DESP: Demixing Cell State Profiles from Dynamic Bulk Molecular Measurements

Ahmed Youssef[1,2], Indranil Paul[2], Mark Crovella[1,3,4], Andrew Emili[1,2,4,5]

Affiliations

[1] Graduate Program in Bioinformatics, Boston University

[2] Center for Network Systems Biology, Boston University

[3] Computer Science Department, Boston University

[4] Faculty of Computing and Data Sciences, Boston University

[5] Knight Cancer Institute, Oregon Health and Science University

Abstract

There is wide interest to determine the dynamic expression of proteins and other molecules that drive phenotypic remodeling in development and pathobiology, but due to technical limitations these systems remain largely unexplored at the resolution of the underlying cell states. Here, we present *DESP*, a novel algorithm that leverages independent readouts of cellular proportions, such as from single-cell RNA-sequencing or cell sorting, to resolve the relative contributions of cell states to bulk molecular measurements, most notably quantitative proteomics, recorded in parallel. We applied *DESP* to an in-vitro model of the epithelial-to-mesenchymal transition and demonstrated its ability to accurately reconstruct cell state signatures from bulk-level measurements of both the proteome and transcriptome providing insights into transient regulatory mechanisms. *DESP* provides a generalizable computational framework for modeling the relationship between bulk and single-cell molecular measurements, enabling the study of proteomes and other molecular profiles at the cell state-level using established bulk-level workflows.

Introduction

Advances in single-cell molecular profiling, most notably single-cell RNA-sequencing (scRNA-Seq), offer a powerful approach for the identification and characterization of heterogenous cell states in dynamic biological processes, such as cell differentiation and tissue development, leading to breakthroughs in emerging domains of biomedical research[1,2]. Related computational methods can define the quantitative transcriptomic profiles underlying these discrete cell states[3]. However, these transcriptomic profiles have not proven to be accurate predictors of the corresponding protein expression patterns due to multiple biological and technical factors that lead to poor correlation between the cognate mRNA and protein levels of genes, preventing accurate inference of protein abundance using gene expression values such as those derived from scRNA-Seq[4,5,6]. This represents an important gap in our understanding of the specialized molecular programs that define cell states in heterogenous biological contexts.

The proteome occupies a crucial position in the functional landscape of dynamic cell states, as changes in protein expression underly biochemical processes driving phenotypic responses and altered cellular functions. In contrast to rapid advances in single-cell nucleic acid profiling, experimental challenges limit the vast majority of existing proteome studies to bulk measurements of mixed cell populations, mixing together protein abundance contributions from the individual constituent cell states and obscuring the native cellular contexts of the proteome[7,8,9] (Fig. 1a). Going beyond this global birds-eye view generated by standard proteomics datasets to resolve the foundational cell-level proteomes is crucial to improve mechanistic understanding of dynamic cell states with potential to enhance the prioritization of susceptible cell subpopulations for targeted therapeutics. While the ideal scenario is to directly quantify cellular proteomes, current state-of-the-art single-cell proteomics technologies are limited to measuring fewer than 1,000 proteins per cell[10,11]. Emerging technologies aim to address the technical challenges associated with single-cell proteomics but significant limitations remain, including the extent of proteome coverage and quantifying functional information such as macromolecular interactions and post-translational modifications[10,11].

As a generalizable solution to this challenge, we present a novel demixing algorithm titled *DESP* (Demixing Cell State Profiles in Bulk Omics) that leverages information on cell state proportions derived from single-cell techniques to deconvolve standard bulk proteomics measurements of heterogenous cell mixtures by distinguishing the contributions of the underlying

56 cell states comprising the mixture. *DESP*'s mathematical model is designed to circumvent the poor
57 mRNA-protein correlation that represents a major challenge for multi-omics integration efforts.
58 Rather, our approach leverages the ability of single-cell readouts, such as scRNA-Seq or FACS,
59 to identify the proportions of molecularly-distinct cell states within dynamic heterogeonous
60 samples for which bulk proteomics measurements were generated. *DESP*'s model combines these
61 observed changes in cell state proportions with the corresponding variations in bulk-level protein
62 abundance to estimate the relative protein abundance levels associated with each cell state. While
63 we demonstrate *DESP*'s utility using proteomics data in this paper, it can be applied to demix any
64 type of bulk 'omics data, such as metabolomics or phopshoproteomics.

65 We applied *DESP* to a multi-omics time-course dataset investigating the biomolecular
66 changes of pre-cancerous cells undergoing TGF-Beta-induced epithelial-to-mesenchymal
67 transition (EMT) in-vitro. EMT is a complex biological process by which epithelial cells lose their
68 adhesion and gradually differentiate into mobile mesenchymal cells and is of particular importance
69 to the biomedical community due to its integral role in wound healing, tissue development, and
70 cancer progression[12]. We generated bulk quantitative proteomics and scRNA-Seq data generated
71 at eight consecutive timepoints representing MCF10A mammary epithelial cells undergoing
72 EMT[13], and we show that *DESP* is able to demix the corresponding cellular proteomes and recover
73 expected patterns of canonical EMT markers. Additionally, *DESP* goes further by identifying the
74 proteomic profile of transient cell states key to the differentiation process that were not detected
75 at the transcriptomics level and whose signal is significantly diluted in the bulk proteomics due to
76 the confounding presence of other cell states in the mixture. These results both demonstrate a
77 compelling use case of *DESP* and also shed light on the transient nature of intermediate molecular
78 mechanisms driving EMT.

Results

DESP algorithm overview

DESP is a tool for deconvoluting bulk proteomics data to the cell state-level by estimating the contributions of the underlying cell states to the observed bulk measurements. Fig. 1b shows an overview of *DESP* as applied to proteomics data as an example. *DESP* combines bulk proteomics data with single-cell data, such as scRNA-Seq, acquired for a set of biological samples representing different physiological contexts. Examples of potential applications for *DESP* include time-course experiments, cross-tissue 'omics atlases, or studies comparing disease subtypes. The single-cell data is used simply to delineate the proportions of the cell states that make up the samples under study and their quantitative similarities to each other. The resultant cell state composition matrix is fed into *DESP* alongside corresponding bulk protein measurements (Fig. 1c), where the objective is to find the most likely possible combination of cell state-level protein abundances whose sum would lead to the observed bulk measurements. The output of *DESP* is a matrix containing the predicted protein levels for each of the pre-defined cell states, effectively demixing the measured bulk proteomics matrix into the underlying cell state proteomes (Fig. 1d).

The premise of *DESP* is that the bulk proteome can be modeled as the sum of the underlying single-cell proteomes. Near-complete sequencing coverage of current single-cell technologies like scRNA-Seq make it routinely feasible to obtain the identity and proportions of cell states in given biological samples[1,2], and advances in mass spectrometry make acquiring comprehensive measurements of bulk proteomes similarly feasible[1,2]. We thus sought to integrate these readily-available and comprehensive data types within a computational framework that could predict the experimentally-challenging single-cell proteomes. *DESP* achieves this goal by applying the *bulk-as-sum-of-single-cells* principle through modeling the measured global proteome as the product of the observed cell state proportions and the unknown corresponding cell state proteomes, providing a simple but effective formulation that can be solved mathematically to estimate the cell state proteomes (Fig 1c).

The *DESP* model solves a typically under-determined problem where the number of features is less than the number of samples. For example, an experiment that surveys a handful of cancer subtypes will likely contain dozens of cell states involved, as such giving many possible cell-level proteomes that could lead to the observed bulk proteome. We thus apply a set of both

110 biologically- and mathematically-motivated constraints to guide the algorithm to the optimal
111 solution (Fig 1c). One key constraint leverages cell state correlations in the input single-cell data
112 to maintain the global structure of the cell states within *DESP*'s predictions by preserving the
113 original state-state similarities. A detailed description of the constraints can be found in the
114 Methods section.

115 *DESP* does not aim to define cell states or their proportions from 'omics data, but rather it
116 predicts the proteome of the pre-defined cell states of interest based on the bulk proteomics data.
117 The utility of *DESP* thus lies in the presence of established routine workflows for identifying the
118 global proteome and cell state composition of heterogenous samples, effectively bypassing the
119 lack of methods for deep measurement of individual cell proteomes. It is worth noting that *DESP*
120 is agnostic of the technique used to identify cell state proportions and thus does not directly use
121 the single-cell measurements to infer the protein levels of the given cell states, allowing *DESP* to
122 avoid issues regarding relying on correlations between different data types like RNA-Seq and
123 proteomics.

124 ***Multi-omics profiling of EMT***

125 We applied *DESP* to a time-course study examining pre-cancerous cells undergoing TGF-
126 Beta-induced EMT[13]. Bulk quantitative proteomics and scRNA-Seq data for over 5,600 genes
127 were generated in parallel for eight consecutive timepoints as MCF10A human mammary
128 epithelial cells underwent the EMT differentiation process (Fig 2a). Analysis of the scRNA-Seq
129 data using the *Seurat* tool[14] revealed ten discrete cell states among the 1,913 captured cells, with
130 intermediate states emerging from the predominantly epithelial cultures before transforming into
131 the mesenchymal end state over the timecourse (Fig 2c,d). The availability of both bulk protein
132 measurements and single-cell proportion information for the samples along with the dynamic
133 nature of the samples' cellular compositions renders it an excellent testbed for *DESP* (Fig. 2f). The
134 dataset was thus used for algorithm validation first and as a case study of a real-life application of
135 *DESP* (Fig 2g,h).

136 ***Mathematical validation***

137 To validate the algorithm and assess performance, we investigated *DESP*'s ability to
138 recover the RNA profiles of individual EMT cell states from corresponding bulk RNA profiles.
139 We utilized the scRNA-Seq data only for the mathematical validation because it represents the
140 'true' experimentally-derived single-cell profiles to compare against *DESP*'s bulk-inferred

141 predictions. Since *DESP*'s predictions are made at the resolution of cell states and not single cells,
142 we constructed representative cell state profiles from the scRNA data by averaging the expression
143 of each gene in each cell state's cells to represent the ground truth[13]. Fig. 3a shows a schematic
144 illustrating the validation strategy, where a 'pseudobulk' RNA matrix is first created by summing
145 the dynamic scRNA measurements within each timepoint. We then used *DESP* to demix these
146 pseudobulk profiles to the level of the ten cell states present in this dataset, as guided by the
147 scRNA-derived changes in cell state proportions across the timepoints. These predicted cell state
148 profiles were then compared to the ground truth profiles.

