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 foundational resolution of the underlying cell states. Here, we present *DESP*, a novel algorithm that leverages independent estimates of cell state 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

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 face technical challenges including the difficulty of capturing individual cells from mixtures while maintaining proteome integrity, the small amount of extracted protein, and the infeasibility of efficiently scaling up mass spectrometry runs comparable to sequencing methods[10,11]. Emerging technologies aim to address these challenges but significant limitations remain, including the extent of proteome coverage and the inability to quantify functional information such as macromolecular interactions and post-translational modifications[10,11].

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.

As a generalizable computational solution to this challenge, we present a novel demixing algorithm titled *DESP* (Demixing Cell State Profiles in Bulk Omics) that leverages information on

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

67 We applied *DESP* to a multi-omics time-course dataset investigating the molecular changes
68 of pre-cancerous cells undergoing TGF-Beta-induced epithelial-to-mesenchymal transition (EMT)
69 in-vitro[12,13]. EMT is a complex biological process by which epithelial cells lose their adhesion and
70 gradually differentiate into mobile mesenchymal cells and is of particular importance to the
71 biomedical community due to its integral role in wound healing, tissue development, and cancer
72 progression[12]. We analyzed bulk quantitative proteomics and scRNA-Seq data generated at eight
73 consecutive timepoints representing MCF10A mammary epithelial cells undergoing EMT[13], and
74 we show that *DESP* is able to identify the corresponding cellular proteomes and recover expected
75 patterns of canonical EMT markers. Additionally, *DESP* goes further by identifying the proteomic
76 profile of transient cell states key to the differentiation process that were not detected at the
77 transcriptomics level and whose signal is significantly diluted in the bulk proteomics due to the
78 confounding presence of other cell states in the mixture. These results demonstrate a compelling
79 use case of *DESP* while also shedding light on the transient nature of intermediate molecular
80 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, optionally, 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 samples is less than the number of cell states. For example, an experiment that surveys a handful of cancer subtypes will likely contain dozens of cell states involved, as such giving many possible

112 cell-level proteomes that could lead to the observed bulk proteome. We thus apply a set of both
113 biologically- and mathematically-motivated constraints to guide the algorithm towards the optimal
114 solution in a deterministic manner (**Fig. 1c**). One notable constraint leverages correlations between
115 cell state profiles in the input single-cell data to maintain the global cell state structure within
116 *DESP*'s predictions by preserving the original state-state similarities. Notably, *DESP*'s robustness
117 to input parameters allows it to efficiently make predictions without requiring parameter tuning
118 (**Extended Fig. 1**). A detailed description of the mathematical model can be found in the Methods
119 section.
120 *DESP* does not aim to define cell states or their proportions from 'omics data, but rather it
121 predicts the proteome of pre-defined cell states of interest based on the bulk proteomics data. The
122 utility of *DESP* thus lies in the presence of established routine workflows for identifying the global
123 proteome and cell state composition of heterogenous samples, effectively bypassing the lack of
124 methods for deep measurement of individual cell proteomes. It is worth noting that *DESP* is
125 agnostic of the technique used to identify cell state proportions and thus does not directly use the
126 single-cell measurements to infer the protein levels of the given cell states, allowing *DESP* to avoid
127 issues regarding poor correlations between different data types like RNA-Seq and proteomics.
128
129 ***Multi-omics profiling of EMT***
130 We applied *DESP* to a time-course study examining pre-cancerous cells undergoing TGF-
131 Beta-induced EMT[13]. Bulk quantitative proteomics and scRNA-Seq data for over 5,600 genes
132 were generated in parallel for eight consecutive timepoints as MCF10A human mammary
133 epithelial cells underwent the EMT differentiation process (**Fig 2a**). Analysis of the scRNA-Seq
134 data using the *Seurat* tool[14] revealed ten discrete cell states among the 1,913 captured cells, with
135 intermediate states emerging from the predominantly epithelial cultures before transforming into
136 the mesenchymal end state over the timecourse (**Fig 2c,d**). The availability of both bulk protein
137 measurements and single-cell proportion information for the samples along with the dynamic
138 nature of the samples' cellular compositions renders it an excellent testbed for *DESP* (**Fig. 2f**).
139 This dataset was thus used for both algorithm validation and as a case study of a real-life
140 application of *DESP* (**Fig 2g,h**).
141
142

143 ***Mathematical validation using pseudobulk RNA data***

144 To validate the algorithm and assess performance, we investigated *DESP*'s ability to
145 recover the RNA profiles of individual EMT cell states from corresponding bulk RNA profiles.
146 We utilized the scRNA-Seq data, and not the proteomics, for the mathematical validation as it
147 represents the 'true' experimentally-derived single-cell profiles to compare against *DESP*'s bulk-
148 inferred predictions. Since *DESP*'s predictions are made at the resolution of cell states rather than
149 single cells, we constructed representative cell state profiles from the scRNA data by averaging
150 the expression of each gene in each cell state's cells to represent the ground truth[27]. **Fig. 3a** shows
151 a schematic illustrating the validation strategy, where a 'pseudobulk' RNA matrix is first created
152 by summing the scRNA measurements within each of the eight timepoints. We then used *DESP*
153 to demix these pseudobulk profiles to the level of the ten cell states present in this dataset, as
154 guided by the scRNA-derived changes in cell state proportions across the timepoints (**Methods**).
155 These predicted cell state profiles were then compared to the ground truth profiles.

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

169 Cell state assignments by tools such as *Seurat* can be imprecise. Hence, to ensure the
170 robustness of *DESP* to variations in the number of input cell states, we repeated the validation with
171 numbers of input cell states varying between 5 and 20 and found that while the prediction error
172 expectedly increased with the number of states, it remained within a modest error range (**Fig. 3e**).

173 We also show that *DESP* is robust to a wide range of its two input parameters with negligible
174 effects on performance (**Extended Fig. 1**).
175 Since *DESP* performs predictions on a feature-by-feature basis, we also examined whether
176 the accuracy of predictions varied with the expression of particular genes and observed that more
177 cross-cell stable 'housekeeping' genes had lower prediction errors than the more variable genes,
178 while expression magnitudes were not influential (**Extended Fig. 2a,b**). In a similar vein, we
179 examined the relationship between the prediction errors and cell state properties and found that the
180 larger and more homogenous states had the lowest prediction errors (**Extended Fig. 2c**).
181
182 ***Mathematical validation using the Human Cell Atlas***
183 To further validate the effectiveness of *DESP*, an additional case study was conducted
184 using scRNA-Seq data obtained from the Human Cell Atlas[15]. The Human Cell Atlas provides a
185 comprehensive collection of scRNA-Seq data, encompassing 48,852 genes expressed across 24
186 human tissues. This dataset serves as a global reference atlas of human cell expression profiles,
187 offering an ideal opportunity to assess *DESP*'s performance on large-scale datasets (**Fig. 4a**).
188 For the validation process, we focused on the 36 most prevalent cell types, as determined
189 by the number of tissues in which each cell type was present. Initially, we constructed a matrix
190 that represented the proportion of each cell type within each tissue (**Fig. 4b**). Subsequently, we
191 evaluated *DESP*'s ability to reconstruct the single-cell profiles by applying it to a 'pseudobulk'
192 dataset. This dataset was generated by summarizing the scRNA-Seq data at the tissue level by
193 averaging the expression of each gene within each cell type, and multiplying these averaged
194 expressions by the corresponding cell type proportions across tissues (**Methods**).
195 We performed a qualitative evaluation of *DESP*'s performance by combining the real and
196 predicted cell type profiles into one matrix and constructing a UMAP plot that visualizes these
197 profiles in two dimensions, revealing strong overlap between the corresponding real and predicted
198 cell type profiles within this reduced dimensional space (**Fig. 4c**). As a quantitative assessment,
199 we computed the relative prediction error on a gene-by-gene basis by summing the absolute
200 difference between each gene's real and *DESP*-predicted averaged cell type profiles divided by
201 the real cell-type profiles. The resultant median per-gene error was 15% and the mean was 20%
202 (IQR 6 – 30%) (**Fig. 4d**). We repeated the same analysis for each cell type separately, finding a
203 median per-cell type error of 7% and mean of 10% (IQR 2 – 13%) (**Fig. 4e**). In our experiment, it

