Machine Learning Predictions Surpass Individual mRNAs as a Proxy of Single-cell Protein Expression

Josephine Fisher[1], Jo.Fisher2@gilead.com (corresponding author)

Oliver Wood[1], Oliver.Wood@gilead.com

Samuel Bullers[1], Sam.Bullers@gilead.com

Lynne Murray[1], lynnemurray@hotmail.com

Li Li[2], Li.Li@gilead.com

Matthew A. Jackson-Wood[1], Matt.JacksonWood@gilead.com

Institutional Affiliation:

1. *Gilead Sciences, 9400 Oxford Business Park, Garsington, Oxford OX4 2HN*
2. *Gilead Sciences, 333 Lakeside Dr, Foster City, CA 94404, USA*

Abstract

Background

Expansive repositories of single-cell RNA-seq data are now available for large-scale transcriptomics analysis at single-cell resolution. However, proteomics data are relatively limited in scale and sensitivity. These data are often analysed assuming that mRNA abundance reflects the expression of their cognate proteins. However, post-transcriptional and translational regulation make mRNA an inadequate proxy for protein. High sparsity in low abundance mRNAs from single-cell transcriptomics data further complicates the extrapolation of protein expression levels. Although methods for single-cell surface protein quantification exist, they incur additional technical steps at greater expense and have yet to see wide-spread adoption. Computational approaches for protein imputation from scRNAseq data have been published, which learn transcriptome-wide patterns that predict protein expression. These models can then be applied to infer surface protein expression on RNA-seq only data, to increase the utility of existing data repositories.

Results

We tested 8 such methods and compared the accuracy of predictions between approaches, and against cognate mRNAs as a direct proxy. Predictions from the trained models outperformed the use of mRNA expression as a proxy. We identify notable cases of cell surface proteins with very poor correlation with their mRNA that were predicted very successfully by imputation using the whole transcriptome. We find cell type signatures are a major determinant of predicted protein levels and, as such, prediction methods require representative training data.

Conclusions

These results reiterate that mRNA level is not a reliable predictor of cell surface protein expression, and that whole-transcriptome informed imputation can improve protein estimations given appropriately trained models.

Keywords: machine, learning, protein, prediction, transcriptomics, proteomics, single-cell, CITE-seq

Background

There has been widespread adoption of methods to profile transcriptomics at the single-cell level (scRNAseq) (1, 2). This has resulted in the publication of an abundance of scRNAseq data sets, including several high-profile projects to generate atlases of human health and disease (3-5) . While scRNAseq is a powerful tool, it should be used with respect to its limitations. Limited read depth, transcript bias, and low capture efficiency can produce highly sparse scRNAseq data, with lower abundance genes often failing to be detected (6). Complexity of post-transcriptional and translational regulation means that mRNA transcript abundance may not reflect functional protein levels. The relationship between mRNA and protein expression could also differ by cell type and state.

New technologies seek to expand options for single cell profiling to measure multiple modalities simultaneously, some of which quantify protein expression alongside mRNA. These include CITE-seq stategies, such as AbSeq from BD Bioscience and Total-Seq from Biolegend, both antibody-based methods that can measure the expression of 100s of cell surface proteins (2, 7-10) . In all three, oligonucleotide barcoded antibodies are used to bind and detect cell-surface protein. Subsequently the cell can be lysed, and protein abundance measured alongside the mRNAs by sequencing. Such data presents an opportunity to quantify the relationship between mRNA transcript quantities and surface protein expression. However, the relative novelty of these approaches and challenge of generating and profiling oligo conjugated antibody panels means that uptake remains comparatively limited. As an alternative to direct measurement a multimodal reference may be used to train a machine learning (ML) model and used to impute protein levels on the large volume of scRNA-seq data already available.

Various methods have recently been published to infer protein levels from scRNAseq using models trained on a multimodal reference. The package Seurat, commonly used for single-cell data

processing, is capable of graph-based reference mapping to assign protein expression based upon cell-cell distances (11); sciPENN, totalVI, BABEL, scMMT and cTPnet all apply some variation of a neural network to generate predicted protein expression values (12-15); the method scLinear uses a comparatively simple approach, applying dimensional reduction followed by linear regression (16). A related unsupervised approach, SPECK, uses reduced rank reconstruction for protein prediction directly from RNA expression alone, clustering each gene separately without using a multimodal reference or whole-transcriptome context (17). To date there has been no independent comparison of the accuracy of these prediction methods and their value over using mRNA abundance alone as an estimator. Additionally, inter-method comparisons of performance when tested on different features and tissues are lacking. To address this, we present a comprehensive evaluation of 8 prediction methods, seeking to define their strengths and limitations.

Protein expression predictions were made on PBMC CITE-seq data from healthy donors, split into training and test partitions (11). We measured correlation between predicted and measured expression values and compared between methods and against the performance of the cognate mRNA as a proxy alone. We also evaluated the main factors driving predictive power by training and testing methods on cell type restricted partitions of the data and by testing the ability of models trained on healthy PBMCs to predict on unseen data from tumour tissue. Overall, we found transcriptome-wide predictions of cell surface protein expression outperformed using the proteins mRNA in isolation. Baseline expression level and variability between cell types were major determinants of a protein's predictability. Our findings highlight that mRNA can be a poor representative of cell surface protein abundance, and that there is sufficient information in the wider transcriptome to make a more accurate inference of protein expression. However, current methods require a well-matched multi-modal reference, which may limit their applicability.

Results

Gene mRNA abundances are an inconsistent proxy of surface protein expression

101 Prior to testing prediction methods, we measured the naïve correlation between proteins and their
102 mRNA abundances as a baseline expectation of correspondence. For this we utilised a large CITE-seq
103 data set profiling PBMCs from healthy volunteers from a vaccine trial, published by Hao et al (11)(Fig.
104 1A). This has been used extensively in demonstrations of multimodal single cell analyses (18-22). It is
105 comprised of 161764 cells, classified into 31 functional types. The diversity of cell types and protein
106 features further supports generation of large independent training and test data sets in latter
107 analyses.

108 Of the 228 proteins measured, 205 could be directly aligned with a single gene measured in the RNA
109 data (excluding non-specific and control antibodies). We computed Pearson correlation between the
110 measured protein and its corresponding mRNA (Supp. Fig. 1). We identified several cases where the
111 expected expression profile as estimated by RNA was dramatically different from the expression
112 observed by antibody measurement. A pronounced example was the tissue-residency marker CD103,
113 showing extracellular protein detected only in a narrow subset of T-cells despite uniform abundance
114 of RNA counts across all immune cell types (Fig. 1B). Of the 205 protein-mRNA pairs, 14 proteins
115 have strong [0.7,1] Pearson correlation with normalised RNA abundance, 36 have moderate [0.4,0.7)
116 correlation, 127 have weak [0,0.4) correlation, and 28 have negative correlation (Fig. 1C).

117 Proteins that are well approximated by RNA include several canonical immunophenotyping markers,
118 such as CD3, CD16, CD14 and CD20. However, some markers of lymphocyte activation and cell state,
119 such as CD25, showed poor correlation with RNA in circulating PBMCs (Fig. 1C). Some of these cases
120 align with known biology; isoforms such as CD45RA and CD45RO are heavily controlled by post-
121 transcriptional processes such as alternative splicing, so aren't expected to align well with PTPRC
122 expression(23). Other discrepancies have less obvious causes, and our results suggest a low
123 probability that any given gene will have RNA measurements well-correlated with protein expression.
124 To address this difficulty, we explore alternative metrics by which RNA can be used as an estimator
125 for extracellular protein, computing either the percentage of RNA positive cells or the pseudobulk
126 count over a cell-type (Fig1. C). Both approaches can recover a well-correlated, population-level
127 estimate of median protein expression, generating similar results between the two methods.

