologically relevant integration of pan-cancer gene expression profiles from cancer cell lines, PTXs, and patient tumors simultaneously. MOBER can be used to guide the selection of cell lines and patient-derived xenografts and identify models that more closely resemble patient tumors. We applied it to integrate transcriptomics data from 932 cancer cell lines, 434 PTXs, and 11,159 patient tumors from TCGA, MET500, and CMI, without relying on cancer type labels. We developed a web application that provides a valuable resource to help researchers select preclinical models with greatest transcriptional fidelity to clinical tumors. MOBER can be broadly applied as a batch effect removal tool for any transcriptomics datasets, and we made the method available as an open-source Python package.

RESULTS

The MOBER method

MOBER is an adversarial conditional variational autoencoder (VAE) that generates biologically informative gene expression embeddings robust to confounders (Fig. 1). MOBER consists of two neural networks (see Materials and Methods for more details). The first is a conditional VAE (*30*, *31*) that is optimized to generate embeddings that can reconstruct the original input. The second is an adversarial neural network (aNN) (*32*) that takes the embedding generated by the VAE as an input and tries to predict the origin of the input data. The VAE consists of an encoder that takes as an input a gene expression profile of a sample and encodes it as a distribution into a low dimensional latent space and a decoder that takes an embedding sampled from that distribution and the origin sample labels to reconstruct the gene expression profile from it. The goal is to learn an embedding space that encodes as much information as possible on the input samples while not encoding any information on the origin of the sample. To achieve this, we train the VAE and aNN simultaneously. The VAE tries to successfully reconstruct the data while also preventing the aNN from accurately predicting the data source. This way, during training, the VAE and the aNN will converge and reach an equilibrium, such that the VAE will generate an embedding space that is optimally successful at input reconstruction and the aNN will only randomly predict the origin of the input data from this embedding. In other words, the VAE will converge to generating an embedding that contains no information about the origin of the input data, while the aNN will converge to a random prediction performance. During the MOBER training process, we use one-hot representation of the origin of the input samples (either TCGA, CCLE, PTX, MET500, or CMI) and provide that information to the decoder. This enables the VAE to decode the data embeddings conditionally and reconstruct the expression data as if the sample was coming from another source type. To project a transcriptomics profile from one origin (e.g., preclinical) into another (e.g., clinical), after the model is trained, we can pass that transcriptomics profile through the VAE and simply change the one-hot vector informing on the sample origin to the desired one (e.g., CCLE to TCGA to decode cell lines as if they were TCGA tumors).

Global pan-cancer alignment of transcriptional profiles from cancer cell lines, PTXs, and patient tumors

We analyzed the transcriptomics profiles from 932 CCLE cell lines, 434 PTXs, 10,550 patient tumors from TCGA, 406 metastatic tumors from MET500 (*33*), and 203 breast tumors from CMI (*16*). Integrating these datasets by performing dimensionality reduction with two-dimensional Uniform Manifold Approximation and Projection (UMAP) reveals a clear separation of samples based on their origin (Fig. 2A). As expected, there is a global separation between cell lines, xenografts, and patient tumors. In addition, there are still strong batch effects between CMI, TCGA, and MET500 datasets, despite these samples all being derived from patient biopsies.

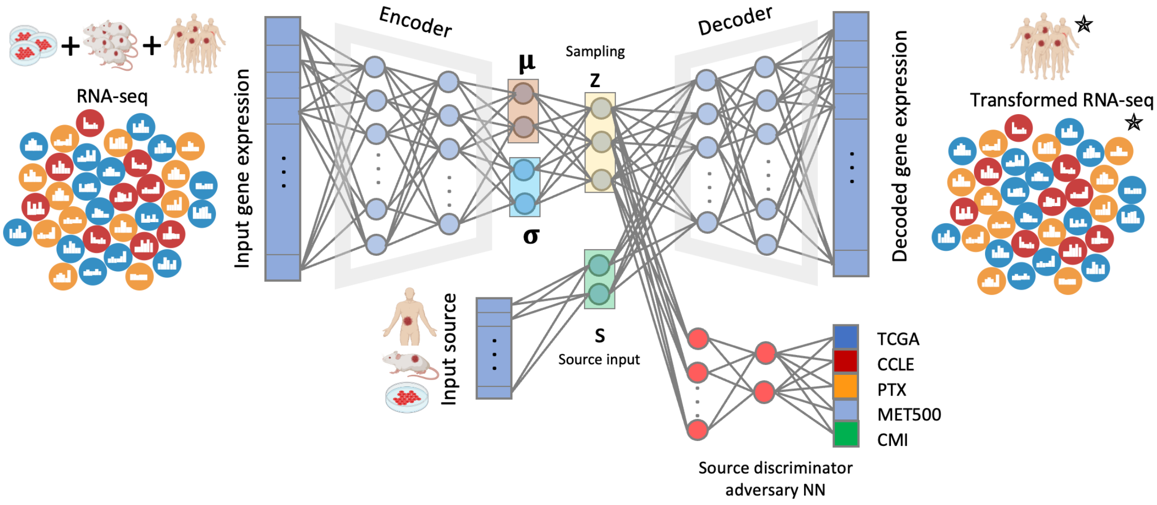

Fig. 1. MOBER architecture. MOBER consists of two neural networks: one conditional variational autoencoder (VAE) and one source discriminator neural network that is trained in adversarial fashion. The encoder takes as an input a gene expression profile and encodes it into a latent space, and the decoder takes a sampling from the latent space and reconstructs the gene expression profile from it. The source discriminator adversary neural network takes the sampling from the latent space and tries to pre-dict the origin of the input data. Parts of this figure were created with BioRender.com.