149 Several metrics were used to compare the real and predicted cell state profiles, henceforth
150 referred to as 'centroids', to determine whether *DESP*'s predictions were reasonable
151 approximations of the ground truth. To visualize the similarity of the predicted and real centroids
152 within the global structure of the single-cell data, we created a UMAP projection of the scRNA
153 data with the centroids overlayed showing strong overlap between the real and predicted centroids
154 (Fig. 3b). We also quantitively assessed *DESP*'s ability to recover the true centroids by computing
155 the average Euclidean distance between each state's constituent single-cell measurements to their
156 real centroid compared to that of the predicted centroid, again showing that the predictions fall
157 within the correct intra-cell state range (Fig. 3c). Mapping the real and predicted centroids in a
158 one-to-one manner based on minimizing their Euclidean distances showed that *DESP* correctly
159 assigned centroid labels with a near-zero total distance (Fig. 3d). Additionally, the mean per-gene
160 relative prediction error was only 15% (IQR 6–17%), suggesting reasonable approximations of the
161 real data even down to the level of individual genes (Fig. 3e).

162 Cell state assignments by tools such as *Seurat* can be imprecise. Hence, to ensure the
163 robustness of *DESP* to variations in the number of input cell states, we repeated the validation with
164 numbers of input cell states varying between 5 and 20 and found that while the prediction error
165 expectedly increased with the number of states, it remained within a modest error range (Fig. 3d).
166 We also showed that *DESP* is robust to a wide range of its two input parameters with negligible
167 effects on performance (Extended Fig. 1).

168 Since *DESP* performs predictions on a feature-by-feature basis, we also examined whether
169 the accuracy of predictions varied with the expression of particular genes and observed that more
170 cross-cell stable 'housekeeping' genes had lower prediction errors than the more variable genes,
171 while expression magnitudes were not influential (Extended Fig. 2). In a similar vein, we examined

172 the relationship between the prediction errors and cell state properties and found that the larger
173 and more homogenous states had the lowest prediction errors (Extended Fig. 2).
174 In the supplement, we describe similar results of these validation experiments on a single-
175 cell proteomics dataset, demonstrating that *DESP* can successfully model the relationship between
176 the bulk and single-cell layers of different types of 'omics data (Extended Fig. 3), as well as
177 validating *DESP*'s performance on significantly larger-scale data using the Human Cell Atlas[15]
178 (Extended Fig. 5).
179 ***Comparison with experimentally-derived profiles***
180 To further benchmark *DESP*, we analyzed single-cell proteomics data from an analogous
181 model of EMT generated by the SCoPE2[10] platform which quantified >1,700 proteins across 420
182 cells at three key transition timepoints[16]. The similarity of this independent experimental model to
183 our EMT multi-omics case study combined with the presence of more than 1,000 overlapping
184 proteins between them make it the ideal for comparing *DESP*'s computationally-predicted profiles
185 to independently-acquired experimentally-derived ones.
186 We first used a k-nearest neighbor strategy to impute missing values in the single-cell
187 proteomics data followed by k-means clustering to reveal groups of proteomic-related cells
188 (Methods). The clusters were labeled as Epithelial (E), Intermediate (I), or Mesenchymal (M)
189 based on their temporal dynamics and expression of classical EMT markers (Fig. 4a, Extended
190 Fig. 4). To avoid potentially-confounding issues with technical experimental differences, our
191 benchmarking strategy focused on comparing the three cell states by examining whether the
192 protein fold-changes derived from *DESP*'s predicted profiles resembled the experimentally-
193 derived fold-changes more than those derived from the scRNA data. We thus compared the mean
194 gene/protein fold-changes between cell-states across the scRNA-Seq and single-cell proteomics
195 data and *DESP*'s cell-state proteomics predictions (Fig. 4b,c).
196 For this benchmark, we selected the proteins with high correlation (Pearson $R^2 > 0.5$)
197 between the bulk proteomics and a 'pseudobulk' matrix constructed by summing the single-cell
198 proteomics data in each timepoint to address technical variation between the single-cell and bulk
199 proteomics measurements (Methods). We also focused our comparison on the genes that differ the
200 most (Pearson $R^2 < -0.5$) between the bulk proteomics and pseudobulk scRNA, representing the
201 genes that are the best at distinguishing the RNA and protein levels in this data and thus most
202 appropriate for benchmarking *DESP*. Computing the Pearson correlation between the resultant

203 fold-changes showed that those from *DESP* are not only significantly correlated ($p < 0.05$) with
204 those from the single-cell proteomics data, but are also significantly closer to the single-cell
205 proteomics than the scRNA-derived fold-changes, suggesting that *DESP*'s predictions agree with
206 the experimentally-derived profiles (Fig. 4d).

207 ***Detection of transient EMT mechanisms using DESP***

208 We applied *DESP* to demix the EMT dynamic bulk proteomics data by separating the
209 contributions of the ten cell states to each of the eight timepoint bulk measurements (Methods).
210 The output allowed us to contextualize the cross-timepoint proteome changes according to the
211 underlying cell states and characterize the distinct protein programs that occur during EMT that
212 are lost in the bulk measurements as elaborated on below.

213 To evaluate the extent to which *DESP*'s predictions provided information beyond what the
214 bulk proteomics or scRNA data provided alone, we compared the results of performing differential
215 expression analysis between the cell states from the *DESP* predictions versus those two original
216 data types by themselves and found that *DESP* detected hundreds of markers for the EMT stages
217 that were not deemed differential in the other techniques (Extended Fig. 4). Several of these
218 markers are known to be involved in EMT, such as TPM2 in the E state and CD44 in the M state.

219 Gene set enrichment analysis also revealed more than 200 pathways that were significantly
220 enriched among *DESP*-identified markers but not revealed by the other standard approaches,
221 demonstrating *DESP*'s ability to glean novel insight on the involved molecular mechanisms rather
222 than just recapitulating what existing methods would have found (Fig. 5a). These pathways
223 included both canonical EMT-related pathways such as signaling, migration, and wound healing,
224 as well as potentially novel biological processes that had not been associated with EMT before,
225 such as cellular stress response, spliceosome assembly, and protein transport (Fig. 5b).

226 The intermediate cell states are of high interest due to their transient nature and expected
227 influential role in driving EMT. We thus sought to investigate if unique insights could be gleaned
228 by *DESP* on the transient molecular mechanisms governing EMT. Notably, eight proteins,
229 including three encoded by histone genes, were identified as being key to the transition due to their
230 relative enrichment in the predicted intermediate cell state proteome compared to their diluted
231 signal in the bulk proteomics data, representing a signal that likely would have been missed if
232 analyzing bulk proteomics alone (Fig. 5d).

233 Of these eight proteins, H2AFX stands out as a marker (99th percentile) of the predicted
234 intermediate proteomics profile yet with low abundance (9th percentile) in the corresponding
235 scRNA profile suggesting extensive post-transcriptional regulation. H2AFX is a core histone
236 protein variant that both contributes to chromatin remodeling and is post-translationally modified
237 during EMT[17,18]. Notably, H2AFX is hyper-phosphorylated during the earliest timepoints of
238 EMT[13], suggesting a change in this protein's functional activity during the initial intermediate cell
239 states (Fig. 5e).

240 The STRING database[19] shows that H2AFX and two other *DESP*-identified markers are
241 functionally-linked with nine proteins annotated as EMT-associated in the GO[20] and MSigDB
242 Hallmark genesets[21]. One of these associated proteins is EZH2, which is part of the PRC2
243 chromatin remodeling complex that performs transcriptionally-inhibitory histone methylation
244 (Fig. 5c). EZH2 has emerged in recent years as a promising target for cancer treatments[22,23], with
245 a FDA-approved drug (Tazemetostat) targeting EZH2 as a cancer treatment[24], and another EZH2-
246 targeting compound currently in clinical trials[25]. In contrast to its target H2AFX, EZH2 shows
247 elevated phosphorylation during later timepoints of the experiment, suggesting the activation of a
248 delayed regulatory 'switch' after three days of TGF-Beta induction.