128 Examining changes in group distributions, we see these metrics can improve correlation values for
129 strong and moderately correlated RNA-protein pairs. However, they largely fail to produce high
130 correlation values for proteins with negative or low-valued correlation with the naïve mRNA proxy.
131 Additionally, these metrics cannot be used where single-cell context is important and there remains
132 a set of 102 proteins that fail to achieve correlation ≥ 0.7 using approximation by RNA counts,
133 percentage positivity, or pseudobulk count (Supp. Fig. 2). CD103 is one such example, remaining
134 resistant to these approaches due to fundamental disagreement between RNA and protein
135 measurements.

136

137 **Cell surface protein prediction is comparable across most transcriptome wide modelling methods**

138 Motivated by the insufficiency of protein estimation directly from source mRNA, we evaluated 8
139 published approaches for predicting protein abundances using the full transcriptomic profiles of
140 cells. Again, performance was tested on the Hao et al CITE-seq dataset (11). We took a randomly
141 sampled 20% fraction as the test (i.e. query) data and trained each method on the remaining 80%.
142 Pearson correlation between the predicted and measured protein expression values was used as the
143 primary evaluation metric, henceforth referred to as the prediction correlation.

144 All methods showed bimodality in prediction correlation across features, indicating a discrete set of
145 features that are more challenging to impute than others (Fig. 2B). We sought to identify what
146 characteristics might make a protein easier or more difficult to impute and found a linear association
147 between the mean prediction correlation and total number of counts of a protein (Fig. 2C). Proteins
148 with high standard deviation also trend towards having better average prediction correlation. Thus,
149 features with strong and distinct signal in the data appear easier for prediction methods to impute.

150 For many proteins there is low variability in prediction correlation across the different approaches,
151 indicating a uniform ability across methods to impute expression (Fig2A). A notable exception is the
152 unsupervised strategy SPECK, which does not utilise protein data in its training. SPECK consistently
153 produced predictions with lower correlation than other methods. In several cases SPECK produced

near zero prediction correlation for features which were well-predicted by other methods, such as CD4 and CD9 (Fig2A). SPECK also failed to produce a prediction for some features owing to a lack of RNA detection in the training data, likely due to dropout.

Most methods produced similarly correlated predictions and are not easy to distinguish by accuracy alone. We therefore additionally scored usability of the software and thoroughness of the associated publication, applying an adapted version of the framework presented in evaluations of single-cell integration and trajectory inference methods (24, 25). For comparison the scores are aggregated, with more detail given in Sup. Fig 3. All tested methods are described in peer reviewed journals, with software freely available from a version-controlled repository. Usage tutorials exist for all methods, although with varying levels of detail and relevance to protein imputation. With this context, an ideally performing method would produce high imputation accuracy, have a well-documented and accessible software package, and a thoroughly interrogative publication. These qualities vary across methods, and we were able to highlight discriminating points between them (Fig. 2D). BABEL is presented as being primarily designed for imputation of ATAC-seq peak data. The documentation is therefore opaque on how to implement protein prediction, and although the approach performs well it is challenging to use. In the cases of Seurat, totalVI and scMMT the publication is most concerned with validating alignment of modalities, with protein imputation left as an auxiliary application. The vignettes do detail the software usage well, so although the methods themselves aren't deeply interrogated for efficacy they are simple to execute. Overall, in terms of usability and prediction outcomes, the best-scoring methods are cTPnet, sciPENN, scLinear and scMMT. These are accessible software packages that apply well-presented methods and achieve high prediction correlation values.

We also assessed the speed/scalability of all methods on ascending partitions of the data, running with access to identical computational resources. The data was split into 6.125%, 12.5%, 25%, 50% and 100% fractions, and methods were trained according to the previously described 80:20 training and test split. Median prediction correlation and RMSE over all features was measured, to identify any changes in performance corresponding to training data size. Resource usage and runtime were

181 also evaluated. We found most methods produced predictions with correlation stable to changes in
182 training data size (Supp. Fig. 4). An exception is scMMT, which showed dramatically reduced
183 performance at data size lower than 50% (approx. 80K cells). Regarding resource usage, we find
184 neural-net based methods to be most demanding. BABEL and scMMT used the most RAM during
185 training (Fig. 2D). The five neural-net based methods were also the highest consumers of CPU
186 resources. cTPnet runtime appears to scale particularly poorly with size of training data, although
187 this may be possible to mitigate by adjusting the training patience parameter.

188

189 **Whole transcriptome gene expression predicts protein abundance better than source mRNA alone**

190 Across all tested methods there were many proteins that consistently showed improved prediction
191 correlation with genome wide imputation compared to direct estimation from cognate mRNA (Fig.
192 3A). Overall, ML imputation maintains comparable median prediction correlation to aggregate RNA
193 metrics and produces no negative correlation values. Of the 205 proteins for which cognate RNA data
194 is available, 70 fail to achieve moderate prediction correlation >0.4 by either direct RNA estimation
195 or aggregate RNA measures (Supp. Table 1); 23 of these achieve mean prediction correlation >0.4
196 when using ML methods.

197 The utility of protein imputation is highlighted by the feature CD110, a thrombopoietin receptor
198 commonly detected on platelets (1). Platelets contain no nucleus, and typically have only a small
199 amount of legacy mRNA transferred from megakaryocytes (26). The CD110 transcript MPL is
200 undetected in the Hao et al. test data rendering direct estimation impossible, however CD110
201 protein is enriched on the platelet surface. Using the full, albeit limited, landscape of platelet RNA,
202 ML prediction methods can learn the transcriptional signature of platelets and recovers CD110
203 enrichment in the population (Fig. 3B). As a result, average prediction correlation across methods for
204 CD110 is strong, at 0.75. Other examples benefiting from imputation include protein isoforms, which
205 show a marked improvement in correlation using ML methods. Both CD45RA and CD45RO are
206 predicted with much greater accuracy by the tested methods than by estimation from PTPRC mRNA
207 alone, achieving mean correlation values of 0.76 and 0.67 respectively (Fig. 3C). Proteins that are

208 well correlated with mRNA abundance such as CD20 and CD16 remain accurately predicted by ML
209 methods, with some also showing a slight improvement in correlation. Splitting the data by cell type
210 there were a small number of features that were better represented by mRNA than predictions (the
211 most in dendritic cells) but for the majority of features the predictions improved upon mRNA in all
212 cell types (Fig. 3D).

213 Only one protein showed reduced average prediction correlation with ML methods (CD301 reduced
214 by 0.025). A subset of 70 proteins were poorly estimated by both normalised RNA-proxy and ML
215 methods having prediction correlation <0.4 in both cases (Supp. Tab. 1). In some cases, these low
216 values are expected; secreted proteins such as galectin-9 are unlikely to be well-captured by CITE-seq
217 measurements targeting the cell-surface. Other weaker predictions may be due to low abundance or
218 low variability of the proteins in the training data (Supp. Tab. 3). Overall, our results show ML
219 imputation to either match or improve upon the naïve prediction correlation produced by
220 normalised mRNA expression.

221

222 **Prediction accuracy depends strongly upon transcriptional signatures of cell type**

223 Given that the proteins best predicted were often the most variable across the data set we
224 investigated if cell type, a likely prominent source of variation, was an influential factor in ML
225 inference. As an initial indicator of this we identified specific examples where the high correlation of
226 prediction was absent when looking within an individual cell type compared to across all cells (Fig.
227 4A); i.e. methods may be good at predicting high level differences in protein expression between cell
228 types but poor at quantifying the finer resolution differences in expression between very similar
229 cells. Indeed, when we compared the prediction correlation as measured across all cells to that
230 within individual cell types across all proteins, we noted a pronounced and consistent drop in
231 accuracy (Fig. 4B).

