Deep Domain Adversarial Neural Network for the Deconvolution of Cell Type Mixtures in Tissue Proteome Profiling

Abstract

Deconvolution is a computational method for the determination/resolution of cell type proportions from bulk RNA sequence data. This technique is frequently used for the analysis of divergent cell types in tumor tissue samples. However, deconvolution technology is still in its infancy for the analysis of cell types using proteomic data sets due to challenges with repeatability/reproducibility, variable reference standards and the issues relating to single-cell proteomic reference data sets. Here, we developed a novel deep learning-based deconvolution method (scpDeconv) tailored to single-cell proteomic data sets. scpDeconv uses an autoencoder to leverage the information from bulk proteomic data to improve the quality of single-cell proteomic data, and employs a domain adversarial architecture to bridge the single-cell and bulk data distributions and transfer labels from single-cell data set to bulk data set. Extensive experiments validated the performance of scpDeconv in the deconvolution of proteomic data from various species/sources and produced using different proteomic technologies. This method should find broad applicability to areas including tumor microenvironment interpretation and clinical diagnosis/classification.

Introduction

The quantitative differentiation of cell types in a tumor microenvironment provides critical information for chemo- and immune-therapy, which is of important clinical and prognostic significance[1–3]. Bulk sequencing technology generally refers to lysing and sequencing all cells in a particular tissue, organ or organ system, and measures the average transcription/expression of genes/proteins in a cell mixture and, thus, ignores the cellular heterogeneity[4]. The development of single-cell (transcriptome/proteome) sequencing technology allows the isolation of single cells from tissue(s) for analysis[5,6]. Despite the advantages of this technology, applying it to large-scale cohorts to explore cellular heterogeneity in tumor microenvironments is impractical due to the relatively high cost and incompatibility to formalin-fixed or paraffin-embedded tissue samples[7,8]. Therefore, some computational methods have been developed to infer cell type proportions from transcription/expression profiles of genes from bulk tissue samples[7–10], providing a relatively low-cost and convenient way of studying tumor microenvironments in a clinical context.

The deconvolution of cell types from transcriptomic data from bulk tissue samples is feasible [4,8,11], and generally assumes that the mixture of cells can be modelled as a system of linear weights equations, therefore non-negative matrix factorization can be applied to estimate cell type proportions[8]. However, in reality, multiple single-cell subpopulations within each cell type do not always share exactly the same transcription/expression profile or marker genes due to the influence of cell cycle, cell state, microenvironment and background in sequence data, where variability in these features prevents high quality modelling using a linear distribution assumption. For this reason, alternative deep-learning-based methods were proposed following these steps: collecting scRNA-seq data and their corresponding cell type labels; training deep learning model with labelled scRNA-seq data as reference; and then inferring the corresponding cell type proportions of bulk data with the trained model[9,12,13]. However, these deep learning models were prone to overfitting on the reference datasets and failed on unseen bulk data from another source[9]. As a remedy, some methods were proposed to finetune the model on new datasets with known cell type proportions to improve performance[9]. However, the finetuning strategy hinders these methods' generalization and is infeasible in real scenarios since clinical bulk data usually lack reliable cell type proportions beforehand.

Existing deconvolution methods only focused on transcriptomics or genomics, but no attention has been paid to proteomics. However, there is a need to develop proteome-based deconvolution algorithms due for the following reasons: (1), proteins are the final executor of cell functions, and proteomics has irreplaceable advantages in biological and clinical studies ranging from understanding biological mechanisms to discovering drug targets [14]; (2), the concentration of most proteins cannot be accurately represented by the transcription/expression level of corresponding genes due to complex post-transcriptional regulation, RNA/protein degradation and post-translation modification[14,15];(3), there is substantial demand for deconvolution tools for the proteomic re-analysis of microenvironments within cancers to explore and understand tumorigenesis[16].

The advent of single-cell proteomics has greatly expanded both the number and types of proteins detected, which has laid the foundation for the deconvolution of bulk proteomic data[17–19]. Compared with transcriptomic data, there are a number of challenges for the analysis of proteomic data sets – which include significant background noise, poor data quality and marked variation depending on analysis run and/or technology. Other challenges relate to existing deconvolution methods being designed for transcriptomic data rather than proteomic data, and include: (i) distinct distributions between proteomic data and transcriptomic data, with most existing deconvolution methods for transcriptomic data being based on an assumed distribution, such as a Poisson or negative binomial distribution, which are recognized as unsuitable for proteomic data[4,11]; (ii) huge mismatch between types of proteins detected by

single cell proteomics and bulk proteomics[20,21], while these mismatched proteins are ignored by existing deconvolution methods; (iii) for single-cell proteomic data and bulk proteomic data, "batch effects" (variability among distributions) are pronounced, which is rarely considered by existing deconvolution methods[22].

Here, we tackled these challenges by developing a novel deep learning architecture. Based on the framework of mix-up, autoencoder and domain-adversarial neural network, this method can circumvent generating huge simulation datasets, making full use of single-cell proteomic data efficiently, impute the missing low abundant proteins with autoencoder, and establish a deconvolution module by domain adversary training, thereby improving the model's performance in a real-world scenario. The deconvolution approach developed gives new insights and potential to re-analyze clinical tissue proteomic data, reveal tumor microenvironment from the proteome level, and discover new clinical diagnostic markers.

Results

Overview of the deconvolution architecture

The overall framework of scpDeconv is illustrated in Figure 1. Firstly, for deconvolution task, we randomly mix-up single-cell proteomic data to construct pseudo-bulk data with corresponding cell type proportions as reference data, and use pseudo-bulk or real-bulk proteomic data as target data to be deconvoluted.

Then an autoencoder model is applied to impute low-abundant proteins in reference data by borrowing information from target data. To be specific, this autoencoder consists of an encoder, a predictor and a decoder, uses both reference data and target data as input, and is trained with the reconstruction loss calculated between the raw input and the reconstructed output and the prediction loss calculated between the ground truth ratios and the predicted cell type proportions for reference data. In this way, by joint training on labelled reference data and unlabelled target data, we can close the gap between reference data and target data, improve the quality of reference data, and impute the low abundant proteins which are missing in reference data.

Finally, the imputed reference data and raw target data are sent to a shared encoder and the extracted features are passed through a domain discriminator for generative adversarial training, thereby facilitating the learning of domain-invariant features by the encoder. A deconvolution head is employed to predict cell type proportions from the extracted domain-invariant features. In this way, this method can efficiently

improve the domain generalizability to overcome diverse distributions from single-cell proteomic data and bulk proteomic data.

Quantitative evaluation of deconvolution performance with proteome profiling

To quantitatively evaluate scpDeconv's performance on cell type mixtures as well as on cell cycle state mixtures, we collected 8 independent single cell proteomic datasets (including 2 murine cell lines datasets with 3 shared cell types, 5 human cell lines datasets with 4 shared cell types, and 1 human breast atlas dataset with 6 cell types) from up-to-date technologies, and 2 bulk proteomic datasets designed for cell cycles (including melanoma cell cycle dataset, and monocyte cell cycle dataset) (Supplementary Table 1). We constructed pseudo-bulk proteomic data with known cell type composition as ground truth ratios from these single-cell or pure-bulk proteomic data to rigorously quantify the performance of scpDeconv, and then compared with four representative and state-of-the-art deconvolution methods developed for transcriptomic data, involving deep learning based method (Scaden)[23], linear regression-based method (MuSiC)[4], Bayes-based method (BayesPrism)[24] and distribution hypothesis based method (DestVI)[25] (Figure 2a, Extended Data Figure 1a). Evaluating by Lin's concordance correlation coefficient (CCC), Root Mean Square Error (RMSE), and Pearson product-moment correlation coefficient (*r*), scpDeconv significantly surpassed Scaden, MuSiC, BayesPrism and DestVI on a broad range of proteomic datasets, due to the rationale design of imputation to overcome the missing low abundant proteins in reference single-cell proteomic data and the domain adversary training tailored for distribution differentiation between reference data and target data. In the contrast, compared with human breast atlas data from the same source, the performance of Scaden dropped in the target data with unseen distributions, such as murine cell line data from two different sources, indicating its limited generalizability (Figure 2a, Extended Data Figure 1a). The performance of MuSiC was hindered by its heavy dependency on gene weights' consistency across subjects, which might be highly inconsistent between reference and target data. BayesPrism had poor performance due to relying on parameter estimation which required high-quality data and was unsuitable for current single-cell proteomics. By reason of the rigorous assumption of distribution on transcriptomics, DestVI obtained the worst performance and was hard to migrate to proteomics.

The human breast atlas is a large-scale clinical dataset with scRNA-seq data and CyTOF data sharing 33 genes/proteins and 6 cell types. To compare the distribution between transcriptomics and proteomics, we selected paired scRNA-seq data(A) and CyTOF data(A) from the same donor(A) and CyTOF data(B) from another donor(B). As shown in Figure 2b and Extended Data Figure 1b, there are significant

differences between the distributions of scRNA-seq data(A) and CyTOF data(A), although they were sampled from the same donor. While the profiles of CyTOF data(A) and CyTOF data(B) are highly similar. Furthermore, scpDeconv, Scaden and MuSiC achieved better performance on deconvoluting pseudo-bulk proteomic data generated by CyTOF data(B) with CyTOF data(A) as reference than that with scRNA-seq data(A) as reference, while DestVI performed better on deconvoluting proteomics with transcriptomics as reference due to its distribution hypothesis tailored for transcriptome, and MuSiC obtained the worst performance on both experiments, indicating the necessity of deconvoluting bulk proteomic data with single-cell proteomic data as reference (Figure 2c-d and Extended Data Figure 1c).

We also extended scpDeconv's application to *Mus musculus*. The murine cell lines datasets contain two independent datasets generated from two advanced single-cell proteomic technologies (N2 and nanoPOTs) with 3 common cell types. These two datasets have less overlap in protein types (Figure 2e) and different distributions (Extended Data Figure 1d). In this case, we used mixed N2 dataset as reference data and mixed nanoPOTS dataset as target data to validate scpDeconv's performance on samples from different sources (Figure 2f). Furthermore, with the changing of reference sample numbers and sizes on murine cell lines datasets, scpDeconv achieved stable performance (Figure 2g-h). We also validated the effect of the imputation module, and the performance dropped after removing imputation module on the deconvolution task of murine cell lines data (Figure 2f and Extended Data Figure 1g).

The imputation step designed for proteomic dataset

Step 2 of scpDeconv is specifically designed for proteomics deconvolution, where an autoencoder imputes missing proteins of reference data (especially the key proteins of target data). To systematically analyze the effect of the imputation module, we integrated 3 single cell human cell lines datasets (SCoPE2_Leduc, pSCoPE_Huffman, and pSCoPE_Leduc datasets) from data-independent acquisition analysis mode as the reference group, and 2 single cell human cell lines datasets (T-SCP and plexDIA datasets) from data-dependent acquisition analysis mode as the target group, which suffered from severe batch effects (Figure 3a-b) and strikingly different protein types (Figure 3c, Extended Data Figure 2a). In particular, the reference group missed the highly variable proteins (HVPs) specific in the target group, which are, to some extent, informative for deconvolution task (Figure 3d and Extended Data Figure 2b). Therefore, scpDeconv introduced imputation module in *Step 2* to tackle with this problem. The ablation study on human cell line datasets showed that, with the help of the imputation module, adding input features with those target specific HVPs improved the model's performance, especially in the prediction of PDAC cell line proportion (Figure 3e-g). In addition, we also explored the effect of the number of

target HVPs on deconvolution task by changing them from 400 to 700, and found that they all achieved better performance than deconvolution without adding target specific HVPs, and got the best CCC score and RMSE value when target HVPs' number is 500 (Figure 3e). As shown in Extended data figure 2c-e, after imputation module, missing features within reference data can be complemented on the basis of target data.

Inference of cell cycle states

Besides the cell type deconvolution task, inference of cell cycle composition is also critical for studying biological progress, tumor microenvironment, and cancer therapy[26–28]. Thanks to the ultra-high sensitivity of mass spectrometry, proteomic technologies can detect the subtle changes of functional proteins involved in cell cycles[29,30]. In this regard, we quantitatively evaluated scpDeconv's performance on deconvoluting cell cycle mixtures simulated by bulk proteomic data under pure melanoma/monocyte cell cycle states. To make model focus on cell cycles not cell types, only marker proteins corresponding to certain cell cycles were chosen as input features (Figure 4a-b). Moreover, we noticed that marker proteins of cell cycle states are highly conserved in melanoma cell line and monocyte cell line (Figure 4c), which means cell cycle deconvolution across cell types/lines is feasible. For this reason, we deconvoluted the cell cycle mixtures of melanoma with monocyte mixtures as reference and vice versa. As shown in Figure 4d-e, Extended Data Figure 3, and Extended Data Figure 4, scpDeconv achieved state-of-the-art performance compared to other deconvolution methods for both cell cycle deconvolution experiments.