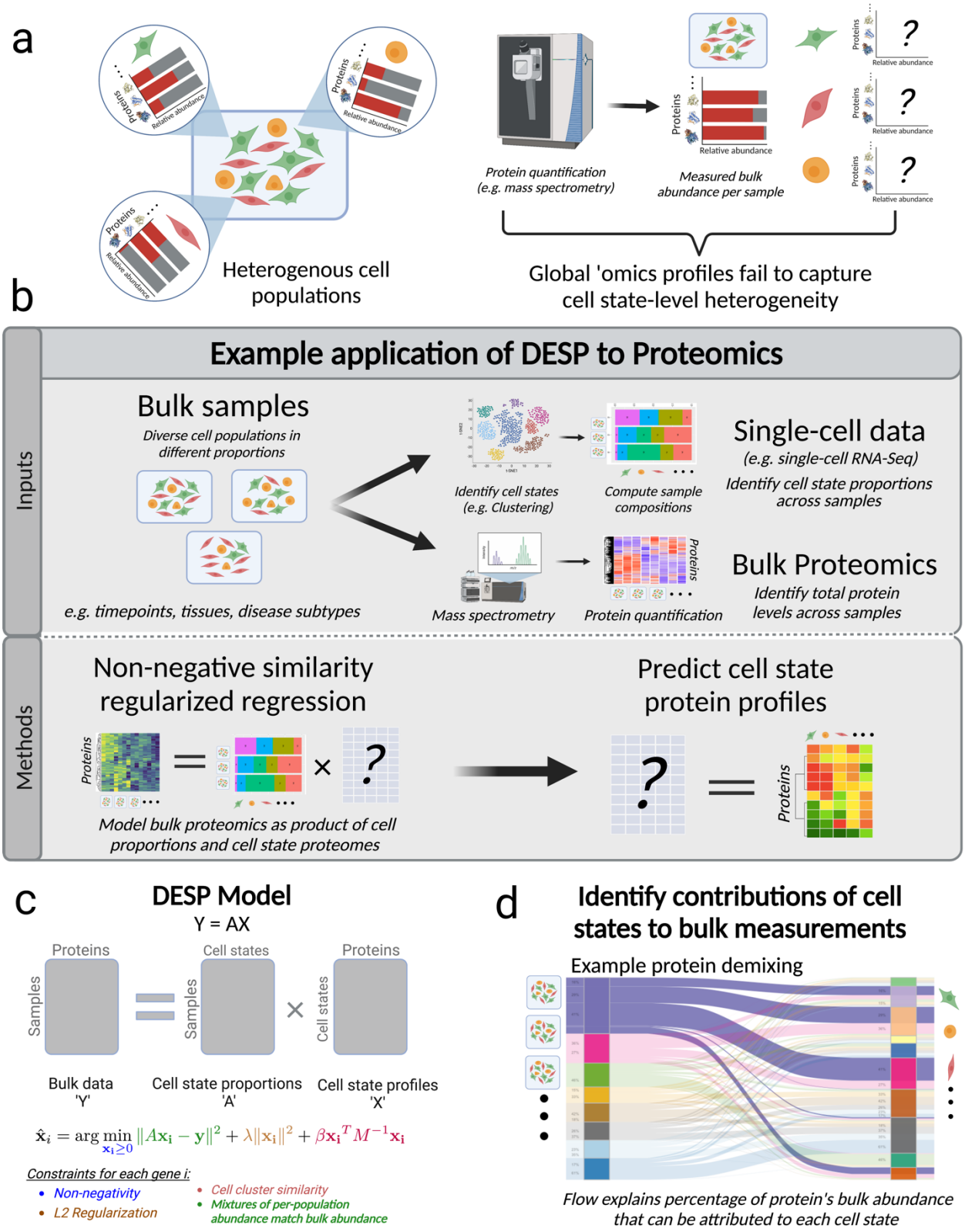
249 Taken together, these findings generated by *DESP*-based proteomic demixing pinpoint
250 specific transient events as plausible mechanisms governing EMT progression despite the RNA-
251 protein discord.

Discussion

Here we present, evaluate, and apply *DESP* as an algorithm for resolving the contributions of cell states to quantitative global proteomics measurements and potentially other bulk 'omics profiles. *DESP* enables researchers to gain insights into dynamic cell state contexts from standard bulk 'omics workflows, helping to integrate underserved 'omics layers like proteomics into the rapidly evolving single-cell analysis toolkit. In contrast to existing 'omics deconvolution algorithms[26], *DESP* does not aim to predict cell state proportions, but instead it predicts the proteome of given cell states based on demixing bulk proteomics data. *DESP* provides a generalizable computational framework for bulk 'omics demixing, with planned future applications including demixing protein interaction networks and spatial transcriptomics.

We validated *DESP* mathematically and obtained independent experimental evidence supporting the validity of *DESP*'s predictions. We also applied *DESP* to tease out the proteome composition of transient intermediate cell states in an in-vitro model of EMT, shining a light on key intermediate molecular mechanisms. *DESP* is publicly-available to the community as an R package (https://github.com/AhmedYoussef95/DESP) and can produce predictions within seconds on datasets containing > 48,000 features.

Figures



$$\hat{\mathbf{x}}_i = \arg\min_{\mathbf{x_i} \geq 0} \|A\mathbf{x_i} - \mathbf{y}\|^2 + \lambda\|\mathbf{x_i}\|^2 + \beta\mathbf{x_i}^T M^{-1}\mathbf{x_i}$$

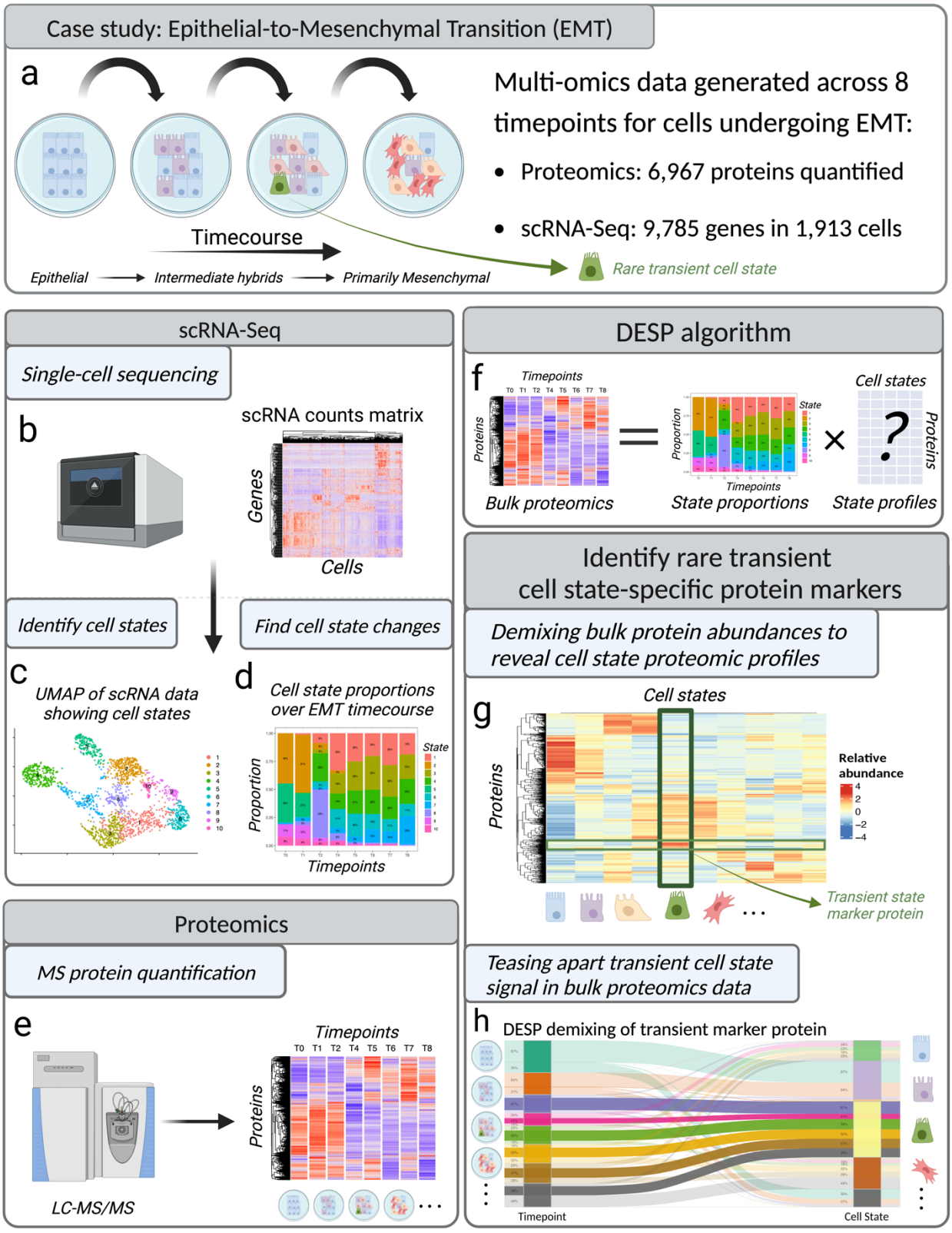
Figure 1: Overview of DESP algorithm

1a: Motivation and rationale behind *DESP*. Standard 'omic techniques such as mass spectrometry-based quantitative proteomics are often limited to bulk measurements of mixed cell populations and fail to resolve sample heterogeneity at the cell-state level.

1b: Analysis workflow. *DESP* takes as an input paired bulk omics data, e.g. proteomics, and matching single-cell data and predicts the contributions of the underlying cell states to the observed bulk measurements.

1c: Mathematical model. *DESP* models the observed bulk profiles, e.g. proteomics measurements, as the sum of unknown underlying cell state profiles. DESP applies several constraints to this formulation to predict the unknown cell state profiles.

1d: Output. Schematic of the *DESP* output showcasing its ability to demix and associate bulk sample measurements to the underlying contributing cell states for any given gene/protein.


Case study: Epithelial-to-Mesenchymal Transition (EMT)
a
Multi-omics data generated across 8 timepoints for cells undergoing EMT:
• Proteomics: 6,967 proteins quantified
• scRNA-Seq: 9,785 genes in 1,913 cells
Timecourse
Epithelial
Intermediate hybrids
Primarily Mesenchymal
Rare transient cell state
scRNA-Seq
Single-cell sequencing
b
scRNA counts matrix
Genes
Cells
Identify cell states
Find cell state changes
c
UMAP of scRNA data showing cell states
d
Cell state proportions over EMT timecourse
Proportion
Timepoints
State
Proteomics
MS protein quantification
e
LC-MS/MS
Timepoints
Proteins
DESP algorithm
f
Timepoints
Proteins
Bulk proteomics
Proportion
State proportions
Cell states
State profiles
Identify rare transient cell state-specific protein markers
Demixing bulk protein abundances to reveal cell state proteomic profiles
g
Cell states
Proteins
Relative abundance
Transient state marker protein
Teasing apart transient cell state signal in bulk proteomics data
h
DESP demixing of transient marker protein
Timepoint
Cell State

282

283

Figure 2: Case study: Epithelial-to-mesenchymal transition (EMT)

2a: Case study: EMT dynamic multi-omics profiling. In this experiment, bulk proteomics and scRNA-Seq data were generated for human MCF10A cells undergoing TGF-Beta-induced EMT across eight consecutive timepoints, generating molecular profiles of transient intermediate states.

2b: Dynamic scRNA-Seq profiles. Heatmap visualizes the scaled relative expression of genes (rows) across cells (columns).

2c: Identification of intermediate cell states occurring during EMT. Clustering the scRNA-Seq data reveals ten intermediate cell states as visualized on UMAP plot.