232 To interrogate this further, we ran protein imputation on two variations of the Hao et al. data (Fig.
233 4C). In the first, each cell type was isolated and used for both training and test partitions as before.

234 This was intended to explore changes in performance with reduced phenotypic variability in the
235 training data. In the second variation, one cell type was withheld as the test set and the prediction
236 models were trained on the remaining cells. This tested the methods ability to infer protein
237 expression on unseen cell types. We refer to the two training variations as within-cell type and held-
238 out cell type respectively. We compared them to the all-cell case, where the entire data was used in
239 both testing and training (Fig 4D).

240 In the within-cell type variant we saw mostly similar prediction correlations as in the all-cells case,
241 with only minor reductions in median prediction correlation for most methods. However, scMMT
242 showed pronounced sensitivity to the restriction of cell type information, producing substantially
243 lower prediction correlation in NK cells, CD8 T cells, B cells and dendritic cells. In the held-out
244 variant, all methods exhibit uniform decline in prediction correlation. When imputing protein on
245 unseen cell types, no methods achieve median prediction correlation exceeding 0.25 across all
246 features. Taken together, our results demonstrate a pronounced sensitivity of reference-based
247 prediction to the composition of the training data.

248 To further investigate the cell type dependence of predictions we explored the features determining
249 model predictions. Given the black-box nature of most of the methods, we focused on scLinear
250 which provides measures of feature importance for its model. We plotted the top 10 most important
251 features predicting a selection of cell type marker proteins and noted distinct mRNAs were important
252 for the prediction of each (Fig. 4E). Some of these were correlated cell type markers that might be
253 expected. For instance, expression of the B-cell marker CD19 was primarily determined by expression
254 of other canonical B-cells markers such as MS4A1 and CD79. The T-cell markers CD3, CD4 and CD8
255 were strongly informed by their corresponding mRNAs and other genes predominantly expressed in
256 T-cells. However, we also observed a small selection of non-obvious genes with high importance for
257 protein imputation but also broad, cross-phenotype, expression. This suggests potential interactions
258 with other genes in the model adding to predictive information. We also observed feature
259 importances aligning with expected biology. For example, protein expression of the tissue residency
260 marker CD103 is informed by a collection of mRNA features with related function: the genes EVL and

261 ITGA4 both have established relevance to immune tissue infiltration (27, 28), and CCL5 is understood
262 to recruit peripheral immune cells to inflammatory sites (29, 30).

263 As a final test of the ability to predict on unseen cell/tissue types, we tested protein prediction using
264 the models trained on PBMC data on scRNAseq data from a non-small cell lung cancer CITEseq data
265 set(31). We found that overall prediction correlation was substantially lower for all reference-based
266 methods, but SPECK maintained a stable but lower median value (Fig. 4F). These results may indicate
267 overfitting of the models to cell type characteristics in the PBMC training data, producing poor
268 accuracy when tested on data of a different biological context.

269

270

271

272

273

274

275

276 **Discussion**

277

278 Our exploration of CITE-seq data confirms that RNA does not always provide an accurate proxy
279 measure for extracellular protein. Many post-transcriptional and post-translational mechanisms
280 could explain the mismatch between intracellular RNA, and protein at the cell surface. Alternative
281 splicing, intracellular trafficking apparatus, protein-protein interaction and decay/secretion rates all
282 influence the quantity of protein maintained on the cell surface (32). This complex chain of events
283 between RNA and protein measurements makes it difficult to determine how any single protein
284 relates to its corresponding RNA. One can instead use an aggregate measure over a population of
285 interest. We found that RNA positivity and pseudobulk expression both produce very similar

286 outcomes, and often provide estimates well-correlated with median protein expression in the target
287 population. However, these summary metrics remove the single-cell granularity of data and will still
288 perform poorly where there is a biological disconnect between mRNA and protein, as with protein
289 isoforms. Our comparison of ML methods suggests that whole transcriptome-based prediction
290 provides a viable alternative to protein estimation and can recover obscured RNA-protein
291 relationships through an understanding of cellular phenotype. Caveats remain regarding the
292 selection of suitable training data and poor performance on a fraction of protein features.

293 The 8 tested ML prediction methods all largely produced similar results, leaving little room for
294 discrimination by accuracy. The exception is SPECK, the only method not to utilise a protein
295 reference. SPECK produced relatively poor prediction correlations in all tests, further emphasizing
296 the limitations of inference directly from sparse RNA data where the target transcript may be
297 completely undetected. The remaining 7 approaches are supervised methods and use a multi-modal
298 protein/mRNA training reference. All produced predictions well-correlated with measurements, with
299 values consistently better than mRNA alone. Regrettably all 7 methods also appear to suffer from
300 overfitting to the PBMC reference, showing reduced performance when tested on tissue data unseen
301 in training. The main differentiating factors between methods were runtime, resource usage, and
302 sensitivity to scale and composition of training data. scLinear and Seurat both maintained
303 performance stable to change in size of the training data, and both methods complete quickly
304 without high computational resource requirements. Of the available options, we recommend these
305 methods as accessible and capable strategies for protein imputation.

306 In our analyses prediction correlations varied strongly across proteins, with individual features having
307 uniformly strong or weak correlations over methods. This suggested some inherent qualities of
308 certain proteins could cause to them be challenging to accurately impute. We found that ML-
309 prediction is not a one-size-fits-all solution for any target. Prediction methods tend to maintain good
310 performance on cell-type defining features already well-approximated by RNA expression, such as
311 CD3 and CD4. They can produce modest improvement in prediction correlation in many other cases
312 with poor mRNA correlation and can rescue intractable features for which RNA is entirely

313 undetected. However, some key immunoregulatory proteins remain poorly represented by both
314 cognate mRNA and prediction methods. The checkpoint receptors BTLA (CD272), TIGIT and CTLA4 all
315 failed to achieve strong average prediction correlation (>0.7) using ML methods but did pass that
316 threshold by RNA aggregation metrics (positivity, pseudobulk) (Supp. Table 1). In such cases we find
317 no reliable, single-cell resolution, estimation strategy. One is forced either to use aggregate metrics
318 or measure the protein expression directly using a method such as CITE-seq.

319 We saw reduced performance when training and testing on different cell type subsets of the PBMC
320 data, and when applying PBMC trained models to predict expression on unseen lung/tumour tissue.
321 This implicated dataset-specific cell type transcriptional signatures as strong determinants of model
322 performance, rather than nuanced learning of consistently relevant translational regulators. Feature
323 importance in scLinear prediction supported this, revealing the genes with most predictive power to
324 have highly differentiated patterns of expression by cell-type. We also observed well-predicted
325 proteins to be those with high counts and variability in the data, generally having a strongly
326 discriminated pattern of expression across cell types. These outcomes suggest a need for training
327 and query data to be of similar phenotypic composition and with similar characteristics of gene
328 expression, to avoid mismatch and misinterpretation of phenotype signatures. This may limit
329 applicability of these methods where well-defined CITE-seq data are not available. Even in such
330 instances it may be possible to design efficient, cost-effective hybrid studies. One could apply CITE-
331 seq to a small subset of samples, and then use ML-imputation to infer protein on higher-throughput,
332 highly multiplexed RNA data that can be generated more affordably (33-35).

333 Accurate quantification by CITE-seq can be challenging in the absence of high quality, validated,
334 antibodies. In such instances, alternate technologies such as mass spectrometry could provide an
335 alternative solution to measure protein without reliance on antibodies. Recent advancements in
336 sample preparation and equipment sensitivity have enabled sensitive single-cell mass spectrometry
337 albeit at a much lower throughput in terms of cell count (36). Going further, a range of new single-
338 cell technologies expanding upon CITE-seq aim to profile intracellular protein and post-
339 transcriptional features, offering new ways to map the pathways between transcription and

340 extracellular protein(37-39). This increasing range of experimental and bioinformatic tools to study
341 protein expression will further reduce the reliance on transcriptomic data as a proxy and could
342 similarly be considered for training models for transfer learning.