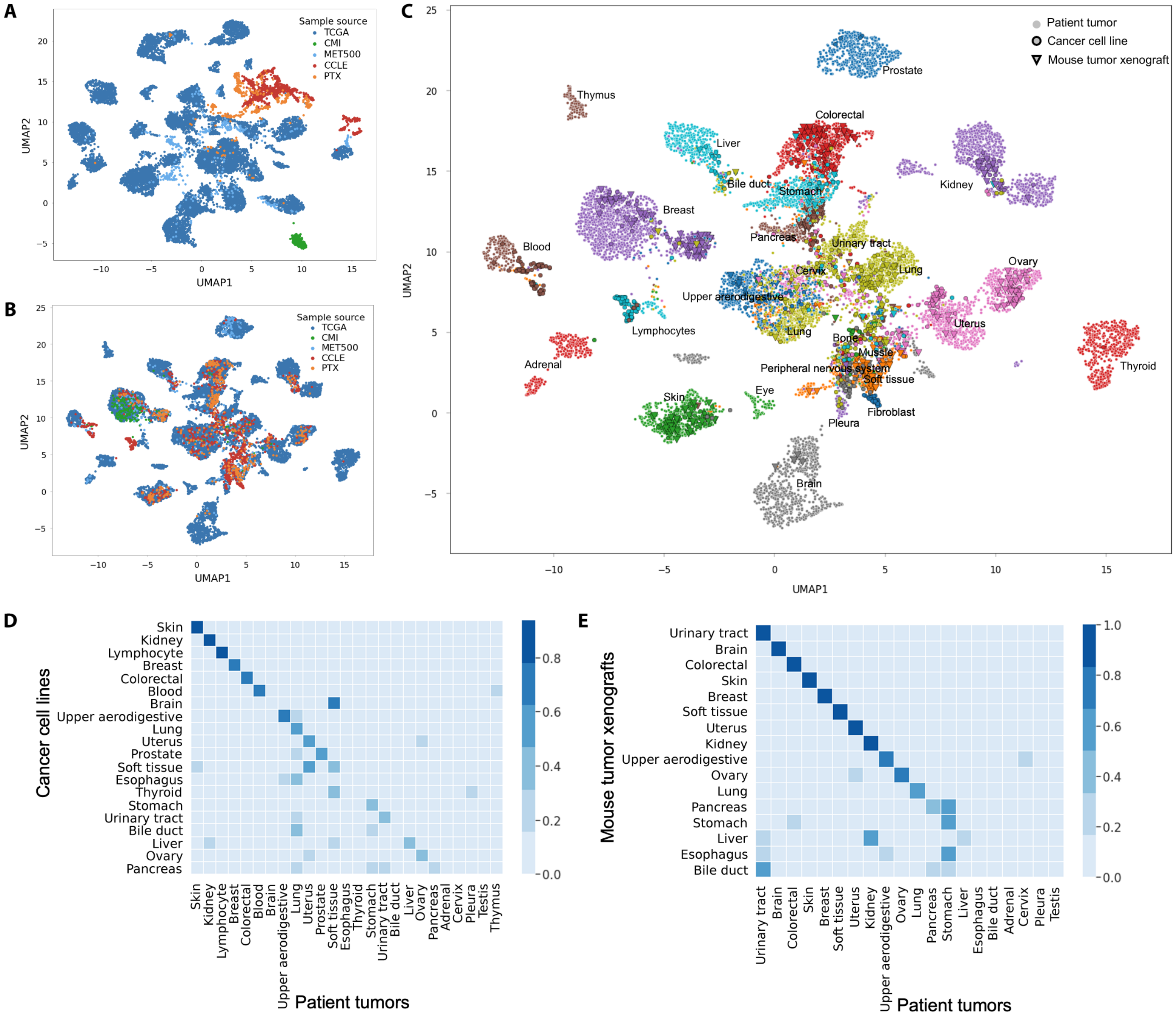

Fig. 2. Global pan-cancer alignment of preclinical and clinical transcriptomes. (**A**) Integration of transcriptional profiles from models and patient tumors by performing Uniform Manifold Approximation and Projection (UMAP) dimensionality reduction, each dot is a sample. (**B**) Integration of transcriptional profiles from models and patient tumors using MOBER, the color corresponds to the sample origin. (**C**) MOBER alignment, where each tumor sample is colored on the basis of cancer indication. (**D**) The proportion of cancer cell lines that are classified as each tumor type using MOBER-aligned data. (**E**) The proportion of PTXs that are classified as each tumor type using MOBER-aligned data. The *x* axis on (D) and (E) shows the TCGA tumor types, and the *y* axis shows the CCLE annotation label (D) and PTX annotation label (E), accordingly.

Aligning the CCLE, PTX, TCGA, MET500, and CMI datasets with MOBER resulted in a well-integrated dataset of transcriptional profiles from cell lines, xenografts, and patient tumors that have been corrected for multiple sources of systematic dataset-specific differences. The UMAP plots using the MOBER-aligned data (Fig. 2, B and C) reveal a map of cancer transcriptional profiles where cell lines, xenografts, and patient tumor samples are largely intermixed, while the biological differences across known cancer types are still preserved (Fig. 2C). The strong batch effects between cell lines, xenografts, and patient tumors were not addressed by applying other widely used batch correction methods, such as ComBat (*21*, *22*), Harmony (*34*), Batch Mean Centering (*35*), and the Regress_Out algorithm as implemented in scanpy (fig. S1) (*36*). Comparisons using simulated data further highlight the superiority of MOBER in data integration, particularly in presence of strong batch effects (see Supplementary Text and figs. S2 to S7).

As illustrated in Fig. 2B, MOBER removes the systematic differences between patient tumors and cancer models, as well as the technical artifacts present in patient tumors coming from different sources (CMI, MET500, and TCGA), producing an integrated cancer expression space with clear clusters composed of mixture of cell lines, xenografts, and patient tumor samples. Figure 2C shows that the aligned expression profiles largely cluster together by disease type, although MOBER does the alignment in a completely unsupervised manner, without relying on any sample annotations such as disease type. We quantified this by classifying the most similar tumor type for each cell line and xenograft model, based on its nearest neighbors among the TCGA tumor samples (see Materials and Methods). We found that, for 73% of the PTX models, the inferred disease type matches the annotated PTX tumor type, while this number goes down to 53% for CCLE models.

A key advantage of MOBER is that it does not assume that any two datasets are necessarily similar to each other, and it does not rely on clinical annotations of individual tumor samples, independently of whether they come from patient biopsies or preclinical models. As a result, the MOBER-aligned expression data can be used to identify which preclinical models have the greatest transcriptional fidelity to clinical tumors and which models are transcriptionally unrepresentative of their respective clinical tumors. Although a high proportion of preclinical models clusters with tumors of the same cancer type, not all cell lines and xenograft models align well with patient tumor samples. Figure 2 (D and E) shows that a significant proportion of cancer models (both PTXs and CCLEs), derived from skin, colorectal, and breast cancer, are faithful representatives of patient tumors, while many models derived from liver, esophagus, and bile duct tend to align with other cancer types. This observation is in agreement with Celligner results on cancer cell lines (*19*). PTX models derived from brain tumors align very well with brain cancer patient biopsies; however, brain cell line models cluster closely to soft tissue patient tumors, but not to clinical brain tumors. This is in line with previous studies that show that in vitro medium conditions cause genomic alterations in brain cell lines that were not present in the original tumors, thus altering their phenotypes (*37*, *38*).

Metastatic tumors from the MET500 dataset tend to cluster together with their corresponding primary tumors from TCGA, although the tissue of biopsy of MET500 tumors is different from the primary site (fig. S8). In 63% of MET500 samples, the inferred disease type matches the annotated primary tumor type. We note that, for 88 samples from the MET500 dataset, the primary site annotation is missing (such samples are shown in black squares in fig. S8). However, the majority of these samples align nicely within TCGA clusters, indicating that the unsupervised pan-cancer alignment with MOBER can be used to infer the primary site for tumors of unknown origin when the transcriptomics profiles are available.

Another key feature of MOBER is that it allows for populations that are only present in one dataset to be aligned correctly in an unsupervised manner. For example, the CMI dataset contains only transcriptional profiles of patient biopsies with metastatic breast cancer. MOBER integrated this dataset very nicely with the other breast patient tumors from TCGA and MET500, as well as breast cancer cell lines and xenograft models (Fig. 2B).