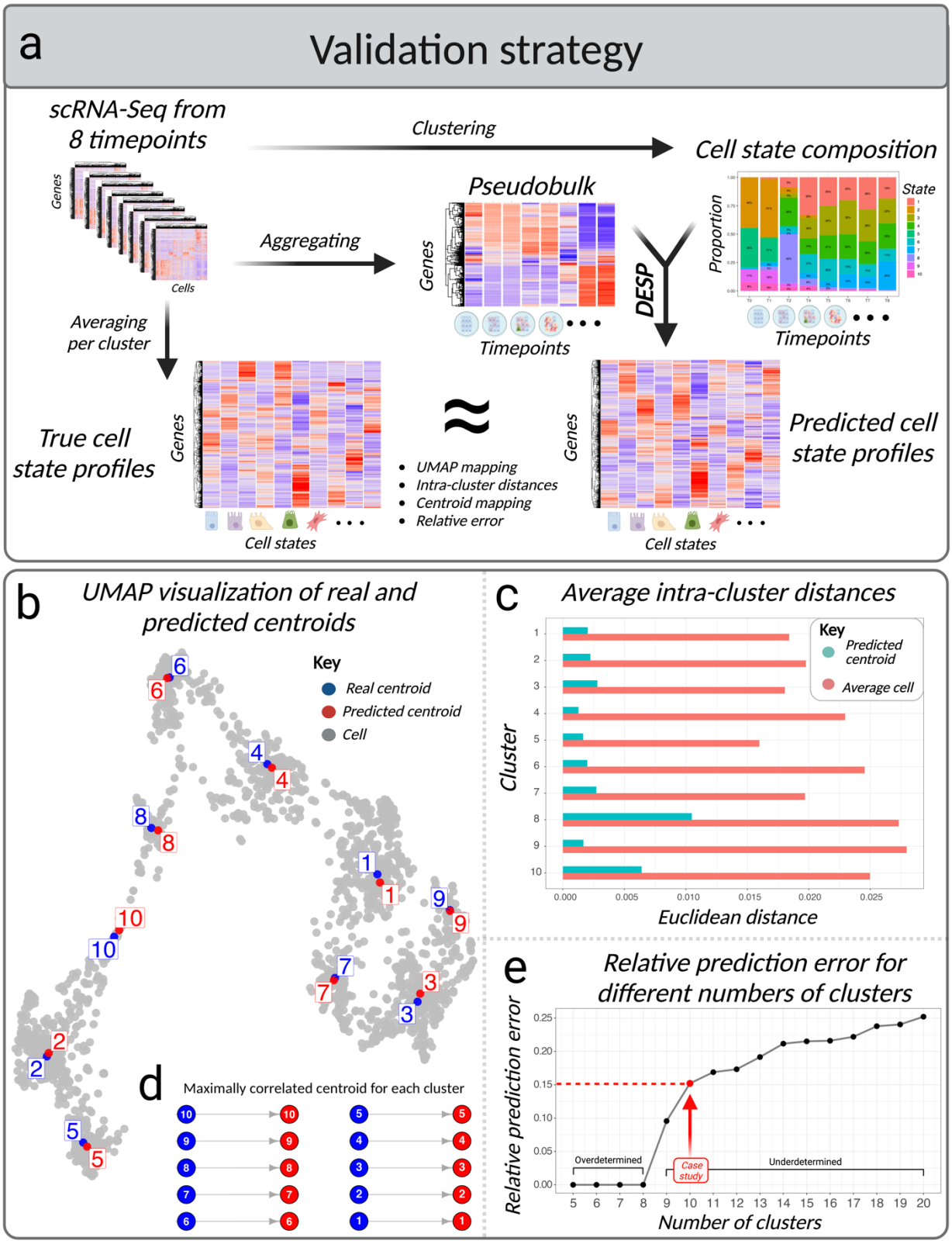
2d: Cell state fluctuations during EMT. Bar plot visualizing cell state composition at each timepoint.

2e: Mass spectrometry-based quantitative proteomics profiles. Heatmap visualizes scaled TMT reporter ion intensity, a proxy for relative abundance, of proteins (rows) across timepoints (columns).

2f: Applying *DESP* to EMT bulk proteomic profiles and scRNA-derived cell state composition matrix to predict cell state-level protein profiles.

2g: Prediction of intermediate cell state profiles using *DESP*. Heatmap visualizes the predicted relative protein abundances for each of the cell states. One putative marker of a transient cell state is highlighted as an example finding from these predictions.

2h: Identifying contribution of individual cell states to timepoint bulk proteomics measurements. Schematic shows *DESP* demixing of a given protein to distinct underlying cell states, highlighting a transient cell state signal that is lost by examining bulk measurements alone due to the presence of confounding signals from multiple cell states within each timepoint.


a
Validation strategy
scRNA-Seq from 8 timepoints
Clustering
Cell state composition
Genes
Cells
Aggregating
Pseudobulk
Genes
Timepoints
DESP
Proportion
State
Timepoints
Averaging per cluster
True cell state profiles
Genes
Cell states
UMAP mapping
Intra-cluster distances
Centroid mapping
Relative error
Genes
Predicted cell state profiles
Cell states
b
UMAP visualization of real and predicted centroids
Key
Real centroid
Predicted centroid
Cell
c
Average intra-cluster distances
Key
Predicted centroid
Average cell
Cluster
Euclidean distance
d
Maximally correlated centroid for each cluster
e
Relative prediction error for different numbers of clusters
Relative prediction error
Overdetermined
Case study
Underdetermined
Number of clusters

306

307

308 *Figure 3: Mathematical validation of DESP*

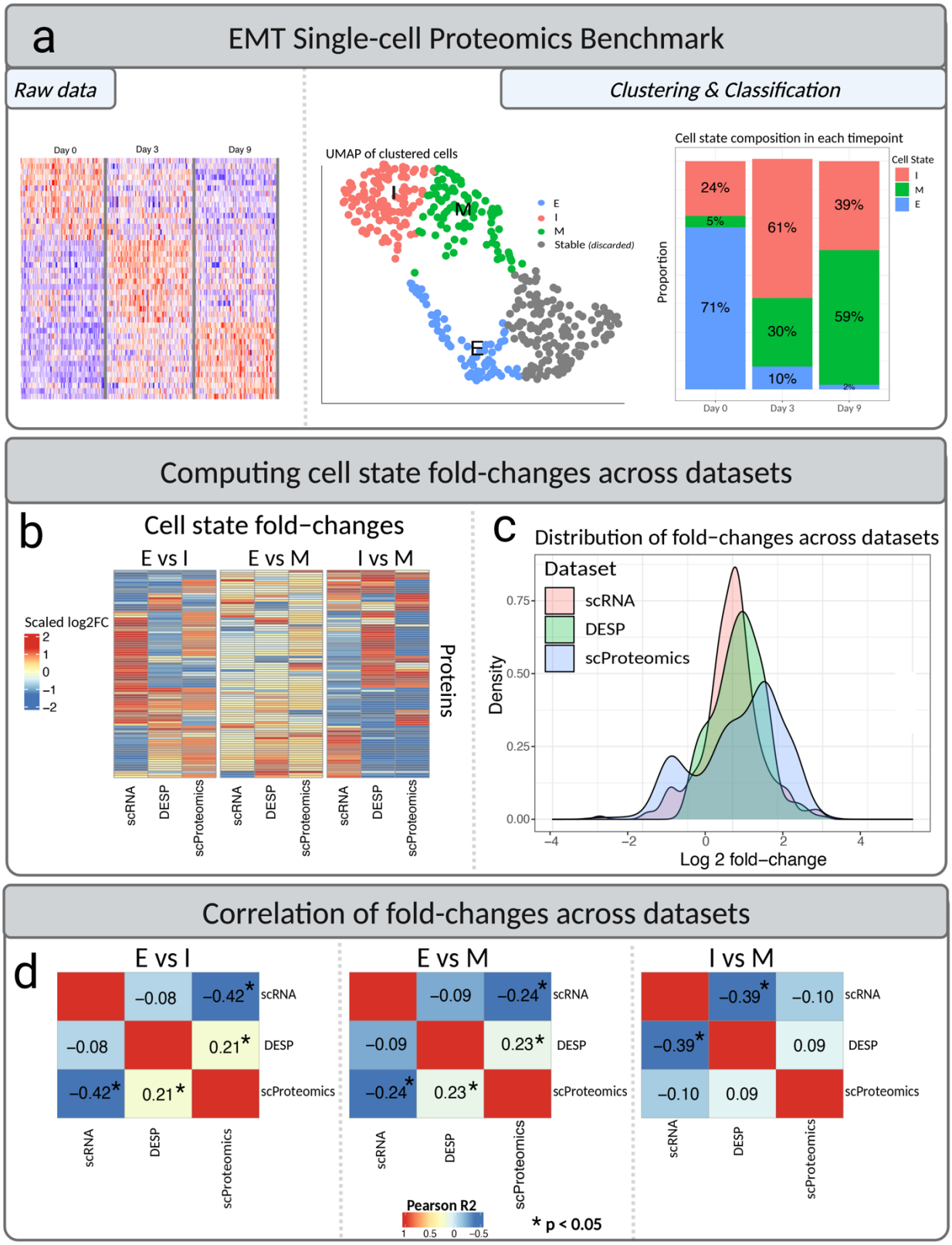
309 3a: Validation strategy. 'Pseudobulk' profiles, generated by summing the scRNA profiles in each
310 timepoint, were demixed by *DESP* guided by the cell state compositions. The predicted *DESP*
311 profiles were compared to the averaged per-cluster gene expression values in the input scRNA
312 data.

313 3b: UMAP showing overlap of the real and predicted cell state profiles, labeled as 'centroids',
314 within the global structure of the scRNA-Seq data.

315 3c: Bar plots showing the average Euclidean distance between each cell state's constituent single-
316 cell measurements to their real centroid compared to that of the predicted centroid.

317 3d: Mapping predicted centroids to closest real centroid based on solving the linear sum
318 assignment problem to minimize total Euclidean distances.