Application on clinical tissue proteomic profiling

Melanoma is a rare but invasive cancer of the skin, with high rates of growth and metastasis[31]. Previous clinical proteomic researches of melanoma samples are limited to marker proteins and phenotypes, hard to look into the tumor microenvironment and the effect of cellular heterogeneity on clinical outcomes[32]. Here, we explored the capacity of scpDeconv to re-analyze the representative clinical proteomic data of 174 metastatic melanoma samples at Grade IV stage and 4884 proteins with high quality[33] (Figure 5a). As shown in Figure 5a and Extended Data Figure 5a, scpDeconv used the most advanced spatial proteomics (Deep Visual Proteomics, DVP) on melanoma tissues as reference, which contained 7 cell classes (vertical growth melanoma, radial growth melanoma, CD146-high melanoma, CD146-low melanoma, in situ melanoma, melanocytes, and stroma) in the tumor microenvironment[34]. Following the procedure of scpDeconv, the proportions of 7 cell classes were estimated from the tissue proteomic profiles of 174 melanoma patients (Figure 5b-c). And obviously, the vertical growth melanoma class was dominant in the

late stage metastasis melanoma samples, while in situ melanoma class was the minority, which is consistent with previous clinical research, that is, the presence of vertical growth phage signifies a melanoma acquires the capacity of metastasis[35]. We further performed survival analysis with the predicted cell classes' proportions for 174 melanoma samples (Figure 5d-f, Extended Data Figure 5b-e). Compared with other cell classes, the proportion of vertical growth melanoma and radial growth melanoma exhibited significant indications for survival risk of melanoma patients. The patients with higher vertical growth melanoma proportions tended to gain lower survival probability, while those with higher radial growth melanoma proportions tended to gain higher survival probability, which is consistent with previous studies, that is, in contrast to melanoma in radial growth stage, melanoma with vertical growth pattern usually accompanies with strong aggressivity and poor prognosis[36,37].

In addition, as BRAF mutation is reported to reflect the tumorigenesis of melanoma, we further combined the indication of predicted cell class proportions with the occurrence of BRAF mutation (Extended Data Figure 6a). In this cohort, we can observe that BRAF mutation's indicator power is much weaker than our predicted vertical/radial growth melanoma proportions (Extended Data Figure 6b). By grouping samples into $BRAF^{(+)}$ (mutation) and $BRAF^{(-)}$ (non-mutation), the predicted vertical or radial growth melanoma proportion showed more significant prognostic indication in $BRAF^{(+)}$ samples (Extended Data Figure 6c-f).

scpDeconv's application on clinical proteome cohort indicated that tumor microenvironment could be reflected by predicted cell type proportions, and demonstrated its potential to be a valuable tool for cancer prognosis.

Discussion

Single-cell proteomic technology is rapidly developing and taking the center stage in biological and clinical researches[20], paving the way for the deconvolution of tissue proteome profiling. By fully considering the inherent characteristics of the recently developed single-cell proteomics technologies, we developed a novel deconvolution method, scpDeconv, tailored for proteomic data. By feat of autoencoder network, scpDeconv's imputation module can fill up missing low-abundant proteins in mixed-up single-cell proteomic data by joint-learning on both mixed-up single-cell proteomic data and target bulk proteomic data. Through domain-adversary training, scpDeconv can extract the domain-invariant latent features of imputed reference data and transfer labels to target data. Rigorous and extensive experiments evaluated scpDeconv's superiority and robustness on cell type deconvolution and inferring cell cycle

states for tissue proteomics. Furthermore, re-analysis of an existing bulk proteomic dataset from clinical melanoma samples showed the clinical application value of scpDeconv in cancer diagnosis and prognosis.

Compared to the existing deconvolution methods tailored for transcriptomics, scpDeconv enjoys four-fold advantages: (1), scpDeconv infers cell type proportions based on joint learning of reference data and target data in a data-driven manner, thereby unrestricted by the prior assumptions of data distribution and able to be flexibly applied to a broad range of proteomic data from different experiments or technology background; (2), scpDeconv brings in domain adaptation strategy, which can efficiently tackle the batch effect and close the distribution gaps between reference data and target data; (3), scpDeconv can impute missing low-abundant proteins in reference data and improve the quality of reference data with the help of imputation module; (4), we have validated that the output of scpDeconv is capable to reflect tumor microenvironment and play roles in disease diagnosis and prognosis. However, there is still limitation of scpDeconv. Reference proteomic data with known cell type composition needed for model training are constructed from single-cell proteomic data with labelled cell types. While single-cell proteomics is still in its infancy and has not been widely applied to a variety of biological samples. The single-cell proteomic data we exhaustively collected in this study only involves a minority of cell types, thereby limiting our method's applicability at present.

This first algorithm, tailored for the deconvolution of proteomic data using single-cell proteomic data as a reference, will have numerous key applications including deconvolution of cell type proportions, estimation of cell cycle states, and re-analysis of the microenvironment of multiple cancers. We hope scpDeconv will benefit the proteomics researcher community for more valuable discoveries in biological and clinical researches from the perspective of proteomics. With the exploration of the publicly available single-cell proteomic atlas in the future, it will be possible to re-analyse existing tissue proteomic data within the broader context. Furthermore, with a set of reliable microenvironment references of a vast range of cancer types, we believe that scpDeconv will be capable of analysing all the components of certain tumor samples at low cost and benefiting for basic medical research.

Methods

The model

The overall model architecture is shown in Figure 1, containing 4 steps: (1) Reference data preparation; (2) Missing features imputation and data quality improvement; (3) Domain-adversarial training and deconvolution; (4) Cell type proportion inference for target samples.

Step 1. Reference data preparation

As a supervised deconvolution method, single-cell proteomic data labelled with cell types are needed to prepare reference data with known cell type fractions. And the bulk proteomic dataset from the same cohort is termed the target data to be deconvoluted. The reference data and target data are usually sourced from different sequencing platforms and experiments, thereby having gaps in distribution. Furthermore, bulk proteomics can usually detect more proteins compared to single-cell proteomics, in which the low abundant proteins are hard to detect.

Mix-up strategy. Mix-up is an efficient way of data augmentation by a random combination of inputs[38]. Here we employed it on single-cell proteomic data to make up pseudo-bulk proteomic data before training. The random combination of single-cell proteomic data increases the data diversity of the reference dataset for training, thereby improving the model's generalizability. In this step, we generated the pseudo bulk proteomic data $\widetilde{X_r}$ with their cell type proportions $Y_r^{(1)}$ as reference data, and then sent it to the autoencoder module for data imputation.

Data alignment. Before model training, we need to align proteins of reference data and target data for model input. We firstly extract the proteins shared between reference and target data. And then we union these shared proteins with HVPs specific in target data as input feature set, since the target data are supposed to contain more proteins than the reference data, and those target specific HVPs may be informative for deconvolution task. The missing proteins of the input feature set within reference data are initialized to zeros and will be imputed in *Step 2*.

Step 2. Missing features imputation and data quality improvement.

Due to the immaturity of single-cell proteomics technology, the mixed-up reference data have strong noises, large-scale missing low abundant proteins, and an obvious distribution gap with the target bulk proteomic data. To compensate for the above obstacles caused by single-cell proteomics technology, we employed an autoencoder module to close the gap between the reference data and the target data, impute missing proteins of the mixed-up reference data, and improve the quality of reference data. As shown in Figure 1b, the target data and the mixed-up reference data generated in *Step 1* are inputted into Encoder1. And then the latent embedding generated by Encoder1 is sent to decoder to reconstruct the input. The loss between the original input and the reconstructed output is MSE loss. The latent embedding is also sent to a deconvolution head to estimate cell type proportions of reference data, and the L1 loss is calculated between the predictions and its known proportions. The formulations of the architecture in *Step 2* are:

$$\mathrm{E1}_r = \mathrm{Encoder1}(\widetilde{X_r})\#(1)$$

$$\widetilde{X_r}' = \mathrm{Decoder}(\mathrm{E1_r})\#(2)$$

$$Y_r^{(2)} = \text{Predictor1}(E1_r)\#(3)$$

$$\text{E1}_t = \text{Encoder1}(X_t)\#(4)$$

$$X_t^{'} = \text{Decoder}(E1_t)\#(5)$$

$$\text{Loss}_{deconvolution1} = \text{loss}_{\text{L1}}\left(Y_r^{(2)}, Y_r^{(1)}\right)\#(6)$$

$$\text{Loss}_{\text{reconstruction}} = \text{loss}_{\text{MSE}}\left(\widetilde{X_r}, \widetilde{X_r^{'}}\right) + \text{loss}_{\text{MSE}}\left(X_t, X_t^{'}\right)\#(7)$$

Where $E1_r$, $E1_t$, $Y_r^{(2)}$, $\widetilde{X_r}$, $\widetilde{X_r^{'}}$, X_t and $X_t^{'}$ are the embedding of mixed up reference data, the embedding of the target data, the output of the predictor1 used for deconvolution, the reference proteomic data, the output of decoder with the embedding of mixed up reference data, the target proteomic data, and the output of decoder with the embedding of the target data, respectively. After the convergence of the training stage, the imputed mixed-up reference data ($\widetilde{X_r^{'}}$) are used as the input of *Step 3*.

Step 3. Domain-adversarial training and deconvolution.

As shown in Figure 1c, the imputed reference data from *Step 2* and the target data from *Step 1* are sent to the Encoder2 in *Step 3*, and the encoded features are sent to the domain discriminator and the deconvolution head. The domain discriminator is employed to discriminate whether the feature is encoded from the reference data or the target data. By joint training of the Encoder2 and the discriminator in this step, the Encoder2 is prone to extracting domain-invariant features, which is suitable for the deconvolution task with the reference data generated from different sources. The deconvolution head consists of 2 fully connected layers and is used for predicting the cell proportions of target data. The Encoder2 and deconvolution head are joint training with the imputed reference data as input and the known proportions as supervision signals. The two kinds of joint training (Encoder2 and discriminator/Encoder2 and deconvolution head) are then executed iteratively until both losses are converged.

As for domain adaptation, we calculate BCE loss and optimize our model's weights by minimizing the distributions between the reference and target domains. The formulations of the architecture in *Step3* are:

$$E2_r = Encoder2\left(\widetilde{X_r^{'}}\right)\#(8)$$

$$D_r = Discriminator(E2_r)\#(9)$$

$$E2_t = Encoder2(X_t)\#(10)$$

$$D_t = Discriminator(E2_t)\#(11)$$

$$Y_r^{(3)} = Predictor2(E2_r)\#(12)$$

$$Loss_{discriminator} = loss_{BCE}\left(D_t, D_t^{'}\right) + loss_{BCE}\left(D_r, D_r^{'}\right)\#(13)$$

$$Loss_{deconvolution2} = loss_{L1}\left(Y_r^{(3)}, Y_r^{(1)}\right)\#(14)$$

Where $E2_r$, $E2_t$, $Y_r^{(3)}$, D_r and D_t are the embedding of imputed reference data by Encoder2, the embedding of the target data by Encoder2, the output of the Predictor2 used for deconvolution, the output of discriminator with the embedding of imputed reference data, and the output of discriminator with the embedding of target data, respectively.

Step 4. Cell type proportion inference for target samples.

As shown in Figure 1d, after domain-adversary training in *Step 3*, the trained Encoder2 and deconvolution head are used to infer cell type proportions of target samples. The formulation of the architecture employed in model inference is:

$$Y_t = Predictor2(E2_t)\#(15)$$

Where $E2_t$ is the output in *Formula* (10) with X_t as input, and Y_t is the final predicted cell type proportions of the target data.

Data

The datasets used in this study are listed in Supplemental Table 1.

Human breast atlas datasets. In the human breast atlas dataset[39], scRNA-seq was performed on 52,681 cells from 16 breast tissues. CyTOF was performed on 38 breast tissues (751,970 cells, 40 protein markers involved in breast development and tumorigenesis). The scRNA-seq and CyTOF datasets have 6 cell types in common (alveolar, hormone sensing, basal, fibroblast, vascular lymphatic, and immune cells).

Human cell line datasets. The human cell line datasets include five different single cell proteomic datasets, generating from T-SCP, plexDIA, pSCoPE_Leduc, SCoPE2_Leduc, and pSCoPE_Huffman, respectively[30,40–43]. These five datasets involve 4 cell types in total, consisting of 1,755 melanoma, 434 Hela cells, 579 PDAC cells, and 163 U-937.

Murine cell line datasets. The murine cell line datasets come from two single-cell proteomic technologies, N2 and nanoPOTS, applied in 3 identical murine cell lines (RAW 264.7, a macrophage cell line; C10, a respiratory epithelial cell line; SVEC, an endothelial cell line)[44,45]. Murine_N2 dataset (Murine cell lines dataset generated by N2) includes 108 cells (36 RAW cells, 36 C10 cells, and 36 SVEC cells) with 1437 proteins high quality detected. Murine_nanoPOTS dataset (Murine cell lines dataset generated by nanoPOTS) includes 72 cells (24 RAW cells, 24 C10 cells, and 24 SVEC cells) with 1032 proteins high quality detected. These two datasets have 762 proteins in common.

Bulk proteomic cell cycle datasets. Pure bulk proteomic datasets of melanoma G1/S/G2 and monocyte G1/S/G2 were generated in the same experiment of pSCoPE_Leduc dataset[43]. The top 20 highly expressed marker proteins of each cell cycle state were selected as model's input features.

Deep Visual Proteomics dataset. Deep Visual Proteomics (DVP) dataset links protein abundance of melanoma samples to complex cellular or subcellular phenotypes while preserving spatial context, involving average 1910 proteins[34]. This dataset contains 7 cell classes (vertical growth melanoma, radial growth melanoma, CD146-high melanoma, CD146-low melanoma, in situ melanoma, melanocytes, and stroma) within the tumor microenvironment of melanoma, defined by spatial location, SOX10 and CD146 staining intensity.

The clinical proteomic dataset of metastatic melanoma. The clinical proteomic dataset of metastatic melanoma was collected from the previous study in 174 melanoma metastasis patients, with 4883 proteins high quality detected[46].

Training details

For all cases, the batch size of training process in stage 2 and stage 3 was set to 50, and the learning rate was set to 0.0001. The other training details, such as pseudo-sample number and pseudo-sample size (cell number in each pseudo-sample) of reference data or target data, number of HVPs identified in target data, and number of training epochs used in each case can be found in Supplemental Table 2.
Furthermore, in the case of murine cell lines data, to explore scpDeconv's stability, we set different reference sample numbers (2000, 4000, and 6000) with fixed sample size, and different reference sample sizes (5, 10, 15, and 20) with fixed sample number.