Preservation of biological subtype relationships in the MOBER alignment

We next sought to determine whether the MOBER alignment keeps known biological differences between more granular cancer subtypes. Focusing on the breast cancer samples, we determined subtype annotations with the PAM50 method (see Materials and Methods). Figure 3A shows that breast cancer patient biopsies, breast cell lines, and xenograft models primarily cluster together by breast cancer subtype [LumA, LumB, normal, human epidermal growth factor receptor 2 (HER2)–enriched, or basal], with only a few PTX basal subtype models clustering elsewhere (see also fig. S9).

MOBER is fully interpretable by design, which allows us to study the changes in the transcriptomics profiles of preclinical models after their in silico transformation to clinical tumors. In this respect, we

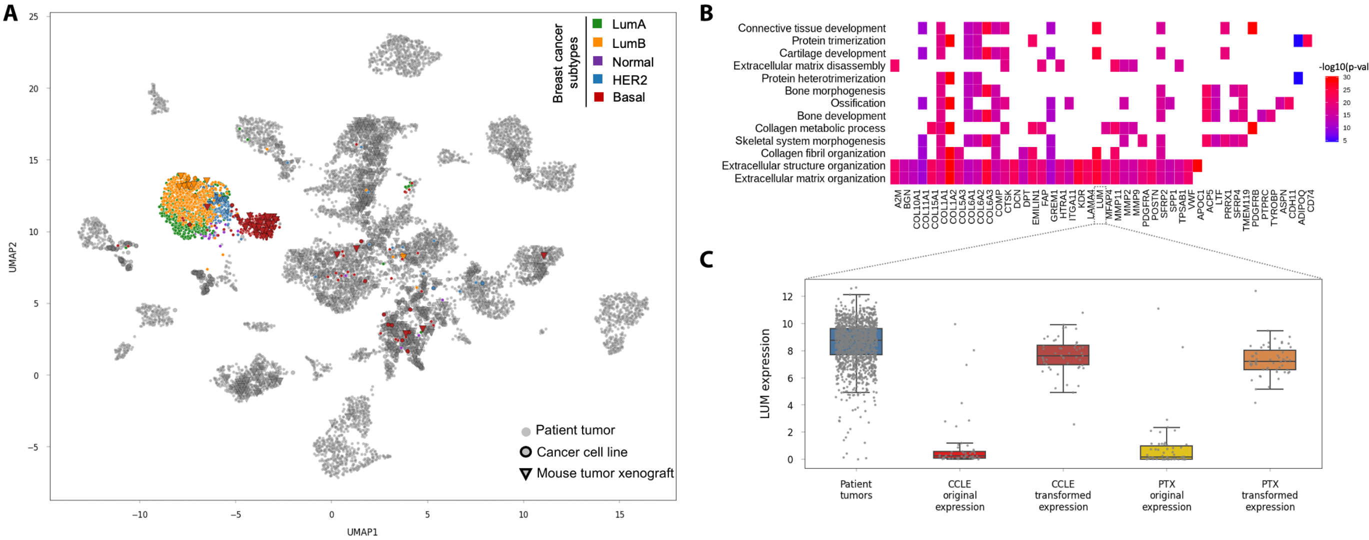

Fig. 3. Alignment of breast cancer subtypes. (**A**) UMAP two-dimensional projection of the MOBER-alignment highlighting breast tumor samples: LumA (green), LumB (orange), normal (purple), HER2-enriched (blue), and basal (red). All other non–breast tumor samples are in gray. (**B**) Genes that were most significantly up-regulated in silico after the alignment of breast cancer cell lines to breast cancer patient biopsies (*x* axis), along with top enriched biological pathways involving the 100 most up-regulated genes (*y* axis). (**C**) Expression values ($\log_2$ counts per million) of a selected gene, *Lumican* (*LUM*), in breast cancer patient tumors (blue), breast cancer cell lines before the alignment (bright red), breast cancer cell lines after the MOBER alignment (dark red), breast xenograft models before the alignment (yellow), and breast xenograft models after the MOBER alignment (orange).

examined the genes that were significantly changed when breast cancer cell lines were aligned to patient biopsies. Figure 3B shows the genes that were most significantly up-regulated after the alignment, along with enriched biological pathways involving these genes. Almost all top enriched biological pathways (e.g., extracellular matrix organization, collagen fibril organization, and connective tissue development) are related to the high presence of stromal tumor microenvironment in breast cancer patient biopsies, which is missing in cell lines (see also fig. S9). Figure 3C illustrates the changes in the expression values of a selected gene, *Lumican* (*LUM*), in preclinical models before and after their in silico transformation. *Lumican* is involved in extracellular matrix organization and connective tissue development processes (among others) and is highly expressed in stromal components in breast cancer patient biopsies, while it has a low expression in preclinical models as they are missing the human stromal component. After transforming the breast preclinical models to resemble breast patient biopsies, we see that the *LUM* expression was increased in the transformed data. Similarly, when transforming blood cancer cell lines to patient biopsies, the most significantly up-regulated genes after the alignment are related to the presence of human immune components in blood cancer patient biopsies that are missing in cell lines (fig. S10, B and C). Together, these results highlight that, during the projection of preclinical models to clinical biopsies, MOBER does not correct genes at random. Instead, it up-regulates in silico the genes that are expressed in the human tumor microenvironment, which is present in clinical biopsies but missing in preclinical models, while still preserving the key biological information on tumor subtypes at a very granular level.

Information transfer between cell line and patient tumor datasets

Next, we demonstrate how the MOBER-transformed gene expression profiles of preclinical models into clinical tumors can be used in other studies where we seek to translate preclinical biomarkers to patients. The Broad Institute recently published a large metastasis map dataset (MetMap) (*15*), determining the metastatic potential of ~500 human cancer cell lines, thus enabling the metastatic patterns of cell lines to be associated with their genomic features. Here, we sought to identify transcriptomics features that are associated with high or low metastatic potential in human cancer cell lines. We built machine learning (ML) models that take as input gene expression profiles of cancer cell lines and try to predict their average metastatic potential toward five different organs, as provided by the MetMap dataset (Fig. 4A; see Materials and Methods). Next, we used the

