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

Fang Wang[1,2,#], Fan Yang[2,#], Longkai Huang[2], Jiangning Song[2,3], Robin B. Gasser[4], Ruedi Aebersold[5,*], Guohua Wang[1,*], Jianhua Yao[2,*]

1, School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001, China.

2, AI Lab, Tencent, Shenzhen, 518000, China.

3, Biomedicine Discovery Institute and Department of Biochemistry and Molecular Biology, Monash University, Melbourne, VIC 3800, Australia.

4, Department of Veterinary Biosciences, Melbourne Veterinary School, The University of Mel-bourne, Parkville, VIC 3010, Australia.

5, Institute of Molecular Systems Biology, Department of Biology, ETH Zürich, Zürich, Switzerland.

[#]These authors contributed equally: Fang Wang, Fan Yang.

*These authors jointly supervised this work: Ruedi Aebersold, Guohua Wang, Jianhua Yao.

Ruedi Aebersold (aebersold@imsb.biol.ethz.ch)

Guohua Wang (ghwang@hit.edu.cn)

Jianhua Yao (jianhuayao@tencent.com)

Abstract

Cell type deconvolution is a computational method for the determination/resolution of cell type proportions from bulk sequencing data, 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 due to challenges with repeatability/reproducibility, variable reference standards and the lack of single-cell proteomic reference data. Here, we developed a novel deep learning-based deconvolution method (scpDeconv) tailored to proteomic data. 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 to bulk data. Extensive experiments validated the performance of scpDeconv in the deconvolution of proteomic data produced from various species/sources and different proteomic technologies. This method should find broad applicability to areas including tumor microenvironment interpretation and clinical diagnosis/classification.

Introduction

The quantitative partitioning of cell types and the estimation of their relative proportions in a tumor sample is of important clinical and prognostic significance. E.g. it can provide critical information for chemo- and immune-therapy[1–3]. Bulk sequencing technologies generally refer to lysing and sequencing all cells in a particular tissue, organ or organ system, and the subsequent measurement of the average abundance of genes/proteins. Bulk sequencing thus ignores the cellular heterogeneity of the sample[4]. The development of single-cell (transcriptome/proteome) sequencing technologies allows the isolation of single cells from tissue(s) for analysis and measurement of the respective molecules[5,6]. Despite the conceptual advantages of this technology, it has been challenging to apply it to explore cellular heterogeneity in tumor microenvironments in large sample cohorts due to the relatively high cost and incompatibility with samples that are formalin-fixed or paraffin-embedded tissue[7,8]. Therefore, computational methods have been developed to infer cell types and their proportions from transcript expression profiles from bulk tissue samples[7–10], providing a relatively low-cost and convenient way of studying the cellular composition of tumor microenvironments in a clinical context.

The deconvolution of cell types for transcriptomic data derived from bulk tissue samples[4,8,11] generally assumes that the mixture of cells can be modelled as a system of linear weight equations, therefore non-negative matrix factorization can be applied to estimate cell type proportions[8]. However, in reality, the expression profiles within each cell type are heterogeneous e.g. due to different cell cycle and biochemical states of tested cells, the microenvironment and background in sequence data. The resulting

52 feature variability impedes high quality modelling using a linear distribution assumption. For this reason,
53 alternative deep-learning-based methods consisting of the following steps have been proposed: i) collect
54 scRNA-seq data with corresponding cell type labels; ii) train a deep learning model with labelled scRNA-
55 seq data as reference; and iii) infer the corresponding cell type proportions of bulk data with the trained
56 model[9,12,13]. In reality, these deep learning models were shown to be prone to overfitting on the reference
57 datasets and failed on unseen bulk data from another source[9]. As a remedy, some methods were proposed
58 to finetune the model on new datasets with known cell type proportions to improve performance[9].
59 However, the finetuning strategy limits the methods' generalization and is impractical in real scenarios
60 since clinical bulk data usually lack reliable cell type proportions beforehand.

61 To date existing deconvolution methods only focused on transcriptomics or genomics. No attention has
62 been paid to deconvolving bulk proteomic data despite of an urgent need for algorithms that perform this
63 task. The proteome provides critical biological and clinical information ranging from understanding
64 biological mechanisms to discovering drug targets[14]. Furthermore, the concentration of most proteins
65 cannot be accurately represented by the transcription/expression level of corresponding genes due to
66 complex post-transcriptional regulation, RNA/protein degradation and post-translation modification[15–17].
67 Meanwhile, most proteins are present at over 1000-fold higher copy numbers than the corresponding
68 transcripts and express much wider dynamic range of abundance than transcripts, and therefore show
69 higher potential to distinguish cells and cell states[15,18] and to provide unique biochemical insights. As a
70 consequence, single cell proteomic data with annotated cell type labels are necessary for deconvolving
71 bulk proteomic samples.

72 By far the most commonly used single-cell proteomic technologies (e.g., CyTOF, FACS) are mainly
73 based on molecularly tagged antibodies which can detect a few to dozens of surface proteins or marker
74 proteins in parallel in a high number of cells[22]. The limited number of proteins detected, however limits
75 CyTOF data as a basis for cell deconvolution. With the advent of mass spectrometry-based single-cell
76 proteomic technologies (e.g., SCoPE), the number of proteins that can be detected has significantly
77 increased to up to 1000-3000 intracellular proteins and thus expanded both the number and types of
78 proteins detected[19–21,23,24]. However, the processing of sufficient numbers of cells to generate data to
79 support cell type deconvolution from bulk data remains challenging. Further, compared with
80 transcriptomic data, single cell proteomic data to train and benchmark deconvolution algorithms face a
81 number of additional challenges. These include significant background noise in single cell proteomic
82 data, poor data quality and marked variation depending on analysis run and/or technology, and limited
83 proteome coverage. Furthermore, the existing deconvolution methods were designed for transcriptomic
84 data and are not directly applicable to proteomic data, because i) different distributions and value ranges

between protein abundance and transcript expressions[15]. Specifically, most existing deconvolution methods for transcriptomic data are based on an assumed distribution, such as a Poisson or negative binomial distribution which are recognized as unsuitable for proteomic data[4,11], ii) much lower coverage of proteome detected by single cell proteomics, compared with bulk proteomics[25,26], a situation that is not considered by existing deconvolution methods, and 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[27].

Here, to tackle these challenges, we propose a novel deep-learning-based method, scpDeconv, to estimate cell type proportions for tissue proteomic data, with single cell proteomic data from similar cohorts as reference. Based on the framework of mix-up strategy, autoencoder and domain-adversarial neural network, scpDeconv can circumvent the need to generate huge simulation datasets, make efficient use of single-cell proteomic data, impute the missing low-abundance proteins with autoencoder, and establish a deconvolution module by domain adversary training, thereby improving the model's performance, particularly in proteomic data and real-world scenarios. The deconvolution approach developed will give 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 algorithm scpDeconv

In this study, we developed scpDeconv to infer cell proportions from bulk proteomic data with single cell proteomic data as reference. To fit for the characteristics of proteomic data, we blended mix-up strategy, autoencoder and domain adversary neural network to assemble the overall framework of scpDeconv (Figure 1).

In stage 1 of scpDeconv (Figure 1a), we collected single cell proteomic datasets annotated with labelled cell types from the same species as the analyzed bulk tissue proteomic data (termed target data to be deconvolved). We then randomly mixed up single-cell proteomic data to construct pseudo-bulk data with corresponding cell type proportions as reference data.

In stage 2 of scpDeconv (Figure 1b), we applied an autoencoder model to impute missing low-abundance proteins in the 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 jointly 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 low-abundance proteins which are undetected in single cell proteomic data.

In stage 3 of scpDeconv (Figure 1c), 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, the method can efficiently improve the domain generalizability to overcome diverse distributions from single-cell proteomic data and bulk proteomic data.

Finally, in stage 4 of scpDeconv (Figure 1d), the corresponding cell type proportions of target proteomic samples are estimated by deep neural network trained in stage 3, and then utilized for re-analysis of biological mechanism or tumor microenvironment of specific tissue samples.

Overall, scpDeconv is designed for a real scenario in which bulk proteomic data from clinical or biological specimens are deconvolved to estimate the proportions of cell types constituting the tissue. This is accomplished by the use of prior single cell proteomic data from particular species and tissues and the annotated cell types as reference. It can be expected that with the present advances in the single cell proteomic technologies, a rapidly increasing number of high quality single cell proteomic data from various species and tissues will become available, thus increasing the reach of the presented method.