Comparison of methods

Scaden. Scaden is a deep learning-based deconvolution algorithm that uses RNA-seq data and paired scRNA-seq data to infer the cellular composition of tissues[9]. The basic architecture of Scaden is a neural network with 4 fully connected layers which takes the pseudo bulk RNA-seq data as input and the predicted cell type fraction as output, using L1 loss during training to minimize the error between the predicted cell fractions and the real cell fractions. The final Scaden model is ensembled by three best-performing models with different parameter combinations.

MuSiC. MuSiC uses cell-type specific gene expression from scRNA-seq data to deconvolute bulk RNA-seq data in tissues[4]. MuSiC first reweighs genes based on cross-subject variance (informativeness), and then uses a recursively clustering strategy to close related cell types. The proportions of cell types are estimated by non-negative least squares regression based on a series of assumptions.

BayesPrism. BayesPrism is a Bayesian deconvolution method for cancer tissues depending on patient-derived scRNA-seq as prior information[24]. The Bayesian algorithm models prior distribution learned from the paired scRNA-seq data and then infers joint posterior distribution of cell proportions as the deconvolution results.

DestVI. DestVI is a Bayesian model that learns cell type specific profiles and latent variations using a conditional deep generative model[25]. The method assumes that transcriptomic data follows a negative binomial distribution to model the reference scRNA-seq data and spatial RNA-seq data.

Evaluation experiments

Cell type mixture decomposition with proteome profiling.

As for testing the deconvolution performance with reference data and target data from different modalities, we deconvoluted the simulated bulk proteomic data constructed from CyTOF data with scRNA-seq data in the same cohort as reference or with CyTOF data in the same cohort as reference in the human breast atlas dataset, splitting the dataset into reference and target data by donor.

There are apparent batch effects between different proteomic technologies and a huge gap between the distributions of bulk proteomic data and single-cell proteomic data. Therefore, we evaluated scpDeconv's deconvolution performance on the simulated bulk proteomic data constructed from murine_nanoPOTS dataset with the mixed-up murine_N2 dataset as reference to mimic the situation where the bulk/single cell proteomic data come from totally different proteomic technologies and different experiments.

Furthermore, we employed the integrated human cell line datasets to evaluate scpDeconv's performance on integrated multi-centre datasets. The Melanoma, PDAC, U-937, and Hela cell lines appeared in different experiments were selected for testing.

Systematic analysis.

To explore the efficiency of the imputation of missing proteins in *Step 2*, we executed cell type deconvolution experiments on the integrated human cell line datasets with adding target specific HVPs or without adding target specific HVPs, and also compared deconvolution performance between model with imputation stage and model without imputation stage. For the experiments without adding target specific HVPs, we only used the overlap proteins from both reference and target datasets as input features. For the experiments with adding target specific HVPs, we changed the number of HVP in target data from 400 to 700.

Application

Cell cycle inference with proteome profiling.

The cell cycles reflect the replication process of cancer cells and can be represented by the subtle changes of cell-cycle-relevant proteins. We collected the bulk proteomic data of melanoma or monocyte samples with pure cell subpopulations within G1, S, and G2 stage. Then the melanoma cell cycle mixtures were constructed from the melanoma cell subpopulations within G1, S, and G2 stages, and monocyte cell cycle mixtures were constructed from the monocyte subpopulations within G1, S, and G2 stages. We deconvoluted the melanoma cell cycle mixtures with monocyte pure cell cycle subpopulations as reference, and deconvoluted the monocyte cell cycle mixtures with melanoma pure cell cycle subpopulations as reference. In this way, we can estimate scpDeconv's performance on the cell cycle inference task across cell types/cell lines.

Investigation of microenvironment and clinical outcomes with clinical tissue proteome profiling.

The establishment of scpDeconv tailored for cell type deconvolution on proteomic data opens up an opportunity to mine the tumor microenvironment of previously released clinical tissues with provided bulk proteomic data. We deconvoluted the proteomic data from metastasis melanoma cohort and analysed the role of cell type proportions in cancer prognosis.

Evaluation metrics

The linear concordance between ground truth ratios and predicted proportions of cell types or cell states is estimated by Lin's concordance correlation coefficient (CCC), Root Mean Square Error (RMSE), and Person product-moment correlation coefficient (r). The formulas are as follows,

$$CCC(y,\hat{y}) = \frac{2r\sigma_y\sigma_{\hat{y}}}{\sigma_y^2 + \sigma_{\hat{y}}^2 + \left(\mu_y - \mu_{\hat{y}}\right)} \#(16)$$

$$RMSE(y,\hat{y}) = \sqrt{avg(y-\hat{y})^2} \#(17)$$

$$\mathrm{r}(\mathrm{y},\hat{\mathrm{y}}) = \frac{\mathrm{cov}(\mathrm{y},\hat{\mathrm{y}})}{\sigma_{\mathrm{y}}\sigma_{\hat{\mathrm{y}}}} \#(18)$$

Where y are the ground truth cell type (or cell cycle) proportions, $\hat{y}$ are the prediction of cell type (or cell cycle) proportions, σ_y is the standard deviation of y, $\sigma_{\hat{y}}$ is the standard deviation of $\hat{y}$, $cov(y,\hat{y})$ is the covariance of y and $\hat{y}$, and μ_y, $\mu_{\hat{y}}$ respectively are the mean of y and $\hat{y}$.

Statistical analysis

The group comparisons are estimated by the two-sided Wilcoxon test. Kaplan-Meier survival analysis is conducted to assess if the predicted cell type proportions are associated to outcomes (overall survival) in the clinical cohort. The P values are calculated using the log-rank test and are considered statistically significant if they are two-tailed less than 0.05.

Biological analysis

Highly variable proteins are selected by Scanpy[47]. The Gene Ontology (GO) and pathway enrichment are performed using DAVID platform[48].

Data Availability

All data used in this study are publicly available and the usages were fully illustrated in the Method section.

Code Availability

The codes are implemented in Python and released on https://github.tencentailab/scpDeconv with detailed instruction. The comparison methods are implemented under their official guidance and the reference codes are listed in the Method section.

Acknowledgements

Author Contributions

F.Y. and J.Y. conceived and designed the project. F.W. developed the algorithm and conducted experiments under the supervision of F.Y. and J.Y.. F.W. and F.Y. analysed the results. F.Y. and F.W. write the manuscript. F.W. finish the figure under the guidance of F.Y. and J.Y.. J.Y., G.W., J.Q. revised the manuscript. L.H. gave suggestion for the domain adaptation and improvement of the manuscript. J.S. gave suggestion for the project. All authors reviewed and approved the manuscript. F.W. and F.Y. contributed equally.

Competing Interests

The authors declare no competing interests.

Figure Legends

Figure 1. The architecture of our proposed method.

a, Stage 1: preparation of training data. The mix-up process that combines the proteome profiling of different cell types by random proportions.

b, Stage 2: domain transfer. The encoder-decoder scheme to close the gap between the reference data and the target data, and imputation of the mixed up single-cell proteomic data generated in **Stage 1**. Reconstruction loss is applied for model training.

c, Stage 3: deconvolution model training. The domain-adversary training architecture of encoder. The domain discriminator is used to discriminate whether the latent feature is encoded from the mixed up reference data or from the target bulk proteomic data. The deconvolution head is used to predict the proportions of cell type.

d, Stage 4:inference stage. Target bulk proteomic data are inputted to the trained encoder and the deconvolution head to infer the proportions of cell types.

Figure 2. Overall performance comparation of cell type (or cell cycle state) mixture deconvolution.

a, CCC (top) and RMSE values (bottom) of scpDeconv, Scaden, MuSiC, BayesPrism, and DestVI on human breast atlas datasets, murine cell lines datasets, human cell lines datasets, and cell cycle datasets. Boxplots show the median (center lines), interquartile range (hinges), and 1.5 times the interquartile range (whiskers).

b, Correlation plots between donor A's proteome and donor A's transcriptome (left) or donor A's proteome and donor B's proteome (right) within 6 cell types studied in human breast atlas datasets. AV: alveolar, HS: hormone sensing, BA: basal, FI: fibroblast, VA: vascular lymphatic, IM: immune cells.

c, Scatter plots of true (y axis) and predicted cell type proportions (x axis) by deconvoluting pseudo-bulk proteomic data from donor B with single-cell transcriptomic data from donor A as reference (top) or with single-cell proteomic data from donor B as reference (bottom) in human breast atlas datasets.

d, CCC values of scpDeconv and the comparison methods by deconvoluting pseudo-bulk proteomic data from donor B with single-cell transcriptomic data from donor A as reference (triangle) or with single-cell proteomic data from donor B as reference (circle) in human breast atlas datasets.

e, Venn diagram showing the overlap of proteins between two murine cell line datasets (N2 dataset and nanoPOTS dataset).

f, Scatter plots of true (y axis) and predicted proportions (x axis) of three cell types (C10, SVEC, and RAW) in murine cell lines datasets using scpDeconv.

g, CCC values of scpDeconv on murine cell lines datasets with different reference sample numbers in stage 1.

h, CCC values of scpDeconv on murine cell lines datasets with different reference sample sizes in stage 1.

Figure 3. Deconvolution performance on integrated human cell lines datasets with different HVP numbers.

a, Integrated t-SNE plots colored by data sources (left panel) and by cell types (right panel) of the reference group (SCoPE2_Leduc, pSCoPE_Huffman, and pSCoPE_Leduc datasets).

b, Integrated t-SNE plots colored by data sources (left panel) and by cell types (right panel) of the target group (T-SCP and plexDIA datasets).

c, Venn diagram showing the overlap of proteins from five different human cell lines datasets.

d, Venn diagram showing the overlap between shared proteins in all datasets and highly variable proteins (HVPs) in target group. Heatmap showing the protein expression of specific HVPs in the target group.

e, CCC (left panel) and RMSE values (right panel) of scpDeconv with different target HVP numbers on integrated human cell line datasets.

f, Scatter plots of true (y axis) and predicted proportions (x axis) of four cell types (Melanoma, PDAC, U-937, and Hela cell) in human cell lines datasets using scpDeconv without adding target specific HVPs.

g, Scatter plots of true (y axis) and predicted proportions (x axis) of four cell types (Melanoma, PDAC, U-937, and Hela cell) in human cell lines datasets using scpDeconv with setting target HVP number as 500.

Figure 4. Inference of cell cycle across cell lines.

a, Heatmap of cell cycle specific marker proteins in melanoma and monocytes.

b, GO enrichment of marker proteins of G1, S, and G2 phages.

c, Heatmap showing the correlation of marker proteins' abundance in melanoma and monocytes.

d, Scatter plots of true (y axis) and predicted proportions (x axis) of three cell cycle states (G1, S, and G2) by deconvoluting monocytes (mon) pseudo samples with melanoma (mel) sample as reference (top) and vise versa (bottom) using scpDeconv.

e, CCC (top) and RMSE values (bottom) of scpDeconv and the comparison methods by deconvoluting monocytes pseudo samples with melanoma samples as reference (circle) and vise versa (triangle).

Figure 5. Application on clinical tissue proteomic data of metastasis melanoma.

a, The pipeline of applying scpDeconv on deconvoluting clinical tissue proteome of metastasis melanoma patients with melanoma DVP dataset as reference.

b, Heatmap of proteomic profile (top) and histogram of predicted cell class proportions (bottom) for each sample in clinical melanoma dataset.

c, Distribution of predicted cell class proportions of melanoma samples. Box plots show the median (center lines), interquartile range (hinges), and 1.5 times the interquartile range (whiskers).

d, KM plot of the predicted proportions of vertical growth melanoma for 174 melanoma patients.

e, KM plot of the predicted proportions of radial growth melanoma for 174 melanoma patients.

f, KM plot of the predicted proportions of in situ melanoma for 174 melanoma patients.

Extended Data Figure 1. Overall performance comparation of cell type (or cell cycle state) mixture deconvolution.

a, Pearson's correlation values of scpDeconv, Scaden, MuSiC, BayesPrism, and DestVI on human breast atlas datasets, murine cell lines datasets, human cell lines datasets, and cell cycle datasets. Boxplot show the median (center lines), interquartile range (hinges), and 1.5 times the interquartile range (whiskers).

b, Heatmap of transcriptomic/proteomic profile from donor A and proteomic profile from donor B on 33 shared markers (x axis) and 6 cell types (y axis) studied in human breast atlas datasets. AV: alveolar, HS: hormone sensing, BA: basal, FI: fibroblast, VA: vascular lymphatic, IM: immune cells.

c, RMSE values of scpDeconv and the comparison methods by deconvoluting pseudo-bulk proteomic data from donor B with single-cell transcriptomic data from donor A as reference (triangle) or with single-cell proteomic data from donor B as reference (circle) in human breast atlas datasets.

d, t-SNE plot showing the distribution of N2 dataset and nanoPOTS dataset.

e, RMSE values of scpDeconv on murine cell lines datasets with different reference sample numbers in stage 1.

f, RMSE values of scpDeconv on murine cell lines datasets with different reference sample sizes in stage 1.

g, Scatter plots of true (y axis) and predicted proportions (x axis) of three cell types (C10, SVEC, and RAW) in murine cell lines datasets after removing stage 2 of scpDeconv.

Extended Data Figure 2. Systematical analysis of missing feature imputation in human cell lines datasets.

a, Venn diagram showing the overlap of proteins from different sources in the reference group (left) or target group (middle), and the overlap between reference groups' shared proteins and target groups' shared proteins (right) in the human cell lines datasets.

b, Heatmap showing the z-score of target specific HVPs in the target group.

c, Heatmap showing the raw expression of all proteins and missing features in reference pseudo-bulk data after stage 1 of scpDeconv. Proteins and samples were ordered by hierarchical clustering on raw expression.

d, Heatmap showing the imputed expression of all proteins and missing features in reference pseudo-bulk data after stage 2 of scpDeconv. Proteins and samples kept the same order with above.

e, Heatmap showing the raw expression of all proteins in the target data. Proteins kept the same order with above and samples were ordered by hierarchical clustering.