343 **Conclusions**

344 Our study highlights the inadequacy of mRNA abundance as a proxy of protein expression in single-
345 cell data. We have evaluated the utility of transcriptome-wide ML based protein prediction as an
346 alternative approach and define the limitations of these methods. Accurate estimation of cell surface
347 protein expression is of great importance for the definition of cell types, study of disease pathology,
348 and identification of therapeutic targets. Thus, adoption of models to more accurately estimate cell
349 surface protein expression could prove valuable to improve the utility of large existing repositories of
350 scRNA-seq data. Focus should be given to identifying and generating suitable reference data sets to
351 learn from and the development of more advanced, globally applicable, foundational models
352 associating cell transcriptomes with cell surface protein measures.

353 **Methods**

354 **Data and preprocessing**

355 We explored performance of protein imputation methods on two CITE-seq datasets. The first is the
356 Hao et al. PBMC data
357 (GSE164378[https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE164378]). This dataset
358 describes blood samples obtained from eight volunteers enrolled in an HIV vaccine trial, taken at
359 baseline and at days 3 and 7 following vaccine administration. Whole transcriptome RNA profiles are
360 measured on 161,764 cells. A panel of 228 antibodies was used to profile surface proteins on all
361 cells. This data was downloaded as a Seurat object provided by the New York Genome Center
362 [https://atlas.fredhutch.org/nygc/multimodal-pbmc/], used with no further processing or cell
363 filtering. We retained the log-normalized RNA counts and CLR normalized protein counts. The second
364 CITE-seq dataset used is the non-small cell lung cancer data presented by Leader et al. RNA counts
365 are provided under accession

366 GSE154826[https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE154826], and protein data is
367 available via the github repository [https://github.com/effiken/Leader_et_al]. The data contains RNA
368 profiles for 361,929 cells, taken from tumour and non-involved human lung tissue. Of these, protein
369 was measured on a subset of 230,820 cells. For our analyses we used protein data for 82 antibodies,
370 taken from distinct panels measured on different patient subsets. For any protein-specific analyses
371 (correlation computation, etc.) we include only cells on which the protein was measured. To align
372 with our other analyses, we applied CLR normalization to the protein counts in this dataset.

373 **Aggregate RNA metrics**

374 In our evaluation of correspondence between RNA and protein, we explore the value of using
375 percent positivity and pseudocount estimators. Percent positivity is computed simply as the
376 proportion of cells in a group in which the transcript was detected. Pseudobulk counts were
377 computed using the Seurat function *AggregateExpression()*. RNA counts were summed over donor
378 and cell-type. The final pseudobulk value per cell-type was taken as the mean value over donors.

379 **Protein Imputation**

380 We ran 8 protein prediction methods using a custom Nextflow pipeline [link to code repository when
381 available] (40) . Code for model training and prediction was sourced from author-provided github
382 repositories or vignettes for each method. Arguments controlling prediction strategy were set to
383 default according to vignettes or example code. Prior to prediction the data was partitioned into a
384 training and test set, taking an 80:20 split, selecting cells at random. Predictions methods were
385 trained and tested on the same partitions. Predictions made on the test data were evaluated against
386 the measured protein by computation of Pearson correlation and RMSE. All imputation methods
387 were left to apply built-in normalization as defined in the published software, and were provided
388 unnormalized RNA counts as input.

389 **Evaluating Prediction Correlation**

390 When evaluating baseline correlation between mRNA and protein in the Hao et al. data, we first had
391 to determine suitable alignment between RNA and protein features. Where antibodies had no

392 sensible, direct mRNA equivalent (isotype controls, TCR binding antibodies) they were dropped from
393 correlation analysis. Where correlation was not possible to evaluate due to high sparsity and RNA
394 dropout, we set prediction value to zero to facilitate comparison with other methods. RNA-protein
395 correlations were computed only on cells from the test set, for parity with ML-prediction
396 correlations. Where prediction correlation is evaluated over a set of features, we take the median
397 correlation value. Where prediction correlation is evaluated over ML methods we take the mean
398 correlation value.

399

400 **Prediction Benchmarking**

401 To assess comparative performance of prediction methods Nextflow configuration was used to
402 ensure each prediction process had access to identical memory and CPU resources, fixed at 240GB
403 and 32 cores. The trace and execution reports generated by Nextflow were used to report runtime
404 and resource usage for each method.

405 To explore how prediction results change with differences in training data, we ran the methods under
406 three separate variations.

407

408 1. The data was subsampled into 6.125%, 12.5%, 25%, 50% and 100% fractions. Following this, it was
409 split into a training and test set in an 80:20 ratio. Prediction methods were then trained and tested.
410 This approach was intended to measure differences in performance as the size of the training sample
411 increases.

412 2. For each cell type group in the set B, CD4 T, CD8 T, NK, Mono and DC the selected type was entirely
413 held out as the test set. Methods were trained on the remaining cells, and prediction correlation
414 measured on the held-out cells. This approach is intended to assess ability to accurate predict
415 protein on unseen cell types.

416 3. For each cell type group in the set B, CD4 T, CD8 T, NK, Mono and DC the selected type was taken
417 as the entire data and split into a training and test set in an 80:20 ratio. All other cells were

418 discounted. Prediction methods were trained and tested on only one cell type at a time. This method
419 was intended to measure capacity to detect within-cell type variation in protein expression.

420 **Feature Importance in scLinear**

421 While not possible for all prediction methods, we were able to extract a measure of feature
422 importance from the linear models generated by scLinear. As described by the package authors,
423 feature importance is calculated as the Jacobian matrix of the predicted ADT values, with respect to
424 the input RNA matrix. This is computed as the product of the Jacobians for the singular value
425 decomposition, linear regression and z-score normalisation components of the model. Further
426 details of computation are given by Hanhart et al. (16). Genes with large, positive importance value
427 are understood to be influential over prediction outcomes.

428 **Usability Scoring**

429 We measured usability with an adapted version of a framework previously applied to benchmark
430 methods for single-cell integration and pseudotime trajectory inference (24, 25). Methods were
431 evaluated according to 11 criteria (Supp. Table 2) which measure both how accessible the software is
432 to use, and how thoroughly the publication evaluates prediction accuracy against other approaches.
433 The first two criteria are simply that the method be available on a version-controlled repository, and
434 that the underlying code is freely available and executable on an open-source platform. Four
435 conditions are used to evaluate the quality of available software tutorials. These are that a tutorial
436 be freely available, that it shows the imputation method in different scenarios, that it shows how to
437 run the method in a non-native language, and that running the tutorial produces no notable errors.
438 The documentation of function inputs, outputs and intended use is also scored. Finally, a set of four
439 criteria evaluate the publication. The first is simply that the method is published in a peer reviewed
440 journal. The others assess the number of datasets on which the method is tested, the diversity of
441 robustness assessments, and breadth of comparison with other protein imputation strategies.

442 **Declarations**

443 <u>Availability of data and materials</u>

444 Two published CITE-seq datasets were used to evaluate imputation method performance. The first is
445 the Hao et al. PBMC data
446 (GSE164378[https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE164378]). The second is non-
447 small cell lung cancer data presented by Leader et al. RNA counts are provided under accession
448 GSE154826[https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE154826], and protein data is
449 available via the github repository [https://github.com/effiken/Leader_et_al]. The Nextflow pipeline
450 and accompanying scripts used for method execution and evaluation are available in the associated
451 github repository [https://github.com/fisherj2/protein_imputation].

