Universitat Oberta
de Catalunya

uoc.edu

Protein Abundance Prediction In Bulk And Single-Cell Transcriptomics

Antonio Rodríguez Romera
Master in Bioinformatics and Biostatistics

Area 3

Izaskun Mallona González

Laura Calvet Liñan

June 8th 2021

FINAL PROJECT'S CARD

Title:	*Protein Abundance Prediction In Bulk And Single-Cell Transcriptomics*
Author:	*Antonio Rodríguez Romera*
Consultant:	*Izaskun Mallona González*
PRA:	*Laura Calvet Liñan*
Submission date:	*06/2021*
Studies:	*Master in Bioinformatics and Biostatistics*
Area:	*Area 3*
Language:	*English*
Credits:	*15*
Key words:	*Single-cell, Transcriptomics, Proteomics*

Abstract (220 words):

Omic technologies are invaluable tools to study the biological organisation of organisms. In the field of transcriptomics, scientific advancements have enabled the development of extraordinarily sensitive techniques that can measure complete transcriptomes from single cells. On the other hand, several attempts have been made to measure single cell proteomes. However, most of them lack the coverage and the sensitivity of transcriptomic techniques.

Although transcriptomic tools are widely used, several studies have shown that RNA and proteins are not sufficiently correlated to act as proxies for each other.

In this thesis we have explored the use of the Protein/RNA ratios to improve RNA and protein correlations. Using a bulk proteogenomic dataset we expanded the work of previous authors and showed that this ratio can be used to impute protein levels from transcriptomic abundances in several human tissues. Importantly, this strategy was independent of the tissue composition and was also applicable for cell surface proteins.

Using recently published CITE-seq atlases we explored for the first time this approach in single-cell data. Our results showed that Protein/RNA ratios can better predict protein levels in single cell data when they are computed from CITE-seq datasets compared to bulk-data-calculated ratios. Interestingly, protein prediction performed well using correction factors computed from a different experiment, suggesting that this approach can be generalised to other single cell datasets.

Resumen del trabajo (242 palabras):

Las tecnologías ómicas son una herramienta indispensable para el estudio de la organización biológica de los organismos. Avances científicos en transcriptómica han permitido el desarrollo de técnicas extraordinariamente sensibles que miden transcriptomas completos en células individuales. Por otro lado, aunque se han realizado varios intentos para medir el proteoma de una sola célula la mayoría de ellos carecen de la sensibilidad o cobertura de las técnicas transcriptómicas.

Aunque las técnicas transcriptómicas son ampliamente usadas, varios estudios han demostrado que ARN y proteínas no están lo suficientemente correlacionados como para actuar uno en representación del otro.

En este trabajo hemos explorado el uso de las ratios proteína/RNA para mejorar la correlación entre estas medidas. Usando un atlas proteo-genómico hemos expandido el trabajo de otros autores y demostrado que esta ratio se puede usar para imputar niveles de proteína a partir del transcriptoma en varios tejidos humanos. Además, esta corrección es independiente de la composición del tejido y es aplicable a proteínas de superficie.

Utilizando atlas de CITE-seq hemos explorado por primera vez esta estrategia en conjuntos de datos *single-cell*. Nuestros resultados muestran que las ratios RNA/proteína predicen mejor los niveles de proteína en datos *single-cell* cuando se han estimado a partir de datos CITE-seq en comparación con ratios estimados con datos *bulk*. Además, fue posible predecir niveles de proteína usando ratios calculados a partir de un experimento distinto, lo que sugiere que esta estrategia se puede generalizar a otros conjuntos de datos *single-cell*.

Table of Contents

1 Abstract 1

2 Introduction 1

2.1 Context and project justification 1
2.2 Project goals 5
2.3 Methodology and approach 6
2.4 Workplan 7
2.5 Brief description of results and products obtained. 10
2.6 Brief description of the rest of the project's chapters 11

3 State of the art 11

4 Methods 13

4.1 Data collection 13
4.1 Cell surface proteins selection in bulk datasets 13
4.2 Single Cell Data pre-processing 14
4.3 Dimensionality reduction 14
4.4 Principal components selection 15
4.5 Cell Clustering 15
4.6 Cell type annotation 15
4.7 Non-linear dimensionality reduction 15
4.8 Pseudo-bulk RNA counts 16
4.9 Computing and applying the Correction Factor 16

5 Results 17

5.1 Correction Factor in Bulk data 17
5.2 Correction factor in single cell data 22

6 Discussion 31

7 Conclusions 34

7.1 Conclusions 34
7.2 Future perspectives 34
7.3 Planning Compliance 35

8 Glossary 36

8.1 Definition of relevant terms used in the project 36
8.2 Other acronyms used in the project 36

9 Bibliography 37

10 Supplementary Materials I

10.1 Supplementary methods I
10.2 Supplementary figures II

List of Figures

Fig. 1. Grantt Chart 9

Fig. 2 Protein and RNA distributions across tissues. 17

Fig. 3 Principal Component Analysis plots 18

Fig. 4 Uniform Manifold Approximation and Projection plots......................... 19

Fig. 5 Study of RNA and Protein correlations.. 21

Fig. 6 PCA and JackStraw results 23

Fig. 7 tSNE and UMAP plots... 26

Fig. 8 Pseudo-bulk strategy results.. 27

Fig. 9 Study of CF from bulk data. 28

Fig. 10 Study of CF from single cell data 30

Fig. 11 Study of CF from single cell data across studies. 31

List of Tables

Table 1 RNA and ADT makers. 25

1 Abstract

Omic technologies are invaluable tools to study the biological organisation of organisms. In the field of transcriptomics, scientific advancements have enabled the development of extraordinarily sensitive techniques that can measure complete transcriptomes from single cells. On the other hand, several attempts have been made to measure single cell proteomes. However, most of them lack the coverage and the sensitivity of transcriptomic techniques.

Although transcriptomic tools are widely used, several studies have shown that RNA and proteins are not sufficiently correlated to act as proxies for each other.

In this thesis we have explored the use of the Protein/RNA ratios to improve RNA and protein correlations. Using a bulk proteogenomic dataset we expanded the work of previous authors and showed that this ratio can be used to impute protein levels from transcriptomic abundances in several human tissues. Importantly, this strategy was independent of the tissue composition and was also applicable for cell surface proteins.

Using recently published CITE-seq atlases we explored for the first time this approach in single-cell data. Our results showed that Protein/RNA ratios can better predict protein levels in single cell data when they are computed from CITE-seq datasets compared to bulk-data-calculated ratios. Interestingly, protein prediction performed well using correction factors computed from a different experiment, suggesting that this approach can be generalised to other single cell datasets.

2 Introduction

2.1 Context and project justification

Cells are the structural and functional building blocks of organisms. Inside biological systems, different cell types and molecules form complex regulatory networks that define the physiological state of the tissue they constitute [1]. Scientific advancements over the past decade have led to the development of high-throughput technologies that allow scientists to study this cellular interplay in detail [1].

2.1.1 Single Cell Transcriptomics and Proteomics

RNA sequencing (RNA-seq) is a high-throughput technique that reads and quantifies RNA transcripts, thereby providing a wide view of the transcriptional landscape in biological samples [2]. This technology has become an indispensable tool in molecular biology, especially for the study of differentially expressed genes in a broad range of organisms. However, traditional RNA-seq “sums-up” the unique transcriptomes of thousands of cells, masking the cellular and molecular complexity inherent to any biological sample [2], [3].

In this sense, modifications in the traditional RNA-seq protocols such as the introduction of pre-amplification steps and the establishment of multiplexing strategies like droplet encapsulation have enable the development of single cell RNA sequencing (scRNA-seq) [3]. This increasingly popular technique surveys the transcriptome at the single cell level, allowing scientist to resolve questions that were unattainable with bulk RNA-seq.

scRNA-seq can reveal the uniqueness of individual cells. As a result, this technique has been used to create comprehensive cellular atlases in mouse [4] and human [5]. Single cell transcriptomics also provides valuable insight for diagnosis and treatment of human diseases. This technique has been used to untangle the heterogeneity of the immune system, identifying new immune cell subtypes in health and disease [6]. Moreover, in cancer research scRNA-seq has proven useful to study intra- and inter-tumour variability, shedding light on the importance of different cell subsets in treatment response [7].

Unlike transcriptomics, proteomics does not currently have a state-of-the-art method that can be applied to single cells. Mass spectrometry, the main technique of proteomics, can detect and quantify near-complete proteomes, but such experiments require typically tens of thousands of cells [8]. Furthermore, with current multiplexing techniques such as tandem mass tags (TMT), only around 20 barcodes can be used [8].

Antibody-based approaches are being increasingly applied to protein profiling in single cells. Mass cytometry is a variation of classic flow cytometry that uses antibodies conjugated with heavy-metal ions tags. These ions can be subsequently detected by time-of-flight mass spectrometry, which allows profiling of about 100 different protein targets in single cells [8], [9].

Other antibody-based approaches such as Cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) and RNA expression and protein sequencing (REAP-seq) enable parallel quantification of RNA and surface proteins in single cells [10], [11]. Both methods rely on DNA barcoded antibodies to label and measure surface protein levels. After cells are incubated with the antibodies, the barcodes (antibody derived tags, ADTs) can be processed alongside RNA transcripts, producing a proteomic and transcriptomic readout that can be sequenced. Both CITE-seq and REAP-seq use a similar approach and they only differ in how the barcode is conjugated to the antibody [9].