Extended Data Figure 3. Performance of cell cycle state inference by deconvoluting monocytes pseudo samples with melanoma samples as reference.

a, Scatter plot of true (y axis) and predicted cell cycle proportions (x axis) of all stages, G1 stage, S stage, and G2 stage for monocytes pseudo samples (from left to right) by scpDeconv with melanoma samples as reference.

b, Scatter plot of true (y axis) and predicted cell cycle proportions (x axis) of all stages, G1 stage, S stage, and G2 stage for monocytes pseudo samples (from left to right) by Scaden with melanoma samples as reference.

c, Scatter plot of true (y axis) and predicted cell cycle proportions (x axis) of all stages, G1 stage, S stage, and G2 stage for monocytes pseudo samples (from left to right) by MuSiC with melanoma samples as reference.

Extended Data Figure 4. Performance of cell cycle state inference by deconvoluting melanoma pseudo samples with monocytes as reference.

a, Scatter plot of true (y axis) and predicted cell cycle proportions (x axis) of all stages, G1 stage, S stage, and G2 stage for melanoma pseudo samples (from left to right) by scpDeconv with monocytes samples as reference.

b, Scatter plot of true (y axis) and predicted cell cycle proportions (x axis) of all stages, G1 stage, S stage, and G2 stage for melanoma pseudo samples (from left to right) by Scaden with monocytes samples as reference.

c, Scatter plot of true (y axis) and predicted cell cycle proportions (x axis) of all stages, G1 stage, S stage, and G2 stage for melanoma pseudo samples (from left to right) by MuSiC with monocytes samples as reference.

Extended Data Figure 5. Survival analysis of the predicted cell class proportions of clinical melanoma samples.

a, Three cases showing the scpDeconv's pipeline on deconvoluting clinical tissue proteome of metastasis melanoma patients.

b, KM plot of the predicted proportions of CD146-high melanoma for 174 melanoma patients.

c, KM plot of the predicted proportions of CD146-low melanoma for 174 melanoma patients.

d, KM plot of the predicted proportions of melanocytes for 174 melanoma patients.

e, KM plot of the predicted proportions of stroma for 174 melanoma patients.

Extended Data Figure 6. Comprehensive survival analysis by combining BRAF mutation and the predicted cell class proportions of clinical melanoma samples.

a, Distribution of predicted cell class proportions of melanoma samples grouped by BRAF mutation. Boxplots show the median (center lines), interquartile range (hinges), and 1.5 times the interquartile range (whiskers).

b, Hazard ratios of BRAF mutation, vertical growth melanoma and radial growth melanoma in melanoma patients.

c, KM plot of the predicted proportions of vertical growth melanoma for melanoma samples without BRAF mutation.

d, KM plot of the predicted proportions of radial growth melanoma for melanoma samples without BRAF mutation.

e, KM plot of the predicted proportions of vertical growth melanoma for melanoma samples with BRAF mutation.

f, KM plot of the predicted proportions of radial growth melanoma for melanoma samples with BRAF mutation.

References

1. Galon, J. *et al.* Type, density, and location of immune cells within human colorectal tumors predict clinical outcome. *Science (1979)* **313**, 1960–1964 (2006).

2. Ino, Y. *et al.* Immune cell infiltration as an indicator of the immune microenvironment of pancreatic cancer. *British Journal of Cancer 2013 108:4* **108**, 914–923 (2013).

586 3. Shang, B., Liu, Y., Jiang, S. J. & Liu, Y. Prognostic value of tumor-infiltrating FoxP3+ regulatory T
587 cells in cancers: a systematic review and meta-analysis. *Scientific Reports 2015 5:1* **5**, 1–9 (2015).

588 4. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-
589 subject single-cell expression reference. *Nature Communications 2019 10:1* **10**, 1–9 (2019).

590 5. Qian, J. *et al.* A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by
591 single-cell profiling. *Cell Research 2020 30:9* **30**, 745–762 (2020).

592 6. Azizi, E. *et al.* Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor
593 Microenvironment. *Cell* **174**, 1293-1308.e36 (2018).

594 7. Frishberg, A. *et al.* Cell composition analysis of bulk genomics using single-cell data. *Nature*
595 *Methods 2019 16:4* **16**, 327–332 (2019).

596 8. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with
597 digital cytometry. *Nature Biotechnology 2019 37:7* **37**, 773–782 (2019).

598 9. Menden, K. *et al.* Deep learning-based cell composition analysis from tissue expression profiles.
599 *Sci Adv* **6**, (2020).

600 10. Cable, D. M. *et al.* Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature*
601 *Biotechnology 2021 40:4* **40**, 517–526 (2021).

602 11. Patrick, E. *et al.* Deconvolving the contributions of cell-type heterogeneity on cortical gene
603 expression. *PLoS Comput Biol* **16**, e1008120 (2020).

604 12. Lv, X. *et al.* Deconvolute Gene Expression Based on Deep Learning in Scrna-Seq. *Proceedings -*
605 *2021 17th International Conference on Computational Intelligence and Security, CIS 2021* 70–74
606 (2021) doi:10.1109/CIS54983.2021.00023.

607 13. Lin, Y. *et al.* DAISM-DNNXMBD: Highly accurate cell type proportion estimation with in silico data
608 augmentation and deep neural networks. *Patterns* **3**, 100440 (2022).

609 14. Marguerat, S. *et al.* Quantitative Analysis of Fission Yeast Transcriptomes and Proteomes in
610 Proliferating and Quiescent Cells. *Cell* **151**, 671–683 (2012).

611 15. Gygi, S. P., Rochon, Y., Franza, B. R. & Aebersold, R. Correlation between Protein and mRNA
612 Abundance in Yeast. *Mol Cell Biol* **19**, 1720–1730 (1999).

613 16. Edwards, N. J. *et al.* The CPTAC data portal: A resource for cancer proteomics research. *J*
614 *Proteome Res* **14**, 2707–2713 (2015).

615 17. Strack, R. Spatial proteomics with subcellular resolution. *Nature Methods 2022 19:7* **19**, 780–780
616 (2022).

617 18. Petelski, A. A. *et al.* Multiplexed single-cell proteomics using SCoPE2. *Nature Protocols 2021*
618 *16:12* **16**, 5398–5425 (2021).

619 19. Woo, J. *et al.* High-throughput and high-efficiency sample preparation for single-cell proteomics
620 using a nested nanowell chip. *Nature Communications 2021 12:1* **12**, 1–11 (2021).

621 20. Perkel, J. M. Single-cell proteomics takes centre stage. *Nature* **597**, 580–582 (2021).

622 21. Doerr, A. Single-cell proteomics. *Nature Methods 2018 16:1* **16**, 20–20 (2018).

623 22. Vanderaa, C. & Gatto, L. Replication of single-cell proteomics data reveals important
624 computational challenges. *https://doi.org/10.1080/14789450.2021.1988571* **18**, 835–843 (2021).

625 23. Menden, K. *et al.* Deep learning–based cell composition analysis from tissue expression profiles.
626 *Sci Adv* **6**, (2020).

627 24. Chu, T., Wang, Z., Pe'er, D. & Danko, C. G. Cell type and gene expression deconvolution with
628 BayesPrism enables Bayesian integrative analysis across bulk and single-cell RNA sequencing in
629 oncology. *Nature Cancer 2022 3:4* **3**, 505–517 (2022).

630 25. Lopez, R. *et al.* DestVI identifies continuums of cell types in spatial transcriptomics data. *Nature*
631 *Biotechnology 2022 40:9* **40**, 1360–1369 (2022).

632 26. Warrener, R. *et al.* Tumor cell-specific cytotoxicity by targeting cell cycle checkpoints. *The FASEB*
633 *Journal* **17**, 1–21 (2003).

634 27. Li, J. & Stanger, B. Z. Cell Cycle Regulation Meets Tumor Immunosuppression. *Trends Immunol* **41**,
635 859–863 (2020).

636 28. Schwartz, G. K. & Shah, M. A. Targeting the cell cycle: A new approach to cancer therapy. *Journal*
637 *of Clinical Oncology* **23**, 9408–9421 (2005).

638 29. Brunner, A.-D. *et al.* Ultra-high sensitivity mass spectrometry quantifies single-cell proteome
639 changes upon perturbation. *Mol Syst Biol* **18**, e10798 (2022).

640 30. Leduc, A., Huffman, R. G. & Slavov, N. Droplet sample preparation for single-cell proteomics
641 applied to the cell cycle. doi:10.1101/2021.04.24.441211.

642 31. Balch, C. M. *et al.* Final Version of 2009 AJCC Melanoma Staging and Classification. *Journal of*
643 *Clinical Oncology* **27**, 6199 (2009).

644 32. Betancourt, L. H. *et al.* The Human Melanoma Proteome Atlas—Complementing the melanoma
645 transcriptome. *Clin Transl Med* **11**, e451 (2021).

646 33. Beck, L. *et al.* Clinical proteomics of metastatic melanoma reveals profiles of organ specificity and
647 treatment resistance. *Clinical Cancer Research* **27**, 2074–2086 (2021).

648 34. Mund, A. *et al.* Deep Visual Proteomics defines single-cell identity and heterogeneity. *Nature*
649 *Biotechnology 2022 40:8* **40**, 1231–1240 (2022).

650 35. Crowson, A. N., Magro, C. M. & Mihm, M. C. Prognosticators of melanoma, the melanoma report,
651 and the sentinel lymph node. *Modern Pathology 2006 19:2* **19**, S71–S87 (2006).

652 36. Prognosis and survival for melanoma skin cancer | Canadian Cancer Society.
653 https://cancer.ca/en/cancer-information/cancer-types/skin-melanoma/prognosis-and-survival.

654 37. Ciarletta, P., Foret, L. & Amar, M. ben. The radial growth phase of malignant melanoma: multi-
655 phase modelling, numerical simulations and linear stability analysis. *J R Soc Interface* **8**, 345
656 (2011).

657 38. Zhang, H., Cisse, M., Dauphin, Y. N. & Lopez-Paz, D. mixup: Beyond Empirical Risk Minimization.
658 Preprint at https://github.com/facebookresearch/mixup-cifar10. (2022).

659 39. Gray, G. K. *et al.* A human breast atlas integrating single-cell proteomics and transcriptomics. *Dev*
660 *Cell* **57**, 1400-1420.e7 (2022).

661 40. Brunner, A. *et al.* Ultra-high sensitivity mass spectrometry quantifies single-cell proteome
662 changes upon perturbation. *Mol Syst Biol* **18**, (2022).

663 41. Derks, J. *et al.* Increasing the throughput of sensitive proteomics by plexDIA. *Nature*
664 *Biotechnology 2022* 1–10 (2022) doi:10.1038/s41587-022-01389-w.

665 42. Huffman, R. G. *et al.* Prioritized single-cell proteomics reveals molecular and functional
666 polarization across primary macrophages. *bioRxiv* 2022.03.16.484655 (2022)
667 doi:10.1101/2022.03.16.484655.

668 43. Leduc, A., Huffman, R. G. & Slavov, N. Droplet sample preparation for single-cell proteomics
669 applied to the cell cycle. doi:10.1101/2021.04.24.441211.

670 44. Dou, M. *et al.* High-Throughput Single Cell Proteomics Enabled by Multiplex Isobaric Labeling in a
671 Nanodroplet Sample Preparation Platform. *Anal Chem* **91**, 13119–13127 (2019).

672 45. Woo, J. *et al.* High-throughput and high-efficiency sample preparation for single-cell proteomics
673 using a nested nanowell chip. *Nature Communications 2021 12:1* **12**, 1–11 (2021).

674 46. Beck, L. *et al.* Clinical Proteomics of Metastatic Melanoma Reveals Profiles of Organ Specificity
675 and Treatment Resistance. doi:10.1158/1078-0432.CCR-20-3752.

676 47. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: Large-scale single-cell gene expression data
677 analysis. *Genome Biol* **19**, 1–5 (2018).

678 48. Sherman, B. T. *et al.* DAVID: a web server for functional enrichment analysis and functional
679 annotation of gene lists (2021 update). *Nucleic Acids Res* **50**, W216–W221 (2022).