319 3e: Relative per-gene prediction error when varying the number of cell clusters input to *DESP*,
320 highlighting the chosen 10-cluster scenario. Standard error bars are not visible due to small error
321 range.


a
EMT Single-cell Proteomics Benchmark
Raw data
Clustering & Classification
Day 0
Day 3
Day 9
UMAP of clustered cells
I
M
E
E
I
M
Stable (discarded)
Cell state composition in each timepoint
Cell State
I
M
E
Proportion
24%
5%
71%
61%
30%
10%
39%
59%
2%
Day 0
Day 3
Day 9
Computing cell state fold-changes across datasets
b
Cell state fold−changes
E vs I
E vs M
I vs M
Scaled log2FC
2
1
0
−1
−2
Proteins
scRNA
DESP
scProteomics
c
Distribution of fold−changes across datasets
Dataset
scRNA
DESP
scProteomics
Density
0.75
0.50
0.25
0.00
−4
−2
0
2
4
Log 2 fold−change
Correlation of fold-changes across datasets
d
E vs I
−0.08
−0.42*
scRNA
−0.08
0.21*
DESP
−0.42*
0.21*
scProteomics
E vs M
−0.09
−0.24*
scRNA
−0.09
0.23*
DESP
−0.24*
0.23*
scProteomics
I vs M
−0.39*
−0.10
scRNA
−0.39*
0.09
DESP
−0.10
0.09
scProteomics
Pearson R2
1
0.5
0
−0.5
* p < 0.05

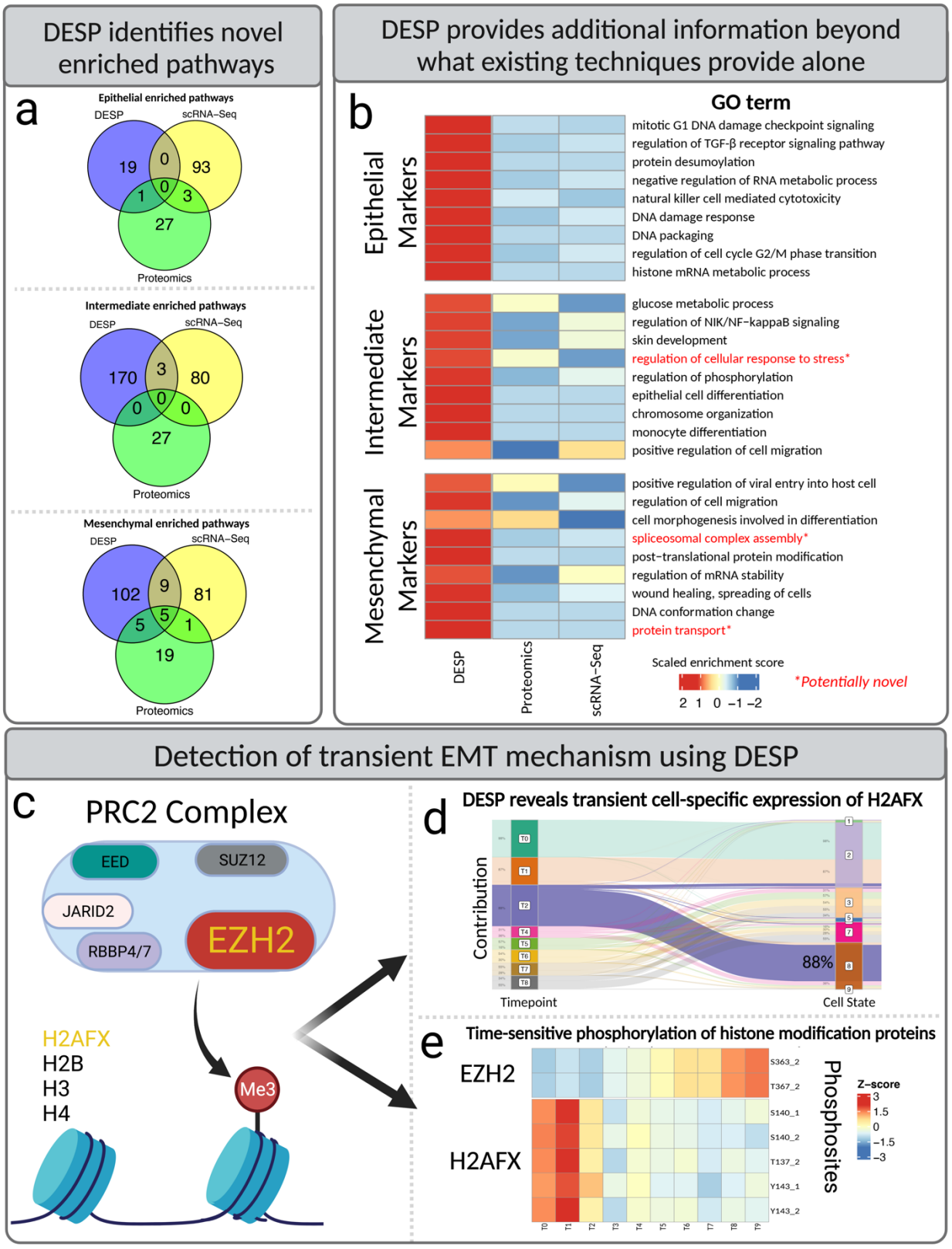
324 *Figure 4: Comparing DESP predictions to experimentally-derived profiles*

325 4a: Dynamic EMT single-cell proteomics data used for biological validation.

326 4b: Heatmap of computed protein relative abundance fold-changes between cell states in single-
327 proteomics data, scRNA data, and bulk-demixed *DESP* predictions.

328 4c: Density plot showing distribution of computed fold-changes across the three datasets.

329 4d: Pearson correlation of the protein relative abundance fold-changes between cell-states in each
330 dataset.


DESP identifies novel enriched pathways
a
Epithelial enriched pathways
DESP
scRNA–Seq
Proteomics
19
0
93
1
0
3
27
Intermediate enriched pathways
170
3
80
0
0
0
27
Mesenchymal enriched pathways
102
9
81
5
5
1
19
DESP provides additional information beyond what existing techniques provide alone
b
GO term
Epithelial Markers
mitotic G1 DNA damage checkpoint signaling
regulation of TGF-β receptor signaling pathway
protein desumoylation
negative regulation of RNA metabolic process
natural killer cell mediated cytotoxicity
DNA damage response
DNA packaging
regulation of cell cycle G2/M phase transition
histone mRNA metabolic process
Intermediate Markers
glucose metabolic process
regulation of NIK/NF-kappaB signaling
skin development
regulation of cellular response to stress*
regulation of phosphorylation
epithelial cell differentiation
chromosome organization
monocyte differentiation
positive regulation of cell migration
Mesenchymal Markers
positive regulation of viral entry into host cell
regulation of cell migration
cell morphogenesis involved in differentiation
spliceosomal complex assembly*
post-translational protein modification
regulation of mRNA stability
wound healing, spreading of cells
DNA conformation change
protein transport*
DESP
Proteomics
scRNA–Seq
Scaled enrichment score
2 1 0 −1 −2
*Potentially novel
Detection of transient EMT mechanism using DESP
c
PRC2 Complex
EED
SUZ12
JARID2
RBBP4/7
EZH2
H2AFX
H2B
H3
H4
Me3
d
DESP reveals transient cell-specific expression of H2AFX
Contribution
88%
Timepoint
Cell State
e
Time-sensitive phosphorylation of histone modification proteins
EZH2
H2AFX
S363_2
T367_2
S140_1
S140_2
T137_2
Y143_1
Y143_2
Phosphosites
Z-score
3
1.5
0
−1.5
−3

331

332

333 *Figure 5: Detection of transient EMT mechanisms using DESP*

334 5a: Venn diagrams showing overlap of significantly-enriched pathways within each EMT stage
335 detected by scRNA-Seq data alone, bulk proteomics data alone, and *DESP* predictions.

336 5b: Heatmap of relative enrichment scores of selected GO terms among cell stage markers that
337 were enriched among *DESP*'s predictions compared to results obtained from scRNA-Seq or bulk
338 proteomics alone.

339 5c: PRC2 core histone methylation complex. The EZH2 and H2AFX proteins are highlighted
340 based on their inferred influential transient profiles detected by *DESP*.

341 5d: *DESP* demixing of H2AFX across transient cell states, with intermediate timepoint T2
342 highlighting the time-sensitive disproportionate contribution by transient cell state 8.

343 5e: Heatmap of temporal phosphorylation of EZH2 and H2AFX in global quantitative
344 phosphoproteomics analysis of the same EMT samples.

Methods

DESP algorithm

Background

We formalize *DESP* as recovery of cell state profiles under the mixing equation $AX = Y$, with mixing matrix A (samples × states) applied to unobserved state profiles X (states × genes) yielding observed bulk data Y (samples × genes).

Each pair of corresponding columns of X and Y is an independent equation; in what follows we denote a column of X as x and the corresponding column of Y as y.

This problem is challenging when there are fewer samples than states, leading to an ill-posed linear inverse problem. To constrain the solution toward meaningful and biologically likely profiles, *DESP* applies a series of regularizations as described next.

Nonnegativity

First, *DESP* imposes nonnegativity on the solutions. This leads to the nonnegative least squares problem:

$$\hat{x} = \underset{x \geq 0}{argmin} \, \|Ax - y\|^2.$$

which is equivalent to the quadratic program:

$$\hat{x} = \underset{x \geq 0}{argmin} \; -2y^T Ax + x^T A^T Ax.$$

Tikhonov Regularization

Next, *DESP* incorporates Tikhonov regularization (ridge regression) to favor small-norm solutions:

$$\hat{x} = \underset{x \geq 0}{argmin} \, \|Ax - y\|^2 + \lambda_x \|x\|^2$$

We use the notation λ_x to indicate that the proper amount of regularization can be chosen based on the properties of x. (As shown in Extended Figure 2, statistical properties of individual gene

369 expressions are correlated with estimation accuracy.) However, our results are based on using the
370 same value of λ_x for all genes. As shown in Extended Figure 1, using a single value of λ_x across
371 all genes yields good recovery in general with results that are relatively insensitive to the specific
372 value of λ_x (chosen within a broad range).

373 Expressed as a quadratic program this becomes:

374 $$\hat{x} = \underset{x \geq 0}{argmin} \; -2y^T Ax + x^T(A^T A + \lambda_x I)x$$

375 ***Similarity Regularization***

376 Finally, we make the following observation: the process used to construct A may have access to
377 additional information. For example, when scRNA clusters are used to define the cell states
378 leading to the cluster mixtures in A, *DESP* can extract useful information from the cluster RNA
379 expression profiles. We emphasize that this information can be useful even though *no assumption*
380 *is made* about any particular relationship between, for example, RNA expression levels and protein
381 abundance.