204 took only 6 seconds to perform the demixing for all 48,852 genes on a standard laptop machine,
205 demonstrating *DESP*'s ability to swiftly produce accurate predictions on large-scale datasets.
206
207 ***Performance comparison between transcriptomics and proteomics***
208 *DESP* provides a mathematical framework for charting the relationship between single-
209 cell and bulk omics datasets which depends on the assumption that the bulk data can be modeled
210 as the sum of the single-cell data. To ensure that this relationship holds for different types of omics
211 data, we tested the scenario where both datasets were obtained from the same samples, we applied
212 *DESP* to matched single-cell transcriptome and protein data generated by the CITE-Seq
213 technology from the Hao et al. study[33]. This dataset comprised 161,764 peripheral blood
214 mononuclear cells (PBMCs) isolated from 8 individuals at three timepoints after receiving an HIV
215 vaccine. The authors grouped the cells into 31 annotated immune cell types. Notably, there were
216 187 genes that had both RNA and protein measurements available for analysis.
217 We used *DESP* to independently demix two pseudobulk datasets created using the same
218 underlying cell type composition matrix: the aggregated RNA data and the aggregated protein data
219 per sample (**Methods**). Comparing the 31 *DESP*-predicted cell type profiles to the true ones
220 measured by CITE-Seq revealed a median per-gene relative prediction error of 27% (IQR 21-34%)
221 for the protein data, compared to 46% (IQR 28-70%) for the RNA data (**Fig. 4f**). These results
222 indicate that *DESP* can predict single-cell protein measurements from transcriptomics-derived cell
223 type structures with comparable, even superior, accuracy to its predictions for transcriptomics data.
224 These observations align with previous research highlighting the consistency of high-level cell
225 type assignments between RNA and protein measurements, despite the discordance in per-gene
226 correlations between RNA-Seq and proteomics data[11,33].
227
228 ***Detection of transient EMT mechanisms using DESP***
229 Applying *DESP* to demix the bulk EMT proteomics profiles allowed us to contextualize
230 the cross-timepoint proteome changes according to the underlying cell states and characterize the
231 distinct protein programs that occur during EMT that are lost in the bulk measurements as
232 elaborated upon below.
233 To evaluate the extent to which *DESP*'s predictions provided information beyond what the
234 bulk proteomics or scRNA data provided alone, we compared the results of performing differential

235 expression analysis between the cell states from the *DESP* predictions versus those two original
236 data types by themselves and found that *DESP* detected hundreds of markers for the EMT stages
237 that were not deemed differential in the other techniques (**Extended Fig. 3**). Several of these
238 markers are known to be involved in EMT, such as TPM2 in the E state and CD44 in the M state.
239 Gene set enrichment analysis also revealed more than 200 pathways that were significantly
240 enriched among *DESP*-identified markers but not revealed by the other standard approaches,
241 demonstrating *DESP*'s ability to glean novel insight on the involved molecular mechanisms rather
242 than just recapitulating what existing methods would have found (**Fig. 6a**). These pathways
243 included both canonical EMT-related pathways such as signaling, migration, and wound healing,
244 as well as potentially novel biological processes that had not been associated with EMT before,
245 such as cellular stress response, spliceosome assembly, and protein transport (**Fig. 6b**).
246 The intermediate cell states are of high interest due to their transient nature and expected
247 influential role in driving EMT. We thus sought to investigate if unique insights could be gleaned
248 by *DESP* on the transient molecular mechanisms governing EMT. Notably, eight proteins,
249 including three encoded by histone genes, were identified as being key to the transition due to their
250 relative enrichment in the predicted intermediate cell state proteome compared to their diluted
251 signal in the bulk proteomics data, representing a signal that likely would have been missed if
252 analyzing bulk proteomics alone (**Fig. 6d**).
253 Of these eight proteins, H2AFX stands out as a marker (99th percentile) of the predicted
254 intermediate proteomics profile yet with low abundance (9th percentile) in the corresponding
255 scRNA profile suggesting extensive post-transcriptional regulation. H2AFX is a core histone
256 protein variant that both contributes to chromatin remodeling and is post-translationally modified
257 during EMT[17,18]. Notably, H2AFX is hyper-phosphorylated during the earliest timepoints of
258 EMT[13], suggesting a change in this protein's functional activity during the initial intermediate cell
259 states (**Fig. 6e**).
260 The STRING database[19] shows that H2AFX and two other *DESP*-identified markers are
261 functionally-linked with nine proteins annotated as EMT-associated in the GO[20] and MSigDB
262 Hallmark genesets[21]. One of these associated proteins is EZH2, which is part of the PRC2
263 chromatin remodeling complex that performs transcriptionally-inhibitory histone methylation
264 (Fig. 5c). EZH2 has emerged in recent years as a promising target for cancer treatments[22,23], with
265 an FDA-approved drug (Tazemetostat) targeting EZH2 as a cancer treatment[24], and another EZH2-

266 targeting compound currently in clinical trials[25]. In contrast to its target H2AFX, EZH2 shows
267 elevated phosphorylation during later timepoints of the experiment, suggesting the activation of a
268 delayed regulatory 'switch' after three days of TGF-Beta induction.

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

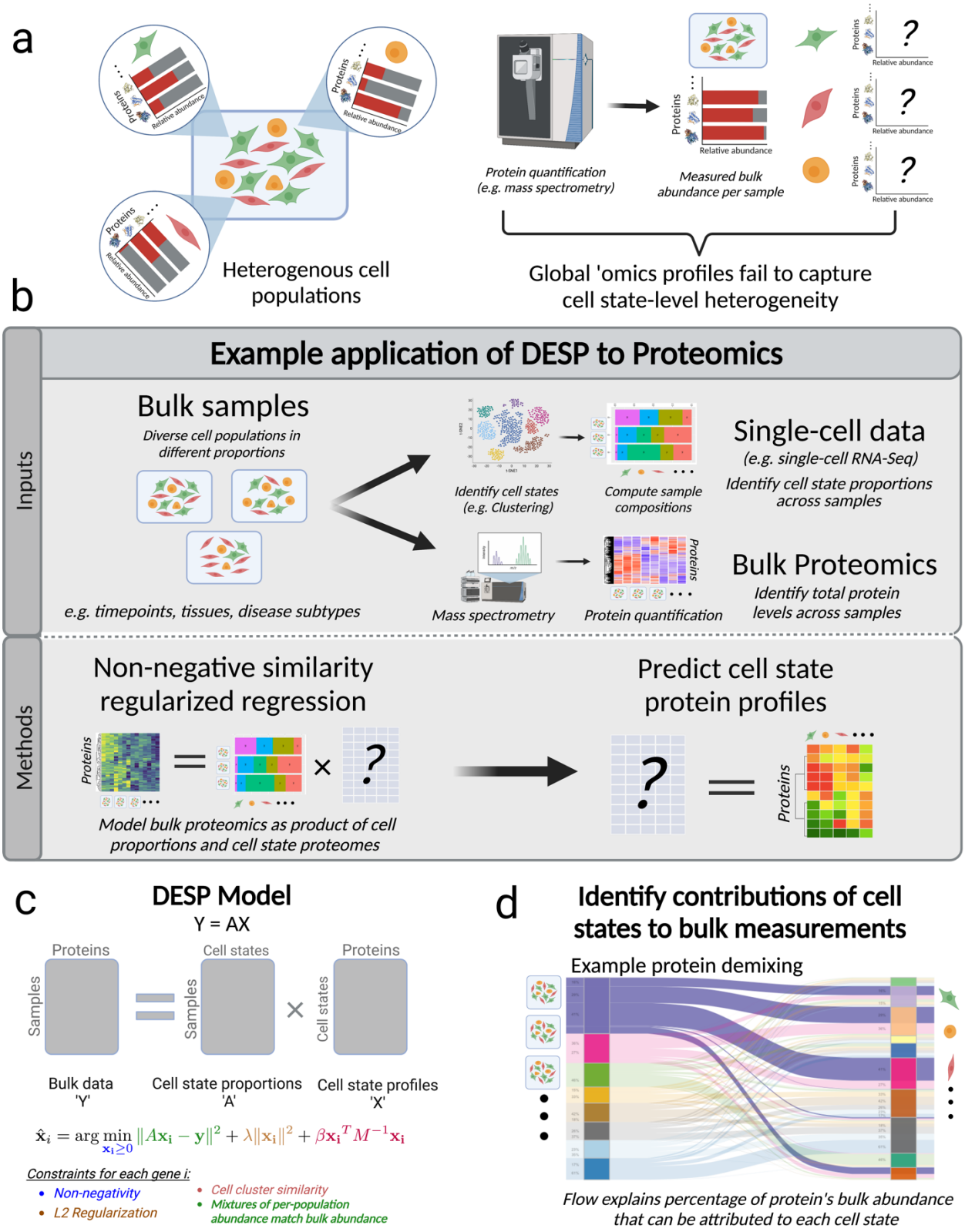
Discussion

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. 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.

We note that as a new type of 'demixing' algorithm, *DESP* does not aim to define, nor does it require, the optimal cell state structures for the bulk data to be demixed. Rather, *DESP* solves the problem of mapping bulk omics measurements, such as proteomics, to a pre-defined cell state structure of interest to the user, independent of how that structure was originally defined. *DESP* provides a generalizable computational framework for bulk omics demixing, with planned future applications including demixing protein interaction networks and spatial transcriptomics data.

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 the bulk data. *DESP* also offers a generalizable framework applicable to all types of omics data unlike existing single-cell inference algorithms, such as TCA[35] and BMIND[36], that are tailored to specific data types. To facilitate adoption and usability by the community, *DESP* is available as an R package at (https://github.com/AhmedYoussef95/DESP).