Quantitative evaluation of cell type deconvolution performance

To quantitatively evaluate scpDeconv's performance of cell type deconvolution, we carried out benchmarking calculations by simulating 3 different cases of increasing complexity. These are, i), cell type deconvolution across modalities using human breast atlas datasets consisting of scRNA-seq data and CyTOF data from the same source and involving 6 shared cell types within breast tissue; ii), cell type deconvolution across different data sources using 2 murine cell line datasets with 3 shared cell types generated with different single cell proteomic technologies, and iii), cell type deconvolution using integrated datasets from 5 human cell line datasets acquired by different single cell proteomic technologies from 4 shared cell types. In each case, we carried out the following steps to assess the deconvolution performance. First, we chose one or more single cell proteomic datasets to construct pseudo-bulk proteomic data with known cell type composition as reference data. Second, we also used similar single cell proteomic data in the same case to simulate pseudo-bulk proteomic data with true cell

146 proportions as target data. Third, we employed scpDeconv to predict cell proportions of these target data.
147 Finally, we compared scpDecov with four representative and state-of-the-art deconvolution methods
148 developed for transcriptomic data, involving the deep-learning-based method (Scaden)[28], the linear-
149 regression-based method (MuSiC)[4], the Bayes-based method (BayesPrism)[29] and the distribution-
150 hypothesis-based method (DestVI)[30].

151 As shown in Figure 2a and Extended Data Figure 1a, evaluating by Lin's concordance correlation
152 coefficient (CCC), Root Mean Square Error (RMSE), and Pearson product-moment correlation
153 coefficient (*r*), scpDeconv significantly outperformed Scaden, MuSiC, BayesPrism and DestVI in all the
154 simulated cell type deconvolution cases. This is mainly due to the rational design of imputation to
155 overcome the problem of missing low-abundance proteins in reference single-cell proteomic data, and the
156 domain adversary training tailored for distribution differentiation between reference data and target data.
157 The performance comparisons of the different tools provided further insights, as follows: i), compared to
158 cases where cell type distribution data from the same source were available, exemplified the human breast
159 atlas data, the performance of Scaden dropped when the target data were derived from samples with
160 unseen cell distributions as was the case with murine cell line data from two different sources. This
161 feature limits the general applicability of Scaden; ii), The performance of MuSiC was hampered by its
162 heavy dependency on the consistency of gene weights across subjects, a feature that might be highly
163 inconsistent between reference and target data; iii), BayesPrism showed poor performance due to its
164 reliance on parameter estimation which required high-quality training data that are challenging to
165 generate with present single cell proteomic technologies; vi), due to rigorous assumption about the
166 distribution of transcripts in transcriptomic data, DestVI showed the worst performance and was hard to
167 migrate to proteomics.

168 To compare the performance of prior cell type deconvolution algorithms with scpDeconv with
169 transcriptomic and proteomic data, respectively, we carried out a deep exploration on the human breast
170 atlas data. This is a large-scale clinical dataset where transcript data was generated by scRNA-seq and
171 proteomic data was generated by CyTOF from 6 kinds of cell types present in breast tissue. Specifically,
172 the cell types were alveolar cells, hormone sensing cells, basal cells, fibroblasts, vascular lymphatic cells,
173 and immune cells[31]. 33 genes/proteins were measured in both the transcriptomic and proteomic data. To
174 carry out the comparison under strict conditions, paired scRNA-seq data$^{(A)}$ and CyTOF data$^{(A)}$ from the
175 same donor (A) and CyTOF data$^{(B)}$ from another donor(B) were selected from the human breast atlas
176 dataset. We extracted the mean expression/abundance of 33 genes/proteins and calculated the correlations
177 between gene expression and protein abundances within each cell type. As shown in Figure 2b and

178 Extended Data Figure 1b, the corresponding transcript (scRNA-seq data$^{(A)}$) and protein (CyTOF data$^{(A)}$)
179 patterns showed relatively low correlation, even though they were sampled from the same donor. In
180 contrast, the proteomic profiles from CyTOF data$^{(A)}$ or CyTOF data$^{(B)}$ were highly similar. Furthermore,
181 to explore the benefits of using single cell proteomic data as reference in deconvolving bulk proteomic
182 data, we simulated pseudo-bulk samples by using CyTOF data$^{(B)}$, and then used scpDeconv to deconvolve
183 them under the guidance of scRNA-seq data$^{(A)}$ or CyTOF data$^{(A)}$. In this case, compared with other
184 methods, scpDeconv obtained the best performance on both deconvolution tasks with different reference
185 data (Figure 2c-d and Extended Data Figure 1c). In addition, scpDeconv, Scaden and MuSiC performed
186 better on the deconvolution task with CyTOF data$^{(A)}$ as reference than that with scRNA-seq data$^{(A)}$ as
187 reference. In contrast, DestVI performed better on transcript data compared to CyTOF data due to its
188 heavy reliance on the assumed distribution of transcripts.

189 Overall, these results indicate the necessity and the benefits of deconvolving bulk proteomic data with
190 single-cell proteomic data as reference (Figure 2d and Extended Data Figure 1c).

191 In the case of the murine cell line datasets, we collected two independent single cell proteomic datasets
192 using two advanced technologies (N2 and nanoPOTs) from the 3 common murine cell types (C10, RAW,
193 and SVEC, respectively) grown under standard culture conditions. The two datasets showed a small
194 overlap in protein types (Figure 2e) and different data distributions in terms of protein abundance
195 (Extended Data Figure 1d). To assess scpDeconv's performance on samples from different acquisition
196 methods, we specified the mixed-up N2 dataset as reference data and the mixed-up nanoPOTS dataset as
197 target data. We then executed scpDeconv to estimate proportions of C10, RAW, and SVEC for target data
198 (Figure 2f) and compared the results with those generated by other deconvolution methods. In this test,
199 scpDeconv achieved the best CCC score, RMSE and Pearson's correlation values (Figure 2a, Extended
200 Data Figure 1a). Furthermore, when altering the sample numbers/sizes of the mixed-up N2 dataset in
201 stage 1, scpDeconv maintained stable performance and showed a high level of robustness (Figure 2g-h).
202 In addition, we also validated the effect of stage 2 (imputation module) by removing it from the workflow
203 of scpDeconv. This significantly lowered the performance on the deconvolution task (Extended Data
204 Figure 1g).

205 Performance of imputing missing proteins in cell type deconvolution on human 206 cell line datasets

As mentioned in *Method* part, *Stage 2* of scpDeconv is specifically designed for proteomics deconvolution by developing an autoencoder that imputes missing proteins of reference data (especially the highly variable proteins observed in target data). To systematically analyze the effect of the imputation module on spcDeconv, we carried out an ablation study on missing proteins using the human cell line datasets by changing the extent of missing proteins in reference data. To construct a more complex scenario with large gaps of detected proteins and altered data distributions, we integrated 3 single cell human cell lines datasets (namely SCoPE2_Leduc, pSCoPE_Huffman, and pSCoPE_Leduc) as the reference group (Figure 3a), and 2 additional single cell human cell lines datasets (namely T-SCP and plexDIA) as the target group (Figure 3b). The datasets shared 4 human cell lines but suffered from severe batch effects and strikingly different identified protein types (Figure 3c, Extended Data Figure 2a). In particular, in the reference group, the highly variable proteins (HVPs) specific in the target group were not detected, although they are some extent informative for deconvolution task (Figure 3d and Extended Data Figure 2b). The scpDeconv imputation module in *Stage 2* was added to tackle this problem.

The ablation study on missing proteins showed that the imputation module improved the model's performance to some extent by supplementing input features with the undetected, target-specific HVPs. Performance improvement was particularly noticeable for the prediction accuracy of the proportion of PDAC cells in the sample (Figure 3e-g). In addition, we also explored the magnitude of the effect of target HVPs on the deconvolution task by varying the number of target HVPs between 400 to 700. The results showed that irrespective of the number of HVPs added the system achieved a better performance compared to the performance without target-specific HVPs (Figure 3e). As shown in Extended data figure 2c-e, after the imputation stage, missing features within the reference data can be complemented on the basis of target data.