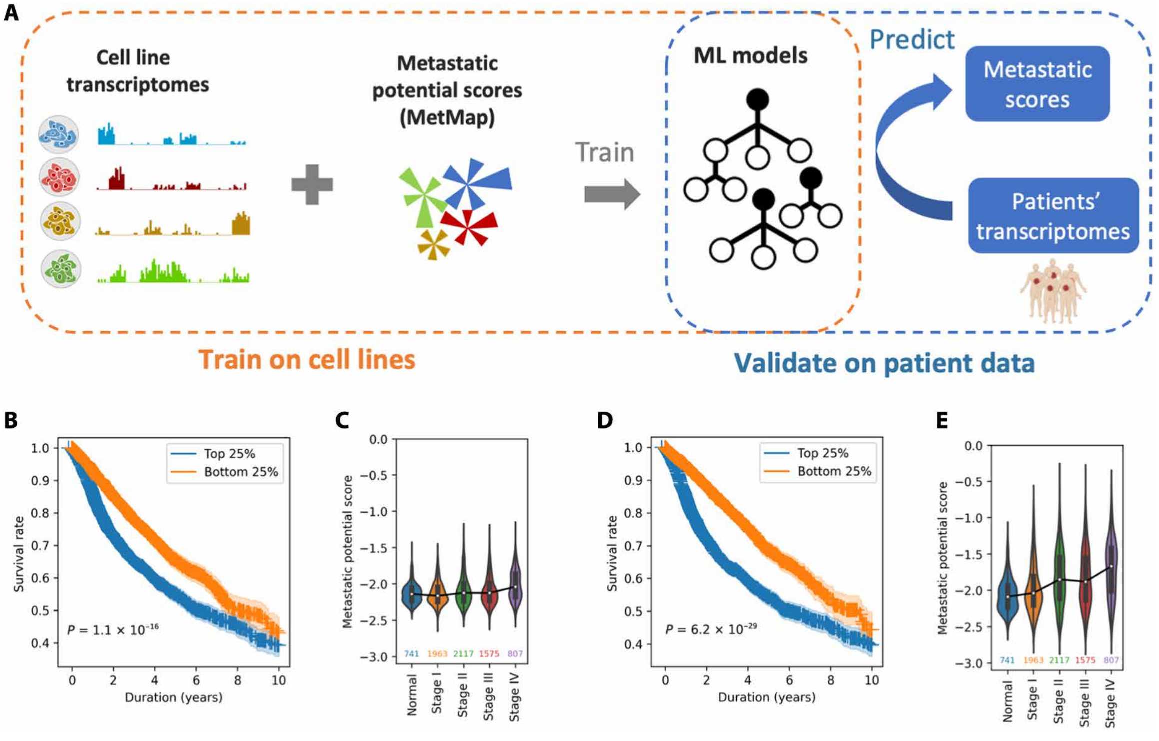

Fig. 4. Associating biomarkers of high/low metastatic potential in human cancer cell lines from MetMap and translating them to patients. (**A**) Using gene expression profiles of cancer cell lines and their experimentally derived metastatic potential scores from the MetMap dataset, we built machine learning (ML) models to predict the metastatic potential of cell lines based on expression data. Then, we translated these models trained on cell lines to patients, trying to predict the metastatic potential of TCGA patient tumors using patients' transcriptomes. Patients whose tumors are predicted to have higher metastatic potential are expected to have worse survival and a more advanced stage of the disease. (**B** and **C**) The association of the predicted metastatic potential with patients' survival and disease stage when ML models are trained on original cell line expression profiles from the CCLE. (B) Difference in survival of TCGA patient tumors for which we predict very high metastatic potential (top 25%, blue) versus low metastatic potential (bottom 25%, orange) with ML models trained on original cell line expression profiles. *P* values are derived from the log-rank test, and shaded areas indicate 90% of confidence intervals. (C) Predicted metastatic potential of TCGA tumors for different clinical stages. No correlation is observed. (**D** and **E**) The same as (B) and (C) but with ML models trained on MOBER-transformed cell line expression profiles to resemble TCGA patients. These models translate better to patient tumors, as evidenced by the improved survival stratification and significant positive correlation between the predicted metastatic potential and disease stage (Spearman's $r = 0.90$, $P = 0.037$)

models trained on cell line expression data to predict the metastatic potential scores of patient tumors from TCGA. We examined whether there is a difference in the survival between patients for which we predict high metastatic potential versus low metastatic potential. In addition, we determined whether there is any correlation between the predicted metastatic scores and the clinical stage of patient tumors. Using the ML models that are trained on original gene expression profiles of CCLE cell lines, we see that there is a significant difference in the survival of TCGA patient tumors for which we predict very high metastatic potential (top 25%) versus low metastatic potential (bottom 25%) ($P = 1.1 \times 10^{-16}$) (Fig. 4B). However, there is no significant association with clinical stage of TCGA patient tumors (Fig. 4C, *P* value of Spearman's correlation is 0.1). Then, we built new ML models that can predict the metastatic potential scores, but, this time, we trained them on gene expression profiles of cell lines that are transformed with MOBER to resemble TCGA tumors. Applying these ML models on TCGA patient tumors, we achieve more significant survival stratification of TCGA patient tumors for which we predict very high metastatic potential (top 25%) versus low metastatic potential (bottom 25%) ($P = 6.2\times 10^{-29}$) (Fig. 4D). In addition, we note that such models predict higher metastatic potential of the late-stage tumors, compared to early-stage tumors (Fig. 4E, Spearman's correlation $r = 0.90$, $P = 0.037$). The same analyses performed separately for each disease type confirm the improved translatability of the ML models when they are trained on MOBER-transformed cell line transcriptomes to resemble patient tumors (fig. S11). The ML models that we used here might be too simple to faithfully infer the metastatic potential of tumors; however, our results demonstrate the utility of using the MOBER-transformed gene expression profiles in finding biomarkers that are better translatable to patients.

DISCUSSION

Preclinical cancer research critically relies on in vitro and in vivo tumor models, such as cell lines and PTXs. However, because of the differences in the growing conditions and the absence of the human stromal microenvironment, these models are often not predictive of the drug response in clinical tumors and do not follow the same pathways of drug resistance (*10*, *39*). Therefore, the identification of the best models for a given cancer type, without relying on annotated disease labels, is critically important. To address this, we developed MOBER, a deep learning–based method that performs biologically relevant integration of transcriptional profiles from various preclinical models and clinical tumors. MOBER can be used to guide the selection of cancer cell lines and patient-derived xenografts and identify models that more closely resemble clinical tumors. We integrated gene expression data from 932 cancer cell lines, 434 PTXs, and 11,159 clinical tumors simultaneously and demonstrate that MOBER can conserve the inherent biological signals while removing confounder information.

We identified pronounced differences across cancer models in how well they recapitulate the transcriptional profiles of their corresponding tumors in patients. Certain cancer types, such as the ones in skin, breast, colorectal, kidney, and uterus, exhibit greater transcriptional similarity between models and patients, while models of cancers of the bile duct, liver, and esophagus have transcriptional profiles that are unlike their patient tumors. There are also notable differences between CCLE and PTX as cancer model systems. While brain and soft tissue CCLEs appear to have diverged from their corresponding patient tumors, brain and soft tissue PTX models have high transcriptional fidelity. Notably, among the PTX models, urinary tract xenografts attain greatest transcriptional similarity to corresponding patient tumors; however, many of the urinary tract CCLEs have drifted away from their labeled tumors. As pointed in previous studies (*19*, *40*), many CCLEs are not classified as their annotated labels. We report that PTX models, on average, show greater transcriptional similarity to patient tumors, as compared to their CCLE counterparts. This could suggest that the lack of immune component is not the main CCLE confounder, but CCLEs likely undergo transcriptional divergence due to the culture condition, high number of passages, and genetic instability (*41*, *42*). In addition, cancer models (both PTXs and CCLEs) could be misannotated because of inaccurate assignment based on unclear anatomical features or mismatch during sampling (*40*, *43*). Our pan-cancer global alignment with MOBER could be used to identify cancers that are underserved by adequate preclinical models and to determine the primary site label for models and clinical tumors where such annotation is missing.