In contrast to using fluorescence or heavy-metal ions, DNA barcodes have high multiplexing capacity, allowing scientist to measure over 200 surface proteins at the single cell level [12]. Noteworthy, CITE-seq and REAP-seq are restricted to cell surface proteins, which reduces the scope of proteins that can be surveyed with these techniques. Very recent efforts in the single cell community have expanded antibody based approaches to intracellular proteins [13], overcoming this limitation. However, CITE-seq and REAP-seq are more commonly used in the literature, as these are longstanding techniques.

2.1.2 Problem to address: current limitations and proposed approach

Findings obtained from scRNA-seq often need to be validated or further explored using wet lab techniques, which require the purification of the cell types identified from complex mixtures. In this regard, Fluorescence-activated cell sorting (FACS) is the method of choice to isolate millions of single cells based on multiple fluorescence parameters [14]. As a result, this purification strategy requires a priori knowledge of a cell-specific markers. In this sense, even though surface markers are already available for many cell types [15], new cell populations described by scRNA-seq analysis still require the identification of said markers using in silico approaches.

Laboratories often use differentially expressed genes that code for cell surface proteins as potential biomarkers for the cell populations they are enriched in [16]. This approach, although simple and widely applicable, assumes that gene expression and protein abundance are linearly correlated. However, several studies have measured RNA along with Protein abundances reporting relative low correlations between them in both bulk and single cell samples [13], [17]. This phenomenon suggests that RNA levels do not necessarily represent the actual protein abundances, making RNA-based marker discovery less reliable.

Although CITE-seq and REAP-seq provide an accurate estimations of surface protein levels, these techniques require a previous selection of the proteins that need to be measured. Moreover, these protocols can be costly and technically challenging to implement and most single cell studies, including Human Cell Atlas project [18], quantify the transcriptome only and do not have cell-matched measurements of relevant surface proteins.

With the rising popularity of single cell transcriptomic technologies, and the shortage of high-coverage single cell proteomics protocols, there is a strong need for the development of techniques that can impute protein levels from the cellular transcriptome.

Reviews on the topic have shown that RNA and Protein correlations can be affected by several factors such as technical biases, temporal and spatial constrains, or biological differences [19]. Previous studies have used pulse labelling techniques to evaluate RNA and Protein turnover dynamics in mammalian cell lines [20]. Using this approach Björn Schwanhäusser and colleagues showed that Protein levels are principally controlled at the level of translation, with other processes such as RNA and Protein degradation having almost negligible effects. More importantly, these authors showed for the first time that it is possible to improve RNA and Protein correlations by adjusting for the translation rate of every gene [20].

We hypothesize this approach can be expanded to scRNA-seq datasets, allowing a more accurate identification of cell surface markers from the transcriptomic data. Therefore, in this final thesis we set out to answer the two following questions: can we use protein translation rates estimated in bulk datasets to predict protein levels in scRNA-seq experiments? and additionally, can we apply this method using only scRNA-seq data?

2.2 Project goals

To address the questions outlined for this project, and to account for the limitations in currently published approaches (see State of the art section) we need to achieve two main goals:

2.2.1 Estimation of a gene-wise RNA correction factor from bulk RNA and Proteomic data.

This objective will focus on the estimation of protein translation rates from bulk RNA and proteomic datasets. For this project, we will refer to protein translation rates as "correction factors" (CF), since that is indeed the use we are making of these values. Two specific objectives are necessary to achieve this goal:

- **To calculate an RNA CF.** Using RNA and protein data from published datasets, an RNA CF will be estimated as described before [21].
- **To evaluate the CF performance.** Before implementing this correction factor in scRNA-seq data, we will assess its efficiency in improving the RNA-Protein correlation. Moreover, we will validate some of the observations reported previously by other authors [21], [22].

2.2.2 Study of the suitability of the correction factor for scRNA-seq data.

The second main goal for this project is to evaluate the performance of the CF in single cell RNA sequencing data. To this end, we propose three specific goals.

- **To apply the correction factor into scRNA-seq data**. Since we require relatively good coverage of both RNA and Protein measurements, we will focus on CITE-seq/REAP-seq datasets as these provide high proteomic coverage. Due to the differences between bulk and single cell transcriptomics several additional steps will need to be applied before of the implementation of the correction factor (See Methods).
- **To evaluate the performance of the CF**. Upon implementation of the CF, it is crucial to evaluate the ability of this approach to impute protein levels from single-cell RNA sequencing data.

- **To evaluate the performance of a CF estimated from single cell data.** Following similar procedures, we will calculate a CF this time based only on CITE-seq data. The performance of this CF will be evaluated within and across single-cell datasets, hence assessing how well this method extrapolates to other data.

2.3 Methodology and approach

It is generally accepted that omics data, and especially single cell omics, require powerful statistical software for their robust and reproducible analysis. In this sense, the most widespread tools used for the scientific community are Python [23] and R [24]. Although both present well stablished pipelines for data analysis, this thesis will be conducted entirely using the R programming language (v 4.0.5) and the Integrated Development Environment RStudio [25]. This decision stems from my relative extensive experience using this software, particularly for the analysis of bulk transcriptomic data.

In this sense, for this project we will make use of pre-processed data, i.e., count matrices (RNA and scRNA sequencing) and protein abundance information. Most databases provide this type of data, reducing time-consuming tasks such as alignment and quality control steps. Consequently, BASH software, commonly used for these pre-processing steps, will not be necessary.

Regarding bulk transcriptomics and proteomics data manipulation, most of the calculations in this thesis can be easily applied using base R functions. Nevertheless, we will use the collection of R packages Tidyverse [26] to aid data manipulation and use clearer syntax.

On the other hand, single cell transcriptomics (and especially single-cell proteo-genomics) require efficient and organised data-manipulation strategies to reduce the time of analysis. In this sense, the R package Seurat (v 4.0.1) [12] provides a suite of functions and objects that facilitate the analysis of single cell multimodal data such as CITE-seq results. Moreover, the package is regularly reviewed and updated, and contains well stablished pipelines as well as comprehensive documentation for scRNA-seq analysis [27].

2.4 Workplan

In this section we provide a detailed description of the fundamental tasks needed to achieve the goals established. Tasks are appended to the goals previously described for a better understanding of the project’s structure. Number of days dedicated to every task are indicated.

1. **Estimation of a gene-wise RNA correction factor for surface proteins.**

   1.1. To calculate an RNA correction factor.

   1.1.1. Task 1 (5d): obtain bulk RNA and Protein data from different tissues. Due to the nature of this analysis, both RNA and Protein data should have been obtained from the same sample or at least from the same biological source in comparable conditions (i.e., same tissue from two healthy subjects).

   1.1.2. Task 2 (5d): Integrate datasets. For this task we will apply some basic data exploration to ensure we are working with high quality data (i.e., values distribution, Principal Components Analysis). We will also filter and prepare RNA and Protein datasets to facilitate the CF calculation.

   1.1.3. Task 3 (3d): RNA CF. An RNA CF will be estimated as described before (see Methods).

   1.2. To evaluate the RNA CF performance.

   1.2.1. Task 1 (7d): Using the CF we will correct the RNA expression so it better correlates with protein abundances.

   1.2.2. Task 2 (10d): Compare predictions between corrected and uncorrected data. Correlation of protein abundance with corrected and uncorrected RNA expression data will be calculated. We will pay special attention to surface proteins as these will be important in subsequent analyses.

2. **Study of the suitability of the correction factor for scRNA-seq data.**

   2.1. **To prepare the scRNA-seq dataset to apply the CF.**

      2.1.1. **Task 1 (5w):** Obtain CITE-seq/REAP-seq data. Looking ahead, to test the performance of this algorithm, datasets with high surface protein coverage are preferred as these enable as to get more robust results. We have allocated several weeks for this task. Since searching for scRNA-seq datasets is not restricted to any previous results, we consider this task can be completed while we work on the first main goal.

      2.1.2. **Task 2 (7d):** Dimensionality reduction. To aid to the visualization of single cell data, we will use dimensionality reduction techniques such as Principal Component Analysis (PCA), t-Distributed Stochastic Neighbour Embedding (t-SNE) or Uniform Manifold Approximation and Projection (UMAP).

      2.1.3. **Task 3 (7d):** Cluster scRNA-seq data. Before cell type identification cells will be clustered according to their expression profiles.

      2.1.4. **Task 4 (7d):** Identify cell types in clusters. Upon clustering, cell types will be identified according to the expression of characteristic cell markers.

      2.1.5. **Task 5 (7d):** Calculate pseudo-bulk RNA. In this task we will calculate the pseudo-bulk RNA profiles of the defined clusters. This step is crucial to avoid single-cell related artifacts in RNA-Protein correlations.

   2.2. **To evaluate the performance of corrected data detecting surface markers.**

      2.2.1. **Task 1 (7d):** To correct pseudo-bulk data using the CF estimated previously.

      2.2.2. **Task 2 (7d):** Comparison of surface protein abundance with corrected RNA values. Finally, the correlation between pseudo-bulk RNA data and cell surface proteins estimations will be computed. This value will be used as an indicator of the algorithm's performance.

2.3. To evaluate the performance of a CF estimated from single cell data.

2.3.1. Task 1 (7d): Estimation of a CF from single cell data. Using the same procedures as in previous tasks we will compute a CF, this time using only single cell data.

2.3.2. Task 2 (7d): Lastly, we will apply this correction factor within and across datasets and estimate its performance.

Figure 1 illustrates a Grantt Chart with the task's organisation throughout the semester. PECs have been included at the top for reference. Orange pegs indicate the milestones completion date (see section below).