Quantitative evaluation of cell cycle deconvolution across cell lines

Besides the cell type deconvolution task, inference of cellular state is also critical for studying biological processes, tumor microenvironment, and cancer therapy[32–34]. Thanks to the ultra-high sensitivity of mass spectrometry, proteomic technologies can detect the subtle changes of functional proteins that determine or correlate with the state of a cell. This is exemplified by different cell cycle states which are characterized by well understood changes in abundance of multiple proteins[35,36]. To quantitatively evaluate scpDeconv's performance for the challenging estimation of the distribution of cell cycle states in bulk proteomic samples, we collected bulk proteomic data with constant and pure cell cycle state and cell type for each sample (namely G1/S/G2-state melanoma, or G1/S/G2-state monocyte samples). We used

these data to simulate pseudo-bulk proteomic data with randomly mixed-up cell states at a constant cell type distribution. Furthermore, to force the model to focus on cell cycle states rather than cell types, only functional proteins (top 20 for each state) corresponding to certain cell cycle states were chosen as input features (Figure 4a-b). By comparing the abundance of these 60 functional proteins between G1/S/G2-state melanoma and G1/S/G2-state monocyte samples, we noticed that functional proteins of cell cycle states were highly conserved between melanoma and monocyte samples (Figure 4c), which means that cell cycle deconvolution across cell types/lines is feasible. Therefore, we used scpDeconv, along with other deconvolution methods to compare their respective performance in estimating cell cycle state proportions for pseudo-bulk data simulated by G1/S/G2-state monocyte samples with G1/S/G2-state melanoma samples as reference (marked as *mel_to_mon*), and vice versa (marked as *mon_to_mel*). As shown in Figure 4d-e, and Extended Data Figure 3-4, compared with other deconvolution methods, scpDeconv achieved superior performance for both cell cycle state deconvolution studies, demonstrating the great potential of the new method for the inference of subtle differences in the distribution of cell cycle states in bulk samples and, by implication, the inference of other cell cycle states, provided proteins characteristic for these states are known and detectable.

Application to clinical tissue proteomics

Melanoma is a rare but invasive cancer of the skin, with high rates of growth and metastasis[37]. Previous clinical proteomic analyses of melanoma samples have focused on the detection of proteins as markers for disease stratification and responsiveness to therapy. Generally, these studies have not considered the tumor microenvironment and the effect of cellular heterogeneity on clinical outcomes[38]. Here, we explored the capacity of scpDeconv to re-analyze representative clinical proteomic data from the bulk analysis of 174 metastatic melanoma Grade IV samples consisting of 4884 high quality protein identifications[39]. 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 data, which contained 7 cell classes (vertical growth melanoma, radial growth melanoma, CD146-high melanoma, CD146-low melanoma, in situ melanoma, melanocytes, and stroma) identified in the tumor microenvironment[40]. Following the procedure of scpDeconv, the proportions of 7 cell classes were estimated from the tissue proteomic profiles of 174 melanoma samples (Figure 5b-c). The results indicated that the vertical growth melanoma cell class was dominant in the late-stage metastasis melanoma samples, while in situ melanoma cell class was the minority, consistent with previous clinical research and indicating that the presence of vertical growth phase signifies that the melanoma acquires the capacity of metastasis[41]. We further performed survival analysis using the predicted cell class proportions

for 174 melanoma samples as input data (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. Consistent with previous studies, these results showed that in contrast to melanoma in radial growth stage, melanoma with vertical growth pattern usually accompanies with strong aggressivity and poor prognosis[42,43].

In addition, as mutation of the BRAF protein 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).

Overall, the application of scpDeconv on a clinical proteome cohort indicated that the 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 is becoming an important component of biological and clinical research[25]. One particularly important use case is the deconvolution of cell types/states from bulk tissue proteome profiling data, a task that has not yet been achieved. By fully considering the inherent characteristics of the recently developed single-cell proteomics technologies, we developed a novel deconvolution method, scpDeconv, that is specifically tailored to proteomic data. By feat of an autoencoder network, scpDeconv's imputation module can impute values for low-abundance proteins missing in single-cell proteomic data. This is achieved 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 superior performance and robustness on cell type deconvolution as well as the inference of cell cycle states from bulk tissue proteomic datasets. 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 to transcriptomics, scpDeconv offers four levels of advantages for proteome data deconvolution: i), scpDeconv infers cell type proportions based on joint-learning of reference data and target data in a data-driven manner and is thereby unrestricted by the prior assumptions of data distribution and compatible with a broad range of proteomic data from different experiments or technology background; ii), scpDeconv introduces a domain adaptation strategy which can efficiently handle batch effects and close the distribution gaps between reference data and target data; iii), scpDeconv can impute missing values of low-abundance proteins in single cell proteomic data and improve the quality of reference data with the help of an imputation module; iv), we have validated that the output of scpDeconv is capable of reflecting cell type distribution in the tumor microenvironment and that the thus identified cell type distributions are useful for disease diagnosis and prognosis. At present, scpDeconv shows some limitations. Reference proteomic data with known cell type composition needed for model training are constructed from single-cell proteomic data with labelled cell types. At present single-cell proteomics is still in its infancy and has not been applied to a broad range of biological samples. The single-cell proteomic data we exhaustively collected in this study only involve a minority of cell types, thereby limiting our method's applicability at present.

This first algorithm tailored for cell type/state deconvolution from proteomic data using single-cell proteomic data as reference will have numerous key applications. These include the deconvolution of cell type proportions, estimation of cell states, exemplified by cell cycle states, and re-analysis of the microenvironment of multiple cancers thus adding additional value to already acquired bulk proteomic datasets. We therefore anticipate that scpDeconv will benefit the proteomics community by supporting the extraction of additional, biologically and clinically important results from bulk proteomic datasets. With the exploration of the publicly available single-cell proteomic atlas, it will be feasible in the future 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 stages: i), reference data preparation; ii), missing features imputation and data quality improvement; iii), domain-adversarial training and deconvolution; iv), cell type proportion inference for target samples.

Stage 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 similar cohort is termed the target data to be deconvolved. The reference data and target data are usually sourced from different sequencing platforms and experiments, thereby having gaps in data distribution and proteome coverage. Bulk proteomics can usually detect much more proteins compared to single-cell proteomics, in which the low-abundance proteins are hard to detect.

Mix-up strategy. Mix-up is an efficient way of data augmentation by a random combination of inputs[44]. Here we employed it on single-cell proteomic data to make up pseudo-bulk proteomic data before training. The random combination of cell type distributions from single-cell proteomic data increases the data diversity of the reference dataset for training, thereby improving the model's generalizability. In this stage, we generated the pseudo bulk proteomic data $\widetilde{X_r}$ with random cell type proportions $Y_r^{(1)}$ as reference data, and then sent it to the autoencoder module of stage 2 for data imputation.

Data alignment. Before model training, we need to align proteins of reference data and target data for model input. We firstly extracted the proteins shared between reference and target data. And then we took the union set of these shared proteins and 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 in the input feature set for reference data are initialized to zeros and will be imputed in *Stage 2*.

Stage 2. Missing features imputation and data quality improvement.

Due to the immaturity of single-cell proteomic technology, the reference data generated by single cell proteomic data have strong noises, large-scale missing low-abundance proteins, and an obvious distribution gap with the target bulk proteomic data. To compensate for the above obstacles caused by single-cell proteomic technology, we employed an autoencoder module to close the gap between the reference data and the target data, impute missing proteins of the reference data, and improve the quality of reference data. As shown in Figure 1b, the target data and the reference data generated in *Stage 1* are inputted into Encoder1. And then the latent embedding generated by Encoder1 is sent to a 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 *Stage 2* are:

$$\mathrm{E1}_r = \mathrm{Encoder1}\left(\widetilde{X_r}\right) \tag{1}$$

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

$$Y_r^{(2)} = \mathrm{Predictor1}(E1_r) \tag{3}$$

$$\mathrm{E1}_t = \mathrm{Encoder1}(X_t) \tag{4}$$

$$X_t^{'} = \mathrm{Decoder}(E1_t) \tag{5}$$

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

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

Where $E1_r$, $E1_t$, $Y_r^{(2)}$, $\widetilde{X_r}$, $\widetilde{X_r^{'}}$, X_t and $X_t^{'}$ are the embedding of the 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 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 reference data ($\widetilde{X_r^{'}}$) are used as the input of *Stage 3*.