680

681

Figure 1

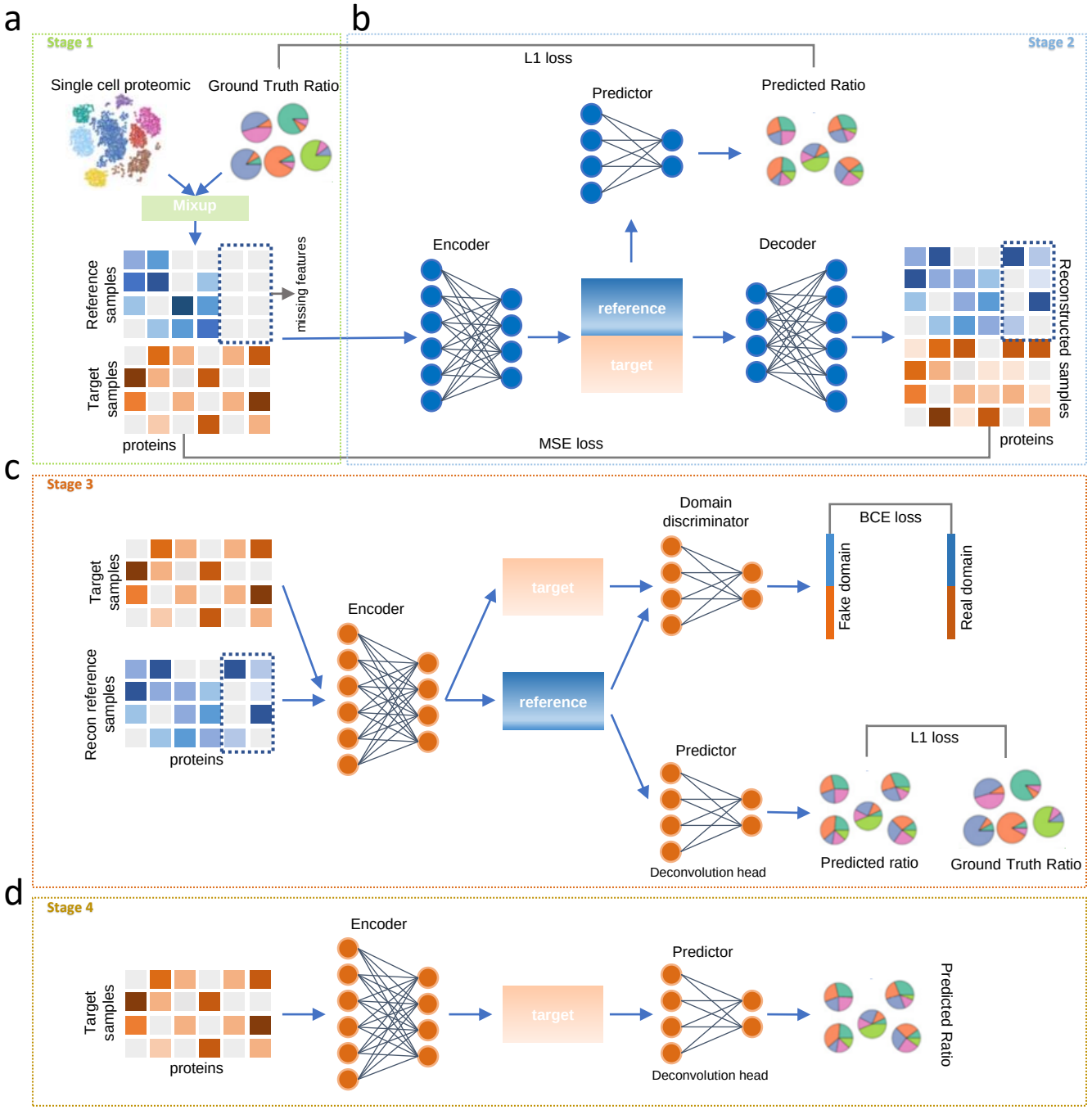

a
b
Stage 1
Stage 2
L1 loss
Single cell proteomic
Ground Truth Ratio
Predictor
Predicted Ratio
Mixup
Reference samples
missing features
Encoder
Decoder
reference
target
Reconstructed samples
Target samples
proteins
MSE loss
proteins
c
Stage 3
Domain discriminator
BCE loss
Target samples
Encoder
target
Fake domain
Real domain
Recon reference samples
reference
proteins
Predictor
L1 loss
Deconvolution head
Predicted ratio
Ground Truth Ratio
d
Stage 4
Encoder
Predictor
Target samples
target
Predicted Ratio
proteins
Deconvolution head

Figure 2

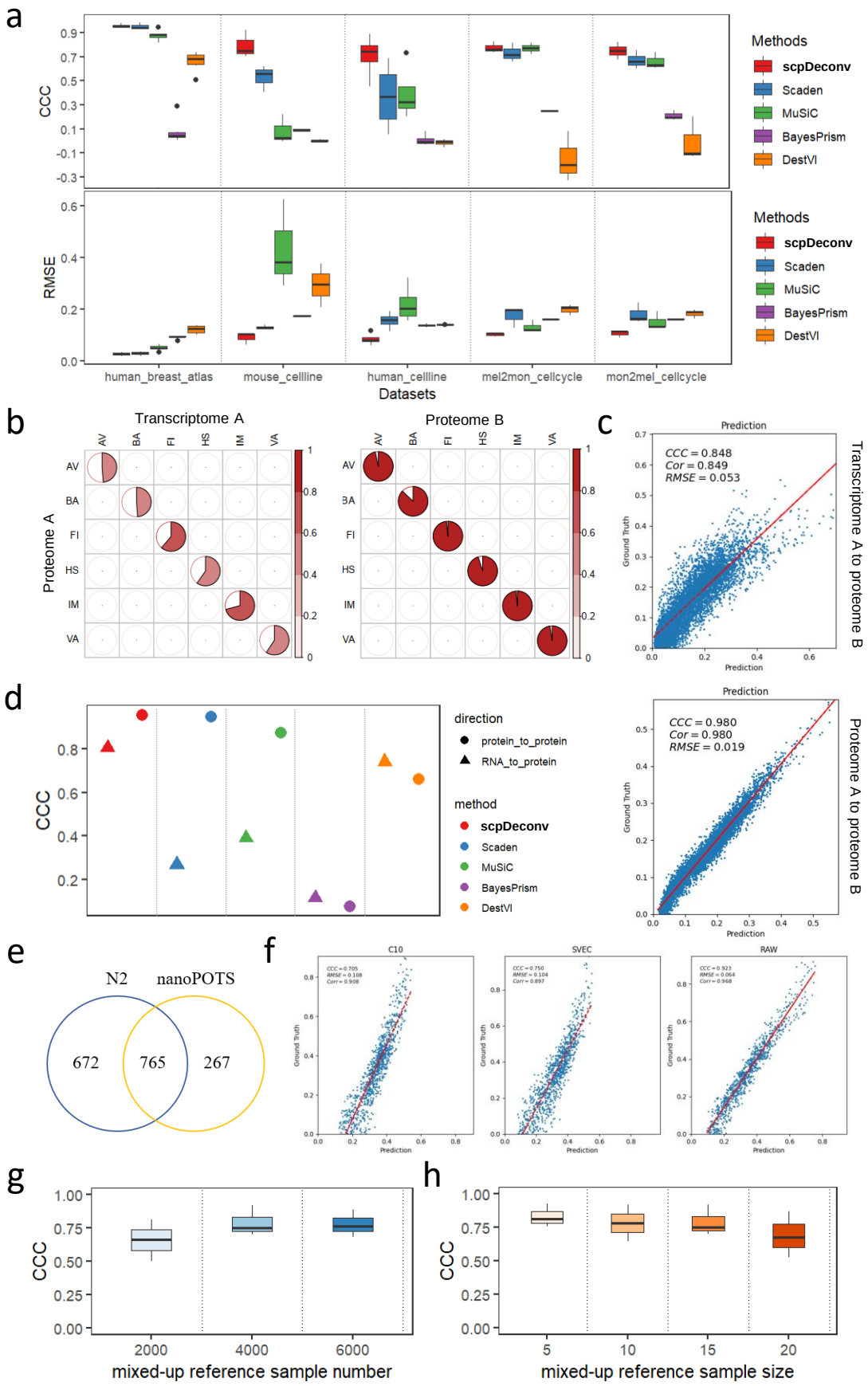

a
CCC
RMSE
Methods
scpDeconv
Scaden
MuSiC
BayesPrism
DestVI
human_breast_atlas
mouse_cellline
human_cellline
mel2mon_cellcycle
mon2mel_cellcycle
Datasets
b
Transcriptome A
Proteome B
Proteome A
AV
BA
FI
HS
IM
VA
c
Prediction
CCC = 0.848
Cor = 0.849
RMSE = 0.053
Ground Truth
Transcriptome A to proteome B
d
CCC
direction
protein_to_protein
RNA_to_protein
method
CCC = 0.980
Cor = 0.980
RMSE = 0.019
Proteome A to proteome B
e
N2
nanoPOTS
672
765
267
f
C10
SVEC
RAW
g
CCC
2000
4000
6000
mixed-up reference sample number
h
5
10
15
20
mixed-up reference sample size

Figure 3

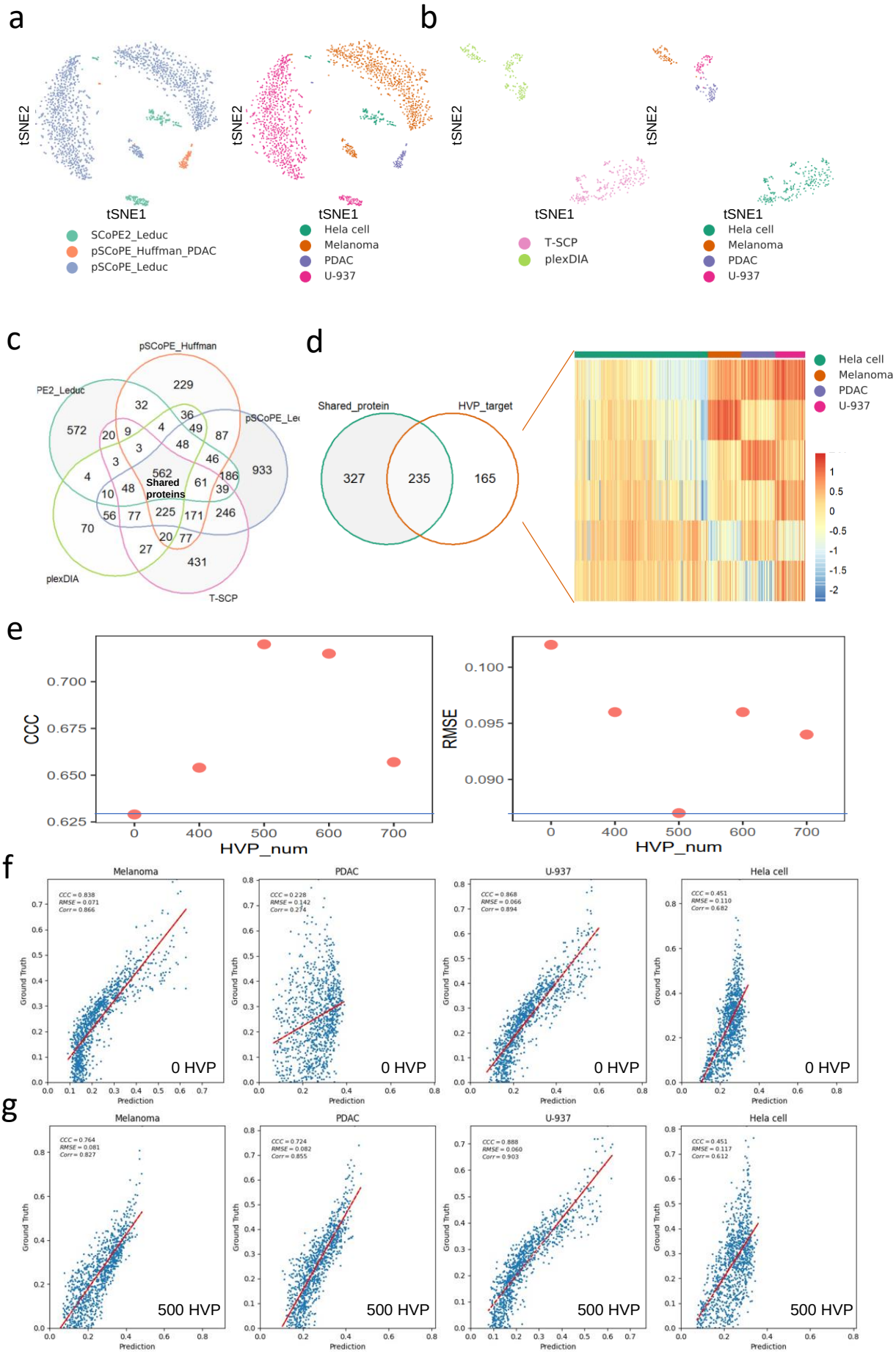

a
tSNE2
tSNE1
SCoPE2_Leduc
pSCoPE_Huffman_PDAC
pSCoPE_Leduc
Hela cell
Melanoma
PDAC
U-937
b
T-SCP
plexDIA
c
pSCoPE_Huffman
PE2_Leduc
pSCoPE_Le
plexDIA
T-SCP
Shared proteins
d
Shared_protein
HVP_target
327
235
165
Hela cell
Melanoma
PDAC
U-937
e
CCC
RMSE
HVP_num
f
Melanoma
PDAC
U-937
Hela cell
Ground Truth
Prediction
0 HVP
g
500 HVP