MOBER is interpretable by design, therefore allowing drug hunters to better understand the underlying biological differences between models and patients that are responsible for the observed lack of clinical translatability. The observed differences vary across disease types, emphasizing the importance of using unsupervised nonlinear approach that enables identification of disease-type-specific variations.

As a batch effect removal method, MOBER offers several advantages compared to previously published methods (*19*, *21*, *44*): It (i) supports integration of multiple datasets simultaneously, (ii) enables transformation of one dataset into another, and (iii) does not make any assumption on the datasets composition. We demonstrate that MOBER can remove batch effects between gene expression datasets even when the cell population representation across datasets is different.

We note that, when transforming preclinical models to resemble clinical tumors, MOBER corrects not only for hidden technical differences but also for differences due to the absence of the tumor microenvironment in preclinical models. While this correction is crucial for aligning preclinical models with clinical tumors and identification of outlier models, this may not be desirable in studies aimed at investigating the inherent biological disparities between models and patient tumors. With the in silico addition of the tumor microenvironment to cell lines, some components that are inherent to cell culture (e.g., extracellular matrix genes in stromal reach tumors) would be affected. However, the differences in these cell components between the different cell lines are still maintained after their in silico transformation to clinical tumors, as evidenced by the meaningful alignment of disease types and subtypes in a completely unsupervised way.

To facilitate the use of MOBER, we made the source code available at https://github.com/Novartis/MOBER and developed an interactive web app available at https://mober.pythonanywhere.com to allow users to explore the MOBER aligned expression profiles coming from cancer models and clinical tumors. In addition, the web app enables the identification of preclinical models that best represent the transcriptional features of a tumor type or even a particular tumor subtype of interest. Future version of MOBER that integrates genetic and epigenetic features of models and patient tumors could potentially enable even more detailed analysis between models and patients.

MATERIALS AND METHODS

The MOBER method and model training details

Each gene expression profile of a sample i is a vector $\mathbf{x}_i$ with length equal to the total number of genes. The input gene expression data x is run through a VAE that uses variational inference to reconstruct the original data in a conditional manner. The encoder estimates the probability density function of the input expression data $Q(z|x)$. Then, a latent vector z is sampled from $Q(z|x)$. The decoder decodes z into an output, learning the parameters of the distribution $P(x|z)$. The loss function is then given by

$$\text{Loss}_{\text{VAE}} = \underbrace{-E_{z\sim Q(z|x)}\left[\log P(x|z)\right]}_{\text{reconstruction error}} + \underbrace{w_{\text{KL}} * \text{KL}[Q(z|x) \parallel P(z)]}_{\text{regularization}}$$

where $\mathbf{x}$ and z indicate the gene expression data and latent space, respectively; Q and P are the estimated probability distributions; E denotes an expectation value; KL is the Kullback-Leibler (*45*) divergence; and w is a weight parameter that determines the importance given to the Kullback-Leibler divergence in the VAE loss function. The first term in the loss function is the reconstruction error (i.e., expected negative log-likelihood of the data sample), and the second term is the Kullback-Leibler divergence between the encoder's distribution $Q(z|x)$ and $P(z)$. In addition to the input expression data, we provide to the decoder an information about the origin of the input sample (in our case CCLE, PTX, TCGA, MET500, or CMI) transformed into a one-hot encoding vector $\mathbf{s}$, that consists of 0's in all cells and a single 1 in a cell used to uniquely identify the input source. This allows for reconstruction of the latent vector z by the decoder in a conditional manner, and it enables projection of one dataset into another.

In addition to training the VAE, we simultaneously train an aNN that acts as a source discriminator. It takes as an input an embedding vector z sampled from the latent space, and tries to predict the source label of the input data (s). This is a multi-class fully connected neural network with negative log-likelihood loss function as given by

$$\text{Loss}_{\text{aNN}} = -E_{Q(z|x)}\left[\log p(s|z)\right]$$

The joint loss is computed as

$$\text{Loss}_{\text{MOBER}} = \text{Loss}_{\text{VAE}} - \lambda * \text{Loss}_{\text{aNN}}$$

where λ is a coefficient that determines the weight that the model gives to the adversarial loss.

In our study, the encoder, decoder, and aNN are designed as fully connected neural networks each with three layers. Each layer of the encoder (and decoder respectively) consisted of 256, 128, and 64 nodes each. We used Scaled Exponential Linear Unit (*46*) activation function between two hidden layers, except the last decoder layer, where we applied Rectified Linear activation Unit (*47*). The last aNN layer had five hidden nodes corresponding to the number of data source classes and softmax activation.

We implemented MOBER using PyTorch (*48*). We set the minibatch size to 1600 and trained it with Adam optimizer (*49*) using a learning rate of 1×10^{-3}. The weight for the KL loss of the VAE was set to 11×10^{-6}, and the weight for the source adversary loss was set to 11×10^{-2}. The best hyperparameter set from numerous possibilities was chosen from a grid search that minimized the joint loss and maximized the clustering performance of models to patients.

Datasets

The RNA-seq gene expression counts for TCGA samples were downloaded from the TCGA portal (https://tcga-data.nci.nih.gov/tcga) (*14*) and the CMI gene expression counts from the Genomic Data Commons portal (*50*), and CCLE (*2*), PTX (*9*), and MET500 (*33*) data were obtained from the corresponding publications. All gene expression data were normalized with Trimmed Mean of M-values (TMM) method using EdgeR (*51*) and then transformed to $\log_2$ counts per million using the edgeR function "cpm," with a pseudocount of 1 added. Gene expression data were subset to 17,167 protein-coding genes that were present in all datasets. The MetMap (*15*) dataset was downloaded from the DepMap portal (https://depmap.org/metmap). We excluded the indications that have less than five samples.

Alignment evaluation

We projected each sample from the CCLE, PTX, CMI, and MET500 datasets to TCGA, by changing the one-hot encoded source information and setting it to TCGA. Then, we decoded the expression data with MOBER.

To evaluate the alignment of preclinical samples to TCGA samples, for each CCLE and PTX sample, we identified the 25 TCGA nearest neighbors in 70-dimensional principal components analysis space. Each preclinical sample was classified as a tumor type by identifying the most frequently occurring tumor type within these 25 nearest neighbors.

The identification of breast cancer molecular subtypes was done using the PAM50 classifier as implemented in the geneFu R package (*52*). The identification of differentially expressed genes comparing the transcriptional profiles before and after their transformation to patient tumors was done with Seurat v3.6 (*53*) using the t-test method. Pathway enrichment analysis was done for the top 100 most differentially expressed genes ordered by their fold change and with an adjusted P value of <0.01 using the clusterProfiler (*54*) package.