Stage 3. Domain-adversarial training and deconvolution.

As shown in Figure 1c, the imputed reference data from *Stage 2* and the target data from *Stage 1* are sent to the Encoder2 in *Stage 3*, and the encoded features are sent to the domain discriminator and the deconvolution head. The domain discriminator is employed to discriminate whether the latent features are encoded from the reference data or the target data. By jointly training the Encoder2 and the discriminator in this *stage*, the Encoder2 is prone to extracting domain-invariant features, which is suitable for the deconvolution task in which the reference data and the target data are from different sources. The deconvolution head consists of 2 fully connected layers and is used for predicting cell type proportions of target data. The Encoder2 and deconvolution head are jointly trained 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 calculated BCE loss and optimized our model's weights by minimizing the distributions between the reference and target domains. The formulations of the architecture in *Stage 3* are:

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

$$D_r = Discriminator(E2_r) \tag{9}$$

$$E2_t = Encoder2(X_t) \tag{10}$$

$$D_t = Discriminator(E2_t) \tag{11}$$

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

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

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

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 deconvolution head, the output of discriminator with the embedding of imputed reference data, and the output of discriminator with the embedding of target data, respectively.

Stage 4. Cell type proportion inference for target samples.

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

$$Y_t = Predictor2(E2_t) \tag{15}$$

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[31], scRNA-seq was performed on 16 breast tissues (52,681 cells) and 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 cells, hormone sensing cells, basal cells, fibroblasts, vascular lymphatic cells, and immune cells). The processed scRNA-seq dataset was downloaded from GEO (accession no. GSE180878) and the processed CyTOF dataset was downloaded from https://data.mendeley.com/datasets/vs8m5gkyfn/1.

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)[24,49]. N2 dataset (Murine cell line dataset generated by N2) includes 108 cells (36 RAW cells, 36 C10 cells, and 36 SVEC cells) with 1,437 high quality protein identifications, and was downloaded from MassIVE data repository (accession no. MSV000086809). nanoPOTS dataset (Murine cell line dataset generated by nanoPOTS) includes 72 cells (24 RAW cells, 24 C10 cells, and 24 SVEC cells) with 1,032 high quality protein identifications, and was downloaded from MassIVE data repository (accession no. MSV000084110). These two datasets have 762 proteins in common.

Human cell line datasets. The human cell line datasets consist of five different single cell proteomic datasets, namely T-SCP, plexDIA, pSCoPE_Leduc, SCoPE2_Leduc, and pSCoPE_Huffman, generated by different state-of-the-art technologies[36,45–48]. T-SCP dataset contains 434 Hela cells and 2501 proteins and was downloaded from PRIDE partner repository (accession no. PXD024043). plexDIA was performed on 155 Melanoma, PDAC, and U-937 cells and detected 1475 high quality protein identifications, and was downloaded from https://scp.slavovlab.net/Derks_et_al_2022. pSCoPE_Leduc dataset (involving 1755 Melanoma and U-937 cells and 2844 proteins) was generated by pSCoPE technology and was downloaded from https://scp.slavovlab.net/Leduc_et_al_2022. SCoPE2_Leduc dataset (involving 163 Hela cell and U-937 cell and 1647 proteins) was generated by SCoPE2 technology and was downloaded from https://scp.slavovlab.net/Leduc_et_al_2021. pSCoPE_Huffman dataset (involving 206 PDAC cells and 1659 proteins) was also generated by pSCoPE technology and was downloaded from https://scp.slavovlab.net/Huffman_et_al_2022. After integration, these five datasets involved 4 cell types in total, consisting of 1,755 melanoma, 434 Hela cells, 579 PDAC cells, and 163 U-937.

Bulk proteomic cell cycle datasets. Bulk proteomic datasets in which each sample has pure and constant cell type and state (G1/S/G2-state melanoma and G1/S/G2-state monocyte samples) were generated in the same experiment as pSCoPE_Leduc dataset[48]. And following the authors' reported procedure, the top 20 highly expressed marker proteins of each cell cycle state (G1, S, G2) 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[40]. This dataset contains 7 cell classes (vertical growth melanoma, radial

448 growth melanoma, CD146-high melanoma, CD146-low melanoma, in situ melanoma, melanocytes, and
449 stroma) within the tumor microenvironment of melanoma, defined by spatial location and SOX10/CD146
450 staining intensity. We downloaded this dataset from the PRIDE partner repository (accession no.
451 PXD023904).

452 *The clinical proteomic dataset of metastatic melanoma.* The bulk proteomic dataset of metastatic
453 melanoma was collected from the previous clinical study, in which mass spectrometry-based proteomic
454 analysis was performed in 174 metastasis melanoma patients, with 4883 high quality protein
455 identifications[50].

456 Training details

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

465 Comparison of methods

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

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

476 *BayesPrism.* BayesPrism is a Bayesian-based deconvolution method for cancer tissues depending on
477 patient-derived scRNA-seq as prior information[29]. The Bayesian algorithm models prior distribution
478 learned from the paired scRNA-seq data and then infers joint posterior distribution of cell type
479 proportions as the deconvolution results.

480 *DestVI.* DestVI is also a Bayesian-based method that learns cell type specific profiles and latent variations
481 using a conditional deep generative model[30]. The method assumes that transcripts profiles follow a
482 negative binomial distribution to model the reference scRNA-seq data and spatial transcriptomic data.

483 Evaluation experiments

484 *Cell type mixture decomposition with proteome profiling.*

485 As for testing the deconvolution performance with the reference data and target data from different
486 modalities, in the simulation case of human breast atlas datasets, we deconvolved the simulated bulk
487 proteomic data constructed from CyTOF data with scRNA-seq data in the same cohort as reference or
488 with CyTOF data in the same cohort as reference, splitting the dataset into reference and target data by
489 donors.

490 There are apparent batch effects between different proteomic technologies and large gaps between the
491 distributions of bulk proteomic data and single-cell proteomic data. Therefore, to evaluate scpDeconv's
492 deconvolution performance on the situation where the bulk/single-cell proteomic data came from totally
493 different proteomic technologies and different experiments, we deconvolved bulk proteomic data
494 simulated by nanoPOTS dataset with the mixed-up N2 dataset as reference.

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

499 To explore the efficiency of the imputation of missing proteins in *Stage 2*, we executed cell type
500 deconvolution experiments on the integrated human cell line datasets with considering target-specific
501 HVPs or without considering target-specific HVPs, and also compared deconvolution performance
502 between model with imputation stage and model without imputation stage. For the experiments without
503 considering target-specific HVPs, we only used the overlap proteins from both reference and target

datasets as input features. For the experiments with considering target-specific HVPs, we changed the number of HVPs identified in target data from 400 to 700.

Application

Cell cycle inference with proteome profiling.

The cell cycle states 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, or G2 stage. Then the melanoma cell cycle mixtures were constructed by mixing up the melanoma cell subpopulations within G1, S, or G2 stage, and monocyte cell cycle mixtures were constructed by mixing up the monocyte subpopulations within G1, S, or G2 stage. Furthermore, to estimate scpDeconv's performance on the cell cycle inference task across cell types/cell lines, we deconvolved the melanoma cell cycle mixtures with monocyte cell cycle profiles as reference, and also deconvolved the monocyte cell cycle mixtures with melanoma cell cycle profiles as reference.

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

The establishment of scpDeconv that is tailored for cell type deconvolution of proteomic data opens up an opportunity to mine the tumor microenvironments of previously released clinical tissues with provided bulk proteomic data. We used scpDeconv to deconvolve the clinical proteomic data from metastasis melanoma cohort and assessed the impact of tumor class proportions on the prognostic determination according the survival analysis.

Evaluation metrics

The linear concordance between ground truth ratios and predicted proportions of cell types or cell states is strictly evaluated by 3 metrics, including Lin's concordance correlation coefficient (CCC), Root Mean Square Error (RMSE), and Person product-moment correlation coefficient (*r*). The formulas are as follows:

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

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

$$r(y, \hat{y}) = \frac{cov(y, \hat{y})}{\sigma_y \sigma_{\hat{y}}} \quad (18)$$

Where y are the ground truth cell type (or cell cycle state) proportions, $\hat{y}$ are the prediction of cell type (or cell cycle state) 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

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[51]. The Gene Ontology (GO) and pathway enrichment are performed using DAVID platform[52].