Figures



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

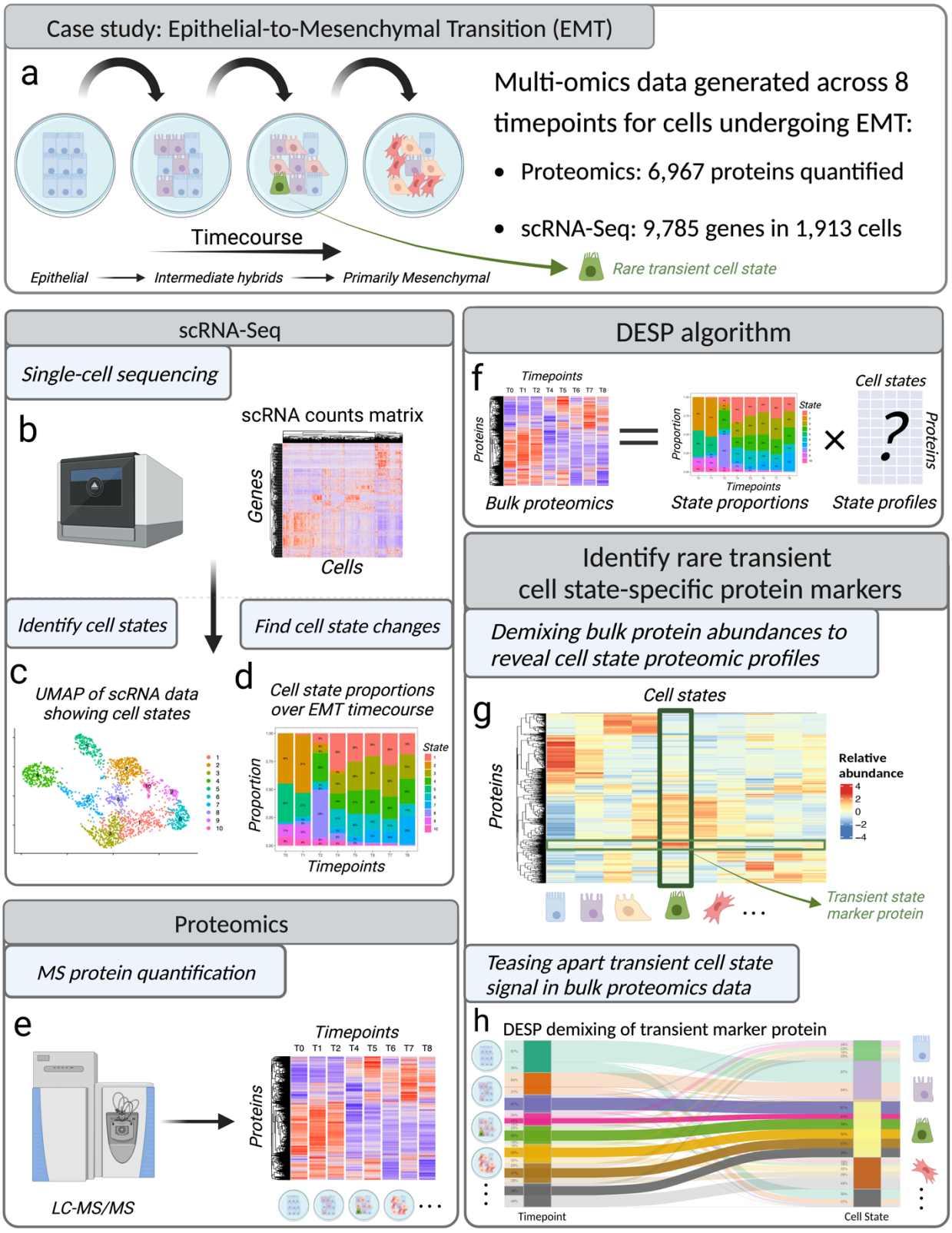
296 *Figure 1: Overview of DESP algorithm*

297 1a: Motivation and rationale behind *DESP*. Standard 'omic techniques such as mass spectrometry-
298 based quantitative proteomics are often limited to bulk measurements of mixed cell populations
299 and fail to resolve sample heterogeneity at the cell-state level.

300 1b: Analysis workflow. *DESP* takes as an input paired bulk omics data, e.g. proteomics, and
301 matching single-cell data and predicts the contributions of the underlying cell states to the observed
302 bulk measurements.

303 1c: Mathematical model. *DESP* models the observed bulk profiles, e.g. proteomics measurements,
304 as the sum of unknown underlying cell state profiles. DESP applies several constraints to this
305 formulation to predict the unknown cell state profiles.

306 1d: Output. Schematic of the *DESP* output showcasing its ability to demix and associate bulk
307 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

308

309

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

311 2a: Case study: EMT dynamic multi-omics profiling. In this experiment, bulk proteomics and
312 scRNA-Seq data were generated for human MCF10A cells undergoing TGF-Beta-induced EMT
313 across eight consecutive timepoints, generating molecular profiles of transient intermediate states.

314 2b: Dynamic scRNA-Seq profiles. Heatmap visualizes the scaled relative expression of genes
315 (rows) across cells (columns).

316 2c: Identification of intermediate cell states occurring during EMT. Clustering the scRNA-Seq
317 data reveals ten intermediate cell states as visualized on UMAP plot.

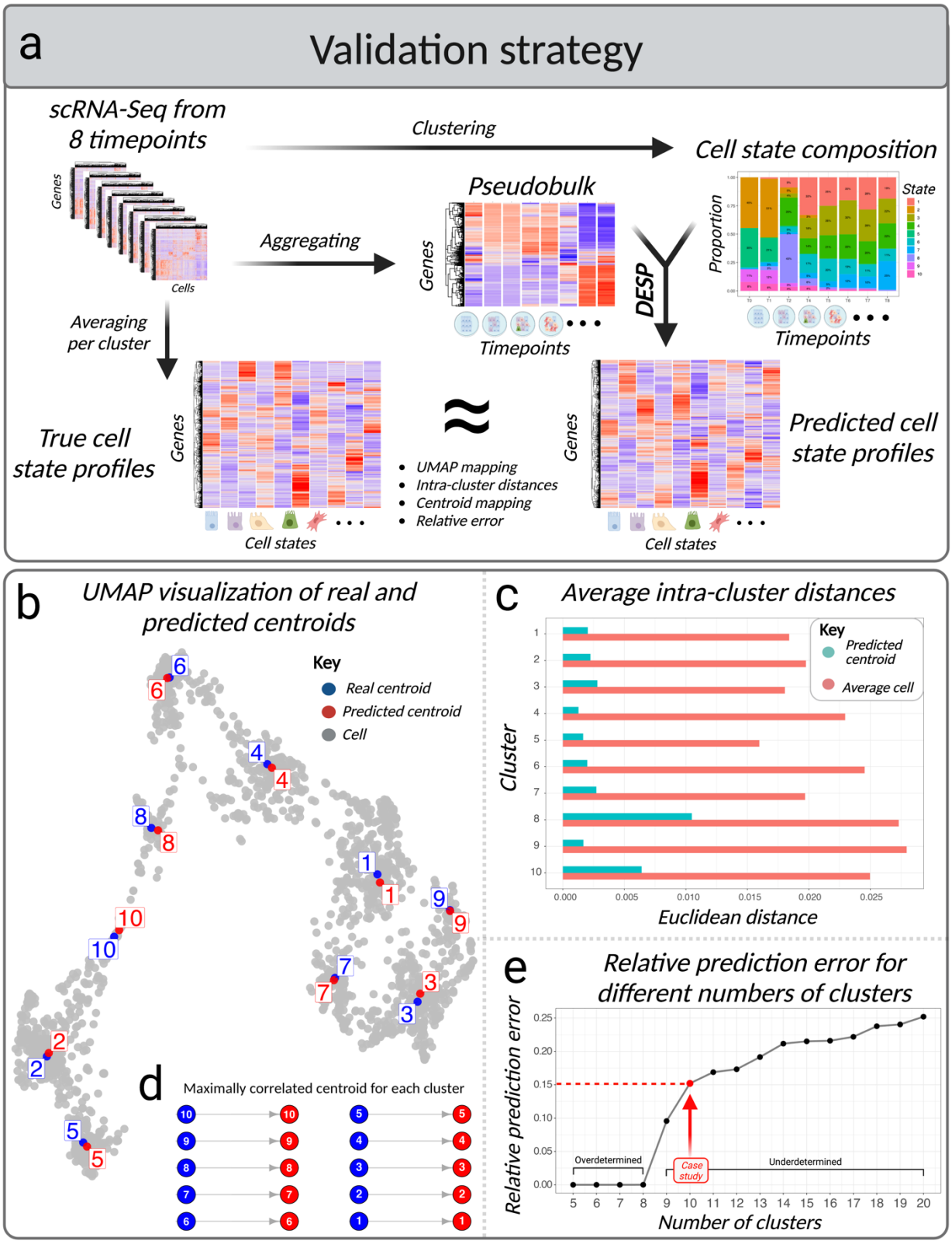
318 2d: Cell state fluctuations during EMT. Bar plot visualizing cell state composition at each
319 timepoint.

320 2e: Mass spectrometry-based quantitative proteomics profiles. Heatmap visualizes scaled TMT
321 reporter ion intensity, a proxy for relative abundance, of proteins (rows) across timepoints
322 (columns).

323 2f: Applying *DESP* to EMT bulk proteomic profiles and scRNA-derived cell state composition
324 matrix to predict cell state-level protein profiles.

325 2g: Prediction of intermediate cell state profiles using *DESP*. Heatmap visualizes the predicted
326 relative protein abundances for each of the cell states. One putative marker of a transient cell state
327 is highlighted as an example finding from these predictions.