Analyses of metastatic potential using MetMap data

We trained random forest models to predict the metastatic potential scores using gene expression values as input, using the scikit-learn (*55*) Python package (0.23.2). The hyperparameters were optimized with grid search strategy using threefold cross validations. Then, the final model was trained using the optimized hyperparameters. The hyperparameter optimized in the model is "max_features." One thousand trees were used, and all other hyperparameters were set as default.

Two different models were trained, using either the original transcriptome or the projected transcriptomes of CCLEs to TCGA patients to predict mean metastatic potential scores of cell lines across five organs (metp500.all5). Then, with each trained model, we predicted the metastatic potential scores for patient tumors from TCGA using the original transcriptome as input. For the survival analysis, the top 25% of samples with highest predicted scores and bottom 25% with the lowest predicted scores were compared. The Kaplan-Meier survival analysis (*56*) was done with the lifelines (*57*) Python package v0.25.4.

Supplementary Materials

This PDF file includes:

Supplementary Text
Figs. S1 to S11

REFERENCES AND NOTES

1. C. R. Ireson, M. S. Alavijeh, A. M. Palmer, E. R. Fowler, H. J. Jones, The role of mouse tumour models in the discovery and development of anticancer drugs. *Br. J. Cancer.* **121**, 101–108 (2019).
2. J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin, A. Reddy, M. Liu, L. Murray, M. F. Berger, J. E. Monahan, P. Morais, J. Meltzer, A. Korejwa, J. Jané-Valbuena, F. A. Mapa, J. Thibault, E. Bric-Furlong, P. Raman, A. Shipway, I. H. Engels, J. Cheng, G. K. Yu, J. Yu, P. Aspesi, M. de Silva, K. Jagtap, M. D. Jones, L. Wang, C. Hatton, E. Palescandolo, S. Gupta, S. Mahan, C. Sougnez, R. C. Onofrio, T. Liefeld, L. MacConaill, W. Winckler, M. Reich, N. Li, J. P. Mesirov, S. B. Gabriel, G. Getz, K. Ardlie, V. Chan, V. E. Myer, B. L. Weber, J. Porter, M. Warmuth, P. Finan, J. L. Harris, M. Meyerson, T. R. Golub, M. P. Morrissey, W. R. Sellers, R. Schlegel, L. A. Garraway, The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).
3. J. P. Gillet, S. Varma, M. M. Gottesman, The clinical relevance of cancer cell lines. *J. Natl. Cancer Inst.* **105**, 452–458 (2013).
4. A. Tsherniak, F. Vazquez, P. G. Montgomery, B. A. Weir, G. Kryukov, G. S. Cowley, S. Gill, W. F. Harrington, S. Pantel, J. M. Krill-Burger, R. M. Meyers, L. Ali, A. Goodale, Y. Lee, G. Jiang, J. Hsiao, W. F. J. Gerath, S. Howell, E. Merkel, M. Ghandi, L. A. Garraway, D. E. Root, T. R. Golub, J. S. Boehm, W. C. Hahn, Defining a cancer dependency map. *Cell* **170**, 564–576.e16 (2017).
5. M. Ghandi, F. W. Huang, J. Jané-Valbuena, G. V. Kryukov, C. C. Lo, E. Robert Mc Donald III, J. Barretina, E. T. Gelfand, C. M. Bielski, H. Li, K. Hu, A. Y. Andreev-Drakhlin, J. Kim, J. M. Hess, B. J. Haas, F. Aguet, B. A. Weir, M. V. Rothberg, B. R. Paolella, M. S. Lawrence, R. Akbani, Y. Lu, H. L. Tiv, P. C. Gokhale, A. de Weck, A. A. Mansour, C. Oh, J. Shih, K. Hadi, Y. Rosen, J. Bistline, K. Venkatesan, A. Reddy, D. Sonkin, M. Liu, J. Lehar, J. M. Korn, D. A. Porter, M. D. Jones, J. Golji, G. Caponigro, J. E. Taylor, C. M. Dunning, A. L. Creech, A. C. Warren, J. M. M. Farland, M. Zamanighomi, A. Kauffmann, N. Stransky, M. Imielinski, Y. E. Maruvka, A. D. Cherniack, A. Tsherniak, F. Vazquez, J. D. Jaffe, A. A. Lane, D. M. Weinstock, C. M. Johannessen, M. P. Morrissey, F. Stegmeier, R. Schlegel, W. C. Hahn, G. Getz, G. B. Mills, J. S. Boehm, T. R. Golub, L. A. Garraway, W. R. Sellers, Sellers, next-generation characterization of the cancer cell line encyclopedia. *Nature* **569**, 503–508 (2019).
6. F. Y. Feng, L. A. Gilbert, Lethal clues to cancer-cell vulnerability. *Nature* **568**, 463–464 (2019).
7. E. R. McDonald III, A. de Weck, M. R. Schlabach, E. Billy, K. J. Mavrakis, G. R. Hoffman, D. Belur, D. Castelletti, E. Frias, K. Gampa, J. Golji, I. Kao, L. Li, P. Megel, T. A. Perkins, N. Ramadan, D. A. Ruddy, S. J. Silver, S. Sovath, M. Stump, O. Weber, R. Widmer, J. Yu, K. Yu, Y. Yue, D. Abramowski, E. Ackley, R. Barrett, J. Berger, J. L. Bernard, R. Billig, S. M. Brachmann, F. Buxton, R. Caothien, J. X. Caushi, F. S. Chung, M. Cortés-Cros, R. S. de Beaumont, C. Delaunay, A. Desplat, W. Duong, D. A. Dwoske, R. S. Eldridge, A. Farsidjani, F. Feng, J. J. Feng, D. Flemming, W. Forrester, G. G. Galli, Z. Gao, F. Gauter, V. Gibaja, K. Haas, M. Hattenberger, T. Hood, K. E. Hurov, Z. Jagani, M. Jenal, J. A. Johnson, M. D. Jones, A. Kapoor, J. Korn, J. Liu, Q. Liu, S. Liu, Y. Liu, A. T. Loo, K. J. Macchi, T. Martin, G. M. Allister, A. Meyer, S. Mollé, R. A. Pagliarini, T. Phadke, B. Repko, T. Schouwey, F. Shanahan, Q. Shen, C. Stamm, C. Stephan, V. M. Stucke, R. Tiedt, M. Varadarajan, K. Venkatesan, A. C. Vitari, M. Wallroth, J. Weiler, J. Zhang, C. Mickanin, V. E. Myer, J. A. Porter, A. Lai, H. Bitter, E. Lees, N. Keen, A. Kauffmann, F. Stegmeier, F. Hofmann, T. Schmelzle, W. R. Sellers, Project DRIVE: A compendium of cancer dependencies and synthetic lethal relationships uncovered by large-scale, deep RNAi screening. *Cell* **170**, 577–592.e10 (2017).
8. Y. H. Huang, C. R. Vakoc, A biomarker harvest from one thousand cancer cell lines. *Cell* **166**, 536–537 (2016).
9. H. Gao, J. M. Korn, S. Ferretti, J. E. Monahan, Y. Wang, M. Singh, C. Zhang, C. Schnell, G. Yang, Y. Zhang, O. A. Balbin, S. Barbe, H. Cai, F. Casey, S. Chatterjee, D. Y. Chiang, S. Chuai, S. M. Cogan, S. D. Collins, E. Dammassa, N. Ebel, M. Embry, J. Green, A. Kauffmann, C. Kowal, R. J. Leary, J. Lehar, Y. Liang, A. Loo, E. Lorenzana, E. Robert McDonald, M. E. McLaughlin, J. Merkin, R. Meyer, T. L. Naylor, M. Patawaran, A. Reddy, C. Röelli, D. A. Ruddy, F. Salangsang, F. Santacroce, A. P. Singh, Y. Tang, W. Tinetto, S. Tobler, R. Velazquez, K. Venkatesan, F. von Arx, H. Q. Wang, Z. Wang, M. Wiesmann, D. Wyss, F. Xu, H. Bitter, P. Atadja, E. Lees, F. Hofmann, E. Li, N. Keen, R. Cozens, M. R. Jensen, N. K. Pryer, J. A. Williams, W. R. Sellers, High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat. Med.* **21**, 1318–1325 (2015).