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 will be released after the publication of this manuscript. The comparison methods are implemented under their official guidance and the reference codes are listed in the Method section.

Acknowledgements

We thank Doctor Chen Li for giving suggestions for the project.

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. wrote the manuscript. F.W. finished the figures under the guidance of F.Y. and J.Y.. J.Y., G.W., R.A. and

R.G. revised the manuscript. L.H. gave suggestions for the domain adaptation and the improvement of the manuscript. J.S. gave suggestions for the project. F.W. and F.Y. contributed equally. All authors reviewed and approved the manuscript.

Competing Interests

The authors declare no competing interests.

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).
3. Shang, B., Liu, Y., Jiang, S. J. & Liu, Y. Prognostic value of tumor-infiltrating FoxP3+ regulatory T cells in cancers: a systematic review and meta-analysis. *Scientific Reports 2015 5:1* **5**, 1–9 (2015).
4. Wang, X., Park, J., Susztak, K., Zhang, N. R. & Li, M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. *Nature Communications 2019 10:1* **10**, 1–9 (2019).
5. Qian, J. *et al.* A pan-cancer blueprint of the heterogeneous tumor microenvironment revealed by single-cell profiling. *Cell Research 2020 30:9* **30**, 745–762 (2020).
6. Azizi, E. *et al.* Single-Cell Map of Diverse Immune Phenotypes in the Breast Tumor Microenvironment. *Cell* **174**, 1293-1308.e36 (2018).
7. Frishberg, A. *et al.* Cell composition analysis of bulk genomics using single-cell data. *Nature Methods 2019 16:4* **16**, 327–332 (2019).
8. Newman, A. M. *et al.* Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology 2019 37:7* **37**, 773–782 (2019).
9. Menden, K. *et al.* Deep learning-based cell composition analysis from tissue expression profiles. *Sci Adv* **6**, (2020).
10. Cable, D. M. *et al.* Robust decomposition of cell type mixtures in spatial transcriptomics. *Nature Biotechnology 2021 40:4* **40**, 517–526 (2021).
11. Patrick, E. *et al.* Deconvolving the contributions of cell-type heterogeneity on cortical gene expression. *PLoS Comput Biol* **16**, e1008120 (2020).
12. Lv, X. *et al.* Deconvolute Gene Expression Based on Deep Learning in Scrna-Seq. *Proceedings - 2021 17th International Conference on Computational Intelligence and Security, CIS 2021* 70–74 (2021) doi:10.1109/CIS54983.2021.00023.

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

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

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

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

594 17. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA
595 Abundance. *Cell* **165**, 535–550 (2016).

596 18. Specht, H. & Slavov, N. Transformative Opportunities for Single-Cell Proteomics. *J*
597 *Proteome Res* **17**, 2565–2571 (2018).

598 19. Strack, R. Spatial proteomics with subcellular resolution. *Nature Methods 2022 19:7* **19**,
599 780–780 (2022).

600 20. Petelski, A. A. *et al.* Multiplexed single-cell proteomics using SCoPE2. *Nature Protocols*
601 *2021 16:12* **16**, 5398–5425 (2021).

602 21. Woo, J. *et al.* High-throughput and high-efficiency sample preparation for single-cell
603 proteomics using a nested nanowell chip. *Nature Communications 2021 12:1* **12**, 1–11
604 (2021).

605 22. Cheung, R. K. & Utz, P. J. CyTOF—the next generation of cell detection. *Nature Reviews*
606 *Rheumatology 2011 7:9* **7**, 502–503 (2011).

607 23. Slavov, N. Unpicking the proteome in single cells:: Single-cell mass spectrometry will help
608 reveal mechanisms that underpin health and disease. *Science* **367**, 512 (2020).

609 24. Woo, J. *et al.* High-throughput and high-efficiency sample preparation for single-cell
610 proteomics using a nested nanowell chip. *Nature Communications 2021 12:1* **12**, 1–11
611 (2021).

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

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

614 27. Vanderaa, C. & Gatto, L. Replication of single-cell proteomics data reveals important
615 computational challenges. *https://doi.org/10.1080/14789450.2021.1988571* **18**, 835–843
616 (2021).

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

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

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

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

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

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

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

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

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

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

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

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

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

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

646 42. Prognosis and survival for melanoma skin cancer | Canadian Cancer Society.
647 https://cancer.ca/en/cancer-information/cancer-types/skin-melanoma/prognosis-and-
648 survival.

649 43. Ciarletta, P., Foret, L. & Amar, M. ben. The radial growth phase of malignant melanoma:
650 multi-phase modelling, numerical simulations and linear stability analysis. *J R Soc Interface*
651 **8**, 345 (2011).

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

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

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

658 47. Huffman, R. G. *et al.* Prioritized single-cell proteomics reveals molecular and functional
659 polarization across primary macrophages. *bioRxiv* 2022.03.16.484655 (2022)
660 doi:10.1101/2022.03.16.484655.

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

663 49. Dou, M. *et al.* High-Throughput Single Cell Proteomics Enabled by Multiplex Isobaric
664 Labeling in a Nanodroplet Sample Preparation Platform. *Anal Chem* **91**, 13119–13127
665 (2019).

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

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

670 52. Sherman, B. T. *et al.* DAVID: a web server for functional enrichment analysis and
671 functional annotation of gene lists (2021 update). *Nucleic Acids Res* **50**, W216–W221
672 (2022).