328 2h: Identifying contribution of individual cell states to timepoint bulk proteomics measurements.
329 Schematic shows *DESP* demixing of a given protein to distinct underlying cell states, highlighting
330 a transient cell state signal that is lost by examining bulk measurements alone due to the presence
331 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

332

333

334 *Figure 3: Mathematical validation of DESP using pseudobulk RNA data*

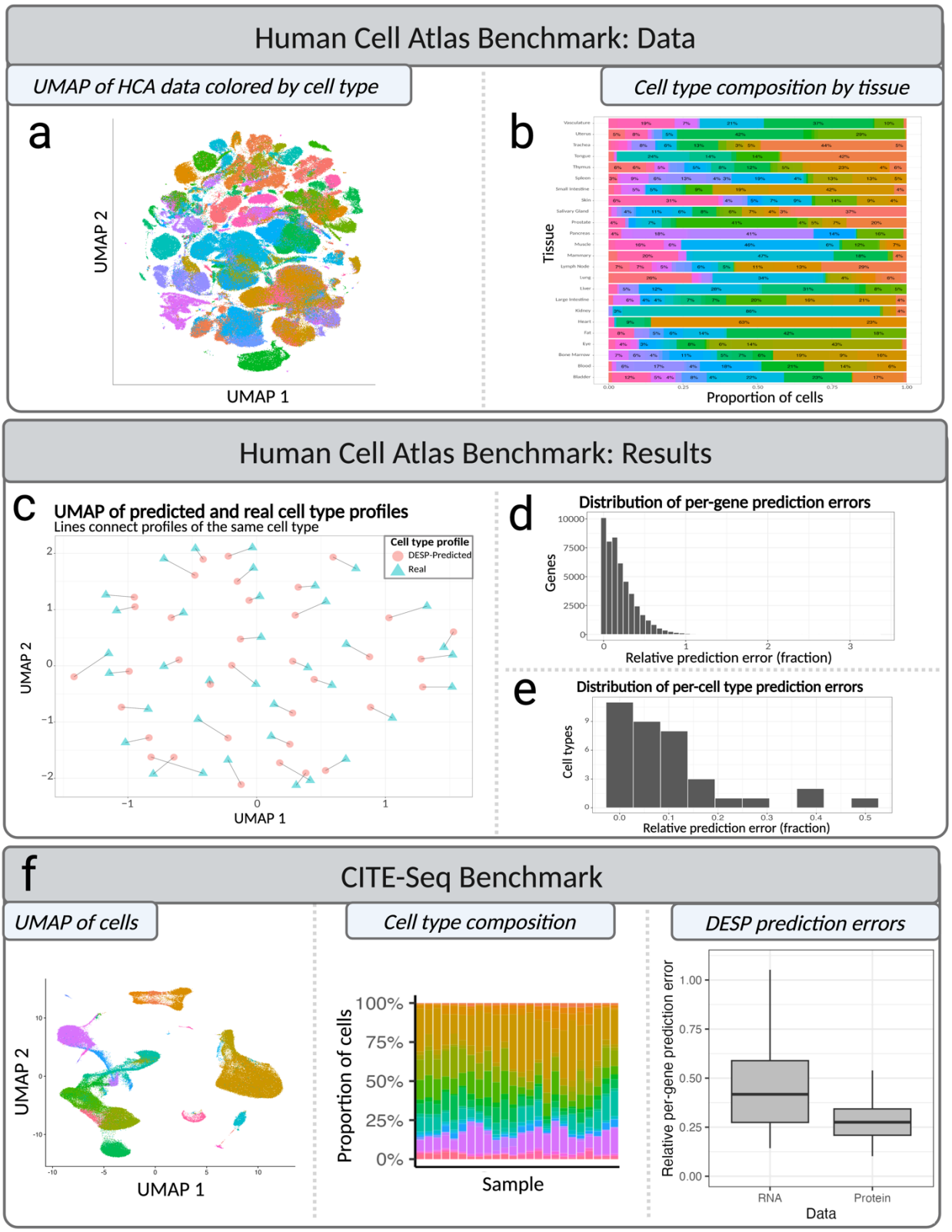
335 3a: Validation strategy. 'Pseudobulk' profiles, generated by summing the scRNA profiles in each
336 timepoint, were demixed by *DESP* guided by the cell state compositions. The predicted *DESP*
337 profiles were compared to the averaged per-cluster gene expression values in the input scRNA
338 data.

339 3b: UMAP showing overlap of the real and predicted cell state profiles, labeled as 'centroids',
340 within the global structure of the scRNA-Seq data.

341 3c: Bar plots showing the average Euclidean distance between each cell state's constituent single-
342 cell measurements to their real centroid compared to that of the predicted centroid.

343 3d: Mapping predicted centroids to closest real centroid based on solving the linear sum
344 assignment problem to minimize total Euclidean distances.

345 3e: Relative per-gene prediction error when varying the number of cell clusters input to *DESP*,
346 highlighting the chosen 10-cluster scenario. Standard error bars are not visible due to small error
347 range.


Human Cell Atlas Benchmark: Data
UMAP of HCA data colored by cell type
a
UMAP 2
UMAP 1
Cell type composition by tissue
b
Tissue
Proportion of cells
Human Cell Atlas Benchmark: Results
c
UMAP of predicted and real cell type profiles
Lines connect profiles of the same cell type
Cell type profile
DESP-Predicted
Real
UMAP 2
UMAP 1
d
Distribution of per-gene prediction errors
Genes
Relative prediction error (fraction)
e
Distribution of per-cell type prediction errors
Cell types
Relative prediction error (fraction)
f
CITE-Seq Benchmark
UMAP of cells
UMAP 2
UMAP 1
Cell type composition
Proportion of cells
Sample
DESP prediction errors
Relative per-gene prediction error
RNA
Protein
Data

350 *Figure 5: Comparing DESP predictions to experimentally-derived profiles*

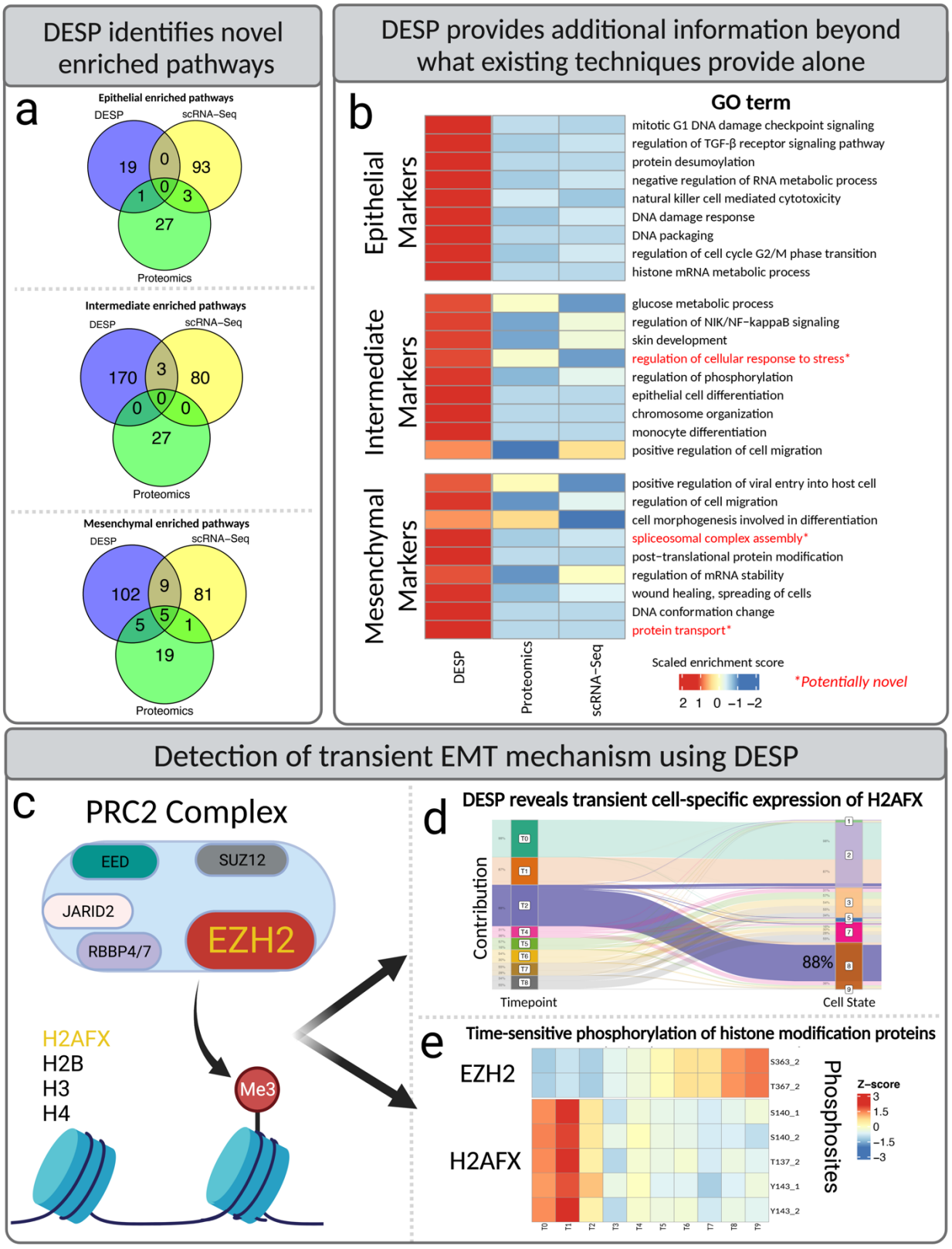
351 5a: Dynamic EMT single-cell proteomics data used for biological validation.