10. A. A. Seyhan, Lost in translation: The valley of death across preclinical and clinical divide–Identification of problems and overcoming obstacles. *Transl. Med. Commun.* **4**, 18 (2019).
11. A. Honkala, S. V. Malhotra, S. Kummar, M. R. Junttila, Harnessing the predictive power of preclinical models for oncology drug development. *Nat. Rev. Drug Discov.* **21**, 99–114 (2022).
12. R. K. Harrison, Phase II and phase III failures: 2013–2015. *Nat. Rev. Drug Discov.* **15**, 817–818 (2016).
13. A. de Weck, H. Bitter, A. Kauffmann, Fibroblasts cell lines misclassified as cancer cell lines. bioRxiv:166199 [Preprint] (2017); https://doi.org/10.1101/166199.
14. K. Tomczak, P. Czerwińska, M. Wiznerowicz, The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. *Contemp. Oncol.* **19**, A68–A77 (2015).
15. X. Jin, Z. Demere, K. Nair, A. Ali, G. B. Ferraro, T. Natoli, A. Deik, L. Petronio, A. A. Tang, C. Zhu, L. Wang, D. Rosenberg, V. Mangena, J. Roth, K. Chung, R. K. Jain, C. B. Clish, M. G. Vander Heiden, T. R. Golub, A metastasis map of human cancer cell lines. *Nature* **588**, 331–336 (2020).
16. N. Wagle, C. Painter, E. M. van Allen, A. J. Bass, E. Anastasio, M. Dunphy, M. McGillicuddy, R. Stoddard, S. Balch, B. Thomas, B. N. Tomson, C. Nguyen, E. Jain, S. Wankowicz, J. Palma, S. Maiwald, E. O. Baker, A. Zimmer, T. Golub, E. Lander, Count me in: A patient-driven research initiative to accelerate cancer research. *J. Clin. Oncol.* **36**, 15 (2018).
17. T. Sørlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, T. Hastie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, T. Thorsen, H. Quist, J. C. Matese, P. O. Brown, D. Botstein, P. E. Lønning, A. L. Børresen-Dale, Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U.S.A.* **98**, 10869–10874 (2001).
18. J. Lapointe, C. Li, J. P. Higgins, M. van de Rijn, E. Bair, K. Montgomery, M. Ferrari, L. Egevad, W. Rayford, U. Bergerheim, P. Ekman, A. M. DeMarzo, R. Tibshiran, D. Botstein, P. O. Brown, J. D. Brooks, J. R. Pollack, Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 811–816 (2004).
19. A. Warren, Y. Chen, A. Jones, T. Shibue, W. C. Hahn, J. S. Boehm, F. Vazquez, A. Tsherniak, J. M. McFarland, Global computational alignment of tumor and cell line transcriptional profiles. *Nat. Commun.* **12**, 22 (2021).
20. E. Blaveri, J. P. Simko, J. E. Korkola, J. L. Brewer, F. Baehner, K. Mehta, S. DeVries, T. Koppie, S. Pejavar, P. Carroll, F. M. Waldman, Bladder cancer outcome and subtype classification by gene expression. *Clin. Cancer Res.* **11**, 4044–4045 (2005).
21. W. E. Johnson, C. Li, A. Rabinovic, Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
22. J. T. Leek, W. E. Johnson, H. S. Parker, A. E. Jaffe, J. D. Storey, The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).
23. C. Virtanen, Y. Ishikawa, D. Honjoh, M. Kimura, M. Shimane, T. Miyoshi, H. Nomura, M. H. Jones, Integrated classification of lung tumors and cell lines by expression profiling. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 12357–12362 (2002).
24. S. Domcke, R. Sinha, D. A. Levine, C. Sander, N. Schultz, Evaluating cell lines as tumour models by comparison of genomic profiles. *Nat. Commun.* **4**, 2126 (2013).
25. K. M. Vincent, L. M. Postovit, Investigating the utility of human melanoma cell lines as tumour models. *Oncotarget* **8**, 10498–10509 (2017).
26. N. Russkikh, D. Antonets, D. Shtokalo, A. Makarov, Y. Vyatkin, A. Zakharov, E. Terentyev, Style transfer with variational autoencoders is a promising approach to RNA-Seq data harmonization and analysis. *Bioinformatics* **36**, 5076–5085 (2020).
27. C. H. Grønbech, M. F. Vording, P. N. Timshel, C. K. Sønderby, T. H. Pers, O. Winther, ScVAE: Variational auto-encoders for single-cell gene expression data. *Bioinformatics* **36**, 4415–4422 (2020).
28. A. W. Lynch, M. Brown, C. A. Meyer, Multi-batch single-cell comparative atlas construction by deep learning disentanglement. *Nat. Commun.* **14**, 4126 (2023).
29. R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, N. Yosef, Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
30. D. P. Kingma, M. Welling, "Auto-encoding variational bayes" in *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings* (2014). https://doi.org/10.48550/arXiv.1312.6114
31. D. J. Rezende, S. Mohamed, D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models" in *31st International Conference on Machine Learning, ICML 2014* (2014), vol. 4.
32. Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-adversarial training of neural networks. *J Mach. Learn. Res.* **17**, 2096–2030 (2016).
33. D. R. Robinson, Y. M. Wu, R. J. Lonigro, P. Vats, E. Cobain, J. Everett, X. Cao, E. Rabban, C. Kumar-Sinha, V. Raymond, S. Schuetze, A. Alva, J. Siddiqui, R. Chugh, F. Worden, M. M. Zalupski, J. Innis, R. J. Mody, S. A. Tomlins, D. Lucas, L. H. Baker, N. Ramnath, A. F. Schott, D. F. Hayes, J. Vijai, K. Offit, E. M. Stoffel, J. S. Roberts, D. C. Smith, L. P. Kunju, M. Talpaz, M. Cieślik, A. M. Chinnaiyan, Integrative clinical genomics of metastatic cancer. *Nature* **548**, 297–303 (2017).
34. I. Korsunsky, N. Millard, J. Fan, K. Slowikowski, F. Zhang, K. Wei, Y. Baglaenko, M. Brenner, P. Loh, S. Raychaudhuri, Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).
35. A. H. Sims, G. J. Smethurst, Y. Hey, M. J. Okoniewski, S. D. Pepper, A. Howell, C. J. Miller, R. B. Clarke, The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets – improving meta-analysis and prediction of prognosis. *BMC Med. Genomics* **1**, 42 (2008).