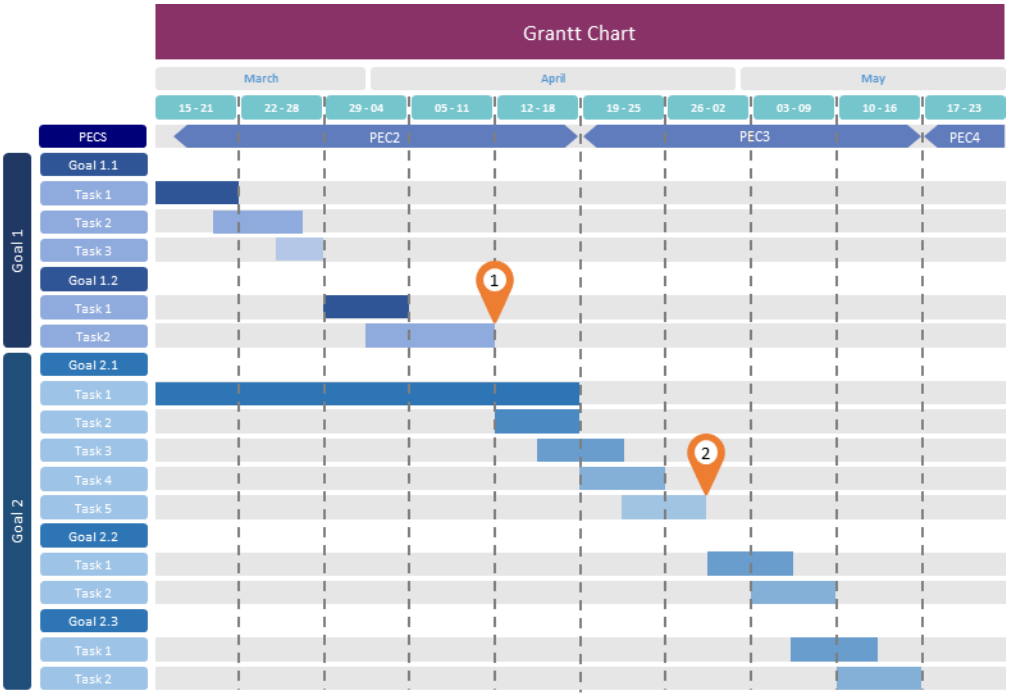


Fig. 1. Grantt Chart

2.4.1 Milestones

Two crucial milestones have been identified in this project, the first one being the estimation of the RNA CF in bulk datasets. Indeed, the obtention of this factor is the main limiting step to start the second main goal. According to the plan stablished, before the end of PEC2 the first milestone should be completed.

The second milestone refers to the processing of scRNA-seq data. As mentioned earlier, several additional steps need to be applied before the CF can be implemented. Indeed, the algorithm cannot be tested before obtaining a fully processed dataset making this a crucial limiting step. As reflected on the chart, this milestone should be complete no later than the end of April and the results should be included in PEC3.

2.4.2 Risk analysis

Absence of suitable data: results obtained in this thesis will be only as good as the data used to produce them. This problem is not expected to affect the first goal as bulk RNA and protein data is widely available from a variety of models, tissues, and conditions [28], [29]. However, as indicated previously, simultaneous quantification of RNA and Protein at the single cell level is technically challenging, especially with high coverage [9]. In this sense, the absence of suitable datasets might pose a critical impediment for the completion of the second goal. To address this problem, sensible time has been designated to review the literature to find the best reports.

Lack of coverage: one of the main characteristics of single cell data is its sparsity. The small amount of RNA recovered per cell affects the number of expressed genes detected by this technique. This could negatively affect our results, especially since we are going to focus on a specific subset of genes such as surface-protein-coding genes. Nevertheless, pseudo-bulk estimations (Goal 2.1, Task 5) should help addressing this problem as combination of cell transcriptomes increases the coverage [19], [30].

2.5 Brief description of results and products obtained.

- **Working Plan:** a comprehensive working detailing the necessary steps to complete the project and the timeline when they should be completed.
- **Report:** results obtained in this final thesis will be summarised in this report, following the department guidelines, along with a critical interpretation of the main conclusions extracted from the data. Also, some background will be provided to introduce the reader on the topic.
- **Final product:** although this algorithm could be implemented in a R package or Shiny application for easy utilization in several datasets, this goal is out of the scope of this thesis. However, for the shake of research transparency all R scripts developed for this project have been upload to an online GitHub repository. https://github.com/AntonioUOC/TFM

- **Slides Presentation:** a series of slides describing the project and the results will also be produced.
- **Project self-examination:** limitations of this project and elements to improve will be evaluated at the end of the project (see conclusions and future perspectives). This will put in perspective the achievements of the project and the future steps in this research area.

2.6 Brief description of the rest of the project's chapters

- **State of the art**: description of the current methods available to address the problem and the contribution of this project to the field.
- **Methods**: comprehensive explanation of the methods employed in this thesis and the reasons why they were used.
- **Results:** all the results obtained throughout the course of the project will be summarised and presented in this chapter.
- **Discussion:** based on the results obtained, in this section we will delineate the main findings in this project.
- **Conclusions:** the final chapter we will summarise the conclusions extracted from this thesis and evaluate the degree of achievement of the proposed goals. Additionally, we will describe future steps that can be followed to advance our knowledge on this topic.

3 State of the art

The Central Dogma of Molecular Biology establishes clear associations between DNA, RNA and Proteins. However, although their sequences are tightly linked by rigid rules, there is no straightforward relationship between the concentration of RNA transcripts and the abundance of their correspondent proteins [19]. In this context, the presence or absence of correlation between RNA and protein levels has been subject of debate for many years. Thanks to the development of transcriptomic and proteomic techniques, there is an increasing body of evidence that shows that proteome and transcriptome abundances are not sufficiently correlated to act as proxies for each other. However, several reports show that the RNA to Protein ratios are well preserved for every gene across different biological sources and can be used to extrapolate protein correlations from their RNA transcripts [20]–[22].

Based on previous works, Wilhelm and colleagues described the use of RNA to Protein ratios to increase the correlation between RNA and Protein abundances, reporting increases in Spearman correlation coefficients from 0.41 to 0.9 after correction [22]. However, in this article RNA and Protein measurements were taken from different biological sources since transcriptomes were obtained from an online repository. In this sense, despite showing an improvement in RNA-Protein correlations, the correction factor estimated does not faithfully represent the RNA translation rates since biological variability across individuals is known to affect the RNA-Protein interrelationship [19].

More recent publications have studied this correction factor using RNA and Protein measurements obtained from the same samples [21]. Although this approach accounts for artifacts derived from inter-sample variability, the correction factor was estimated using mainly cell lines and only a limited number of human tissues (11 different tissues). Cell lines are pivotal for biological research, however cellular models are known to diverge from their parental source, strongly affecting their phenotype and experimental outcomes [31].

Consequently, there is a strong need for a pan-tissue study of RNA to Protein ratios and their potential to correct RNA expression. More importantly, this approach has never been studied at the single cell level, probably due to the lack high-coverage single cell proteomic studies. In this regard, very recent works have reported comprehensive proteo-genomic single-cell datasets covering over 150 cell surface proteins [12], [32], opening the door to expand this approach to single cells.

Building upon previous studies, in this thesis we will use complete proteo-genomic datasets available in the current literature to examine this promising concept, expanding the coverage to 29 different human tissues [29]. Furthermore, we will explore for the first time the potential of this approach in single cell data, paving the way for the creation of robust tools for the prediction of cell surface biomarkers in these studies.

4 Methods

4.1 Data collection

For the first part of the analysis, we have used the data from Dongxue Wang *et al.* [29]. This data set contains a deep atlas of RNA and protein expression of 29 different human tissues, more than doubling the coverage from previous studies.

For the second part of the project, we have utilised a recently presented single cell atlas of the bone marrow of healthy subjects [32]. In particular, the dataset used here contains the RNA expression values of 8285 bone marrow cells from a healthy subject along with the abundance of 197 surface proteins measured by CITE-seq.

4.1 Cell surface proteins selection in bulk datasets

Databases such as UniProt include the keyword “cell membrane” to identify these proteins. However, this annotation also includes proteins attached or integrated in the plasma membrane from the intracellular side, without an extracellular domain [33]. Since we are looking for proteins that can potentially be used as cell markers, not all entries annotated as “cell membrane” in UniProt are of interest.

The gene ontology (GO) database includes the GO term GO:0009897 that specifically characterize proteins at the “*external side of plasma membrane*”. However, only 538 human genes are annotated with this GO term (see Supplementary Method 1). Hence, filtering by this criterion would be far too restrictive for the analysis.

To overcome these limitations Damaris Bausch-Fluck and colleagues developed a machine learning approach to identify cell surface proteins that are at least partially exposed to the extracellular space (referred by the authors as “*surfaceome*”)[33]. By integrating 131 protein features the authors were able to define a human surfaceome of 2886 proteins. To the best of our knowledge, this is the most comprehensive list of human cell surface proteins available in the literature that also meet the criteria for new cell biomarkers identification. Consequently, in this project the term “cell surface genes” refers to the genes reported in this paper and “cell surface proteins” as the proteins coded by said genes.

4.2 Single Cell Data pre-processing

The R package Seurat (v 4.0.1) has been used to process the CITE-seq data [12]. In this case we started the analysis from the RNA and ADTs count matrices with low quality cells and doublets already filtered out.

Gene counts were normalized using the “global-scaling” method “LogNormalize” implemented in the package and that is defined as follow:

$$LogNorm(x_{g,i}) = log\left(\frac{x_{g,i}}{tot_i} \cdot 10000 + 1\right)$$

Where $x_{g,i}$ is the gene count of gene “g” in cell “i” and tot_i is the library size of cell “i”. The scaling factor was left as default (10000). On the other hand, for the surface protein counts, we adopted the relative abundance transformation from Stoeckius et al. [10]:

$$clr(x_i) = \left[ln\frac{x_{1,i}}{g(x_i)}, \ldots, ln\frac{x_{g,i}}{g(x_i)}\right]$$

Where $x_{g,i}$ is the gene count of gene “g” in cell “i” and $g(x_i)$ is the geometric mean of the ADT counts in cell x.