Figure 4

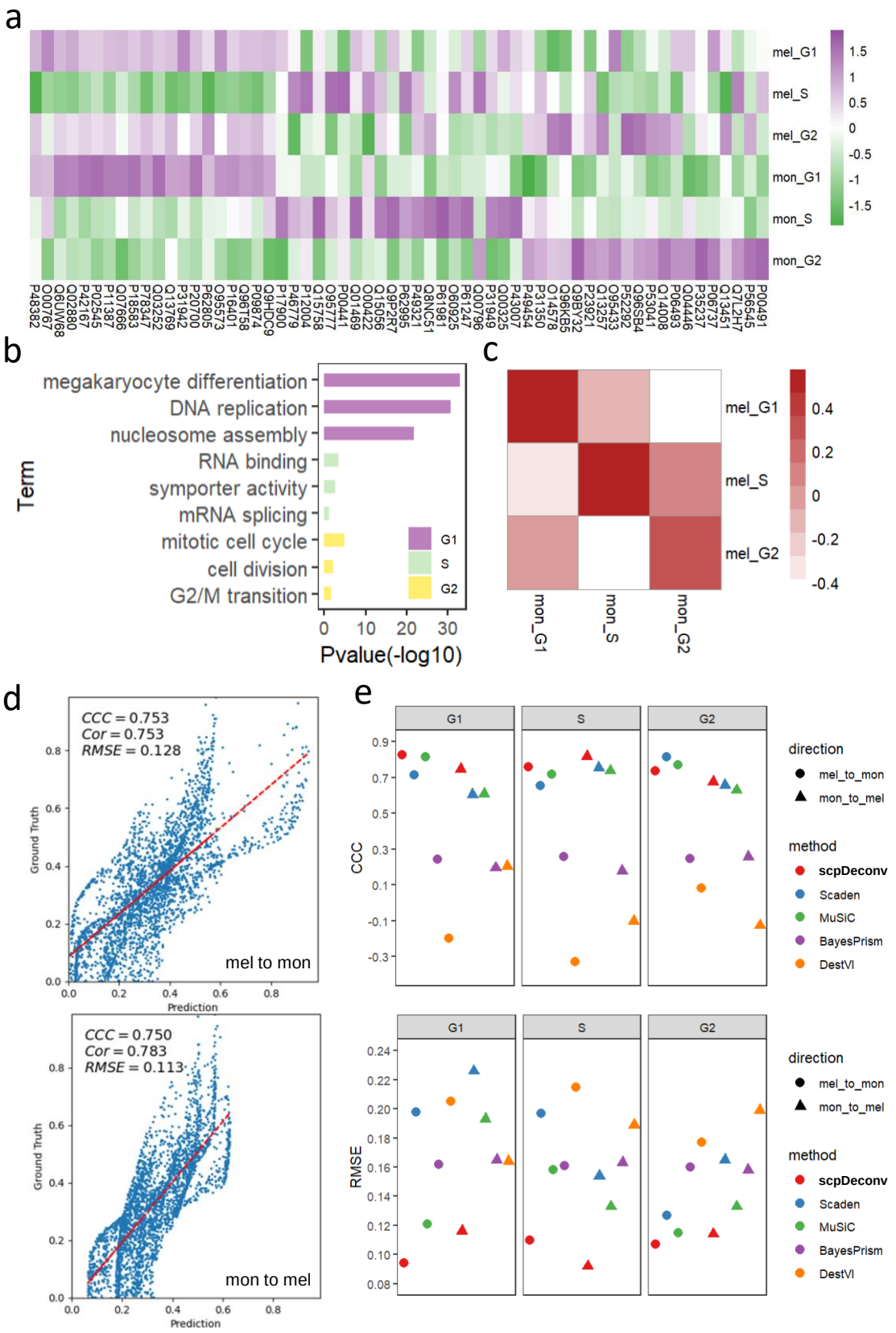
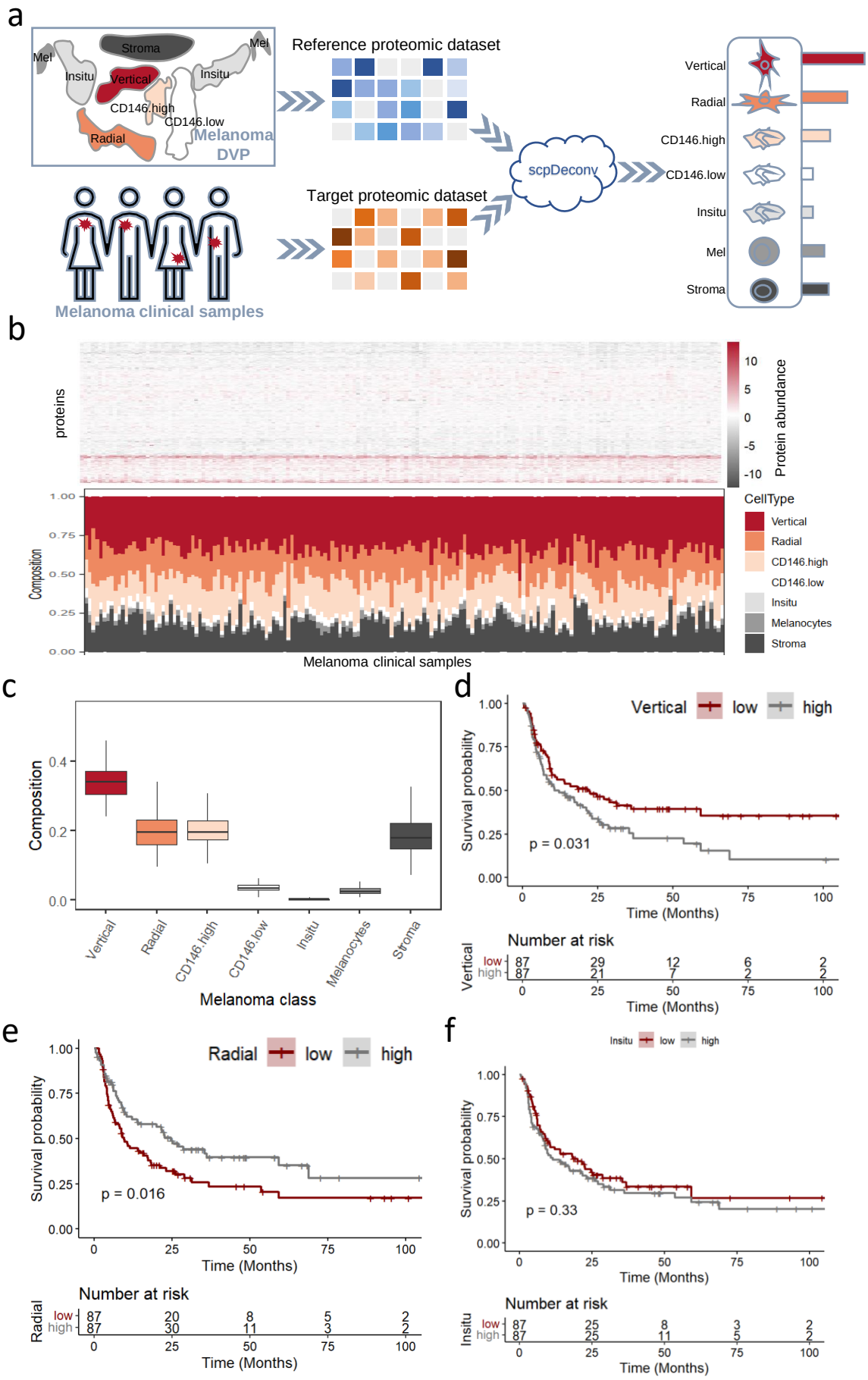
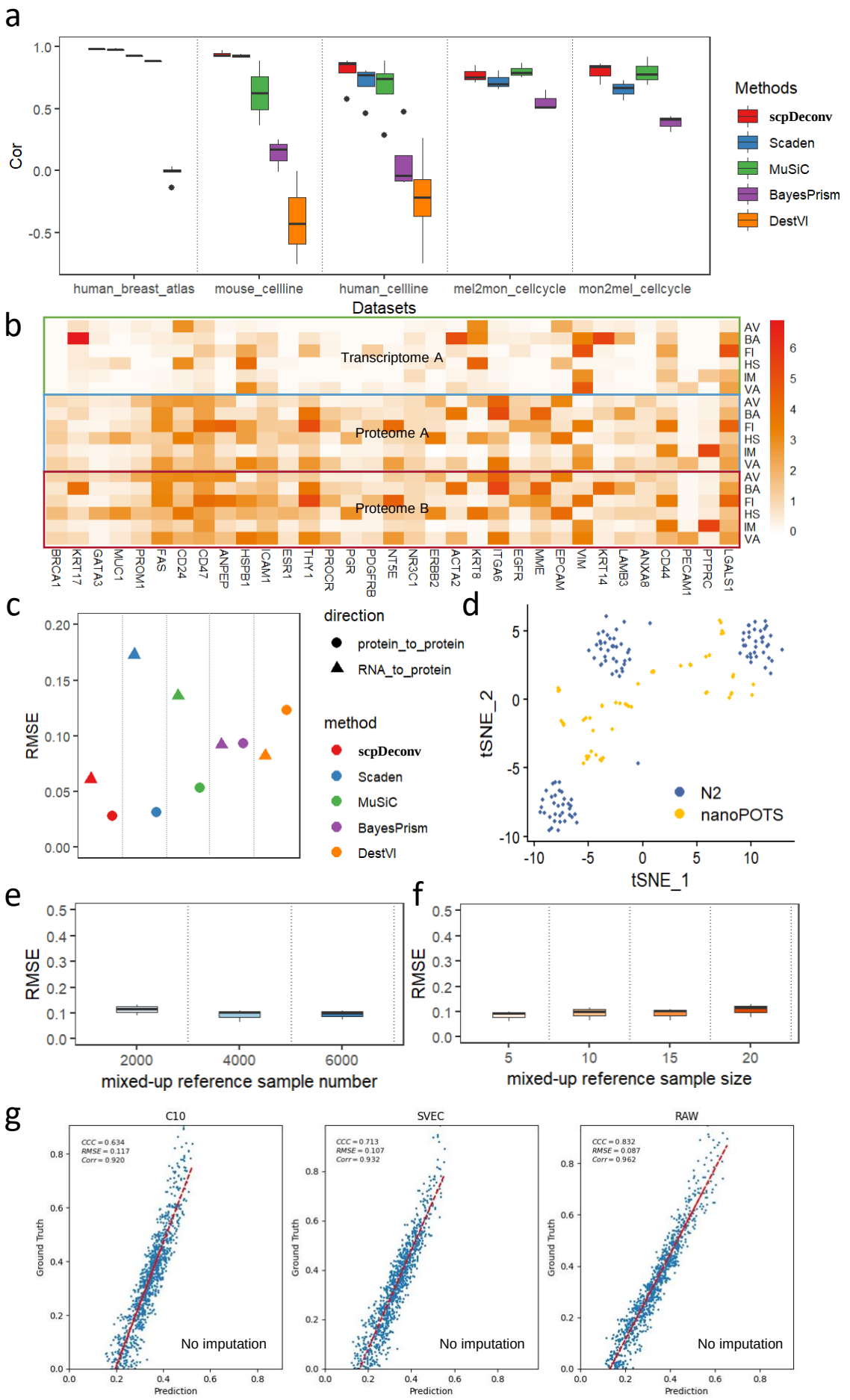

a
mel_G1
mel_S
mel_G2
mon_G1
mon_S
mon_G2
1.5
1
0.5
0
-0.5
-1
-1.5
P48382 O00767 Q8UW68 Q02880 P42167 P02545 P11387 Q07666 P18583 P78347 Q03252 Q13769 P31942 P20700 P62805 P16401 P09874 Q96T58 Q9HDC9 P17900 P46779 P12004 Q15758 O95777 P00441 O01469 O00422 Q15056 Q9P2R7 P62995 P49321 Q8NC51 P61981 O60925 P61247 Q00796 P31949 O00325 P43007 P49454 P31350 O14578 Q96KB5 Q9BY32 P23921 Q13257 O95433 Q96SB4 P52292 P53041 Q14008 P06493 Q04446 P35237 P06737 Q13451 Q7L2H7 P56545 P00491
b
Term
megakaryocyte differentiation
DNA replication
nucleosome assembly
RNA binding
symporter activity
mRNA splicing
mitotic cell cycle
cell division
G2/M transition
G1
S
G2
0
10
20
30
Pvalue(-log10)
c
mel_G1
mel_S
mel_G2
mon_G1
mon_S
mon_G2
0.4
0.2
0
-0.2
-0.4
d
CCC = 0.753
Cor = 0.753
RMSE = 0.128
mel to mon
Ground Truth
Prediction
CCC = 0.750
Cor = 0.783
RMSE = 0.113
mon to mel
Ground Truth
Prediction
e
G1
S
G2
CCC
RMSE
direction
mel_to_mon
mon_to_mel
method
scpDeconv
Scaden
MuSiC
BayesPrism
DestVI

Figure 5

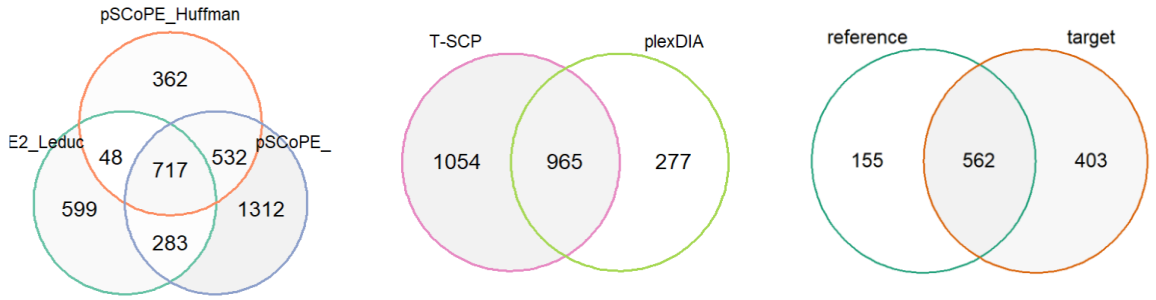

a
Stroma
Mel
Mel
Insitu
Vertical
Insitu
CD146.high
CD146.low
Radial
Melanoma DVP
Reference proteomic dataset
Target proteomic dataset
scpDeconv
Melanoma clinical samples
Vertical
Radial
CD146.high
CD146.low
Insitu
Mel
Stroma
b
proteins
Protein abundance
10
5
0
-5
-10
Composition
1.00
0.75
0.50
0.25
0.00
CellType
Vertical
Radial
CD146.high
CD146.low
Insitu
Melanocytes
Stroma
Melanoma clinical samples
c
Composition
0.4
0.2
0.0
Vertical
Radial
CD146.high
CD146.low
Insitu
Melanocytes
Stroma
Melanoma class
d
Vertical
low
high
Survival probability
p = 0.031
Time (Months)
Number at risk
Vertical
low 87 29 12 6 2
high 87 21 7 2 2
0 25 50 75 100
Time (Months)
e
Radial
low
high
Survival probability
p = 0.016
Time (Months)
Number at risk
Radial
low 87 20 8 5 2
high 87 30 11 3 2
0 25 50 75 100
Time (Months)
f
Insitu
low
high
Survival probability
p = 0.33
Time (Months)
Number at risk
Insitu
low 87 25 8 3 2
high 87 25 11 5 2
0 25 50 75 100
Time (Months)