36. F. A. Wolf, P. Angerer, F. J. Theis, SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
37. P. F. Ledur, G. R. Onzi, H. Zong, G. Lenz, Culture conditions defining glioblastoma cells behavior: What is the impact for novel discoveries? *Oncotarget* **8**, 69185–69197 (2017).
38. J. Gordon, S. Amini, M. K. White, General overview of neuronal cell culture. *Methods Mol. Biol.* **1078**, 1–8 (2013).
39. I. W. Y. Mak, N. Evaniew, M. Ghert, Lost in translation: Animal models and clinical trials in cancer treatment. *Am. J. Transl. Res.* **6**, 114–118 (2014).
40. D. Peng, R. Gleyzer, W. H. Tai, P. Kumar, Q. Bian, B. Isaacs, E. L. da Rocha, S. Cai, K. DiNapoli, F. W. Huang, P. Cahan, Evaluating the transcriptional fidelity of cancer models. *Genome. Med.* **13**, 73 (2021).
41. J. Lee, S. Kotliarova, Y. Kotliarov, A. Li, Q. Su, N. M. Donin, S. Pastorino, B. W. Purow, N. Christopher, W. Zhang, J. K. Park, H. A. Fine, Tumor stem cells derived from glioblastomas cultured in bFGF and EGF more closely mirror the phenotype and genotype of primary tumors than do serum-cultured cell lines. *Cancer Cell* **9**, 391–403 (2006).
42. S. L. Wenger, J. R. Senft, L. M. Sargent, R. Bamezai, N. Bairwa, S. G. Grant, Comparison of established cell lines at different passages by karyotype and comparative genomic hybridization. *Biosci. Rep.* **24**, 631–639 (2004).
43. M. Salvadores, F. Fuster-Tormo, F. Supek, Matching cell lines with cancer type and subtype of origin via mutational, epigenomic, and transcriptomic patterns. *Sci. Adv.* **6**, eaba1862 (2020).
44. A. B. Dincer, J. D. Janizek, S. I. Lee, Adversarial deconfounding autoencoder for learning robust gene expression embeddings. *Bioinformatics* **36**, i573–i582 (2020).
45. S. Kullback, R. A. Leibler, On information and sufficiency. *Annals Math. Stat.* **22**, 79–86 (1951).
46. G. Klambauer, T. Unterthiner, A. Mayr, S. Hochreiter, "Self-normalizing neural networks" in *Advances in Neural Information Processing Systems 30* (NIPS, 2017), pp. 972–981.
47. K. Hara, D. Saito, H. Shouno, "Analysis of function of rectified linear unit used in deep learning" in *International Joint Conference on Neural Networks* (IJCNN, 2015), pp. 1–8.
48. A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, "PyTorch: An imperative style, high-performance deep learning library" in *Advances in Neural Information Processing Systems 32* (NeurIPS, 2019), pp. 8026–8037.
49. D. P. Kingma, J. L. Ba, "Adam: A method for stochastic optimization" in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (2015). https://doi.org/10.48550/arXiv.1412.6980.
50. R. L. Grossman, A. P. Heath, V. Ferretti, H. E. Varmus, D. R. Lowy, W. A. Kibbe, L. M. Staudt, Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* **375**, 1109–1112 (2016).
51. M. D. Robinson, D. J. McCarthy, G. K. Smyth, edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2009).
52. D. Gendoo, N. Ratanasirigulchai, M. Schroder, L. Pare, J. Parker, A. Prat, B. Haibe-Kains, Genefu: An R/Bioconductor package for computation of gene expression-based signatures in breast cancer. *Bioinformatics* **32**, 1097–1099 (2016).
53. T. Stuart, A. Butler, P. Hoffman, C. Hafemeister, E. Papalexi, W. M. Mauck, Y. Hao, M. Stoeckius, P. Smibert, R. Satija, Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902.e21 (2019).
54. G. Yu, L. G. Wang, Y. Han, Q. Y. He, ClusterProfiler: An R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
55. F. Pedregosa, N. A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, F. Pedregosa, G. Varoquaux, A. Gramfort, B. Thirion, P. Prettenhofer, J. Vanderplas, M. Brucher, M. P. E. Duchesnay, A. M. Brucher, M. Perrot, C. F. E. Duchesnay, Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
56. E. L. Kaplan, P. Meier, Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **53**, 457–481 (1958).
57. C. Davidson-Pilon, lifelines: Survival analysis in Python. *J. Open Source Softw.* **4**, 10.21105/joss.01317 (2019).

Acknowledgments: We thank B. Gustafson for help with hosting the web app at https://mober.pythonanywhere.com. We thank Joshua Korn for helpful discussions. G. L. would like to acknowledge the support from Novartis Innovation Postdoctoral Fellowship program. **Funding:** This work was funded by Novartis Biomedical Research. **Author contributions:** Conceptualization: E.Y.D., S.D., R.J., R.G., C.P., and A.K. Methodology: S.D., R.J., E.Y.D., A.S., and C.P. Analyses: S.D., R.J., G.L., A.S., and C.P. Visualizations: S.D., R.J., G.L., and A.S. Software: G.L., S.D., R.J., A.S., and R.G. Project administration: S.D., E.Y.D., and A.K. Supervision: E.Y.D. and A.K. Writing—original draft: S.D. with inputs from G.L. and R.J. Writing—revision: S.D. and R.J. **Competing interests:** During their involvement related to this reported work, all authors were employees at Novartis AG. The authors declare that they have no other competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials or are publicly available. Processed gene expression counts used in the study are available from the respective repositories; TCGA: https://tcga-data.nci.nih.gov/tcga; CMI: https://portal.gdc.cancer.gov/projects/CMI-MBC; PTX: https://figshare.com/articles/dataset/pdxe_sample_annotations_txt/13331072/1; MET500: https://met500.med.umich.edu/downloadMet500DataSets; and CCLE: https://depmap.org/portal/. The MetMap dataset is available at https://ndownloader.figshare.com/files/24009293. The source code of the MOBER method is available at https://github.com/Novartis/MOBER. An interactive web app to explore the MOBER alignment in details is available at https://mober.pythonanywhere.com. In addition, a time-stamped version of the MOBER code, the five simulated datasets, along with Jupyter notebooks to generate the figures and replicate the analyses are published at Zenodo: https://zenodo.org/records/14209839.

Submitted 15 December 2023
Accepted 16 December 2024
Published 17 January 2025
10.1126/sciadv.adn5596