4.3 Dimensionality reduction

Before performing dimensionality reduction on the dataset, we selected the most variable genes as these represent heterogeneous features (i.e., high expression in some genes and low in others).

For bulk transcriptomic and proteomic data, we manually computed the variance of all genes across tissues and selected the 1000 most variable for dimensionality reduction. Regarding scRNA-seq data, we used the function *FindVariableFeatures* from the package Seurat. This function applies a variance-stabilizing transformation to correct for mean-variance relationship that is inherent to single-cell RNA-seq, hence reducing the influence of technical effects while preserving the biological heterogeneity [34].

Having selected the most variable features we performed principal component analysis using the base R function and Seurat's implementation for bulk and single cell data, respectively.

4.4 Principal components selection

Seurat uses the PCA results to perform other analyses such as clustering, which requires the selection of a defined number of components. To select the components that explain a statistically significant proportion of the variance Seurat provides the function "JackStraw". In short, this method performs 1000 PCAs on the input data permuting a small proportion of the genes each time (10% in our case). This way, the Jack Straw method estimates a null distribution of scores for every gene from which a p-value can be computed [35].

4.5 Cell Clustering

For cell clustering we used Seurat's graph-based clustering approach, based on previously published algorithms [36]. Briefly, the function FindNeighbors constructs a k-nearest neighbours graph using the Euclidean distance in the PCA space. The cells in the graph are then iteratively grouped together into highly interconnected communities using a Louvain algorithm with the function FindClusters [37].

4.6 Cell type annotation

Cell type annotation was performed using RNA and cell surface antibody readouts along with the information provided in the original publication [32]. To identify characteristic cell surface markers and expressed genes for every cluster we used FindAllMarkers. This function uses a non-parametric Wilcoxon rank sum test to identify differentially expressed genes and ADTs between one cluster and the rest using the Bonferroni correction to account for multiple-testing errors [38].

4.7 Non-linear dimensionality reduction

In addition to PCA, we implemented two no-linear dimensionality reduction methods to better illustrate the datasets. We applied t-distributed stochastic neighbour embedding (t-SNE) initialised with a subset of components from the PCA analysis. The rationale behind this decision comes from previous benchmarks which showed that PCA initialisation can improve the global structure of the final embedding and increases the reproducibility of outcome [39]. Alternatively, we applied uniform manifold approximation and projection (UMAP) on the datasets as an additional non-linear dimensionality reduction method. Both methods were applied using the package Seurat. The package umap (v0.2.7.0) [40] was used to apply this method to bulk datasets.

4.8 Pseudo-bulk RNA counts

Before applying the CF we computed pseudo-Bulk counts of the clusters using methods previously described in the literature [30]. For this thesis we adopted the following approach: firstly, we selected those clusters with over 400 cells to ensure a good pseudo-bulk approximation. Secondly, we generated the pseudo-bulk profiles by calculating the sum of the transcript counts across all cells per cluster. The pseudo-bulk counts were transformed to counts per million (CPM) using the Bioconductor package EdgeR [41]. Lastly, CPMs were scaled and centred so the mean RNA CPMs matched the RNA average expression of the bulk dataset used to compute the correction factor. The same procedure was applied to the ADT counts.

4.9 Computing and applying the Correction Factor

We can define the correction factor for a gene “g” in tissue “t” as:

$$Cf_{g,t} = \frac{Protein_{g,t}}{RNA_{g,t}}$$

Where $\text{Protein}_{g,t}$ and $\text{RNA}_{g,t}$ are normalised RNA and Protein values for gene “g” in tissue “t” respectively. For bulk data we used the log transformed iBAQ and FPKM values, whereas for single-cell data we used normalised RNA and ADT counts.

This CF can also be interpreted as the translation rate as it estimates the number of proteins produced for every RNA molecule. Using this CF now it becomes possible to predict protein abundance from RNA levels as follow:

$$Protein_{g,t} \approx RNA_{g,t} \cdot Cf_g$$

Using this formula, we corrected the RNA expression values for every gene in all tissues. Note that in each case, the gene-specific CF used for prediction was never estimated an applied in the same sample to avoid overfitting.

5 Results

5.1 Correction Factor in Bulk data

5.1.1 Data exploration

The dataset from Dongxue Wang *et.al*. [29] was already normalized and presented in FPKM (Fragments Per Kilobase Million) and iBAQ (sum of all the peptides intensities divided by the number of observable peptides of a protein [42]) for RNA and protein respectively. Quick inspection of the data showed that RNA and Protein datasets were well normalised across tissues (Fig. 2 A, B). Furthermore, the distribution and dynamic range of transcripts and proteins matched with the described in the literature [17] (Fig. 2 C).

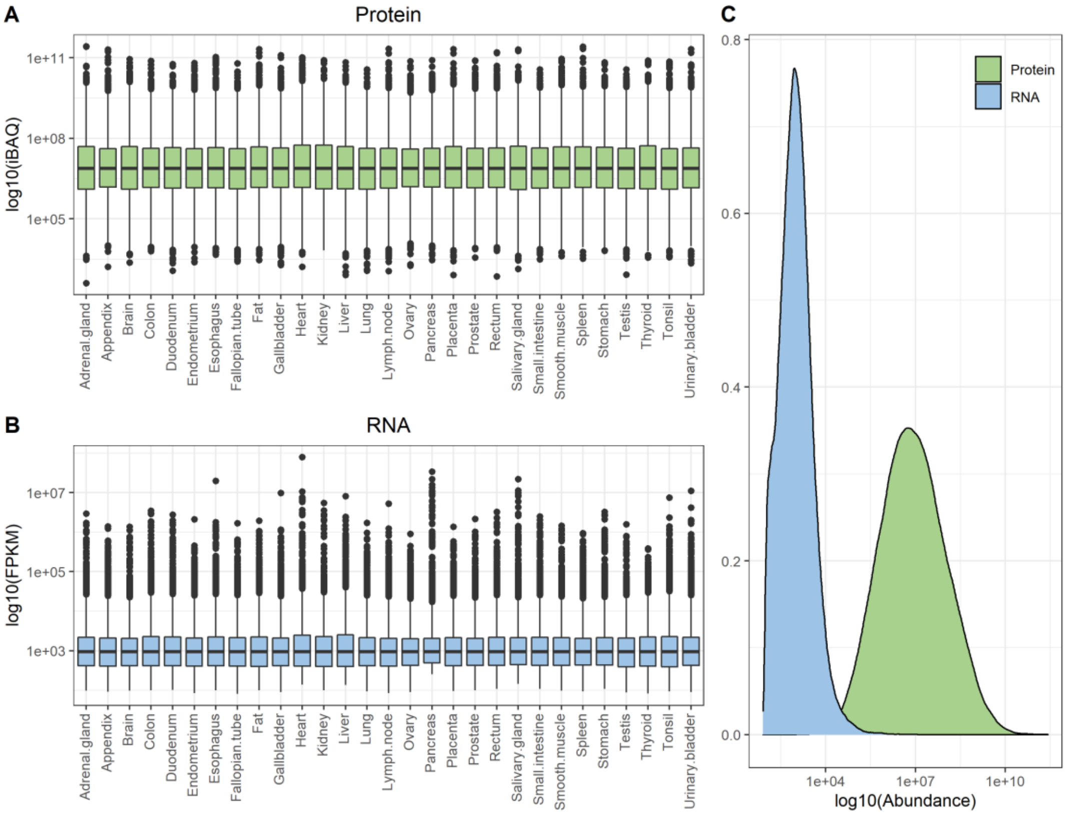


Fig. 2 Protein and RNA distributions across tissues.

Before computing the correction factor, we wanted to test weather RNA and Protein datasets contain information about the tissue's identities, similarities, and differences. To do so we performed principal component analysis (PCA) using the 1000 most variable genes in each dataset.