673

674

675

Figure 1

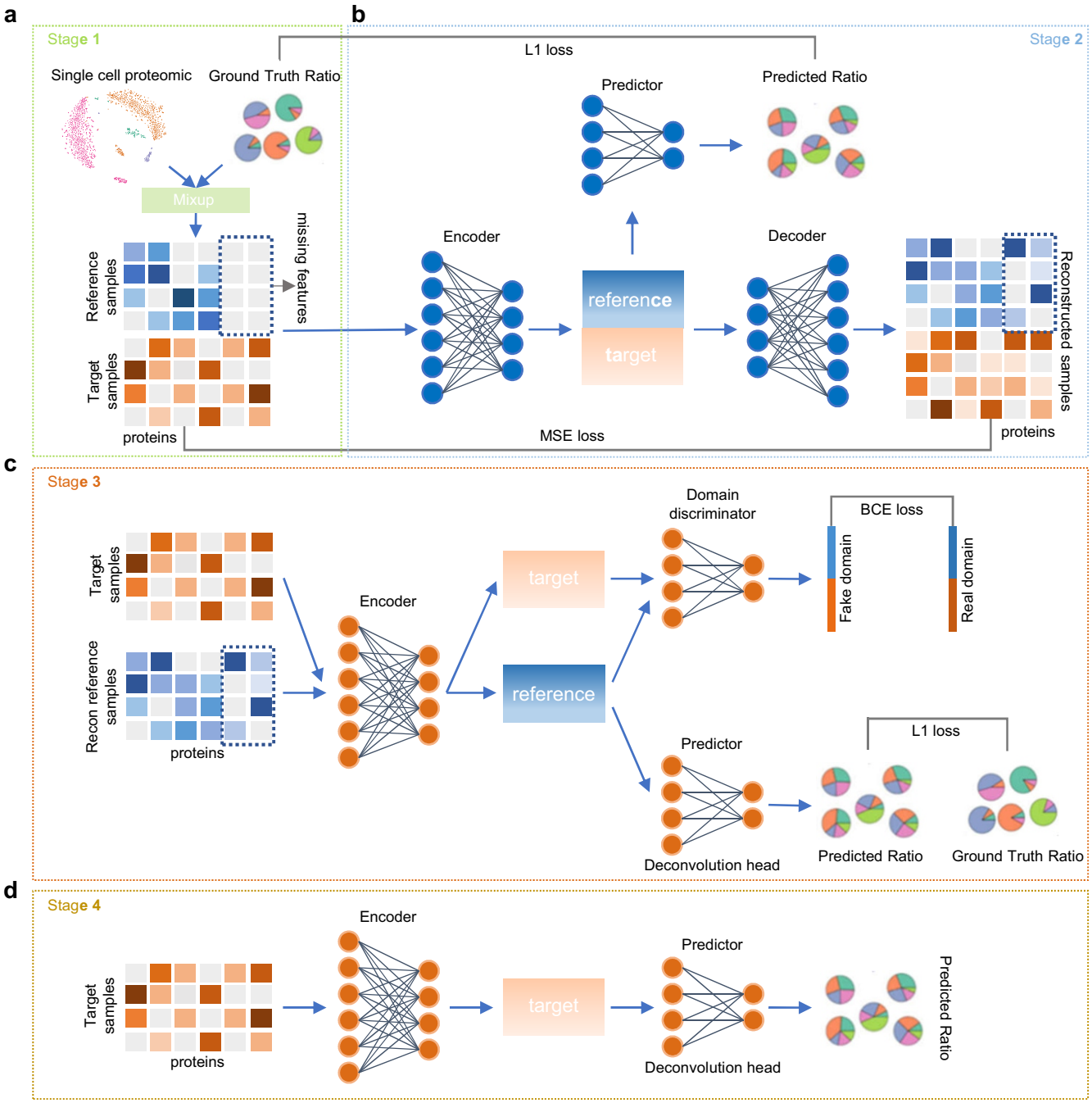

a
b
Stage 1
Stage 2
L1 loss
Single cell proteomic
Ground Truth Ratio
Predictor
Predicted Ratio
Mixup
missing features
Reference samples
Target samples
proteins
Encoder
reference
target
Decoder
Reconstructed samples
proteins
MSE loss
c
Stage 3
Domain discriminator
BCE loss
Target samples
Recon reference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

Figure 1. The architecture of our proposed method.

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

b, Stage 2: missing features imputation and data quality improvement. The encoder-decoder scheme to close the gap between the reference data and the target data and impute the mixed-up single-cell proteomic data generated in Stage 1. Reconstruction loss (MSE loss) and deconvolution loss (L1 loss) are applied for model training.

c, Stage 3: domain-adversarial training and deconvolution. The domain-adversary training architecture of neural network. The domain discriminator is used to discriminate whether the latent features are encoded from the reference data or from the target bulk proteomic data. The deconvolution head is used to predict the proportions of cell types.

d, Stage 4: cell type proportion inference for target samples. Target bulk proteomic data are inputted to the trained encoder and deconvolution head from stage 3 to infer the proportions of cell types.

Figure 2

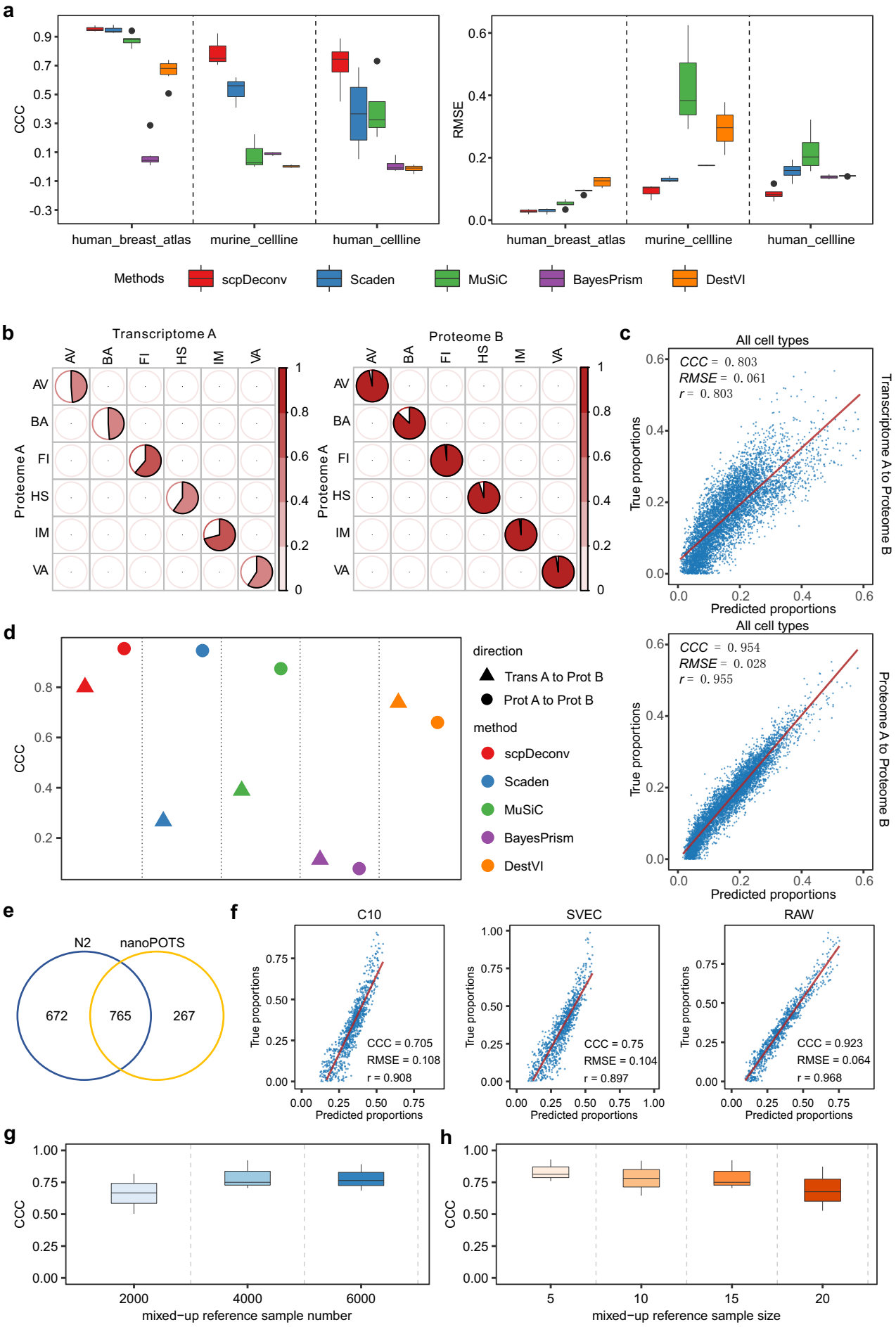

a
CCC
RMSE
human_breast_atlas
murine_cellline
human_cellline
Methods
scpDeconv
Scaden
MuSiC
BayesPrism
DestVI
b
Transcriptome A
Proteome B
Proteome A
AV
BA
FI
HS
IM
VA
c
All cell types
CCC = 0.803
RMSE = 0.061
r = 0.803
True proportions
Predicted proportions
Transcriptome A to Proteome B
CCC = 0.954
RMSE = 0.028
r = 0.955
Proteome A to Proteome B
d
direction
Trans A to Prot B
Prot A to Prot B
method
scpDeconv
Scaden
MuSiC
BayesPrism
DestVI
e
N2
nanoPOTS
672
765
267
f
C10
SVEC
RAW
CCC = 0.705
RMSE = 0.108
r = 0.908
CCC = 0.75
RMSE = 0.104
r = 0.897
CCC = 0.923
RMSE = 0.064
r = 0.968
g
mixed−up reference sample number
h
mixed−up reference sample size

Figure 2. Overall performance comparation of cell type mixture deconvolution.

a, CCC (left) and RMSE (right) values of scpDeconv, Scaden, MuSiC, BayesPrism, and DestVI on the testing cases of the human breast atlas datasets (proteome to proteome), murine cell line datasets, and human cell line 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 cells, HS: hormone sensing cells, BA: basal cells, FI: fibroblasts, VA: vascular lymphatic cells, IM: immune cells.

c, Scatter plots of true (y axis) and predicted cell type proportions (x axis) by deconvolving 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 A as reference (bottom) in the case of human breast atlas datasets.

d, CCC values of scpDeconv and the comparison methods by deconvolving 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 A as reference (circle) in the case of human breast atlas datasets.