382 *DESP* employs this strategy by incorporating a state-state similarity measure. While in principle
383 *DESP* can use any similarity or dissimilarity matrix, in our results *DESP* operates using scRNA
384 data as follows.

385 Define the matrix C to consist of cell state profiles with cell state profiles on the rows and genes
386 on the columns, where entries in C correspond to average gene expression in the cell state.
387 (Processing of scRNA data to form cell state profiles is discussed below.) Rows of C are
388 standardized to zero mean and unit norm, to obtain $\tilde{C}$, which is used to form the correlation matrix
389 of clusters

390 $$M = \tilde{C}\tilde{C}^T$$

391 M is typically invertible due to the independence of cell states, so *DESP* performs similarity
392 regularization by seeking to minimize $x^T M^{-1} x$

393 Hence to include cell state similarity as a form of regularization, *DESP* minimizes:

394 $$\hat{x} = \underset{x \geq 0}{argmin} \|Ax - y\|^2 + \lambda_x \|x\|^2 + \beta x^T M^{-1} x$$

395 which is equivalent to the quadratic program:

396 $$\hat{x} = \underset{x \geq 0}{argmin} - 2y^T Ax + x^T(A^T A + \lambda_x I + \beta M^{-1})x \,. \quad (1)$$

397 As for $\boldsymbol{\lambda_x}$, we find that *DESP* obtains good recovery in general for $\boldsymbol{\beta}$ values across a fairly broad
398 range (Extended Figure 1).

399 ***Implementation***

400 The R function solve.QP from the *quadprog* package can solve any of the above quadratic
401 programs. So for each feature (e.g. gene/protein), *DESP* invokes solve.QP to solve the constrained
402 minimization (1) to estimate cell state profiles. *DESP* also has minimal running time; in our
403 experiments, we found that *DESP* performed predictions for >8,000 features within less than a
404 second on a standard laptop machine.

405 EMT multi-omics experiment summary

406 We used data from (Paul et al., 2021)[13]. In that study, cells of the human mammary epithelial cell-
407 line MCF10A were treated with TGF-Beta to induce epithelial-to-mesenchymal transition (EMT)
408 and samples were aliquoted at ten timepoints for multi-omics profiling. Details on the experiment
409 and the downloadable datasets can be accessed at (Paul et al., 2021)[13].

410 Proteomics data processing

411 Quantitative profiles of 6,967 proteins were generated by nanoLC-MS/MS across ten timepoints
412 (0, 4 hrs, 1–6, 8 & 12 days post-TGF-Beta injection) with three biological replicates per timepoint.
413 Raw MS files were processed using MaxQuant (version 1.6)[32]. Tandem mass spectra were
414 searched against the reference proteome of *Homo sapiens* (Taxonomy ID 9606) downloaded from
415 UniProt in April 2017. Peptides of minimum seven amino acids and maximum of two missed
416 cleavages were allowed, and a false discovery rate of 1% was used for the identification of peptides
417 and proteins. Contaminants and reverse decoys were removed prior to downstream analysis. The
418 average intensity across replicates was computed for each protein in each timepoint for
419 downstream analysis. Timepoints 3 and 9 were not used in this study since they are not present in
420 the scRNA-Seq data generated from these samples and subsequently used in downstream analysis.

421 All data pre-processing of the MaxQuant files was performed using the R statistical software
422 (version 4.1.0).

423 Single-cell RNA-Seq data processing

424 Single-cell RNA-Seq quantified the expression of 9,785 genes in 1,913 cells at eight consecutive
425 timepoints, with roughly 200 cells per timepoint. Details on the experiment and the downloadable
426 data can be found at (Paul et al., 2021)[13]. Low-quality genes detected in less than 5% of all cells
427 were discarded (17 genes). Lowly-expressed genes with less than three counts in at least three cells
428 were also removed (1,240 genes). On average, each cell expressed ~3,600 genes after filtering.
429 Finally, the data was normalized such that each cell's expression values sum to one. Since *DESP*
430 ultimately combines information derived from the scRNA-Seq data with the proteomics data, we
431 filtered the scRNA-Seq data to the set of 5,606 genes for which quantitative bulk protein
432 measurements were also present in this dataset for the identification of co-expressed cell states
433 prior to applying *DESP* to the proteomics data. All processing of the scRNA-Seq data was
434 performed using the R statistical software (version 4.1.0).

435 Identifying intermediate cell states from scRNA-Seq data

436 The cell states in the dataset were identified in an unsupervised manner based on the similarity of
437 gene expression profiles in the scRNA-Seq data. All 1,913 cells from all the timepoints were
438 pooled together for this analysis. The *Seurat* [14] R package was used to identify the cell states using
439 their recommended workflow as follows. The data was normalized by dividing the gene counts in
440 each cell by the total counts in the cell and multiplied by a scale factor of 10,000. The normalized
441 counts were log-transformed after adding a pseudocount of one. The data was then centered by
442 subtracting each gene's expression by its average expression and scaled by dividing the centered
443 expression levels by their standard deviations. The top 2,000 most variable genes were used to
444 perform PCA dimensionality reduction and subsequently construct a k-nearest neighbor graph with
445 k = 20 and 10 principal components used. Finally, the data was clustered using *Seurat*'s
446 FindClusters function with a resolution of 1.1, defining ten co-expressed cell clusters ranging in
447 size from 68 to 303 cells. The choice of ten clusters was made based on manual evaluation of the
448 clustering results at different clustering resolutions. Supplementary Table 1 contains information

449 on the markers enriched among each of these identified clusters. In the rest of the manuscript, we
450 use the term 'cell states' to refer to these ten clusters.
451 A matrix containing the cell state profiles (denoted '*X*') was constructed by averaging the
452 expression of each gene in each cell state. A cell state proportions matrix (denoted '*A')* was also
453 constructed by counting the number of cells from each state in each timepoint. These matrices
454 were used in the downstream validation and application of *DESP* to the EMT multi-omics data.
455 We also tested our approach on different pre-defined numbers of states ranging between five and
456 fifteen by varying the clustering resolution to examine the effect of varying the number of cell
457 populations on *DESP*'s predictions (Extended Fig 1).

458 Each of the ten states identified in the previous step were also classified into one of three
459 EMT stages based on their change in proportions across time: Epithelial (E), Mesenchymal (M),
460 or Intermediate (I). The cell states with their maximum count in the earlier timepoints were labeled
461 *E*, those in the later timepoints were labeled *M*, and those in the middle timepoints being *I*.

462 The following classifications were assigned:

463 • Epithelial states: 2, 5, 9
464 • Intermediate states: 8
465 • Mesenchymal states: 7

466 Any cell states not included in the above list were deemed not to show a distinct cross-timepoint
467 change in proportion pattern that would best fall into one of the three biological states of interest.

468 **Mathematical validation using pseudobulk**

469 Prior to making inferences from the proteomics data, we first investigated
470 *DESP*'s ability to recover the scRNA data from the bulk data at the RNA-level where
471 we have the true single-cell profiles to compare against. To represent the bulk data
472 that *DESP* expects as an input, we constructed a 'pseudobulk' matrix representing
473 the summed gene expression levels across all cell states within each of the eight
474 timepoints. This matrix is constructed as the product of multiplying the cell state
475 proportions matrix '$\boldsymbol{A}$' and the cell state profile matrix '$\boldsymbol{X}$'. The resultant matrix '$\boldsymbol{Y}$'
476 had one value per gene in each timepoint. The cluster similarity matrix '$\boldsymbol{M'}$' was
477 created by computing the Pearson correlation between each pair of cell state profiles,
478 i.e. the rows of the matrix '$\boldsymbol{X'}$', and used to guide the algorithm towards a solution
479 that preserves the relationship between the input cell states. We then use *DESP* to
480 solve the under-determined problem $\boldsymbol{Y} = \boldsymbol{AX'}$ to predict the cell state profile matrix
481 $\boldsymbol{X'}$ as described in the '*DESP* algorithm overview' section. Finally, the predicted cell
482 state profile matrix $\boldsymbol{X'}$ is then compared to the real cell state profile matrix $\boldsymbol{X}$ to assess
483 *DESP*'s performance. We utilized four different methods of assessing the prediction
484 accuracy, each of which is detailed below.

485 We computed the mean per-gene relative absolute error between X and X' by calculating
486 the absolute difference between each gene's real cross-state expression values in X and the
487 predicted ones in X' and taking the mean across all such gene-level errors. This metric was
488 motivated by the fact that *DESP* performs its predictions independently for each gene. This metric
489 was also used to evaluate the robustness of *DESP* to variations in its two input parameters λ and β
490 by examining the results for different values of the parameters (Extended Fig. 1).

491 As a visualization of the similarity of the predicted profiles in X' to the real ones in X within
492 the global structure of the single-cell data, we created a UMAP projection using the R package
493 *uwot* on the original scRNA-Seq gene expression matrix with the addition of the real and predicted

494 cell state profiles, i.e. the rows of X and X'. The profiles were labeled on the subsequent UMAP
495 plots to distinguish them from the individual cells (Fig. 3b).

496 To determine if *DESP*'s predictions are reasonable approximations of the real data, we also
497 compared the predictions to the real data by computing the Euclidean distance between individual
498 cells' measurements and their real state average as opposed to the distance between thee
499 predictions and the same state average. More specifically, we iterated over each of the 1,913 cells
500 in the original single-cell expression matrix and computing the Euclidean distance between the
501 cell's profile and that of its cell state's profile. Next, for each cell state we compute the average
502 distance of its cells to the averaged cell state profile, as well as the distance between the predicted
503 profile and the same corresponding cell state profile. We then created bar plots comparing these
504 two distances side-by-side to determine whether the predicted profiles fall within the correct intra-
505 state range (Fig. 3c).