352 5b: Heatmap of computed protein relative abundance fold-changes between cell states in single-
353 proteomics data, scRNA data, and bulk-demixed *DESP* predictions.

354 5c: Pearson correlation of the protein relative abundance fold-changes between cell-states in each
355 dataset.

356 5d: UMAP plot showing overlap of the real and *DESP*-predicted cell state profiles, labeled as
357 'centroids', within the global structure of the single-cell proteomics data. Inset plot: Results of
358 mapping each predicted centroid profile to the closest real centroid profile in a one-to-one manner
359 based on solving the linear sum assignment problem.

360 5e: Bar plots showing the average Euclidean distance between each cell state's constituent single-
361 cell measurements to their real centroid compared to that of the *DESP*-predicted centroid.


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
Timepoint
Cell State
88%
e
Time-sensitive phosphorylation of histone modification proteins
EZH2
H2AFX
Phosphosites
Z-score
3
1.5
0
−1.5
−3
S363_2
T367_2
S140_1
S140_2
T137_2
Y143_1
Y143_2
T0
T1
T2
T3
T4
T5
T6
T7
T8
T9

364 *Figure 6: Detection of transient EMT mechanisms using DESP*

365 6a: Venn diagrams showing overlap of significantly-enriched pathways within each EMT stage
366 detected by scRNA-Seq data alone, bulk proteomics data alone, and *DESP* predictions.

367 6b: Heatmap of relative enrichment scores of selected GO terms among cell stage markers that
368 were enriched among *DESP*'s predictions compared to results obtained from scRNA-Seq or bulk
369 proteomics alone.

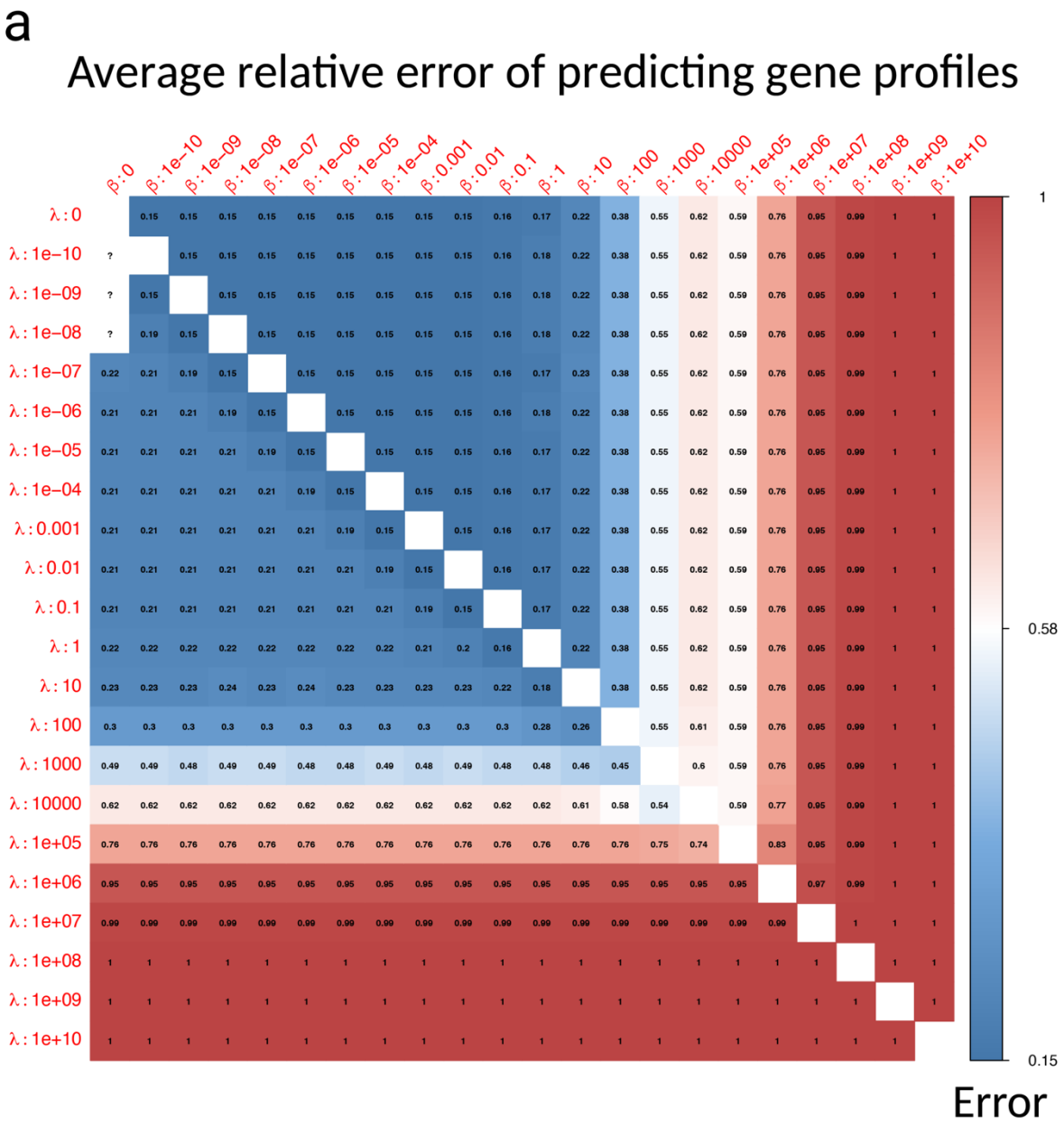
370 6c: PRC2 core histone methylation complex. The EZH2 and H2AFX proteins are highlighted
371 based on their inferred influential transient profiles detected by *DESP*.

372 6d: *DESP* demixing of H2AFX across transient cell states, with intermediate timepoint T2
373 highlighting the time-sensitive disproportionate contribution by transient cell state 8.

374 6e: Heatmap of temporal phosphorylation of EZH2 and H2AFX in global quantitative
375 phosphoproteomics analysis of the same EMT samples.