452

453 Contributions

454 J.F. and M.JW were responsible for design of analyses and manuscript preparation. J.F. conducted
455 computational work and figure compilation. S.B., O.W., L.M. and L.L. contributed to data
456 interpretation and manuscript revision.

457

458 Acknowledgements

459 Christopher Paluch

460

461 Conflicts and Funding

462 All authors are employees of Gilead Sciences.

463

464 **Figure 1. RNA-protein correlation in the Hao et al. CITE-seq data** A) UMAP showing the immune cell
465 types profiled in the Hao et al. PBMC CITE-seq data. B) CD103 expression measured by normalised
466 protein and normalised RNA. C) Density plots showing Pearson correlation of RNA-protein pairs
467 measured by normalised RNA, proportion of RNA positive cells, and by RNA pseudobulk count.

468 Features binned into correlation brackets according to correlation between normalised mRNA
469 expression and measured protein.

470 **Figure 2. Prediction performance and method comparison** A) Heatmap showing Pearson correlation
471 between predicted and measured protein expression values**.** B) Density plots showing distribution of
472 Pearson correlation and RMSE for each method. C) Scatter plots showing average prediction
473 correlation against total protein counts and standard deviation in the training data. D) Overview bar
474 plots providing method summary, showing usability scores for the software and publication (1 being
475 the highest), overall prediction correlation, RAM usage and runtime.

476 **Figure 3. Improvement in prediction correlation using ML methods** A) Histogram of difference in
477 correlation coefficient between mean prediction correlation and gene mRNA correlation. B) UMAP
478 panels highlighting platelet population and showing MPL (CD110) RNA expression against protein
479 expression and scLinear predicted values. C) Gene mRNA correlation plotted against mean prediction
480 correlation across methods, coloured by correlation categories defined in Fig. 1C. Features with large
481 correlation improvement are annotated. D) Correlation scatter as in C), split by cell type.

482 **Figure 4. Sensitivity of prediction performance to training data composition.** A) Scatter plot
483 showing correlation of scMMT prediction with measurements, computed over all cell and over cell-
484 subtypes. B) Boxplot showing mean prediction correlation over methods, compared against gene
485 mRNA correlation. Correlation values measured over all cells, and for each cell type individually. C)
486 Schematic of training and test sets in investigation of cell type sensitivity. D) Panels showing bar plot
487 of overall prediction correlation across all cells, alongside median prediction correlation in each cell-
488 type individually (All Cell Types), when only one cell type is used for both testing and training (Within
489 Cell Type), and when a cell type is held out as the test set (Held-out Cell Type). E) Heatmap of top 10
490 most important RNA features in the scLinear prediction model, shown for a selection of proteins. F)
491 Boxplot comparing prediction correlation in Hao et al. data against prediction correlation with
492 methods trained on Hao et al and tested on Leader et al.

493 **Supplementary Figure 1. The correlation between protein and RNA measurements.** Bar plot
494 showing Pearson correlation between normalised mRNA and measured extracellular protein, for 205
495 features. Bar coloured according to value.

496 **Supplementary Figure 2. Comparison of prediction comparison by different approaches.** Heatmap
497 of prediction correlation values by estimation approach, with mean value taken over ML methods.

498 **Supplementary Figure 3. Evaluation of predictions methods by usability scoring.** Method usability
499 metrics broken down by category, with 1 as the highest score and 0 as the lowest.

500 **Supplementary Figure 4. Performance and resource usage of prediction methods trained on**
501 **ascending fractions of data.** A) Line plots showing median prediction correlation and RMSE for each
502 method, as the data size is varied over subsamples. B) Boxplots showing resource usage and runtime
503 for each method, over subsamples of the data.