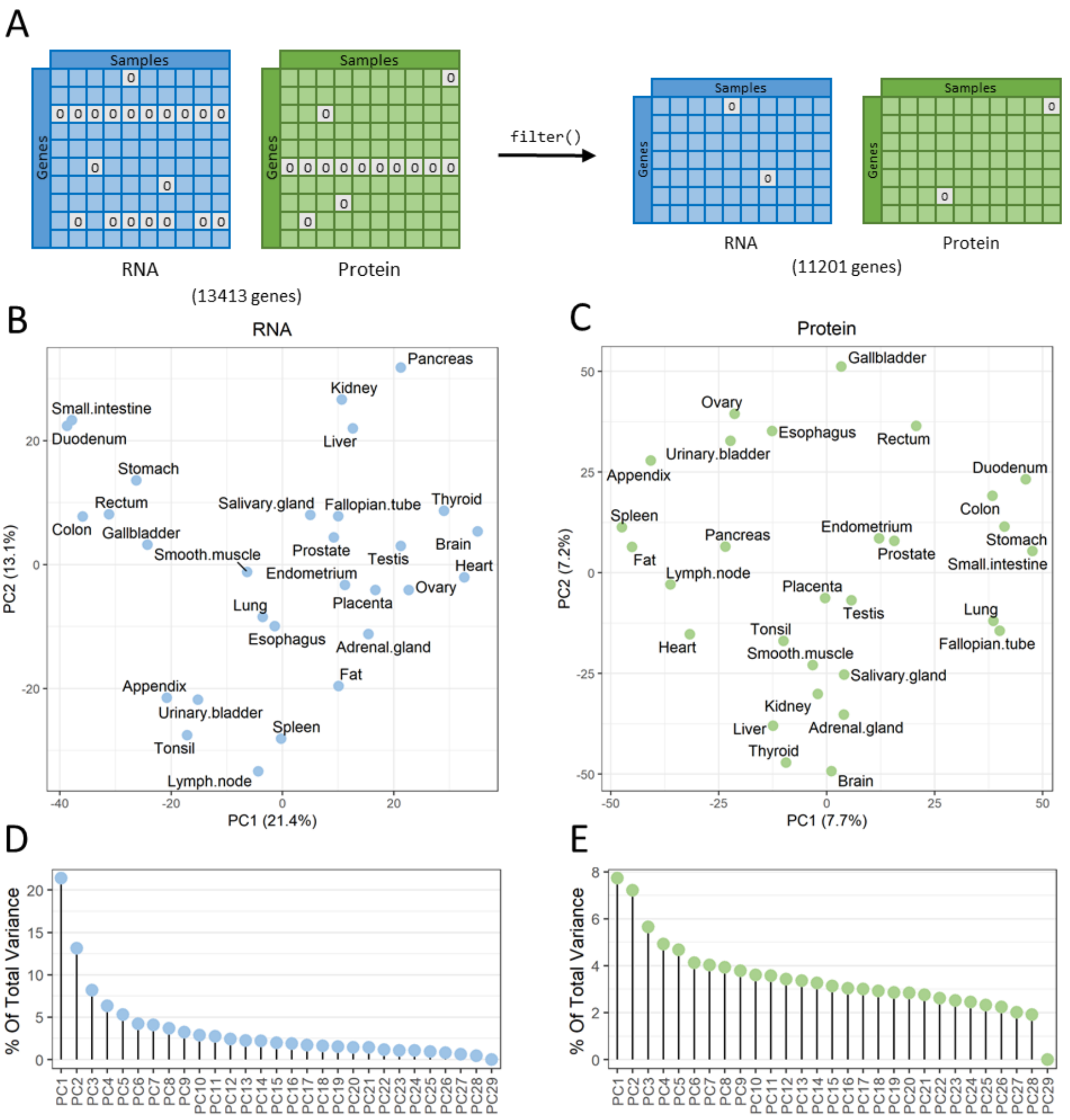


Fig. 3 A, Filtering strategy. B, C, Principal Component Analysis Plots of RNA (B) and Protein (C) datasets. D, E, scree plots of RNA (D) and Protein (E) PCAs.

Transcriptomics PCA showed a roughly homogeneous distribution of the tissues across the first 2 principal components with no evident clusters (Fig. 3 B). Secondary lymphoid tissues such as Spleen, Lymph node and Tonsil gathered close in the PCA plot. Interestingly, the vermiform appendix also localised close to these tissues, consistent with the recently described role of this part of the intestine in human immunity [43]. Tissues with similar cellular composition also localised together such as Duodenum and Small Intestine or Colom and Rectum. In a similar fashion, proteomic PCA did not show any evident clusters (Fig. 3 C).

Overall, the first two principal components accumulate less than 40% of the total variance (35.5% and 14.9% for RNA and Protein PCAs respectively, Fig. 3 C, D).

To account for the low variance captured by PCA and to try improving the datasets low dimensionality representation we also performed UMAP. The number of components to initialise the UMAP was selected using Scree Plots of the components' variances (Fig. 3 D, E). RNA's Scree Plot shows a significant drop in the variance explained across the first 6 components and very little change for the rest. This justifies the selection of the first 6 components for UMAP.

Protein's Scree Plot also showed a sudden drop in the explained variance during the first six components, although PCs7-28 also explained a significant proportion of the variability. We tried starting UMAP using increasing number of components, but we could not note a significant improvement in the dataset representation (Data not shown), hence we decided to also use the first six components.

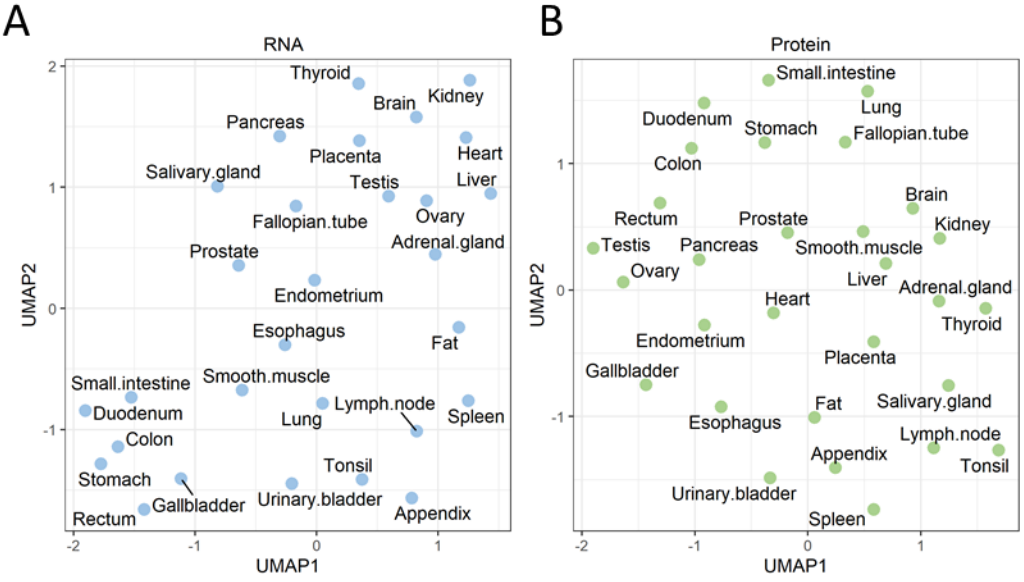


Fig. 4 A, B, Uniform Manifold Approximation and Projection plots of RNA (A) and Protein (B) datasets.

In the final UMAPs no distinct clusters could be observed (Fig. 4). However, tissues with similar cell composition tended to appear close together such as Small Intestine, Duodenum, Colon and Stomach. This could be seen in both RNA and Protein datasets.

5.1.2 Data pre-processing

Both RNA and Protein datasets included a column with the Ensemble ID of the gene to which the expression value (RNA or Protein) was associated to. This list of unique identifiers was used to merge both datasets. Only the genes that were common for both datasets (n = 13413) were kept for the analysis.

After merging, some genes still had null RNA or Protein expression in all tissues or most of them. To avoid potential artifacts in future analyses we only kept those genes that were detected in at least 5 tissues at both Protein and RNA level (Fig. 5 A). After filtering, 11201 genes were retained for analysis.

5.1.3 Correction Factor

After preparing the datasets we then decided to compare the RNA and protein levels in the different tissues. In Fig. 5 C and Fig. S1 the RNA and protein levels for the detected genes are plotted for all the tissues included in the study. A moderate trend can be observed, with Spearman correlation values ranging from 0.42 in the Ovary to 0.58 in the Liver with an average of 0.52 for all the tissues (Fig. 5 G). These results are in line with previous publications showing rather moderate correlation when RNA and protein levels are compared directly [17].

Both RNA and protein levels varied greatly depending on their biological source. However, the Protein/RNA ratio is largely conserved between tissues as shown previously [22]. The study of Protein to RNA ratios in this dataset yield similar conclusions, where the ratio can be shown to be roughly constant across tissues (Fig. 5 H).

Based on this value, we set out to estimate a correction factor that corrects the RNA level for every gene so that it better correlates with its corresponding protein abundance. Note that in each case, the gene-specific CF used for prediction of protein abundance was estimated as the median of the CF for that gene in the other twenty-eight tissues (leave-one-out method), to avoid overfitting (Fig. 5 A, B).