ED Figure 1

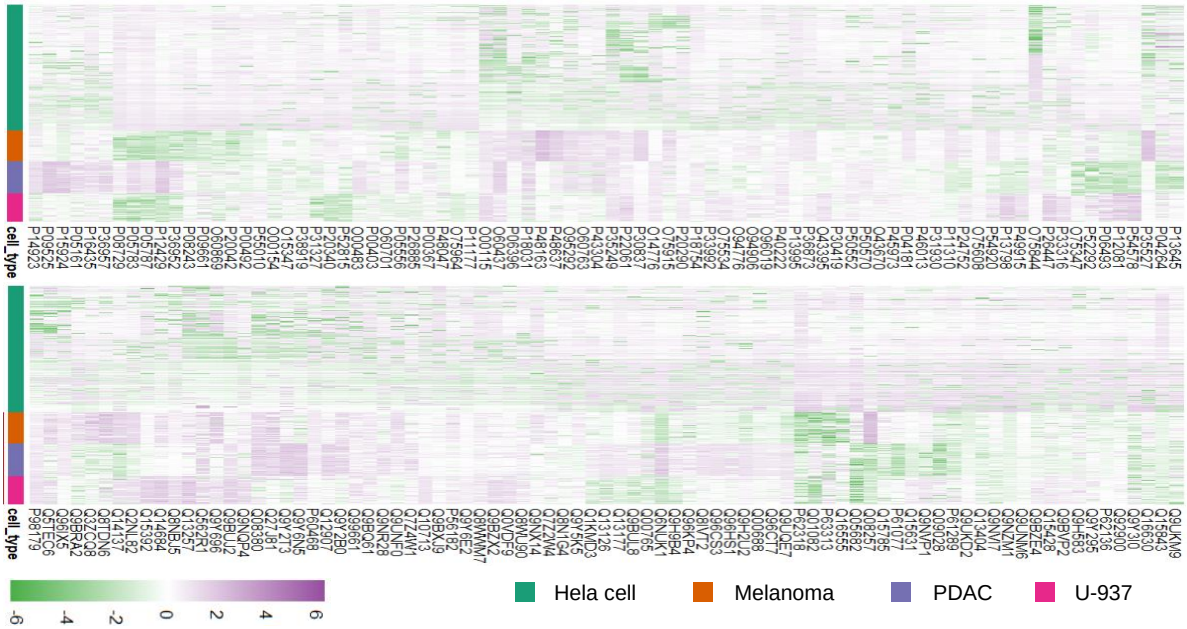

a
Cor
Methods
scpDeconv
Scaden
MuSiC
BayesPrism
DestVI
human_breast_atlas
mouse_cellline
human_cellline
mel2mon_cellcycle
mon2mel_cellcycle
Datasets
b
Transcriptome A
Proteome A
Proteome B
AV
BA
FI
HS
IM
VA
BRCA1
KRT17
GATA3
MUC1
PROM1
FAS
CD24
CD47
ANPEP
HSPB1
ICAM1
ESR1
THY1
PROCR
PGR
PDGFRB
NT5E
NR3C1
ERBB2
ACTA2
KRT8
ITGA6
EGFR
MME
EPCAM
VIM
KRT14
LAMB3
ANXA8
CD44
PECAM1
PTPRC
LGALS1
c
RMSE
direction
protein_to_protein
RNA_to_protein
method
scpDeconv
Scaden
MuSiC
BayesPrism
DestVI
d
tSNE_2
tSNE_1
N2
nanoPOTS
e
RMSE
mixed-up reference sample number
f
RMSE
mixed-up reference sample size
g
C10
SVEC
RAW
CCC = 0.634
RMSE = 0.117
Corr = 0.920
CCC = 0.713
RMSE = 0.107
Corr = 0.932
CCC = 0.832
RMSE = 0.087
Corr = 0.962
Ground Truth
Prediction
No imputation