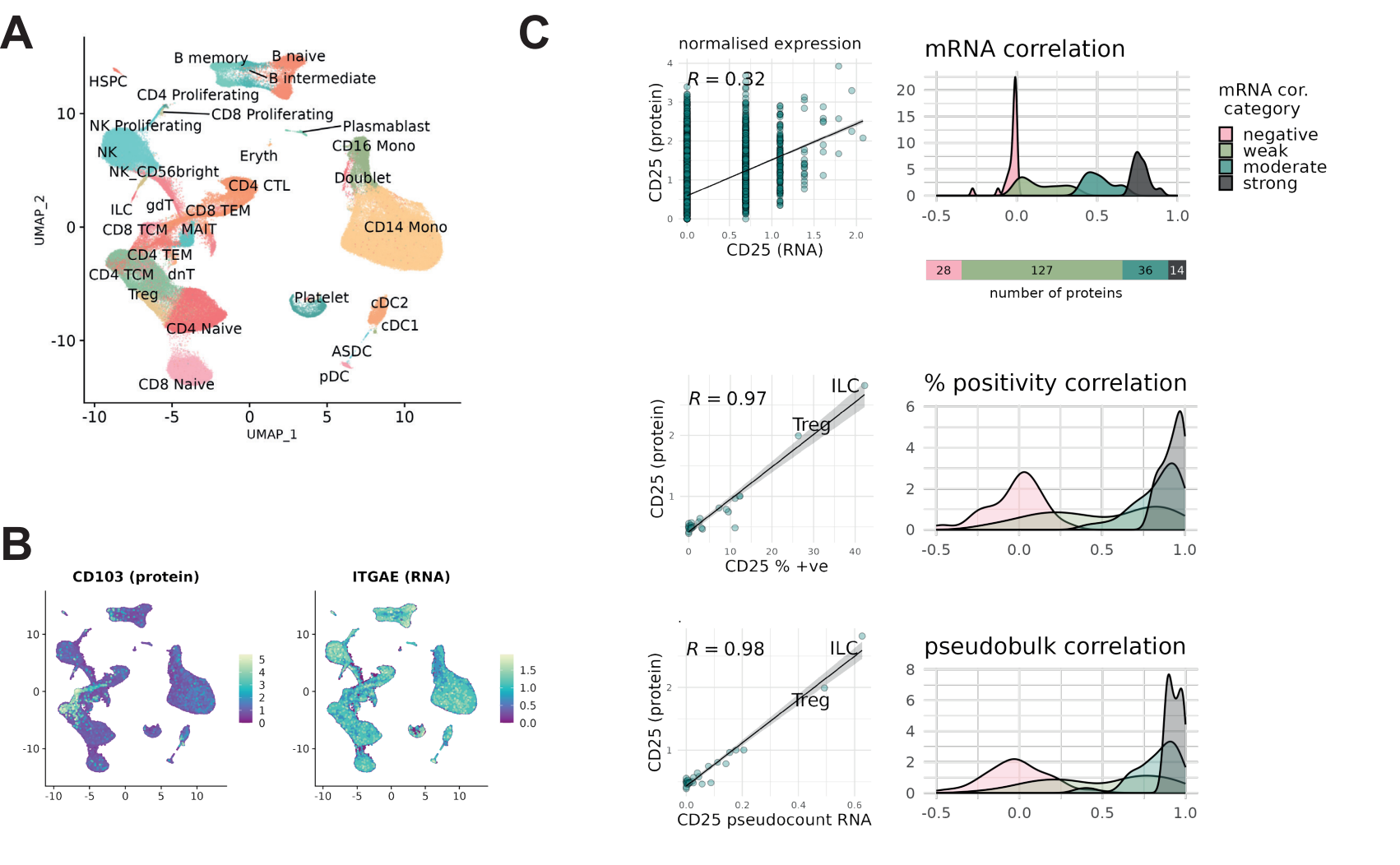
504

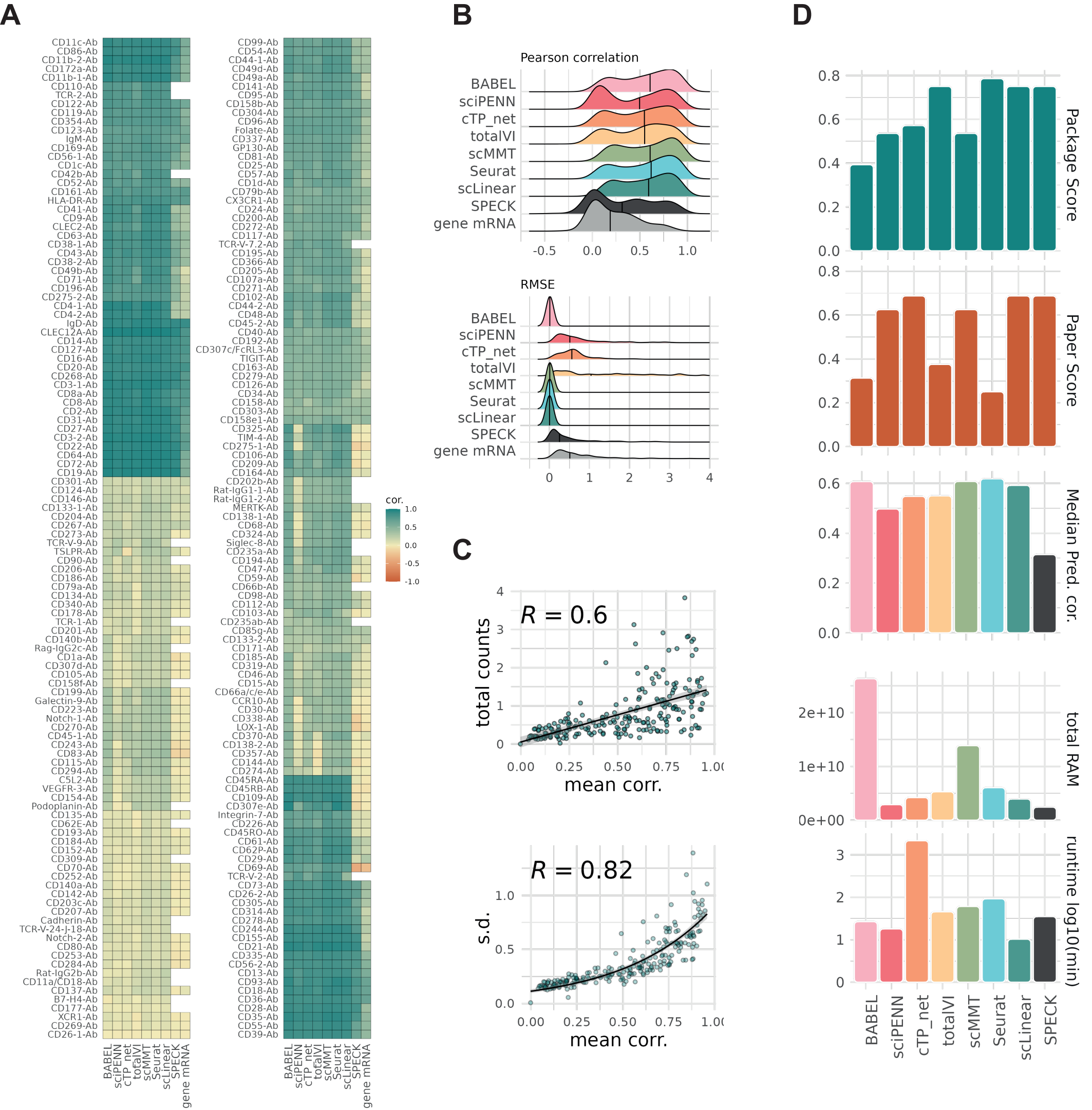
505 1. Ballmaier M, Holter W, Germeshausen M. Flow cytometric detection of MPL (CD110) as a
506 diagnostic tool for differentiation of congenital thrombocytopenias. Haematologica.
507 2015;100(9):e341-4.
508 2. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H,
509 et al. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods.
510 2017;14(9):865-8.
511 3. Rood JE, Wynne S, Robson L, Hupalowska A, Randell J, Teichmann SA, et al. The Human Cell
512 Atlas from a cell census to a unified foundation model. Nature. 2024.
513 4. Sikkema L, Ramirez-Suastegui C, Strobl DC, Gillett TE, Zappia L, Madissoon E, et al. An
514 integrated cell atlas of the lung in health and disease. Nat Med. 2023;29(6):1563-77.
515 5. Wilk AJ, Rustagi A, Zhao NQ, Roque J, Martinez-Colon GJ, McKechnie JL, et al. A single-cell
516 atlas of the peripheral immune response in patients with severe COVID-19. Nat Med.
517 2020;26(7):1070-6.
518 6. Jin K, Li B, Yan H, Zhang XF. Imputing dropouts for single-cell RNA sequencing based on multi-
519 objective optimization. Bioinformatics. 2022;38(12):3222-30.
520 7. Mimitou EP, Lareau CA, Chen KY, Zorzetto-Fernandes AL, Hao Y, Takeshima Y, et al. Scalable,
521 multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. Nat
522 Biotechnol. 2021;39(10):1246-58.
523 8. Peterson VM, Zhang KX, Kumar N, Wong J, Li L, Wilson DC, et al. Multiplexed quantification of
524 proteins and transcripts in single cells. Nat Biotechnol. 2017;35(10):936-9.
525 9. Shahi P, Kim SC, Haliburton JR, Gartner ZJ, Abate AR. Abseq: Ultrahigh-throughput single cell
526 protein profiling with droplet microfluidic barcoding. Sci Rep. 2017;7:44447.
527 10. Swanson E, Lord C, Reading J, Heubeck AT, Genge PC, Thomson Z, et al. Simultaneous
528 trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq.
529 Elife. 2021;10.
530 11. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, 3rd, Zheng S, Butler A, et al. Integrated
531 analysis of multimodal single-cell data. Cell. 2021;184(13):3573-87 e29.