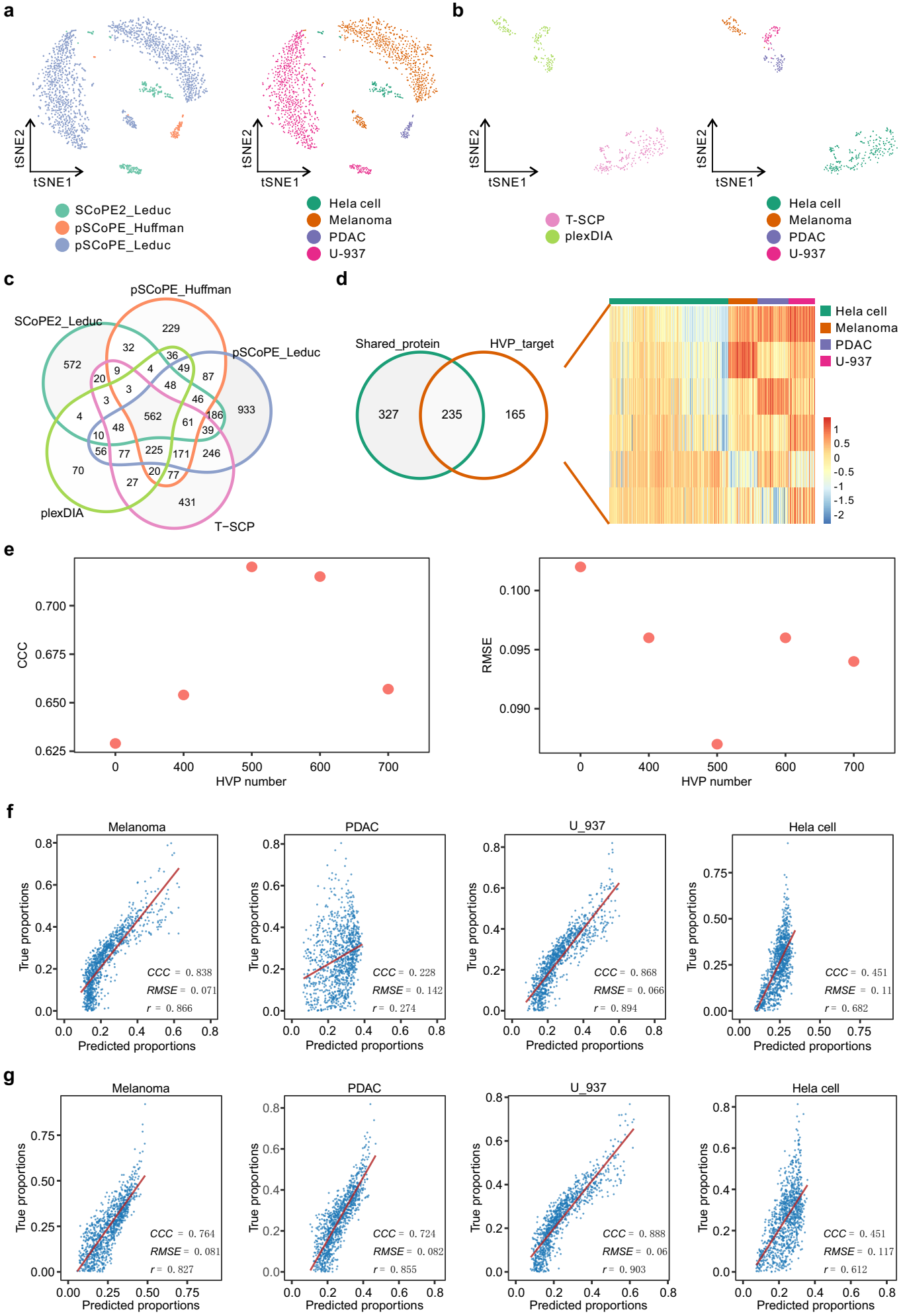
e, Venn diagram showing the overlap of proteins between two murine cell line datasets (N2 dataset and nanoPOTS dataset).

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

g, CCC values of scpDeconv in the case of murine cell line datasets with setting different reference sample numbers in stage 1.

h, CCC values of scpDeconv in the case of murine cell line datasets with setting different reference sample sizes in stage 1.

Figure 3


a
tSNE2
tSNE1
SCoPE2_Leduc
pSCoPE_Huffman
pSCoPE_Leduc
Hela cell
Melanoma
PDAC
U-937
b
T-SCP
plexDIA
c
pSCoPE_Huffman
SCoPE2_Leduc
pSCoPE_Leduc
plexDIA
T-SCP
d
Shared_protein
HVP_target
327
235
165
e
CCC
RMSE
HVP number
f
Melanoma
PDAC
U_937
Hela cell
True proportions
Predicted proportions
CCC = 0.838
RMSE = 0.071
r = 0.866
CCC = 0.228
RMSE = 0.142
r = 0.274
CCC = 0.868
RMSE = 0.066
r = 0.894
CCC = 0.451
RMSE = 0.11
r = 0.682
g
CCC = 0.764
RMSE = 0.081
r = 0.827
CCC = 0.724
RMSE = 0.082
r = 0.855
CCC = 0.888
RMSE = 0.06
r = 0.903
CCC = 0.451
RMSE = 0.117
r = 0.612

712 **Figure 3. Deconvolution performance on integrated human cell line datasets with different**
713 **HVP numbers.**

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

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

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

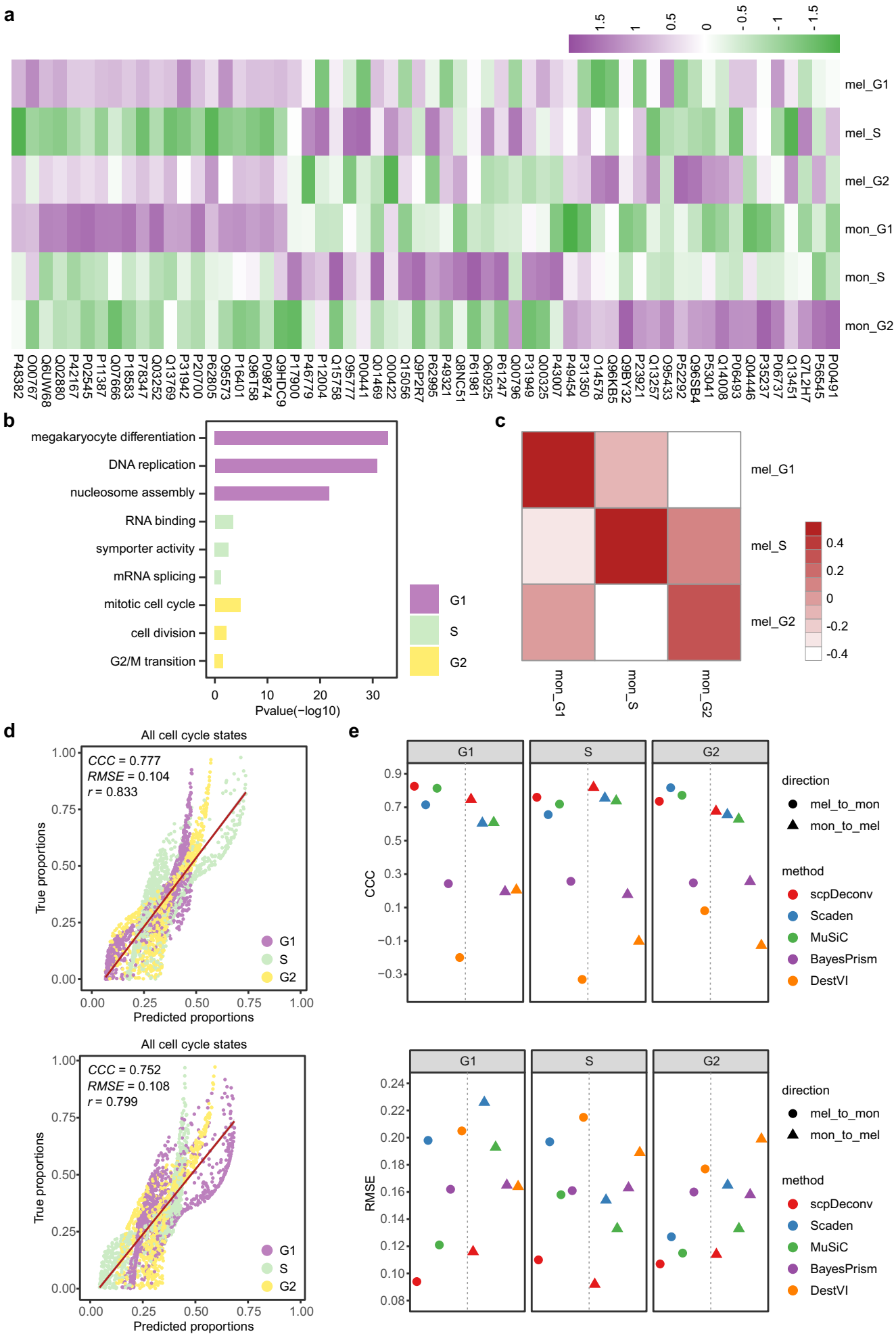
719 d, Venn diagram showing the overlap between shared proteins in all datasets and highly variable proteins
720 (HVPs) identified in target group. Heatmap showing the protein abundance profiles of specific HVPs in
721 the target group.