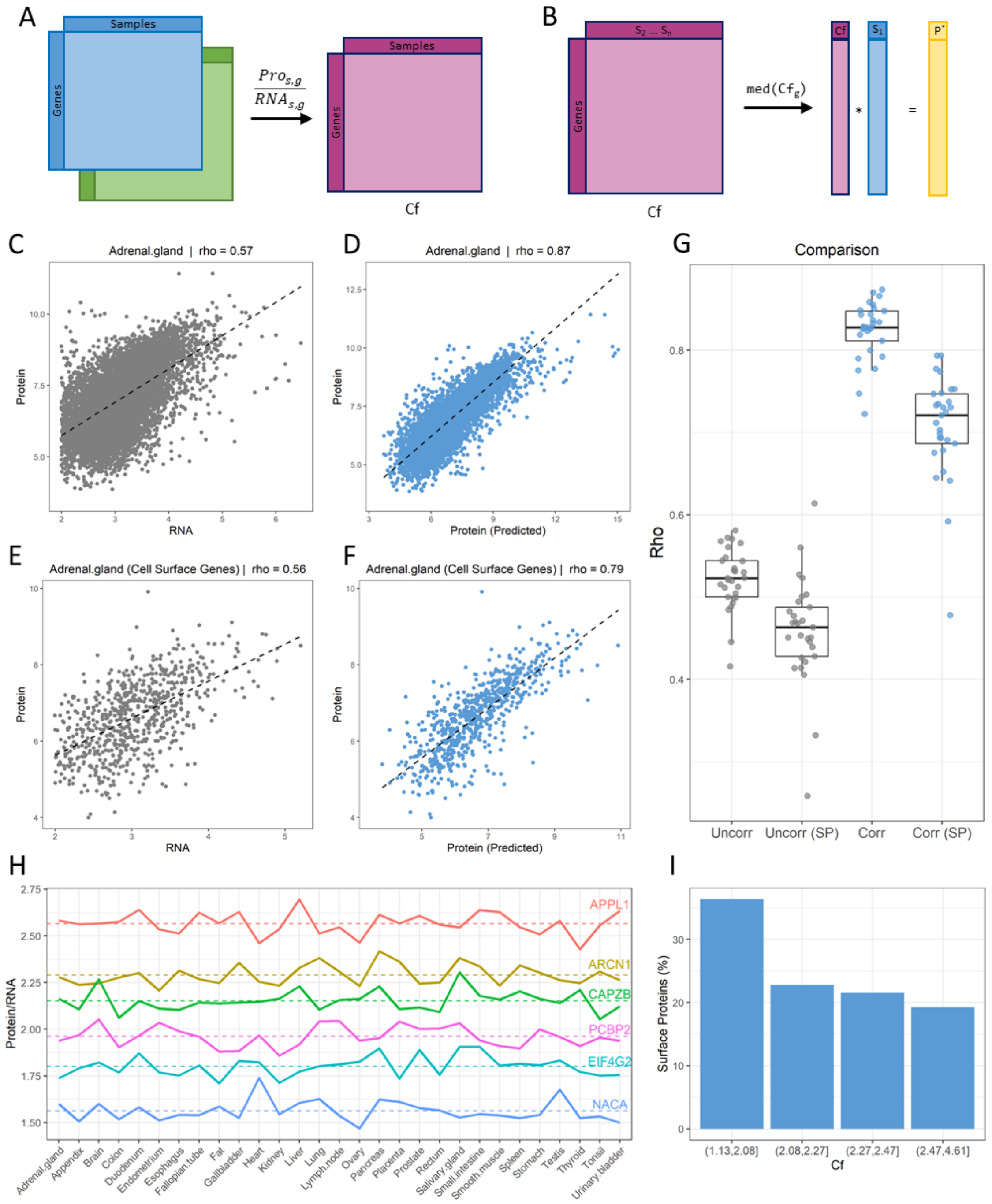


Fig. 5 Study of RNA and Protein correlations. A, B, strategy for computing (A) and applying (B) the correction factor. C-F, RNA and Protein correlation levels in Adrenal Gland before (C, E) and after (D, F) correction for all genes (C, D) and surface protein genes (E, F). G, comparison of the Spearman correlation value between RNA and Protein levels before and after correction for all 29 tissues, SP. Surface Proteins. H, study of the Protein/RNA ratio (CF) across all tissues in 6 representative genes. I, frequency of Cell Surface Genes by binned Cf levels.

As shown in Fig. 5 D, G and Fig. S2, a good correlation can be observed across all the genes in each of the tissues after applying the correction. The Spearman correlation coefficient for corrected data ranged from 0.72 to 0.87 in Ovary and Adrenal Gland respectively with an average of 0.82 for all tissues.

5.1.4 Correction factor in cell surface proteins

Out of 2886 cell surface genes, 1012 were identified in our dataset (31.3%). These genes showed an even distribution in Protein/RNA scatterplots (Fig. 5 E, F, and darker points in Fig. S1, S2). Protein/RNA correlations for surface genes were lower than correlations for all genes both before and after correction. Nevertheless, correlations showed a noticeable improvement after correction with the average Spearman correlation factor increasing from 0.46 to 0.71 before and after correction (Fig. 5 G, paired t-test p.val < 0.001).

Interestingly, genes with low values of CF were enriched in cell surface genes, with 36% of the cell surface genes showing a CF between 1.1 and 2.1 (Fig. 5 I, Pearson's Chi-squared test p.val < 0.001). Nonetheless, 64% of the cell surface genes had a CF greater than 2.1 showing that applying the correction has a great impact in the RNA values for most of these genes.

5.2 Correction factor in single cell data

5.2.1 Data description

We have started the analysis from the count matrices for both RNA and antibody derived tags (ADTs), where the latter is a proxy of the cell surface protein abundances. In particular, this dataset contains the RNA expression values of 8285 bone marrow cells from a healthy subject along with the abundance of 197 surface proteins measured by CITE-seq [32].

5.2.2 Dimensionality reduction and principal components selection

After normalising RNA and ADT count matrices we performed Principal Component Analysis to explore the global structure of the dataset. Additionally, we used the calculated principal components as inputs in subsequent analyses, treating them as “meta-features” that encompass the variance of the data.

We selected the 200 most variable genes using the function *FindVariableFeatures* to compute the PCA. Features were scaled so that gene expression’s means and variances across cells were equal to 0 and 1, respectively.

PCA results (Fig. 6 A) showed moderate separation of the clusters identified (See sections below for more details on cell clustering and labelling). The first principal component (PC1) captured the differences between progenitor cells (i.e., EP/MkP, MP, HSC/MPP) and mature cells (i.e., pDC, B cells, T cells) as well as the variability within the progenitor cells clusters themself. Similarly, the second principal component seemed to represent the variability between the myeloid (i.e., MP, Myelocytes, Monocytes) and lymphoid (i.e., B cells, T cells) compartments.

Overall, the first two PCs recapitulated 38.75% of the total variance which represented the global biological differences between the clusters. Nevertheless, JackStraw analysis of the PC variances showed that the first 14 components retained a significant proportion of the variability at the 5% confident level (Fig. 6 B), suggesting that nonlinear embeddings can improve the dataset representation in lower dimensions.

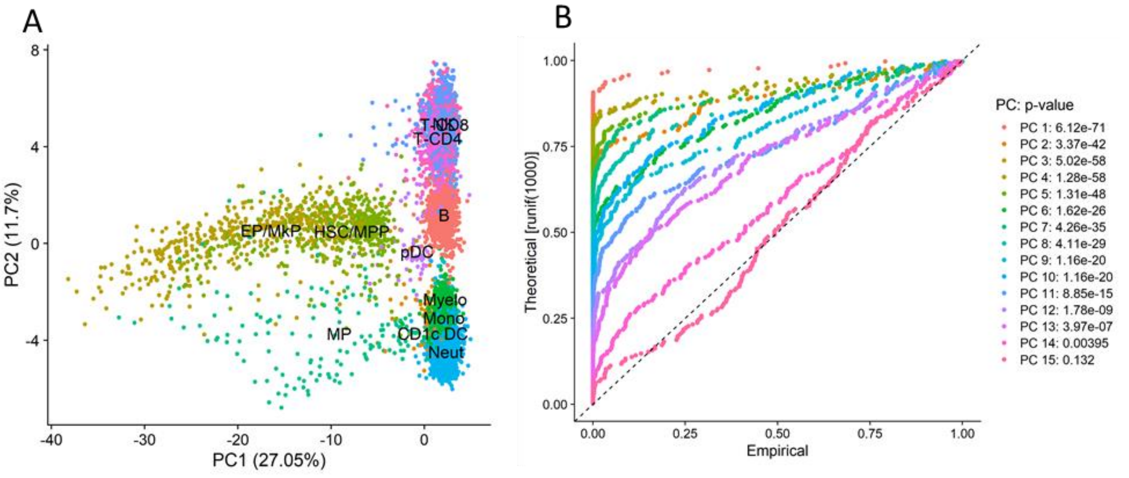


Fig. 6 A, Scatterplot of the first two Principal Components. Cells are coloured by cluster. B, JackStraw results or the first 15 Principal Components.

5.2.3 Cell Clustering

Following PCA application and components selection we set out to identify the main cell clusters in the dataset. To do so we constructed a k-nearest-neighbours graph with the function *FindNeighbors* using the first 14 principal components as defined previously. The cells in the graph were then grouped together in clusters using the function *FindClusters*. The parameter “resolution”, which indirectly controls the number of clusters identified, was set to 0.4. After running both functions, we were able to identify 12 clusters.

5.2.4 Cell type annotation

We identified characteristic cell surface markers and genes for every cluster using Seurat’s function FindAllMarkers. Table 1 shows the five RNAs and ADTs with the highest fold change for each cluster.

Using this information, along with the data provided in the original publication, we identified twelve different haematopoietic cell types and states previously described in the bone marrow [32]. Hematopoietic stem cells and multipotent progenitors (HSC/MPP) were defined via surface expression of CD34 and CD133. On the RNA level, the HSC/MPP cluster is characterized by high expression of CRHBP, NPR3 and PROM.

Erythroid/Megakaryocyte progenitors (EP/MkP) arise from the HSC/MPP cluster and are characterised for the expression of CD34, CD236 and the transferrin receptor CD71. In addition, EP/MkP cells showed characteristic expression of SLC40A1, haemoglobins (HBD) and the coagulation-related protein ITGA2B.

Cells from the myeloid-granulocyte lineage (Myeloid Progenitors (MP), Myelocytes, Neutrophils) showed an increased expression of the myeloid marker CD33. Monocytes showed increased expression of FCGR3A and LST1. On the other hand, Neutrophils showed increased expression of the inflammatory molecules S100A9 and S100A8.

Plasmacytoid dendritic cells (pDCs) displayed expression of characteristic makers such as CF38, CD123 and CD98. Additionally, CD1c dendritic cells were defined by increased expression of Tim3 as well as CD1c.

Regarding the lymphoid compartment, B cells expressed IgD and IgM on the cell surface as well as CD79A and IGHM at the RNA level. T lymphocytes showed characteristic expression of CD3 along with CD8 and CD4 in T-CD8 and T-CD4 lymphocytes, respectively. Lastly, Natural Killer (NK) cells displayed surface markers like CD16 or CD7 as well as high expression levels of the NK granules component NKG7.

Table 1 RNA and ADT makers with the highest fold change for every cluster.