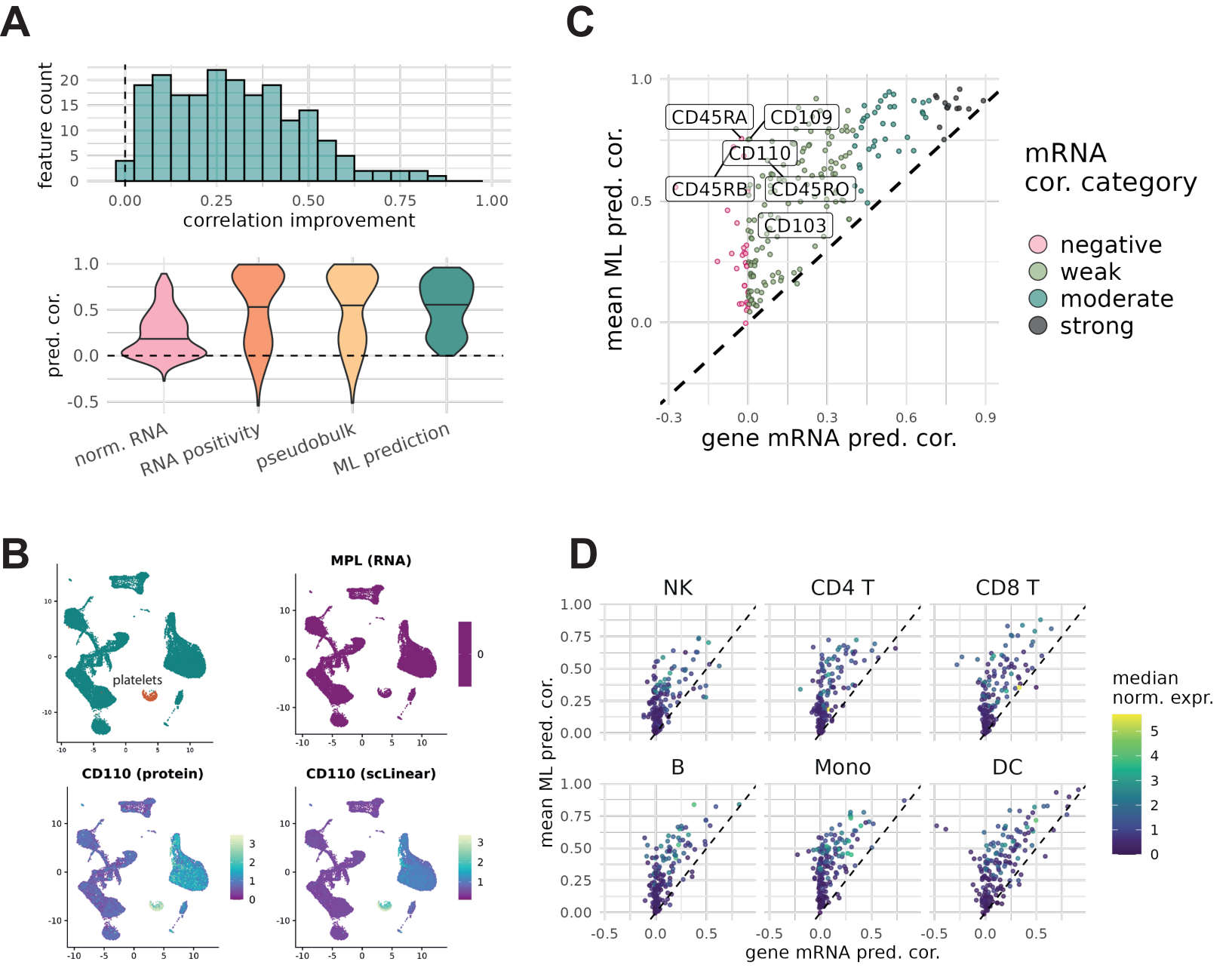
532 12. Gayoso A, Steier Z, Lopez R, Regier J, Nazor KL, Streets A, et al. Joint probabilistic modeling of
533 single-cell multi-omic data with totalVI. Nat Methods. 2021;18(3):272-82.
534 13. Lakkis J, Schroeder A, Su K, Lee MYY, Bashore AC, Reilly MP, et al. A multi-use deep learning
535 method for CITE-seq and single-cell RNA-seq data integration with cell surface protein prediction and
536 imputation. Nat Mach Intell. 2022;4(11):940-52.
537 14. Wu KE, Yost KE, Chang HY, Zou J. BABEL enables cross-modality translation between
538 multiomic profiles at single-cell resolution. Proc Natl Acad Sci U S A. 2021;118(15).
539 15. Zhou Z, Ye C, Wang J, Zhang NR. Surface protein imputation from single cell transcriptomes
540 by deep neural networks. Nat Commun. 2020;11(1):651.
541 16. Hanhart D, Gossi F, Rapsomaniki MA, Kruithof-de Julio M, Chouvardas P. ScLinear predicts
542 protein abundance at single-cell resolution. Commun Biol. 2024;7(1):267.
543 17. Javaid A, Frost HR. SPECK: an unsupervised learning approach for cell surface receptor
544 abundance estimation for single-cell RNA-sequencing data. Bioinform Adv. 2023;3(1):vbad073.
545 18. Vandenbon A, Diez D. A universal tool for predicting differentially active features in single-cell
546 and spatial genomics data. Sci Rep. 2023;13(1):11830.
547 19. He Z, Hu S, Chen Y, An S, Zhou J, Liu R, et al. Mosaic integration and knowledge transfer of
548 single-cell multimodal data with MIDAS. Nat Biotechnol. 2024.
549 20. Marongiu L, Protti G, Facchini FA, Valache M, Mingozzi F, Ranzani V, et al. Maturation
550 signatures of conventional dendritic cell subtypes in COVID-19 suggest direct viral sensing. Eur J
551 Immunol. 2022;52(1):109-22.
552 21. Liu Y, Chen Y, Lu H, Zhong W, Yuan GC, Ma P. Orthogonal multimodality integration and
553 clustering in single-cell data. BMC Bioinformatics. 2024;25(1):164.
554 22. Andreatta M, Berenstein AJ, Carmona SJ. scGate: marker-based purification of cell types from
555 heterogeneous single-cell RNA-seq datasets. Bioinformatics. 2022;38(9):2642-4.
556 23. Westoby J, Artemov P, Hemberg M, Ferguson-Smith A. Obstacles to detecting isoforms using
557 full-length scRNA-seq data. Genome Biol. 2020;21(1):74.
558 24. Luecken MD, Buttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, et al.
559 Benchmarking atlas-level data integration in single-cell genomics. Nat Methods. 2022;19(1):41-50.
560 25. Saelens W, Cannoodt R, Todorov H, Saeys Y. A comparison of single-cell trajectory inference
561 methods. Nat Biotechnol. 2019;37(5):547-54.
562 26. Rowley JW, Schwertz H, Weyrich AS. Platelet mRNA: the meaning behind the message. Curr
563 Opin Hematol. 2012;19(5):385-91.
564 27. Du Y, Zhu L, Li X, Shi H, Jiang W, Zhou W. Immunocyte Infiltration Analysis and
565 Immunohistochemistry Identify EVL as a Potential Prognostic Biomarker for Pancreatic Cancer. J Pers
566 Med. 2023;13(3).
567 28. Mo J, Zhang J, Huang H, Liu C, Cheng Y, Mo Y, et al. The early predictive effect of low
568 expression of the ITGA4 in colorectal cancer. J Gastrointest Oncol. 2022;13(1):265-78.
569 29. Marques RE, Guabiraba R, Russo RC, Teixeira MM. Targeting CCL5 in inflammation. Expert
570 Opin Ther Targets. 2013;17(12):1439-60.
571 30. Schall TJ, Bacon K, Toy KJ, Goeddel DV. Selective attraction of monocytes and T lymphocytes
572 of the memory phenotype by cytokine RANTES. Nature. 1990;347(6294):669-71.
573 31. Leader AM, Grout JA, Maier BB, Nabet BY, Park MD, Tabachnikova A, et al. Single-cell analysis
574 of human non-small cell lung cancer lesions refines tumor classification and patient stratification.
575 Cancer Cell. 2021;39(12):1594-609 e12.
576 32. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and
577 transcriptomic analyses. Nat Rev Genet. 2012;13(4):227-32.
578 33. Rosenberg AB, Roco CM, Muscat RA, Kuchina A, Sample P, Yao Z, et al. Single-cell profiling of
579 the developing mouse brain and spinal cord with split-pool barcoding. Science. 2018;360(6385):176-
580 82.
581 34. Xie Y, Chen H, Chellamuthu VR, Lajam ABM, Albani S, Low AHL, et al. Comparative Analysis of
582 Single-Cell RNA Sequencing Methods with and without Sample Multiplexing. Int J Mol Sci.
583 2024;25(7).
584 35. Tran V, Papalexi E, Schroeder S, Kim G, Sapre A, Pangallo J, et al. High sensitivity single cell
585 RNA sequencing with split pool barcoding. bioRxiv. 2022:2022.08.27.505512.