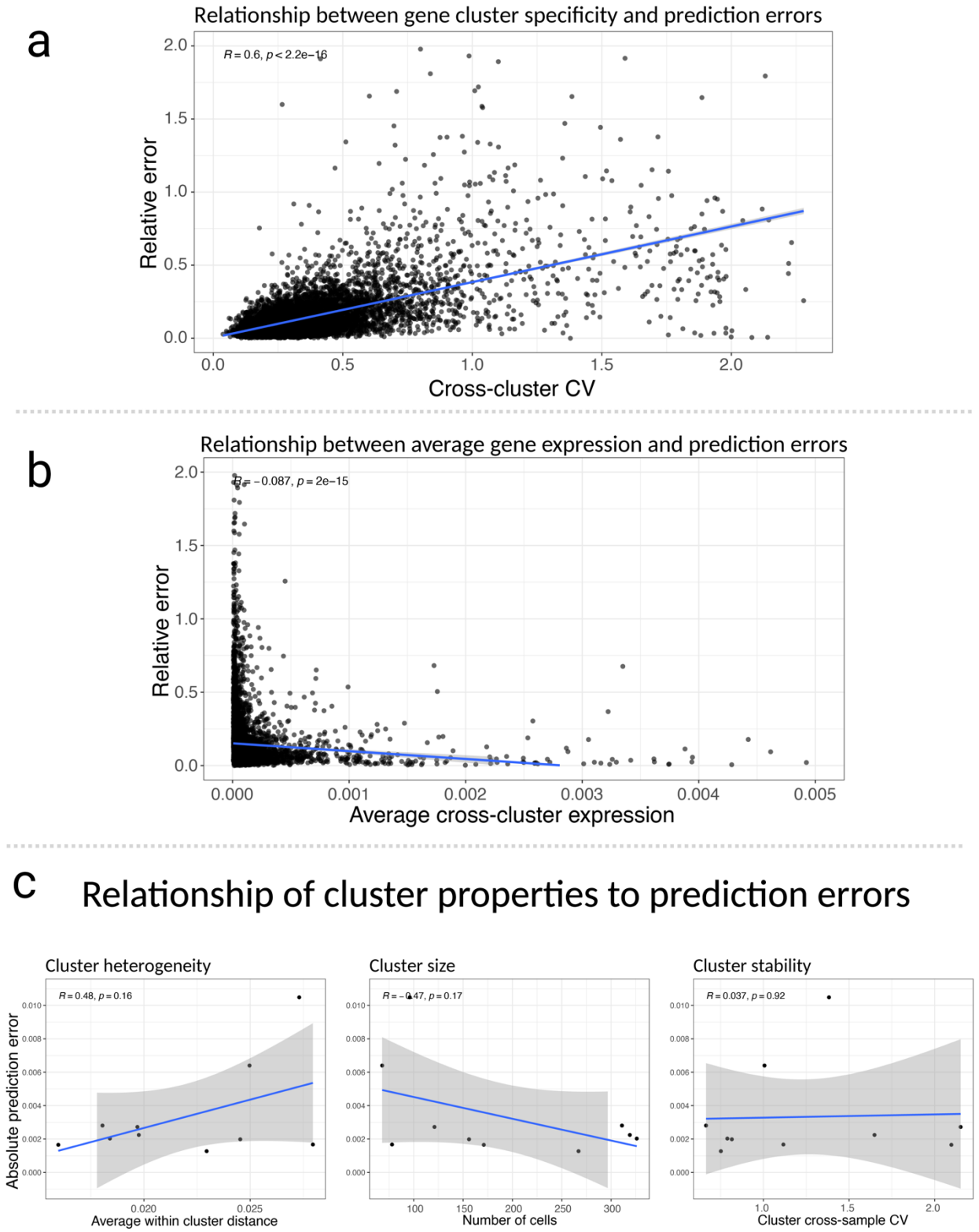
376

Supplementary Figures



379 *Extended Figure 1: Robustness of DESP predictions to input parameters*

380 1a: Mean per-gene relative prediction error for applying *DESP* to EMT 'pseudobulk' data using a
381 wide range of values for the two input parameters λ and β. The error is computed on a feature-by-
382 feature basis by dividing the absolute difference between the real and predicted cell state
383 expression values of each gene by the real expression values. Errors shown in the figure were
384 rounded to 2 decimal points.

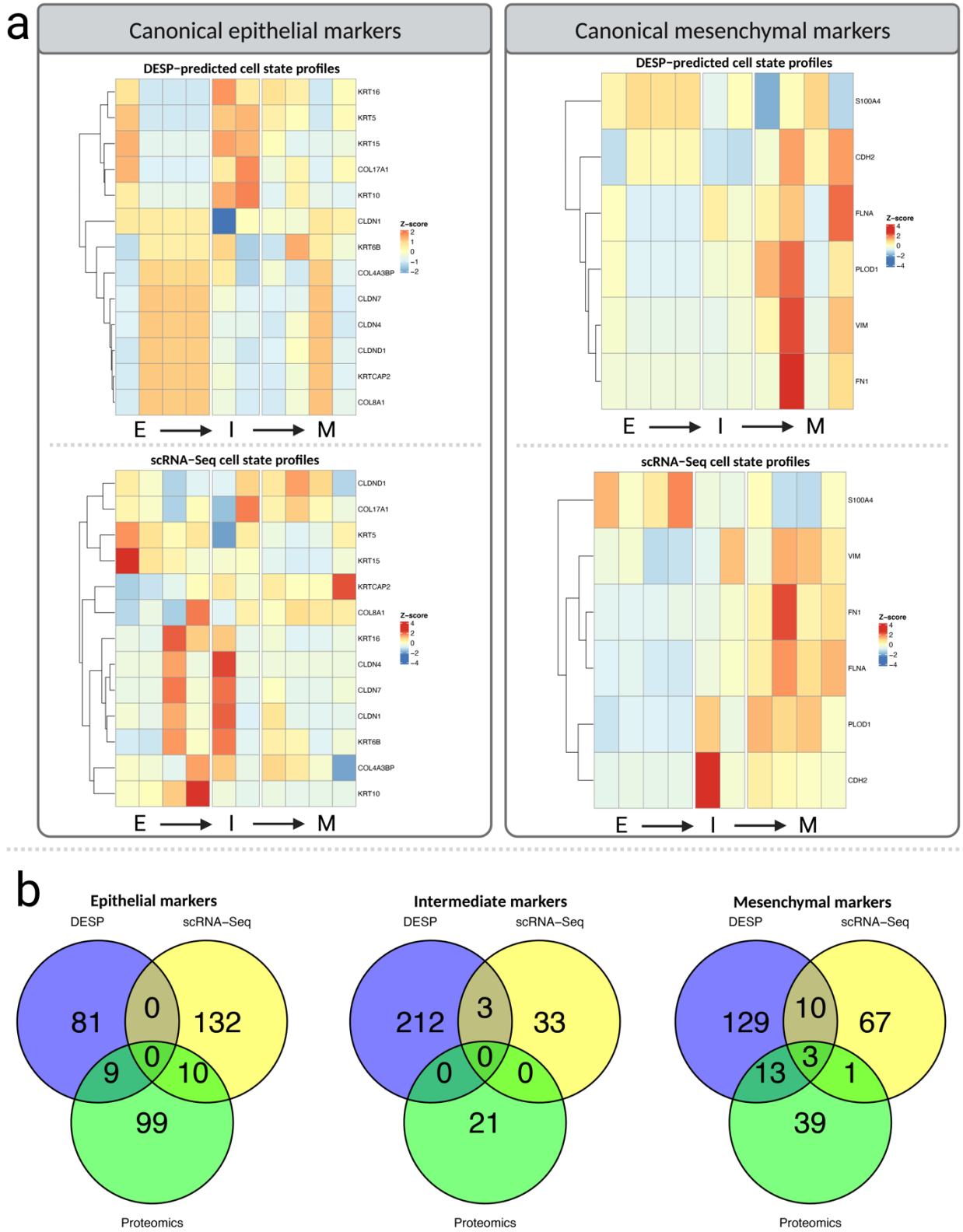

a
Relationship between gene cluster specificity and prediction errors
R = 0.6, p < 2.2e−16
Relative error
Cross-cluster CV
b
Relationship between average gene expression and prediction errors
R = −0.087, p = 2e−15
Relative error
Average cross-cluster expression
c
Relationship of cluster properties to prediction errors
Cluster heterogeneity
R = 0.48, p = 0.16
Absolute prediction error
Average within cluster distance
Cluster size
R = −0.47, p = 0.17
Number of cells
Cluster stability
R = 0.037, p = 0.92
Cluster cross-sample CV

387 *Extended Figure 2: Effect of gene and cluster properties on DESP predictions*

388 2a: Relationship between gene cross-cluster coefficient of variation (CV) and per-gene relative
389 prediction error.

390 2b: Relationship between magnitude of average gene cross-cluster expression and per-gene
391 relative prediction error.

392 2c: Relationship between three mathematical properties of cell clusters (cluster heterogeneity,
393 cluster size, and cluster proportion stability across samples) and corresponding prediction errors.


a
Canonical epithelial markers
DESP−predicted cell state profiles
KRT16
KRT5
KRT15
COL17A1
KRT10
CLDN1
KRT6B
COL4A3BP
CLDN7
CLDN4
CLDND1
KRTCAP2
COL8A1
Z−score
E → I → M
scRNA−Seq cell state profiles
CLDND1
COL17A1
KRT5
KRT15
KRTCAP2
COL8A1
KRT16
CLDN4
CLDN7
CLDN1
KRT6B
COL4A3BP
KRT10
Z−score
E → I → M
Canonical mesenchymal markers
DESP−predicted cell state profiles
S100A4
CDH2
FLNA
PLOD1
VIM
FN1
Z−score
E → I → M
scRNA−Seq cell state profiles
S100A4
VIM
FN1
FLNA
PLOD1
CDH2
Z−score
E → I → M
b
Epithelial markers
DESP
scRNA−Seq
81
0
132
0
9
10
99
Proteomics
Intermediate markers
DESP
scRNA−Seq
212
3
33
0
0
0
21
Proteomics
Mesenchymal markers
DESP
scRNA−Seq
129
10
67
3
13
1
39
Proteomics

396 *Extended Figure 3: Recovery of canonical markers and discovery of novel ones*

397 3a: Expression/abundance patterns of selected canonical EMT markers across cell states in *DESP*'s
398 predicted cell state protein profiles and the original scRNA-Seq data.

399 3b: Overlap of differentially-abundant proteins for each of the three EMT states in the scRNA-Seq
400 data, bulk proteomics data, and *DESP*'s predicted cell state proteomics profiles.

401

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

Expressed as a quadratic program this becomes:

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

Similarity Regularization

Finally, we make the following observation: the process used to construct A may have access to additional information. For example, when scRNA clusters are used to define the cell states leading to the cluster mixtures in A, *DESP* can extract useful information from the cluster RNA expression profiles. We emphasize that this information can be useful even though *no assumption is made* about any particular relationship between, for example, RNA expression levels and protein abundance. *DESP* employs this strategy by incorporating a state-state similarity measure. While in principle *DESP* can use any similarity or dissimilarity matrix, in our results *DESP* operates using scRNA data as follows:

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

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

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

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

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

which is equivalent to the quadratic program:

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

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

Implementation

The R function solve.QP from the *quadprog* package can solve any of the above quadratic programs. So for each feature (e.g. gene/protein), *DESP* invokes solve.QP to solve the constrained

459 minimization (1) to estimate cell state profiles. *DESP* also has minimal running time; in our
460 experiments, we found that *DESP* performed predictions for >8,000 features within less than a
461 second on a standard laptop machine.

462 EMT multi-omics experiment summary

463 We used data from (Paul et al., 2023)[13]. In that study, cells of the human mammary epithelial cell-
464 line MCF10A were treated with TGF-Beta to induce epithelial-to-mesenchymal transition (EMT)
465 and samples were aliquoted at ten timepoints for multi-omics profiling. Details on the experiment
466 and the downloadable datasets can be accessed at (Paul et al., 2023)[13].