506 Finally, we mapped the predicted cell state profiles to the real ones by solving the linear
507 sum assignment problem (LSAP) using the Hungarian method as implemented in the solve_LSAP
508 function in the *clue* [26] R package (Fig. 3d). In summary, the LSAP algorithm expects similarities
509 between cell state profiles as the entries in the input matrix. The idea of the matching is that given
510 two sets A and B, we find the matching that maximizes sum(similarity(a_i, b_j)) where a_i is
511 matched with b_j, and each member of the set is matched with exactly one member of the other
512 set. The algorithm was run twice; once with the Euclidean distance as the similarity metric for the
513 stae profiles where the algorithm minimized the sum of assigned Euclidean distances, and once
514 with the Pearson correlation as the similarity metric where the algorithm maximized the summed
515 correlations across cell state assignments.

516 Applying DESP to demix EMT bulk proteomics data

517 We used *DESP* to demix the EMT bulk proteomics data down to the level of the ten scRNA-
518 defined cell states as described in the '*DESP* algorithm overview' section. The inputs to *DESP*
519 were the processed protein-by-timepoint proteomics matrix, the cell state proportions matrix
520 indicating the proportion of each cell state in each timepoint, and the cell state similarity matrix
521 containing the Pearson correlations between the cell state scRNA profiles. The parameters λ and

522 β were set to 1^{-7} and 1^{-4} respectively based on the tuning results obtained from demixing the
523 'pseudobulk' as described in the previous section. The output was a protein-by-state matrix which
524 was used for the analysis described in the results section of the manuscript.

525 Identifying cell state markers

526 The R package *Seurat*'s[14] FindMarkers function was used to find the genes that distinguish each
527 cell state in the scRNA-Seq data based on a one-vs-all differential expression analysis. A FDR
528 cutoff of 0.05 and log fold-change cutoff of 0.5 were applied to determine the marker genes for
529 each state. These markers are listed in Supplementary Table 1. This analysis was performed once
530 at the level of the ten individual cell states, to confirm the classification of the cell states into the
531 E/I/M stages, and once at the level of the three stages for pathway enrichment and comparing
532 marker genes across datasets. This analysis identified 142 marker genes for the E state, 36 for the
533 I state, and 81 for the M state in the scRNA-Seq data (Supplementary Table 2). For the predicted
534 cell state proteomics profiles from *DESP*, since we have only one measurement per cell state as
535 opposed to the multiple per-cluster cell measurements in the scRNA-Seq data, we define the
536 marker proteins for each cell state as those with an average log2 fold-change greater than 1 between
537 each state and the average of all the other ones (Supplementary Table 1). As with the scRNA-Seq
538 data, this analysis was repeated for the three EMT stages (E, I, and M) instead of the ten cell states
539 after the initial cell states were classified into each of the three stages. This identified 90 marker
540 proteins for the E state, 215 for the I state, and 155 for the M state in the *DESP* predictions
541 (Supplementary Table 2).

542 Identifying timepoint markers in bulk proteomics

543 We grouped the eight timepoints in the bulk proteomics data into the E (first two timepoints), I
544 (middle three timepoints), and M (last three timepoints) stages. Markers of each stage were defined
545 as the proteins with an average log2 fold-change greater than 1 between each stage's timepoints
546 and the average of all the other ones in the bulk proteomics data (i.e. one-vs-all differential
547 expression analysis). The data was log2-transformed and quantile-normalized prior to identifying
548 the markers. This analysis identified 118 marker genes for the E state, 21 for the I state, and 56 for
549 the M state in the scRNA-Seq data. These markers are listed in Supplementary Table 2.

550 ## Identifying intermediate EMT markers

551 To focus on the mechanisms occurring exclusively during the intermediate stage, we decided to
552 focus on the third timepoint (T2) and cell cluster 8. Cell cluster 8 is strongly-associated with this
553 timepoint as it represents 43% of all cells in T2 but only a tiny percentage in the other timepoints.
554 We converted the bulk protein abundances in T2 to proportions of the total T2 abundance and did
555 the same for the predicted protein profile of cluster 8. We then divided these two percentages by
556 each other to find the ratio between them (C8/T2), removing the few cases (~20 proteins) where
557 the protein has an abundance of zero in either dataset. Next, we ranked the proteins based on this
558 ratio, and focused on proteins who have the highest/lowest ratio. The idea is that these proteins are
559 important drivers of the transition since they are relatively important to the transient cluster 8 but
560 whose signal seems diluted in the bulk proteomics based on this ratio, probably due to the presence
561 of other clusters contributing to the bulk signal. The resultant ratios had a mean and median of ~1,
562 and we decided to focus on the 99th percentile, representing the 70 proteins whose ratio was above
563 1.32. In parallel, we also took the 208 proteins defined as upregulated markers of the predicted
564 protein profile of cell cluster 8 based on a one-vs-all differential expression analysis as a second
565 set of intermediate markers. The two analyses described above thus produced two sets of proteins
566 of interest for their inferred role in the transitional phase of EMT, and we found eight proteins that
567 overlapped between them. The results section of the manuscript elaborates on some of the inferred
568 roles of these proteins in the context of EMT.

569 ## Recovering associations with EMT-related genes

570 We curated a set of 320 EMT-associated genes and interrogated the set of predicted intermediate
571 markers for physical and/or functional associations with those genes. The EMT-associated genes
572 came from two sources: 200 genes from the MSigDB Hallmark geneset[21], and 132 genes
573 associated with at least one of the 11 GO Biological Process terms with "epithelial mesenchymal
574 transition" in the title[20]. 12 genes overlapped between the two sets, while none of our predicted
575 markers were present in this set, suggesting potential novelty. We used the STRING database[19] to
576 check for reported physical interactions and/or functional associations between our marker genes
577 and the EMT-associated genes. To cast a wide net, we considered associations reported based on
578 all of STRING's scoring methods, such as RNA co-expression and experimentally-determined

interactions, but filtered them to include only high-confidence interactions (STRING combined score > 70%).

Pathway enrichment analysis

We used the *enrichR*[29] tool to identify biological pathways significantly enriched for each cell state's protein and RNA markers as well as the bulk proteomics timepoint markers based on a hypergeometric test with a p-value cutoff of 0.01. The pathway databases interrogated included GO Biological Process, GO Molecular Function, and GO Cellular Component from the 2021 version of the GO database[20], in addition to MSigDB[21], OMIM[30], and KEGG[31]. We also used the *enrichR* tool to compute and compare the enrichment scores of gene sets of interest across the markers (Fig. 5b, Extended Fig. 5). *enrichR*'s combined score, computed as log(p) * z where p is the Fisher test p-value and z is the z-score for deviation from expected rank, was used for this analysis as a measure of the relative enrichment of a given geneset among a given set of cell state markers. Supplementary Table 3 lists the detected significantly-enriched pathways from the *DESP* predictions, scRNA-Seq, and bulk proteomics data.

Single-cell proteomics data processing

Mass spectrometry-based single-cell proteomics data for cells undergoing TGF-Beta-induced EMT was obtained from (Slavov, Asthagiri et al., *in preparation*)[16]. Their experimental setup consisted of three timepoints (days 0, 3, and 9), and the data has 1,776 proteins x 420 cells, out of which 1,064 proteins overlapped with our EMT case study bulk proteomics and scRNA-Seq data. For all downstream analysis, only this set of ~1,000 genes that overlap between the single-cell proteomics, bulk proteomics, and scRNA-Seq datasets are retained. Notably, 65% of the original single-cell proteomics matrix consists of missing values. Specht et al.[10] proposed a strategy to impute the missing values by k-nearest neighbor imputation using Euclidean distance as a similarity measure between the cells. Briefly, for a given missing value, the expression of that gene in that cell is taken as the mean of its expression in the 3 most similar cells for which its expression is present. The key parameter for this KNN algorithm is the number of neighbors K. We selected K based on a five-fold cross-validation within the training dataset and computing the resultant

606 mean absolute prediction error on the held out non-missing values. The optimal value for *K* was
607 shown to be 10 (Extended Fig. 3) and subsequently used to impute the missing data.

608 Identifying cell states in single-cell proteomics data

609 We clustered the imputed single-cell proteomics data to four clusters using the *K*-means clustering
610 algorithm based on cell-cell similarity in a 2-D UMAP projection of the data. The choice of four
611 clusters was due to its manually-observed improved ability at distinguishing clusters that exhibited
612 noticeable cross-time trends in sample proportion. Extended Fig. 3 visualizes the clusters using a
613 UMAP projection, as well the proportion of each cluster within each timepoint. Based on the
614 change in cluster proportions across time, the clusters were mapped to the Epithelial (*E*),
615 Intermediate (*I*), and Mesenchymal (*M*) states, with one cluster seemingly consistently present
616 across the timecourse and as such was discarded from downstream analyses.

617 Comparing predictions to experimental data

618 Our experimental validation strategy focused on comparing cell state fold-changes by examining
619 whether the fold-changes derived from *DESP*'s prediction resembled the experimentally-derived
620 fold-changes more than those derived from the scRNA data alone. Focusing on the relative fold-
621 changes helps alleviate potential issues caused by technical differences between experiments and
622 pre-processing strategies that would hinder attempts to map cell clusters between the datasets. We
623 computed the gene/protein fold-changes between each pair of cell-states in each of the three single-
624 cell matrices: scRNA-Seq, single-cell proteomics, and *DESP's* single-cell proteomics predictions.
625 Since there are three cell states defined here, this leads to three sets of fold-changes for each
626 dataset: E vs I, E vs M, and I vs M. These fold-changes are computed as log2 fold-changes using
627 *Seurat's* foldChanges function. The resultant fold-changes were centered and scaled prior to
628 downstream analysis. Only the genes/proteins that overlap between the three datasets were used
629 for this analysis.

630 This analysis presented us with two main challenges to overcome:

631 (1) The effects of technical variation between the single-cell and bulk proteomics experiments.

632 (2) Only some genes show a difference between their RNA and protein profiles in this dataset.