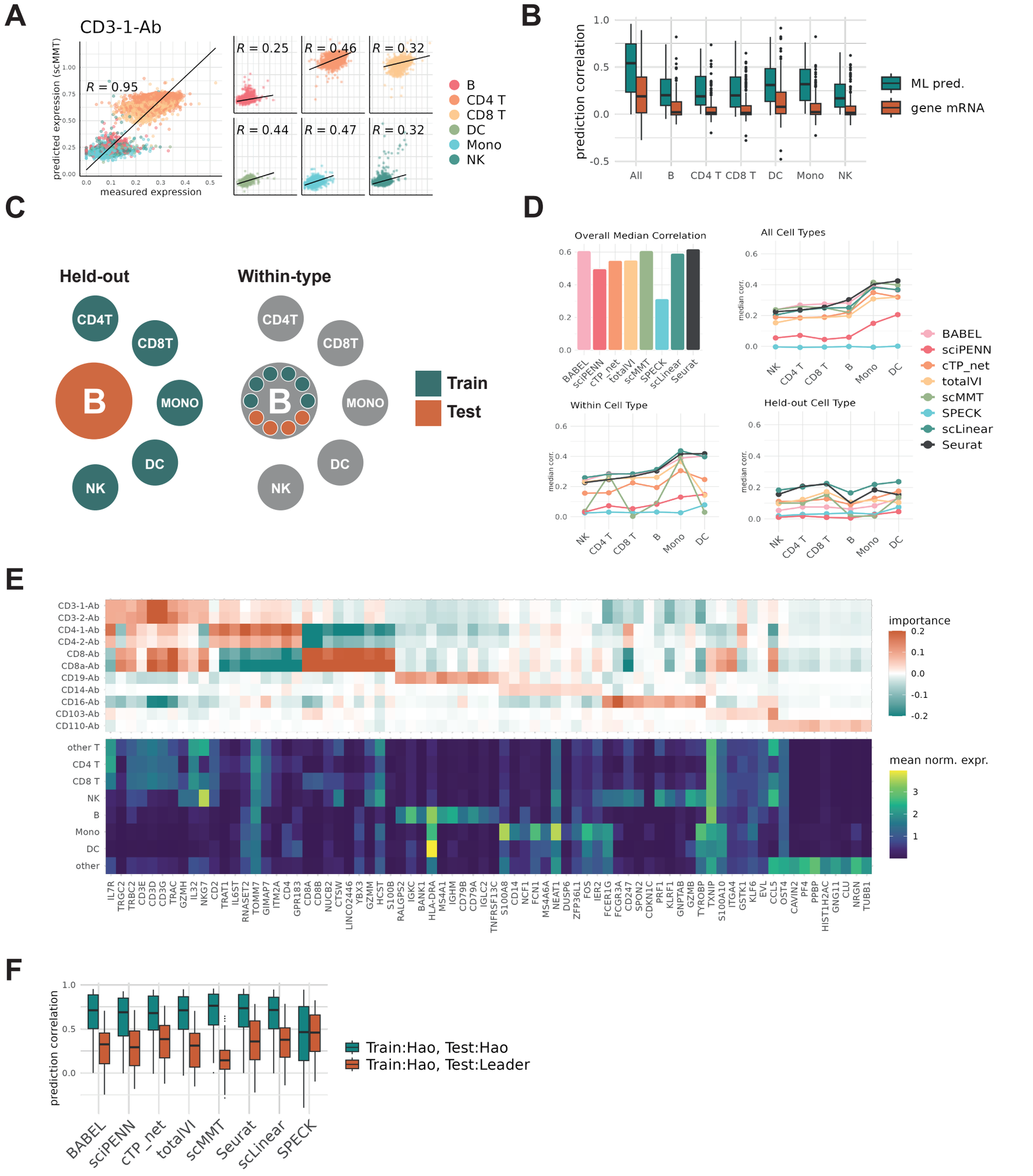
586 36. Bennett HM, Stephenson W, Rose CM, Darmanis S. Single-cell proteomics enabled by next-
587 generation sequencing or mass spectrometry. Nat Methods. 2023;20(3):363-74.
588 37. Chen AF, Parks B, Kathiria AS, Ober-Reynolds B, Goronzy JJ, Greenleaf WJ. NEAT-seq:
589 simultaneous profiling of intra-nuclear proteins, chromatin accessibility and gene expression in single
590 cells. Nat Methods. 2022;19(5):547-53.
591 38. Rivello F, van Buijtenen E, Matula K, van Buggenum J, Vink P, van Eenennaam H, et al. Single-
592 cell intracellular epitope and transcript detection reveals signal transduction dynamics. Cell Rep
593 Methods. 2021;1(5):100070.
594 39. Chung H, Parkhurst CN, Magee EM, Phillips D, Habibi E, Chen F, et al. Joint single-cell
595 measurements of nuclear proteins and RNA in vivo. Nat Methods. 2021;18(10):1204-12.
596 40. Di Tommaso P, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables
597 reproducible computational workflows. Nat Biotechnol. 2017;35(4):316-9.

598


A
B memory
B naive
B intermediate
HSPC
CD4 Proliferating
CD8 Proliferating
NK Proliferating
Plasmablast
NK
Eryth
CD16 Mono
NK_CD56bright
Doublet
CD4 CTL
ILC
gdT
CD8 TEM
CD8 TCM
MAIT
CD14 Mono
CD4 TEM
CD4 TCM
dnT
Treg
Platelet
cDC2
cDC1
CD4 Naive
ASDC
pDC
CD8 Naive
UMAP_2
UMAP_1
B
CD103 (protein)
ITGAE (RNA)
C
normalised expression
R = 0.32
CD25 (protein)
CD25 (RNA)
mRNA correlation
mRNA cor. category
negative
weak
moderate
strong
28
127
36
14
number of proteins
R = 0.97
ILC
Treg
CD25 % +ve
% positivity correlation
R = 0.98
CD25 pseudocount RNA
pseudobulk correlation


A
B
C
D
Pearson correlation
RMSE
BABEL
sciPENN
cTP_net
totalVI
scMMT
Seurat
scLinear
SPECK
gene mRNA
cor.
R = 0.6
total counts
mean corr.
R = 0.82
s.d.
Package Score
Paper Score
Median Pred. cor.
total RAM
runtime log10(min)


A
feature count
correlation improvement
pred. cor.
norm. RNA
RNA positivity
pseudobulk
ML prediction
C
mean ML pred. cor.
gene mRNA pred. cor.
CD45RA
CD109
CD110
CD45RB
CD45RO
CD103
mRNA
cor. category
negative
weak
moderate
strong
B
platelets
MPL (RNA)
CD110 (protein)
CD110 (scLinear)
D
NK
CD4 T
CD8 T
B
Mono
DC
mean ML pred. cor.
gene mRNA pred. cor.
median
norm. expr.


A
CD3-1-Ab
R = 0.95
R = 0.25
R = 0.46
R = 0.32
R = 0.44
R = 0.47
R = 0.32
predicted expression (scMMT)
measured expression
B
CD4 T
CD8 T
DC
Mono
NK
B
prediction correlation
All
B
CD4 T
CD8 T
DC
Mono
NK
ML pred.
gene mRNA
C
Held-out
Within-type
CD4T
CD8T
B
MONO
DC
NK
Train
Test
D
Overall Median Correlation
All Cell Types
Within Cell Type
Held-out Cell Type
median corr.
BABEL
sciPENN
cTP_net
totalVI
scMMT
SPECK
scLinear
Seurat
E
CD3-1-Ab
CD3-2-Ab
CD4-1-Ab
CD4-2-Ab
CD8-Ab
CD8a-Ab
CD19-Ab
CD14-Ab
CD16-Ab
CD103-Ab
CD110-Ab
other T
CD4 T
CD8 T
NK
B
Mono
DC
other
importance
mean norm. expr.
F
prediction correlation
BABEL
sciPENN
cTP_net
totalVI
scMMT
Seurat
scLinear
SPECK
Train:Hao, Test:Hao
Train:Hao, Test:Leader