722 e, CCC (left panel) and RMSE values (right panel) of scpDeconv with different HVP numbers identified
723 in the target group of human cell line datasets.

724 f, Scatter plots of true (y axis) and predicted proportions (x axis) of four cell lines (Melanoma, PDAC, U-
725 937, and Hela cell) in human cell line datasets using scpDeconv without considering target specific
726 HVPs.

727 g, Scatter plots of true (y axis) and predicted proportions (x axis) of four cell lines (Melanoma, PDAC, U-
728 937, and Hela cell) in human cell line datasets using scpDeconv with setting target HVP number as 500.

Figure 4


a
1.5
1
0.5
0
- 0.5
- 1
- 1.5
mel_G1
mel_S
mel_G2
mon_G1
mon_S
mon_G2
P48382 O00767 Q6UW68 Q02880 P42167 P02545 P11387 Q07666 P18583 P78347 Q03252 Q13769 P31942 P20700 P62805 O95573 P16401 Q96T58 P09874 Q9HDC9 P17900 P46779 P12004 Q15758 O95777 P00441 Q01469 O00422 Q15056 Q9P2R7 P62995 P49321 Q8NC51 P61981 O60925 P61247 Q00796 P31949 Q00325 P43007 P49454 P31350 O14578 Q96KB5 Q9BY32 P23921 Q13257 O95433 P52292 Q96SB4 P53041 Q14008 P06493 Q04446 P35237 P06737 Q13451 Q7L2H7 P56545 P00491
b
megakaryocyte differentiation
DNA replication
nucleosome assembly
RNA binding
symporter activity
mRNA splicing
mitotic cell cycle
cell division
G2/M transition
0
10
20
30
Pvalue(−log10)
G1
S
G2
c
mel_G1
mel_S
mel_G2
mon_G1
mon_S
mon_G2
0.4
0.2
0
-0.2
-0.4
d
All cell cycle states
CCC = 0.777
RMSE = 0.104
r = 0.833
True proportions
Predicted proportions
G1
S
G2
All cell cycle states
CCC = 0.752
RMSE = 0.108
r = 0.799
True proportions
Predicted proportions
G1
S
G2
e
G1
S
G2
CCC
direction
mel_to_mon
mon_to_mel
method
scpDeconv
Scaden
MuSiC
BayesPrism
DestVI
RMSE
direction
mel_to_mon
mon_to_mel
method
scpDeconv
Scaden
MuSiC
BayesPrism
DestVI

729 **Figure 4. Inference of cell cycle across cell lines.**

730 a, Heatmap of the abundance of cell cycle specific marker proteins in melanoma or monocyte samples.

731 b, GO enrichment of marker proteins of G1, S, or G2 phases.

732 c, Heatmap showing the correlation of marker proteins' abundance in melanoma and monocyte samples.

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

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

Figure 5

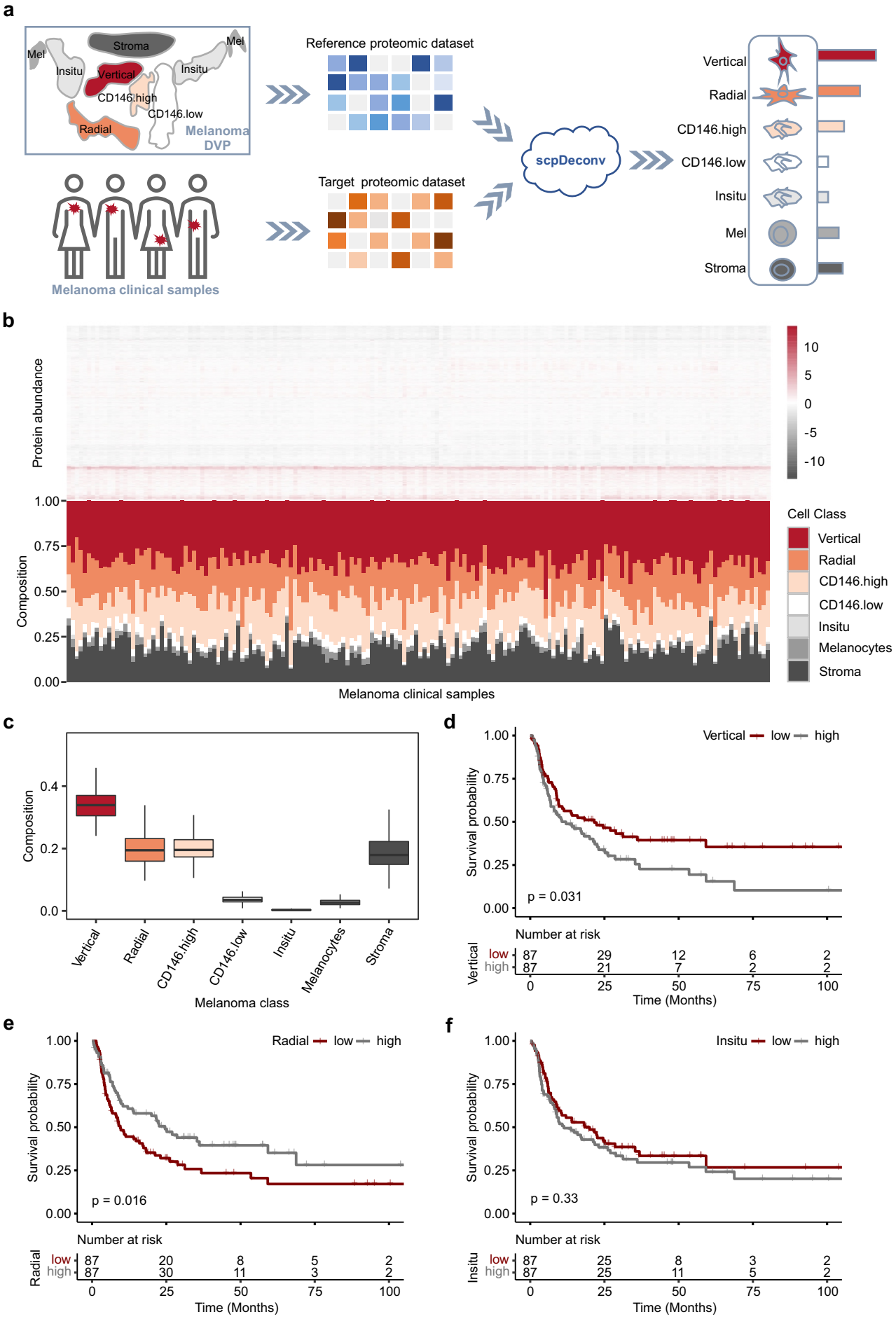

a
Stroma
Mel
Mel
Insitu
Vertical
Insitu
CD146.high
CD146.low
Radial
Melanoma DVP
Melanoma clinical samples
Reference proteomic dataset
Target proteomic dataset
scpDeconv
Vertical
Radial
CD146.high
CD146.low
Insitu
Mel
Stroma
b
Protein abundance
10
5
0
-5
-10
Composition
1.00
0.75
0.50
0.25
0.00
Melanoma clinical samples
Cell Class
Vertical
Radial
CD146.high
CD146.low
Insitu
Melanocytes
Stroma
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
1.00
0.75
0.50
0.25
0.00
p = 0.031
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
1.00
0.75
0.50
0.25
0.00
p = 0.016
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
1.00
0.75
0.50
0.25
0.00
p = 0.33
Number at risk
Insitu
low 87 25 8 3 2
high 87 25 11 5 2
0 25 50 75 100
Time (Months)

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

a, The pipeline of applying scpDeconv on deconvolving 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 clinical 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.