467 Proteomics data processing

468 Quantitative profiles of 6,967 proteins were generated by nanoLC-MS/MS across ten timepoints
469 (0, 4 hrs, 1–6, 8 & 12 days post-TGF-Beta injection) with three biological replicates per timepoint.
470 Raw MS files were processed using MaxQuant (version 1.6)[32]. Tandem mass spectra were
471 searched against the reference proteome of *Homo sapiens* (Taxonomy ID 9606) downloaded from
472 UniProt in April 2017. Peptides of minimum seven amino acids and maximum of two missed
473 cleavages were allowed, and a false discovery rate of 1% was used for the identification of peptides
474 and proteins. Contaminants and reverse decoys were removed prior to downstream analysis. The
475 average intensity across replicates was computed for each protein in each timepoint for
476 downstream analysis. Timepoints 3 and 9 were not used in this study since they are not present in
477 the scRNA-Seq data generated from these samples and subsequently used in downstream analysis.
478 All data pre-processing of the MaxQuant files was performed using the R statistical software
479 (version 4.1.0).

480 Single-cell RNA-Seq data processing

481 Single-cell RNA-Seq quantified the expression of 9,785 genes in 1,913 cells at eight consecutive
482 timepoints, with roughly 200 cells per timepoint. Details on the experiment and the downloadable
483 data can be found at (Paul et al., 2023)[13]. Low-quality genes detected in less than 5% of all cells
484 were discarded (17 genes). Lowly-expressed genes with less than three counts in at least three cells
485 were also removed (1,240 genes). On average, each cell expressed ~3,600 genes after filtering.

Finally, the data was normalized such that each cell's expression values sum to one. Since *DESP* ultimately combines information derived from the scRNA-Seq data with the proteomics data, we filtered the scRNA-Seq data to the set of 5,606 genes for which quantitative bulk protein measurements were also present in this dataset for the identification of co-expressed cell states prior to applying *DESP* to the proteomics data. All processing of the scRNA-Seq data was performed using the R statistical software (version 4.1.0).

Identifying intermediate cell states from scRNA-Seq data

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

A matrix containing the cell state profiles (denoted 'X') was constructed by averaging the expression of each gene in each cell state. A cell state proportions matrix (denoted 'A') was also constructed by counting the number of cells from each state in each timepoint. These matrices were used in the downstream validation and application of *DESP* to the EMT multi-omics data. We also tested our approach on different pre-defined numbers of states ranging between five and fifteen by varying the clustering resolution to examine the effect of varying the number of cell populations on *DESP*'s predictions (Extended Fig 1).

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

The following classifications were assigned:

- Epithelial states: 2, 5, 9
- Intermediate states: 8
- Mesenchymal states: 7

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

Mathematical validation using RNA pseudobulk

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

We computed the mean per-gene relative absolute error between X and X' by calculating the absolute difference between each gene's real cross-state expression values in X and the predicted ones in X' and taking the mean across all such gene-level errors. This metric was motivated by the fact that *DESP* performs its predictions independently for each gene. This metric

544 was also used to evaluate the robustness of *DESP* to variations in its two input parameters λ and β
545 by examining the results for different values of the parameters (Extended Fig. 1).
546 As a visualization of the similarity of the predicted profiles in X' to the real ones in X within
547 the global structure of the single-cell data, we created a UMAP projection using the R package
548 *uwot* on the original scRNA-Seq gene expression matrix with the addition of the real and predicted
549 cell state profiles, i.e. the rows of X and X'. The profiles were labeled on the subsequent UMAP
550 plots to distinguish them from the individual cells (**Fig. 3b**).
551 To determine if *DESP*'s predictions are reasonable approximations of the real data, we also
552 compared the predictions to the real data by computing the Euclidean distance between individual
553 cells' measurements and their real state average as opposed to the distance between thee
554 predictions and the same state average. More specifically, we iterated over each of the 1,913 cells
555 in the original single-cell expression matrix and computing the Euclidean distance between the
556 cell's profile and that of its cell state's profile. Next, for each cell state we compute the average
557 distance of its cells to the averaged cell state profile, as well as the distance between the predicted
558 profile and the same corresponding cell state profile. We then created bar plots comparing these
559 two distances side-by-side to determine whether the predicted profiles fall within the correct intra-
560 state range (**Fig. 3c**).
561 Finally, we mapped the predicted cell state profiles to the real ones by solving the linear
562 sum assignment problem (LSAP) using the Hungarian method as implemented in the solve_LSAP
563 function in the *clue* [26] R package (**Fig. 3d**). In summary, the LSAP algorithm expects similarities
564 between cell state profiles as the entries in the input matrix. The idea of the matching is that given
565 two sets A and B, we find the matching that maximizes sum(similarity(a_i, b_j)) where a_i is
566 matched with b_j, and each member of the set is matched with exactly one member of the other
567 set. The algorithm was run twice; once with the Euclidean distance as the similarity metric for the
568 state profiles where the algorithm minimized the sum of assigned Euclidean distances, and once
569 with the Pearson correlation as the similarity metric where the algorithm maximized the summed
570 correlations across cell state assignments.

571 Applying DESP to demix EMT bulk proteomics data

572 We used *DESP* to demix the EMT bulk proteomics data down to the level of the ten scRNA-
573 defined cell states as described in the '*DESP* algorithm overview' section. The inputs to *DESP*

were the processed protein-by-timepoint proteomics matrix, the cell state proportions matrix indicating the proportion of each cell state in each timepoint, and the cell state similarity matrix containing the Pearson correlations between the cell state scRNA profiles. The parameters λ and β were set to 1^{-7} and 1^{-4} respectively based on the tuning results obtained from demixing the 'pseudobulk' as described in the previous section. The output was a protein-by-state matrix which was used for the analysis described in the results section of the manuscript.

Identifying cell state markers

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

Identifying timepoint markers in bulk proteomics

We grouped the eight timepoints in the bulk proteomics data into the E (first two timepoints), I (middle three timepoints), and M (last three timepoints) stages. Markers of each stage were defined as the proteins with an average log2 fold-change greater than 1 between each stage's timepoints and the average of all the other ones in the bulk proteomics data (i.e. one-vs-all differential

602 expression analysis). The data was log2-transformed and quantile-normalized prior to identifying
603 the markers. This analysis identified 118 marker genes for the E state, 21 for the I state, and 56 for
604 the M state in the scRNA-Seq data. These markers are listed in **Supplementary Table 2**.

605 Identifying intermediate EMT markers

606 To focus on the mechanisms occurring exclusively during the intermediate stage, we decided to
607 focus on the third timepoint (T2) and cell cluster 8. Cell cluster 8 is strongly-associated with this
608 timepoint as it represents 43% of all cells in T2 but only a tiny percentage in the other timepoints.
609 We converted the bulk protein abundances in T2 to proportions of the total T2 abundance and did
610 the same for the predicted protein profile of cluster 8. We then divided these two percentages by
611 each other to find the ratio between them (C8/T2), removing the few cases (~20 proteins) where
612 the protein has an abundance of zero in either dataset. Next, we ranked the proteins based on this
613 ratio, and focused on proteins who have the highest/lowest ratio. The idea is that these proteins are
614 important drivers of the transition since they are relatively important to the transient cluster 8 but
615 whose signal seems diluted in the bulk proteomics based on this ratio, probably due to the presence
616 of other clusters contributing to the bulk signal. The resultant ratios had a mean and median of ~1,
617 and we decided to focus on the 99th percentile, representing the 70 proteins whose ratio was above
618 1.32. In parallel, we also took the 208 proteins defined as upregulated markers of the predicted
619 protein profile of cell cluster 8 based on a one-vs-all differential expression analysis as a second
620 set of intermediate markers. The two analyses described above thus produced two sets of proteins
621 of interest for their inferred role in the transitional phase of EMT, and we found eight proteins that
622 overlapped between them. The results section of the manuscript elaborates on some of the inferred
623 roles of these proteins in the context of EMT.

624 Recovering associations with EMT-related genes

625 We curated a set of 320 EMT-associated genes and interrogated the set of predicted intermediate
626 markers for physical and/or functional associations with those genes. The EMT-associated genes
627 came from two sources: 200 genes from the MSigDB Hallmark geneset[21], and 132 genes
628 associated with at least one of the 11 GO Biological Process terms with "epithelial mesenchymal
629 transition" in the title[20]. 12 genes overlapped between the two sets, while none of our predicted

630 markers were present in this set, suggesting potential novelty. We used the STRING database[19] to
631 check for reported physical interactions and/or functional associations between our marker genes
632 and the EMT-associated genes. To cast a wide net, we considered associations reported based on
633 all of STRING's scoring methods, such as RNA co-expression and experimentally-determined
634 interactions, but filtered them to include only high-confidence interactions (STRING combined
635 score > 70%).