Cluster ID	RNA Markers	ADT Markers
Neutrophils	CD14, S100A9, S100A8, VCAN, FCN1	CD36, CD41, CD282, CD33, CD93
NK Cells	GNLY, TRDC, KLRB1, NKG7, KLRF1	CD7, CD94, CD16, CD45RA, CD122.MIK.BETA3
B Cells	IGHM, CD79A, MS4A1, IGHD, TCL1A	CD73, IgD, CD21, IgM, CD272
Monocytes	FCGR3A, MS4A7, LST1, NR4A1, COTL1	CD16, CD11c, CD31.WM59, CD54, CD371
CD8 T Cells	CD8A, CD8B, TRAC, CD3G, CD3E	CD8, CD8b.2ST8.5H7, CD3, CD2, CD5
CD4 T Cells	IL7R, TRAC, CD27, CD3E, CCR7	CD5, CD4, CD3, CD2, CD27
Erythroid/Mk progenitor	HBD, ITGA2B, SLC40A1, TYMS, STMN1	CD71, CD34, CD43, CD326, CD235a-b
HSC/MPP	SPINK2, CD34, CRHBP, NPR3, PROM1	CD34, CD49b, CD166, CD133, CD110
Myeloid Progenitor	TOP2A, UBE2C, CDC20, AURKB, NUSAP1	CD85k, CD371, HLA.DR, CD33, CD193
CD1c Dendritic Cells	MRC1, FCER1A, CD1C, CST3, AREG	Tim3, CD371, CD141, CD101, CD206
Plasmacytoid DC	MZB1, DERL3, JCHAIN, ITM2C, IL3RA	CD98, CD54, CD123, CD38, CD162
Myelocytes	CD163, CD14, CSF3R, VCAN, FCN1	CD93, CD36, CD116, CD282, CD33

5.2.5 Non-linear dimensionality reduction

We implemented two no-linear dimensionality reduction methods to illustrate the dataset, that is, t-distributed stochastic neighbour embedding (t-SNE) and uniform manifold approximation and projection (UMAP). Both algorithms were initialised with the first 14 PCA components.

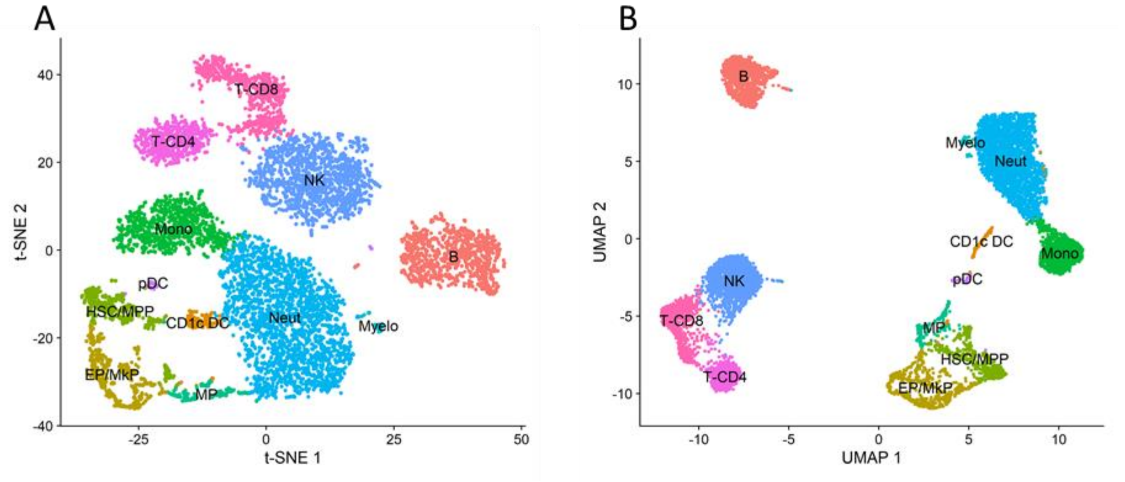


Fig. 7 A, tSNE plot. B, UMAP plot. Neut, neutrophil; NK, natural killer; B, B cell; Mono, monocytes; T-CD8/CD4, T-CD8/CD4+ lymphocytes; EP/MkP, Erythroid/Megakaryocyte Progenitor; HSC/MPP, haematopoietic stem cell/multipotent progenitor; MP, myeloid progenitor; CD1c DC, CD1c dendritic cell; pDC, plasmacytoid dendritic cell; Myelo, myelocyte.

Results of both algorithms showed good agreement with clustering findings (Fig. 7). tSNE embedding had bigger clusters that were closer together. On the other hand, UMAP clusters were smaller and clearly separated by cell lineages such as progenitor cells (EP/MkP, HSC/MPP, MP), T cells (NK, TCD8, TCD4), B cells and myeloid cells (Monocytes, Neutrophils, Myelocytes), highlighting the biological differences among clusters. These results are in line with previous benchmarks which showed that UMAP preserves better the global structure of the data when compared to tSNE [44].

5.2.6 Pseudo-bulk RNA counts

It is expected that cell-to-cell variation captured in scRNA-seq data will affect the RNA to Protein correlation, thus affecting our correction. Important processes such as cell cycle, transcriptional bursting, the delay between transcription and translation and the variation in the influence of external signals can affect this correlation [19]. However, these sources of variation are “averaged out” of the data in bulk RNA-seq experiments, where all cells are assumed to be in the same steady state [19]. Hence, we decided to compute pseudo-Bulk counts of the clusters using methods previously described in the literature [30] (Fig. 8 A).

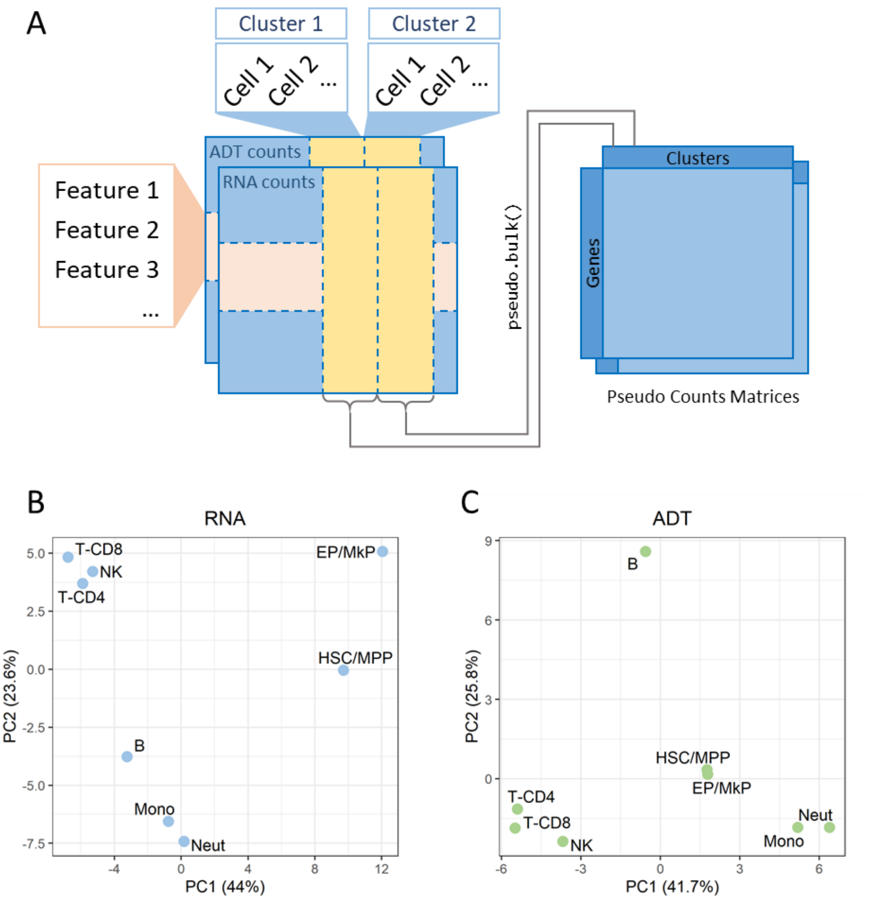


Fig. 8 A, Pseudo-bulk strategy. B, C, PCAs of RNA and ADT's pseudo-bulk profiles.

Principal component analysis of RNA and ADT pseudo-bulk profiles showed that cell types aggregate in a similar way as we observed in UMAP and t-SNE representations of individual cells. In both RNA and ADT's PCAs we could observe clear separations between progenitor cells (EP/MkP, HSC/MPP), myeloid cells (Monocytes, Neutrophils), T cells (T-CD4, T-CD8, NKs) and B cells. These results clearly show that pseudo-pseudo bulk profiles retain the biological information from the single cell clusters.

5.2.7 Merge RNA and ADT datasets

The dataset selected for this analysis had different notation systems for features in the RNA and ADT count matrices. RNAs genes were annotated using gene names whereas ADTs were named using its Cluster of Differentiation (CD) code. To facilitate the datasets integration feature names were substituted for their Ensembl Gene ID since these are stable unique identifiers. The Ensembl IDs corresponding to the gene names were obtained using Biomart [45]. On the other hand, CDs were transformed to Ensembl IDs using the tools from the Hugo Gene Nomenclature Committee's web portal [46].

Using the Ensembl IDs we then filtered and reordered both datasets, so they contained the same features in the same order. In total, 99 genes were selected for further analyses.

5.2.8 Correction factor application

Having merged both datasets we now wanted to investigate the correlation between RNA and their corresponding ADT counts. In a similar way to what we observed on bulk data, RNA and ADT counts sowed a low correlation for all pseudo-bulk profiles (Fig. 9 B, C, Fig. S3 A). Spearman correlation values ranged from 0.33 for T-CD8 cells to 0.57 for HSC/MPP with an average of 0.43 for all cell types.