633 We addressed challenge 1 by selecting the set of proteins that showed high agreement (Pearson
634 $R^2 > 0.5$) between the bulk proteomics dataset and a 'pseudobulk' matrix constructed by summing
635 the single-cell proteomics measurements in each timepoint. We addressed challenge 2 by selecting
636 the set of proteins that differ the most (Pearson $R^2 < -0.5$) between the bulk proteomics data and
637 a 'pseudobulk' matrix constructed by summing the scRNA data in each timepoint, representing
638 the set of genes that are the best at distinguishing the RNA and protein levels in this dataset and
639 are thus most appropriate for our benchmarking analysis. These correlations only factored in the
640 set of proteins and timepoints that overlapped between the two datasets. Ultimately, this led to a
641 set of 92 proteins used in this analysis. We computed the Pearson correlation between the fold-
642 changes derived from the different datasets as a measure of similarity of the results derived from
643 each of the three matrices to determine whether *DESP*'s predicted profiles resembled the
644 experimentally-determined ones more than the input scRNA ones (Fig. 5c). Note: Relaxing the
645 correlation thresholds to 0 instead of +/- 0.5 led to a larger set of 203 proteins (19% of total
646 overlapping proteins) meeting the criteria and had equivalent final results.

647 Phosphoproteomics data processing

648 Phosphoproteomics data for the same set of cells was downloaded from (Paul et al)[13], with three
649 replicates per timepoint. The dataset consists of phosphorylation data for 8,727 phosphosites
650 across 1,776 unique proteins. The intensity was averaged across replicates for each phosphosite.

651 Human Cell Atlas case study

652 Publicly-available scRNA-Seq data was downloaded from the Human Cell Atlas[15]. The data
653 consisted of gene expression values for 58,870 genes across 24 tissues encompassing 177
654 annotated cell types. The following pre-processing steps were applied to the *raw* scRNA-Seq
655 counts:

656 1. Removed genes with zero variance across cells (removed 1,619 genes)
657 2. Removed lowly-expressed genes by retaining genes with more than 3 counts in at least 3
658 cells (removed 8,399 genes)
659 3. Normalized the data such that each cell sums to 1

4. Retained the top 36 cell types based on the number of tissues each cell type is expressed in (led to a minimum of 3 out of 24 tissues per cell type)

To perform *DESP* demixing, we constructed a matrix with the average expression of each gene in each cell type. This matrix represented the true cell type profiles, i.e. the ground truth to compare *DESP*'s predictions against. We also constructed a matrix containing the proportion of each cell type among the cells belonging to each tissue (Extended Fig. 5b). We then applied *DESP* to these two matrices, using a λ parameter value of 0 and β value of 1^{-10}, to predict the expression of the 48,852 retained genes across the 36 cell types.

Acknowledgements

AE and MC acknowledge a generous pilot award from the Rafik B. Hariri Institute for Computing and Computational Science & Engineering at Boston University. We acknowledge the data and advice from the group of Dr. Nikolai Slavov at Northeastern University. We acknowledge the advice and helpful discussions with Dr. Josh Campbell, Dr. Stefano Monti, and Dr. Trevor Siggers at Boston University.

Author Contributions

AY, AE, and MC conceived the project. AY and MC performed the computational analysis. IP provided biological expertise. AY wrote the manuscript. All authors revised the manuscript. AE and MC supervised the project.

References

1. Ogbeide, Silvia, et al. “Into the Multiverse: Advances in Single-Cell Multiomic Profiling.” *Trends in Genetics*, vol. 38, no. 8, 2022, pp. 831–843., https://doi.org/10.1016/j.tig.2022.03.015.
2. Stuart, Tim, and Rahul Satija. “Integrative Single-Cell Analysis.” *Nature Reviews Genetics*, vol. 20, no. 5, 2019, pp. 257–272., https://doi.org/10.1038/s41576-019-0093-7.
3. Chen, Geng, et al. “Single-Cell RNA-Seq Technologies and Related Computational Data Analysis.” *Frontiers in Genetics*, vol. 10, 2019, https://doi.org/10.3389/fgene.2019.00317.
4. Buccitelli, Christopher, and Matthias Selbach. “MRNAs, Proteins and the Emerging Principles of Gene Expression Control.” *Nature Reviews Genetics*, vol. 21, no. 10, 2020, pp. 630–644., https://doi.org/10.1038/s41576-020-0258-4.
5. Fortelny, Nikolaus, et al. "Can we predict protein from mRNA levels?." Nature 547.7664 (2017): E19.
6. Brion, Christian, et al. “Simultaneous Quantification of Mrna and Protein in Single Cells Reveals Post-Transcriptional Effects of Genetic Variation.” *ELife*, vol. 9, 2020, https://doi.org/10.7554/elife.60645.
7. Aebersold, Ruedi, and Matthias Mann. “Mass-Spectrometric Exploration of Proteome Structure and Function.” *Nature*, vol. 537, no. 7620, 2016, pp. 347–355., https://doi.org/10.1038/nature19949.
8. Uhlen, M., et al. “Tissue-Based Map of the Human Proteome.” *Science*, vol. 347, no. 6220, 22 Jan. 2015, pp. 1260419–1260419, 10.1126/science.1260419.
9. Marx, Vivien. “A Dream of Single-Cell Proteomics.” *Nature Methods*, vol. 16, no. 9, 2019, pp. 809–812., https://doi.org/10.1038/s41592-019-0540-6.
10. Specht, Harrison, et al. “Single-Cell Proteomic and Transcriptomic Analysis of Macrophage Heterogeneity Using scope2.” *Genome Biology*, vol. 22, no. 1, 2021, https://doi.org/10.1186/s13059-021-02267-5.
11. Brunner, Andreas-David, et al. “Ultra-High Sensitivity Mass Spectrometry Quantifies Single-Cell Proteome Changes upon Perturbation.” *Molecular Systems Biology*, vol. 18, no. 3, 2022, https://doi.org/10.15252/msb.202110798.
12. Lamouille, Samy, et al. “Molecular Mechanisms of Epithelial–Mesenchymal Transition.” *Nature Reviews Molecular Cell Biology*, vol. 15, no. 3, 2014, pp. 178–196., https://doi.org/10.1038/nrm3758.
13. Paul, Indranil, et al. “A Multi-Tiered Map of EMT Defines Major Transition Points and Identifies Vulnerabilities.” 2021, https://doi.org/10.1101/2021.06.01.446492.
14. Satija, Rahul, et al. “Spatial Reconstruction of Single-Cell Gene Expression Data.” *Nature Biotechnology*, vol. 33, no. 5, 2015, pp. 495–502., https://doi.org/10.1038/nbt.3192
15. Regev, Aviv, et al. “The Human Cell Atlas.” *ELife*, vol. 6, 2017, https://doi.org/10.7554/elife.27041.
16. Slavov, Asthagiri, et al. (in preparation)
17. Segelle, Alexandre, et al. “Histone Marks Regulate the Epithelial-to-Mesenchymal Transition via Alternative Splicing.” *Cell Reports*, vol. 38, no. 7, 2022, p. 110357., https://doi.org/10.1016/j.celrep.2022.110357.
18. Lone, Imtiaz Nisar, et al. “The Role of Histone Variants in the Epithelial-to-Mesenchymal Transition.” *Cells*, vol. 9, no. 11, 2020, p. 2499., https://doi.org/10.3390/cells9112499.
19. Szklarczyk, Damian, et al. “The String Database in 2021: Customizable Protein–Protein Networks, and Functional Characterization of User-Uploaded Gene/Measurement Sets.” *Nucleic Acids Research*, vol. 49, no. D1, 2020, https://doi.org/10.1093/nar/gkaa1074.
20. Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. 2021 Jan 8;49(D1):D325-D334. doi: 10.1093/nar/gkaa1113. PMID: 33290552; PMCID: PMC7779012.
21. Liberzon, Arthur, et al. “The Molecular Signatures Database Hallmark Gene Set Collection.” *Cell Systems*, vol. 1, no. 6, 2015, pp. 417–425., https://doi.org/10.1016/j.cels.2015.12.004.
22. Kim, Kimberly H, and Charles W Roberts. “Targeting EZH2 in Cancer.” *Nature Medicine*, vol. 22, no. 2, 2016, pp. 128–134., https://doi.org/10.1038/nm.4036.
23. Duan, Ran, et al. “EZH2: A Novel Target for Cancer Treatment.” *Journal of Hematology & Oncology*, vol. 13, no. 1, 2020, https://doi.org/10.1186/s13045-020-00937-8.
24. Straining, PharmD, Rachael, and William Eighmy, PharmD. “Tazemetostat: EZH2 Inhibitor.” *Journal of the Advanced Practitioner in Oncology*, vol. 13, no. 2, 2022, pp. 158–163., https://doi.org/10.6004/jadpro.2022.13.2.7.
25. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N, Iynkkaran I, Liu Y, Maciejewski A, Gale N, Wilson A, Chin L, Cummings R, Le D, Pon A,

Knox C, Wilson M. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res. 2017 Nov 8. doi: 10.1093/nar/gkx1037.

26. Avila Cobos, Francisco, et al. "Benchmarking of Cell Type Deconvolution Pipelines for Transcriptomics Data." *Nature Communications*, vol. 11, no. 1, 2020, https://doi.org/10.1038/s41467-020-19015-1.
27. Bilous, Mariia, et al. "Metacells Untangle Large and Complex Single-Cell Transcriptome Networks." *BMC Bioinformatics*, vol. 23, no. 1, 2022, https://doi.org/10.1186/s12859-022-04861-1.
28. Hornik, Kurt. "A Clue for Cluster Ensembles." *Journal of Statistical Software*, vol. 14, no. 12, 2005, https://doi.org/10.18637/jss.v014.i12.
29. Kuleshov, Maxim V., et al. "ENRICHR: A Comprehensive Gene Set Enrichment Analysis Web Server 2016 Update." *Nucleic Acids Research*, vol. 44, no. W1, 2016, https://doi.org/10.1093/nar/gkw377.
30. Amberger, Joanna S., et al. "OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an Online Catalog of Human Genes and Genetic Disorders." *Nucleic Acids Research*, vol. 43, no. D1, 2014, https://doi.org/10.1093/nar/gku1205.
31. Kanehisa, Minoru, et al. "Kegg as a Reference Resource for Gene and Protein Annotation." *Nucleic Acids Research*, vol. 44, no. D1, 2015, https://doi.org/10.1093/nar/gkv1070.
32. Tyanova, Stefka, et al. "The Maxquant Computational Platform for Mass Spectrometry-Based Shotgun Proteomics." *Nature Protocols*, vol. 11, no. 12, 2016, pp. 2301–2319., https://doi.org/10.1038/nprot.2016.136.