ED Figure 2

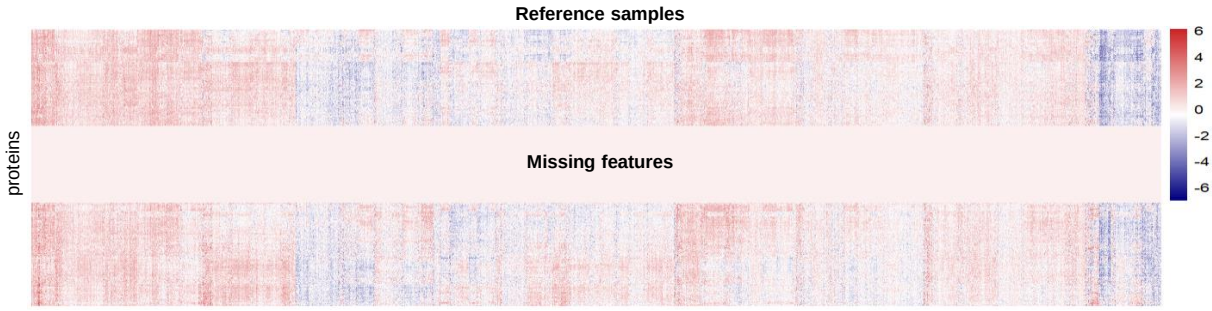
a



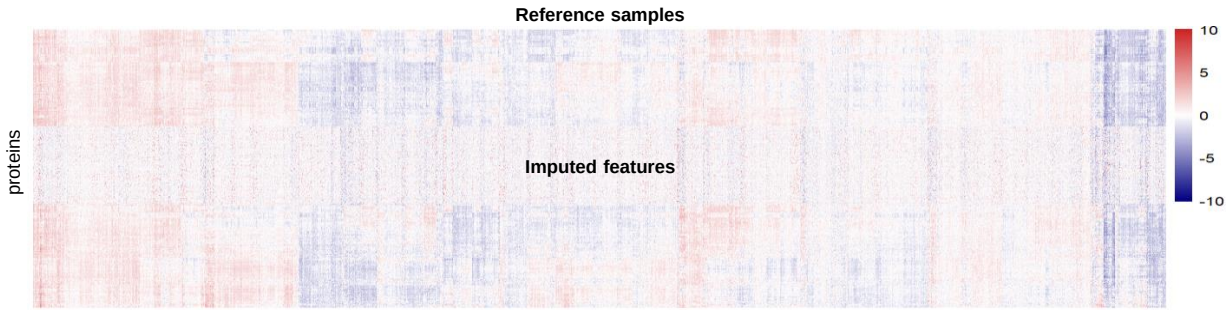
b



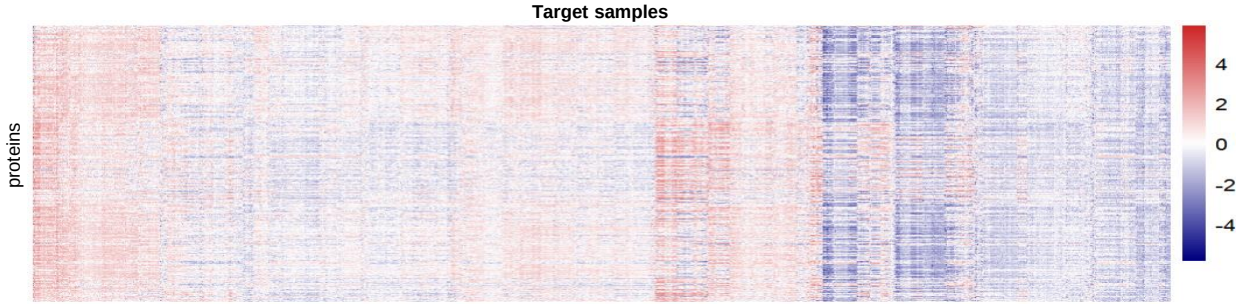
c



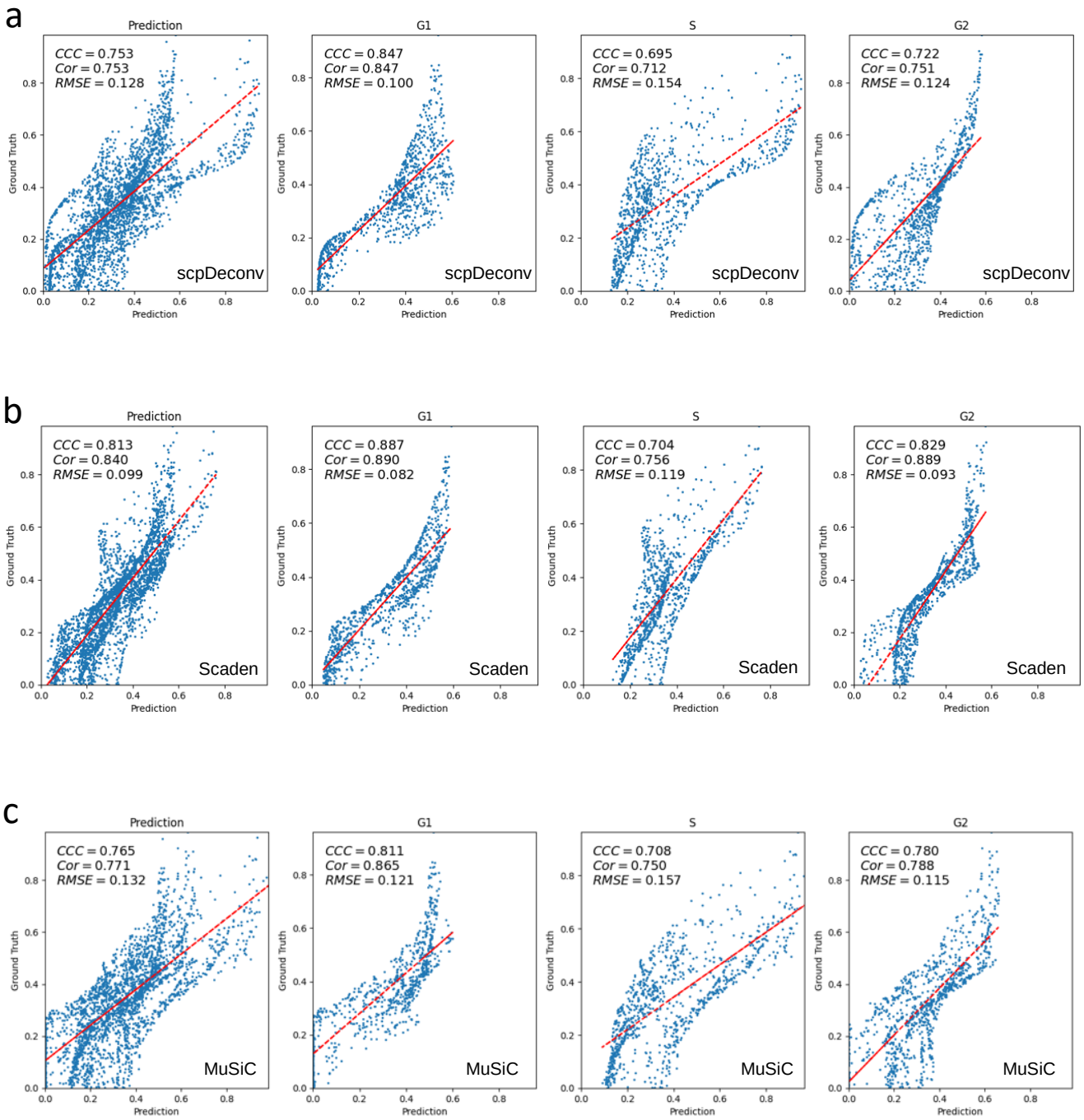
d



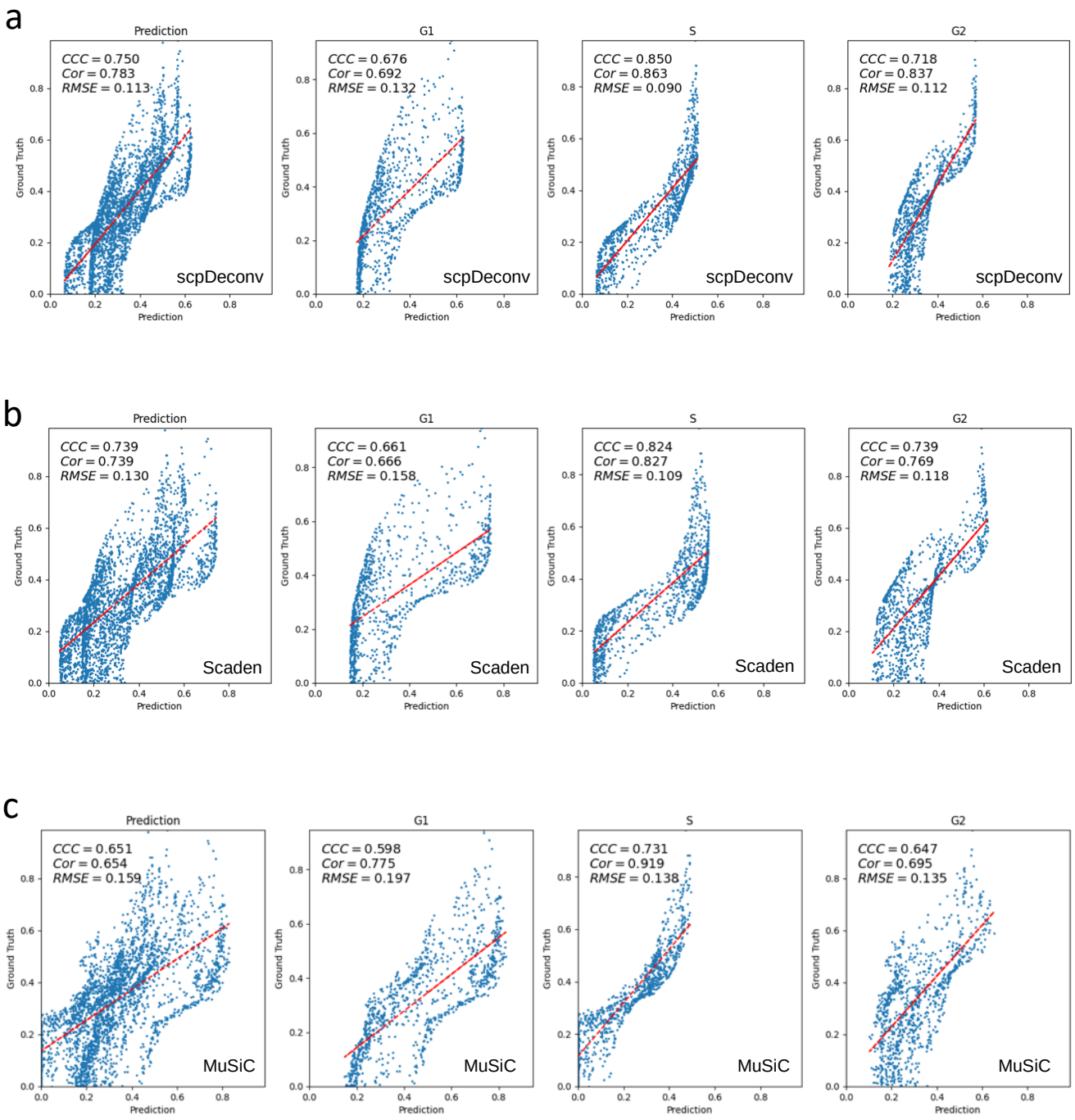
e



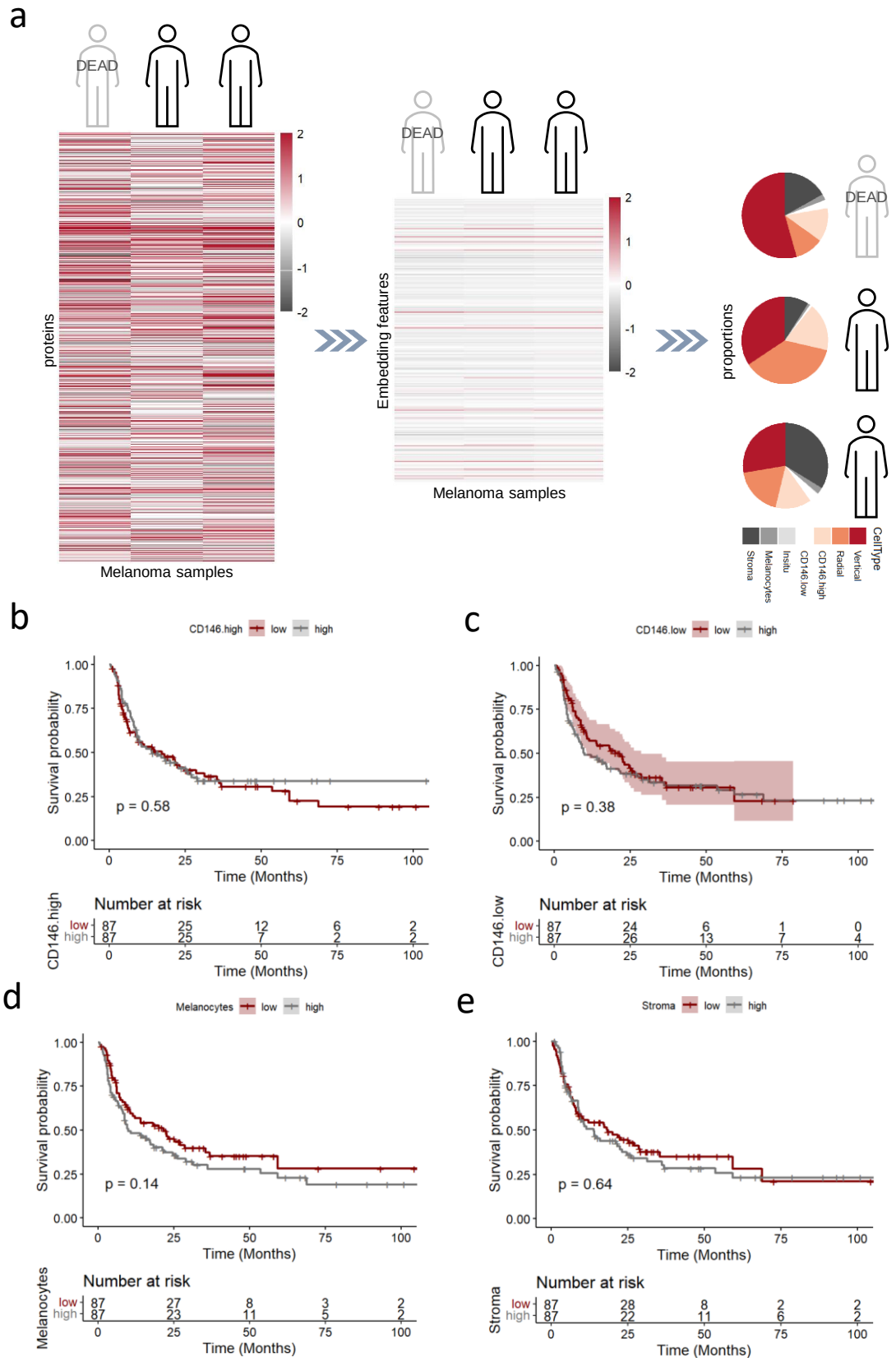
ED Figure 3


a
Prediction
CCC = 0.753
Cor = 0.753
RMSE = 0.128
scpDeconv
G1
CCC = 0.847
Cor = 0.847
RMSE = 0.100
scpDeconv
S
CCC = 0.695
Cor = 0.712
RMSE = 0.154
scpDeconv
G2
CCC = 0.722
Cor = 0.751
RMSE = 0.124
scpDeconv
b
Prediction
CCC = 0.813
Cor = 0.840
RMSE = 0.099
Scaden
G1
CCC = 0.887
Cor = 0.890
RMSE = 0.082
Scaden
S
CCC = 0.704
Cor = 0.756
RMSE = 0.119
Scaden
G2
CCC = 0.829
Cor = 0.889
RMSE = 0.093
Scaden
c
Prediction
CCC = 0.765
Cor = 0.771
RMSE = 0.132
MuSiC
G1
CCC = 0.811
Cor = 0.865
RMSE = 0.121
MuSiC
S
CCC = 0.708
Cor = 0.750
RMSE = 0.157
MuSiC
G2
CCC = 0.780
Cor = 0.788
RMSE = 0.115
MuSiC
Ground Truth
Prediction

ED Figure 4

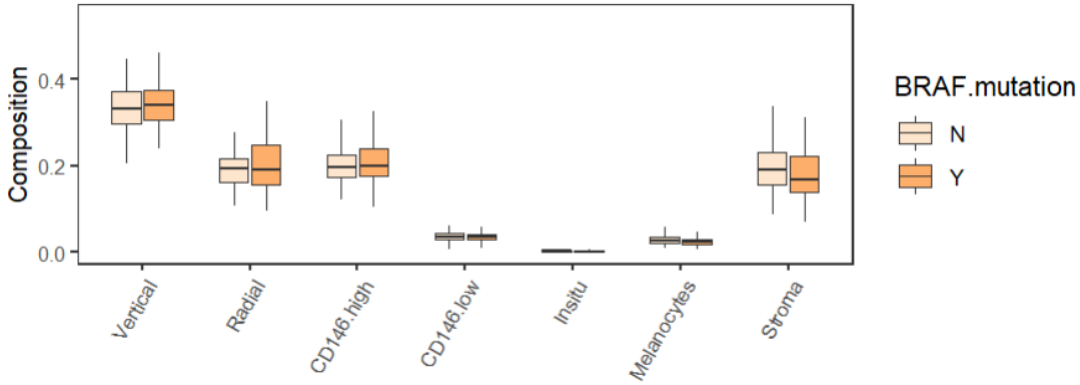

a
Prediction
CCC = 0.750
Cor = 0.783
RMSE = 0.113
scpDeconv
G1
CCC = 0.676
Cor = 0.692
RMSE = 0.132
scpDeconv
S
CCC = 0.850
Cor = 0.863
RMSE = 0.090
scpDeconv
G2
CCC = 0.718
Cor = 0.837
RMSE = 0.112
scpDeconv
b
Prediction
CCC = 0.739
Cor = 0.739
RMSE = 0.130
Scaden
G1
CCC = 0.661
Cor = 0.666
RMSE = 0.158
Scaden
S
CCC = 0.824
Cor = 0.827
RMSE = 0.109
Scaden
G2
CCC = 0.739
Cor = 0.769
RMSE = 0.118
Scaden
c
Prediction
CCC = 0.651
Cor = 0.654
RMSE = 0.159
MuSiC
G1
CCC = 0.598
Cor = 0.775
RMSE = 0.197
MuSiC
S
CCC = 0.731
Cor = 0.919
RMSE = 0.138
MuSiC
G2
CCC = 0.647
Cor = 0.695
RMSE = 0.135
MuSiC
Ground Truth
Prediction

ED Figure 5

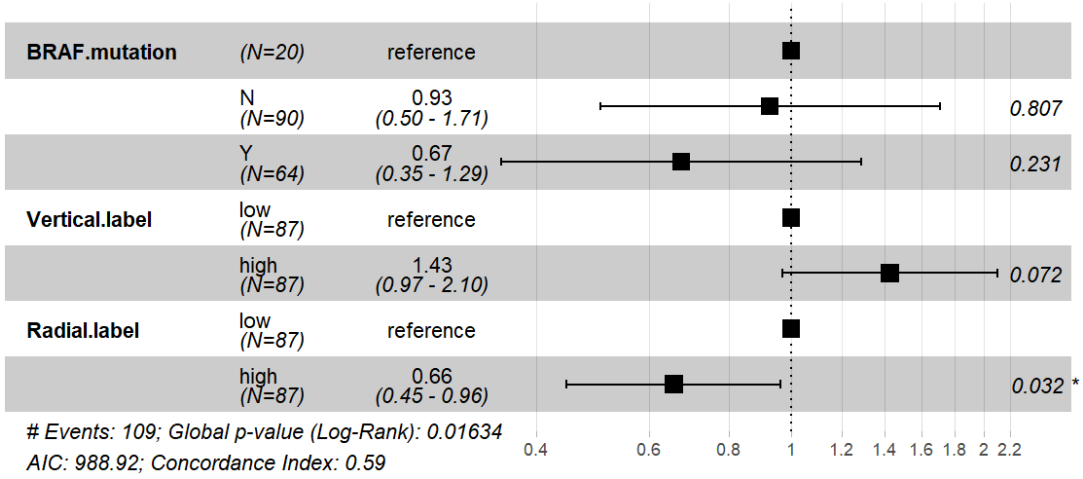

a
DEAD
proteins
Melanoma samples
Embedding features
proportions
CellType
Stroma
Melanocytes
Insitu
CD146.low
CD146.high
Radial
Vertical
b
CD146.high low high
Survival probability
p = 0.58
Time (Months)
Number at risk
CD146.high
c
CD146.low low high
p = 0.38
d
Melanocytes low high
p = 0.14
e
Stroma low high
p = 0.64

ED Figure 6

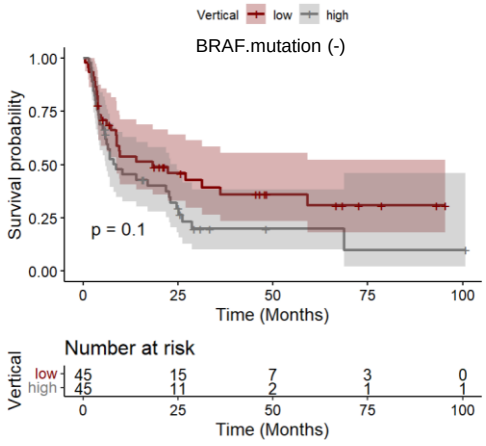
a



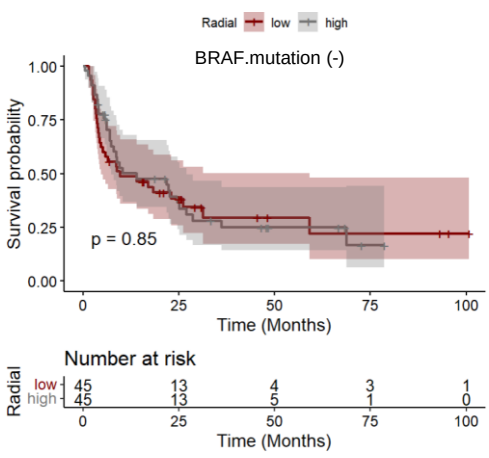
b



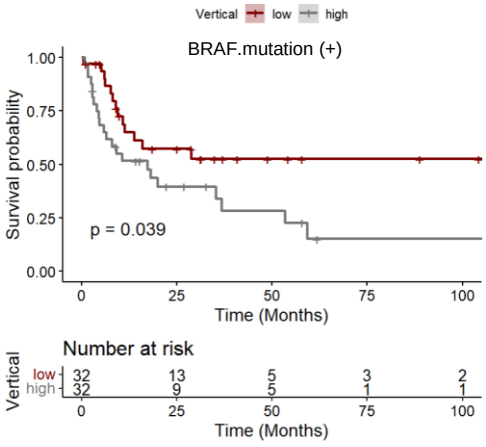
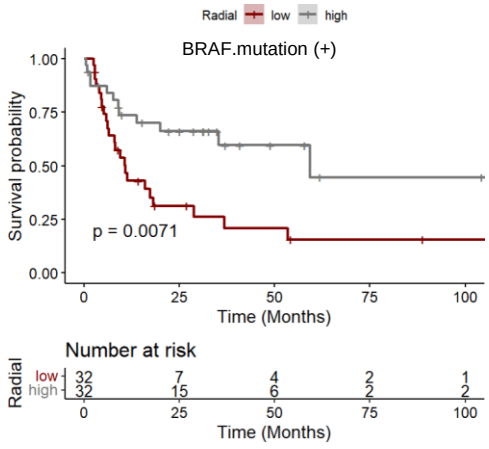
c



d



e



f