636 Pathway enrichment analysis

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

648 Phosphoproteomics data processing

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

652 Human Cell Atlas case study

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

657 1. Removed genes with zero variance across cells (removed 1,619 genes)
658 2. Removed lowly-expressed genes by retaining genes with more than 3 counts in at least 3
659 cells (removed 8,399 genes)
660 3. Normalized the data such that each cell sums to 1
661 4. Retained the top 36 cell types based on the number of tissues each cell type is expressed in
662 (led to a minimum of 3 out of 24 tissues per cell type)

663 To perform *DESP* demixing, we constructed a matrix with the average expression of each gene in
664 each cell type. This matrix represented the true cell type profiles, i.e. the ground truth to compare
665 *DESP*'s predictions against. We also constructed a matrix containing the proportion of each cell
666 type among the cells belonging to each tissue (**Fig. 4b**). We then applied *DESP* to these two
667 matrices, using a λ parameter value of 0 and β value of 1^{-10}, to predict the expression of the 48,852
668 retained genes across the 36 cell types.

669 Analysis of CITE-Seq data

670 Publicly-available CITE-Seq data was downloaded from Hao et al[33]. The dataset consisted of
671 parallel RNA and protein measurements for 161,764 PBMCs isolated from 8 volunteers at three
672 timepoints (days 0, 3, and 7) after receiving an HIV vaccine, with a set of 187 overlapping genes
673 between the RNA and protein measurements. Gene mapping between the RNA and protein
674 measurements were carried out based on common Ensembl identifiers[34]. The authors assigned cell
675 types at 3 hierarchical levels, and we decided to focus our analysis on the intermediate level which
676 grouped the cells into 31 annotated immune cell types. The single-cell data was normalized using
677 the *Seurat* package such that each cell sums to 10,000. We then summarized both the RNA and
678 protein data into the following matrices:

679 • A (cell type composition): Number of cells from each cell type in each sample. This is
680 consistent for both the RNA and protein data.
681 • X (cell type profiles): average expression of each gene/protein in each cell type.
682 • Y (pseudobulk): Matrix multiplication of A and X, representing aggregated single-cell data in
683 each sample.

684 We quantitatively evaluated *DESP*'s ability to predict the cell type profiles in the matrix X given
685 the A and Y matrices, comparing performance for predicting the RNA and protein profiles

686 separately. Since *DESP* predictions are performed on a feature-by-feature basis, we computed the
687 prediction error as the absolute relative difference between the true and predicted cell state values
688 for each feature, and created boxplots comparing *DESP*'s per-gene relative prediction error for the
689 RNA and protein data separately (**Fig. 4f)**.

690

691 # Supplementary Text

692 ***Testing robustness of DESP predictions with different numbers of cell states***

693 In the EMT case study detailed in the manuscript, we identified ten transitional cell states
694 occurring during EMT. These cell states were identified by clustering the single-cell RNA-Seq
695 data to reveal groups of transcriptionally-correlated cells using the *Seurat* R package (**Methods**).
696 Since determining the optimal number of cell states form scRNA-Seq data is not an objective
697 process, we repeated our mathematical validation of *DESP* using different numbers of input cell
698 states to examine the robustness of *DESP*'s predictions.

699 To vary the number of cell states, we used different clustering resolutions within *Seurat*'s
700 FindClusters algorithm followed by repeating the process for constructing a cell state proportions
701 matrix detailing the proportion of each cell state within each of the timepoints. We tested *DESP*
702 using numbers of clusters varying between 5 and 20 and computed mean per-gene relative
703 prediction error for each case (**Fig 3e**). For cases where there were 8 or less cell states, the problem
704 was overdetermined due to the presence of eight timepoints in the data and thus the prediction
705 error was zero. Starting from nine clusters and above, the problem is underdetermined meaning
706 that multiple possible combinations of cell state protein profiles could lead to the observed
707 pseudobulk measurements. As expected, we found that the prediction error increased as the
708 number of cell states increased, with the error ranging from 10-25% for cases with 9-20 clusters
709 respectively.

710 ***Testing robustness of DESP predictions to input parameters***

711 *DESP* expects two input parameters denoted as λ and β (**Methods**). In the EMT case study
712 described in the manuscript, we set the values of these parameters to 1^{-7} and 1^{-4} respectively. To
713 test the robustness of *DESP*'s predictions to variations in the values of the input parameters, we
714 repeated the mathematical validation of *DESP* using combinations of the input parameters ranging
715 from 1^{-10} to 1^{10}, with increasing steps of an order of magnitude each, for each of the two parameters
716 and quantified the mean per-gene relative prediction error for deconvoluting the scRNA-Seq
717 'pseudobulk' data (**Extended Fig. 1)**.

718 We found that the error remained stable at 15% for a wide range of combinations of the
719 input parameters, with predictions made with values of β ranging between 1^{-7} and 0.01 and
720 corresponding values of λ of the same or less value all giving nearly identical results with errors

721 of 15%. Furthermore, the error never exceeded 22% for any combination of the parameters with
722 values between 1^{-10} and 1 for each parameter. The prediction errors were smaller for deconvoluting
723 the pseudobulk single-cell proteomics data (**Fig. 5d,e**), with prediction errors at 2% for any value
724 of β between 1^{-10} and 0.1 and a corresponding λ of the same or less value, and never exceeding
725 16% for any combination of values between 1^{-10} and 1 for either parameter. These results
726 demonstrate the robustness of *DESP* to variations in the input parameters and that there is no need
727 to hypertune the input parameters for different input datasets.

728 ***Effect of gene/protein properties and cell cluster properties on DESP's predictions***

729 *DESP* performs predictions on a feature-by-feature basis. Depending on the input data, these
730 features will typically be genes or proteins. We examined the correlation between mathematical
731 properties of the features in the bulk data and their corresponding prediction error based on
732 applying *DESP* to the scRNA-derived pseudobulk matrix as described in the mathematical
733 validation chapter of the Methods section. To explore the relationship between properties of the
734 genes and their corresponding predicted values, for each of the 8,528 genes in the dataset we
735 computed the following measures:

736 - ***Expression specificity:*** This metric looked at the relative specificity of a given gene's
737 expression to the clusters. The coefficient of variation (CV) of the gene's cluster-specific
738 expression values is computed as the standard deviation of the per-cluster expression
739 values divided by their mean. Genes with a lower CV can be interpreted as being relatively
740 stable 'housekeeping' genes due to the smaller variance in their expressions across clusters.
741 - ***Average expression:*** This metric is concerned with the relative quantity of each gene's
742 transcripts and is simply computed as the mean expression of the gene across clusters.

743 Each of the above properties was correlated to the gene's relative prediction error, defined as
744 the absolute difference between the real and predicted per-cluster values divided by the real value.
745 A Pearson correlation coefficient of 0.6 was observed between the per-gene prediction errors and
746 their expression specificity, indicating that the less variable 'housekeeping' genes tended to have
747 higher prediction accuracy (**Extended Fig. 2a**). Meanwhile, the Pearson correlation coefficient
748 was only -0.01 with the genes' average expression values, suggesting that the magnitude of a
749 gene's expression is less influential to *DESP*'s predictions (**Extended Fig. 2b**).

750 In a similar vein, we computed the Pearson correlation coefficient between the following
751 mathematical properties of the cell clusters and their corresponding prediction errors measured as
752 the Euclidean distance between the predicted centroids and the real ones:

753 - ***Cluster heterogeneity***: For each cluster, the average Euclidean distance between each
754 member cell's scRNA expression profile and that of its corresponding cluster centroid, i.e.
755 the average expression profile of all cells belonging to the cluster.
756 - ***Cluster size***: The number of cells within each cluster.
757 - ***Cluster fluctuation across samples***: The coefficient of variation of each cluster's
758 proportion in each timepoint.

759 We found a positive correlation between cluster heterogeneity and the prediction error
760 (Pearson R^2 = 0.48) indicating that the prediction error was lower for the more homogenous
761 clusters, a negative correlation with size (Pearson R^2 = -0.47) suggesting better prediction for
762 larger clusters, and near-zero correlation with cross-sample cluster fluctuations (Pearson R^2 = 0.04)
763 (Extended Fig. 2c, d, e). These results are potentially confounded however by the fact that cluster
764 size and heterogeneity were strongly negatively-correlated to each other (Pearson R^2 = -0.65).

765

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 generous data and advice from Dr. Nikolai Slavov, Dr. Anand Asthagiri, and Saad Khan 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. “Parallelized multidimensional analytic framework applied to mammary epithelial cells uncovers regulatory principles in EMT.” *Nature Communications*, 14(1), 688, 2023, https://doi.org/10.1038/s41467-023-36122-x.
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.
33. Hao, Yuhan et al. "Integrated analysis of multimodal single-cell data." *Cell* vol. 184,13 (2021): 3573-3587.e29. doi:10.1016/j.cell.2021.04.048
34. Fiona Cunningham et al., Ensembl 2022, Nucleic Acids Research, Volume 50, Issue D1, 7 January 2022, Pages D988–D995, https://doi.org/10.1093/nar/gkab1049
35. Rahmani, E., Schweiger, R., Rhead, B. et al. Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. *Nature Communications* 10, 3417 (2019). https://doi.org/10.1038/s41467-019-11052-9
36. Wang, Jiebiao, Kathryn Roeder, and Bernie Devlin. "Bayesian estimation of cell type-specific gene expression with prior derived from single-cell data." *Genome Research* (2021) 31: 1807-1818.