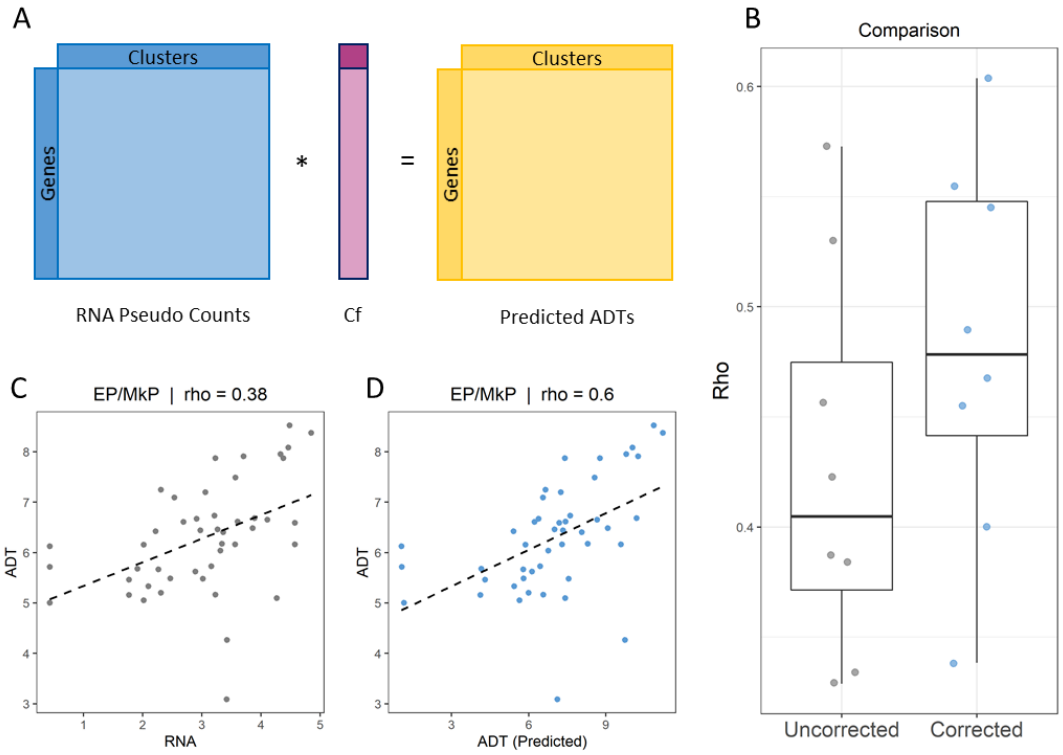


Fig. 9 A, pseudo-bulk profiles correction strategy. B, comparison of the Spearman's correlation value before and after correction for all pseudo-bulk profiles. C, D, RNA and ADT correlation levels in EP/MkP cells before (A) and after (B) correction using the CF estimated with bulk data.

For the next step in our analysis, we wanted to explore whether it would be possible to improve the RNA-ADT correlation using a correction factor computed using bulk transcriptomics and proteomics. To do so, we computed the average

CF for every gene using the data obtained previously and used this value to correct the RNA abundance in our new pseudo-bulk dataset (Fig. 9 A).

The Spearman correlation coefficients between corrected RNA and ADT showed a moderate improvement over the uncorrected data (Fig. 9 B, D, Fig. S3 B). Correlation values ranged from 0.34 to 0.6 for Neutrophils and EP/MkP respectively with an average correlation of 0.48.

5.2.9 Correction factor from single-cell data

Following these results, we asked whether we could estimate a correction factor from single cell data that could better improve RNA-ADT correlations. To do so, we created two new pseudo-bulk profiles called “training” and “test” using the method described above. Training pseudo-bulk profiles were calculated by randomly selecting 70% of the cells from every cluster. The 30% remaining cells were used to compute the test dataset (Fig. 10 A).

Using the training dataset, we estimated the CF for every gene and cell type using the same procedure as before. The correction factor for every gene was computed as the average CF across all cells in the training dataset.

Using this new CF, we then corrected the RNA expression in the testing dataset and re-calculated the RNA-ADT correlations. As shown in Fig. 10 D and Fig. S4 B, after correction a good correlation can be observed across genes for all cell types. The Spearman correlation for corrected data varied from 0.67 to 0.91 for T-CD8 cells and HSC/MPP respectively with an average correlation of 0.80. These correlations were significantly higher than the ones obtained with uncorrected data (paired t-test p-value < 0.001).

5.2.10 Application of the correction factor on a different dataset

Having shown that we can effectively correct RNA expression we wanted to see if we could use this correction factor to improve the RNA-ADT correlation on a different dataset. To test this hypothesis, we used a new dataset also included in the original publication. This new data included RNA abundance of 11252 bone marrow cells from a healthy subject at the single cell level along with 105 ADT measurements. The data was pre-processed following the same steps as before and the results are illustrated in Supplementary Figure 5.

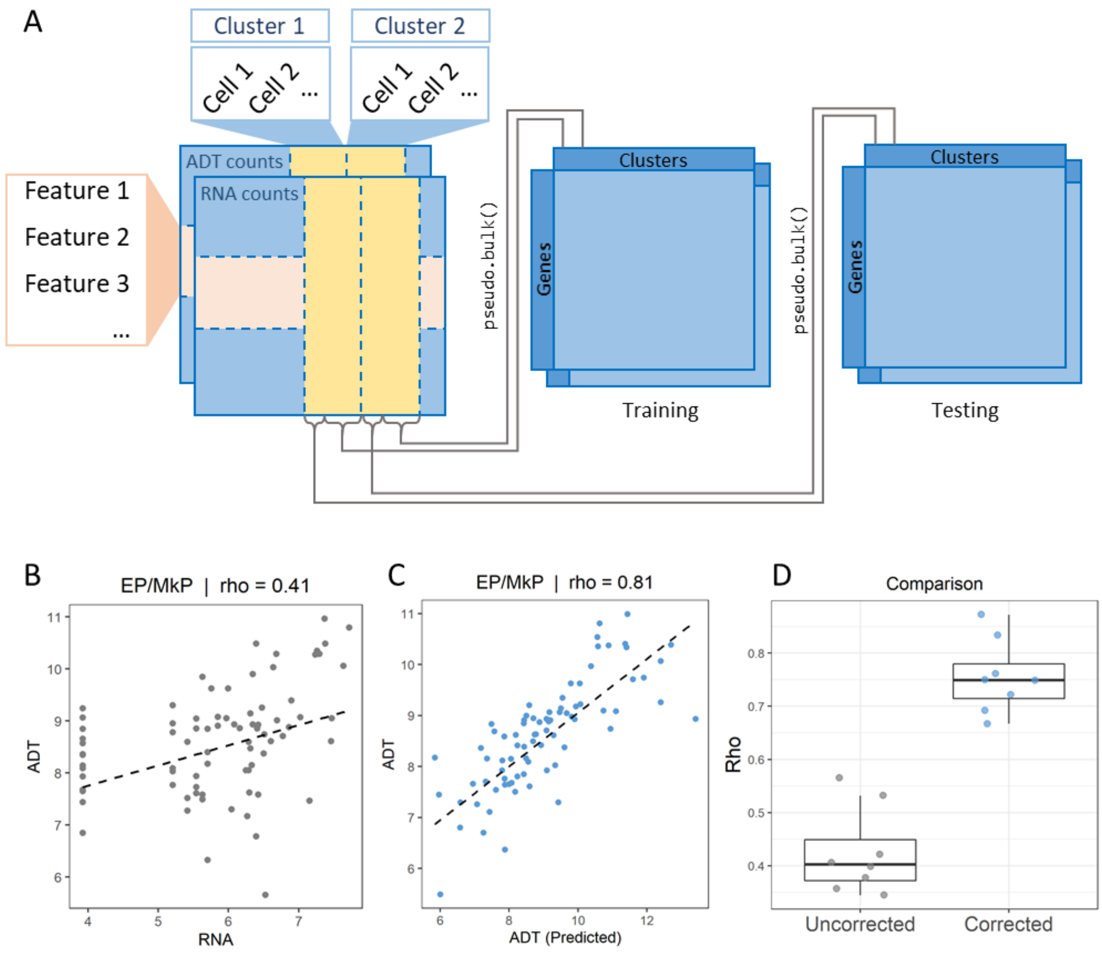


Fig. 10 A, strategy for computing training and testing datasets. B, C, RNA and ADT correlation levels in EP/MkP cells before (A) and after (B) correction using the CF estimated from single cell data. D, comparison of the Spearman's correlation value before and after correction for all pseudo-bulk profiles.

From the processed data we computed the pseudo-bulk profiles as described previously. Using this data, we computed a correction factor. Lastly, we corrected the data using both CFs, the one obtained from the first dataset (CF1) and the one calculated from this one (CF2). When we compare both CFs, we could see that their values were in good agreement (Fig. 11 A), which suggest that RNA-ADT ratios are maintained across samples.

Using the correction factor calculated with the previous dataset (CF1) we were able to substantially increase RNA-ADT correlations (Fig. 11 C, D. Fig. S7). When compared to the uncorrected data, the average Spearman correlation factor increased from 0.4 to 0.67 using CF1 (paired t-test p-value < 0.001). On the other hand, the correction factor computed using the second dataset (CF2) achieved even greater RNA-ADT correlations with an average correction of 0.76 (Fig. 11 C, E. Fig. S8). These results show that the correction factor can be successfully applied across CITE-seq datasets although with certain loss of power. Nonetheless, the correction factor calculated using CITE-seq data clearly outperformed the one computed using bulk data.

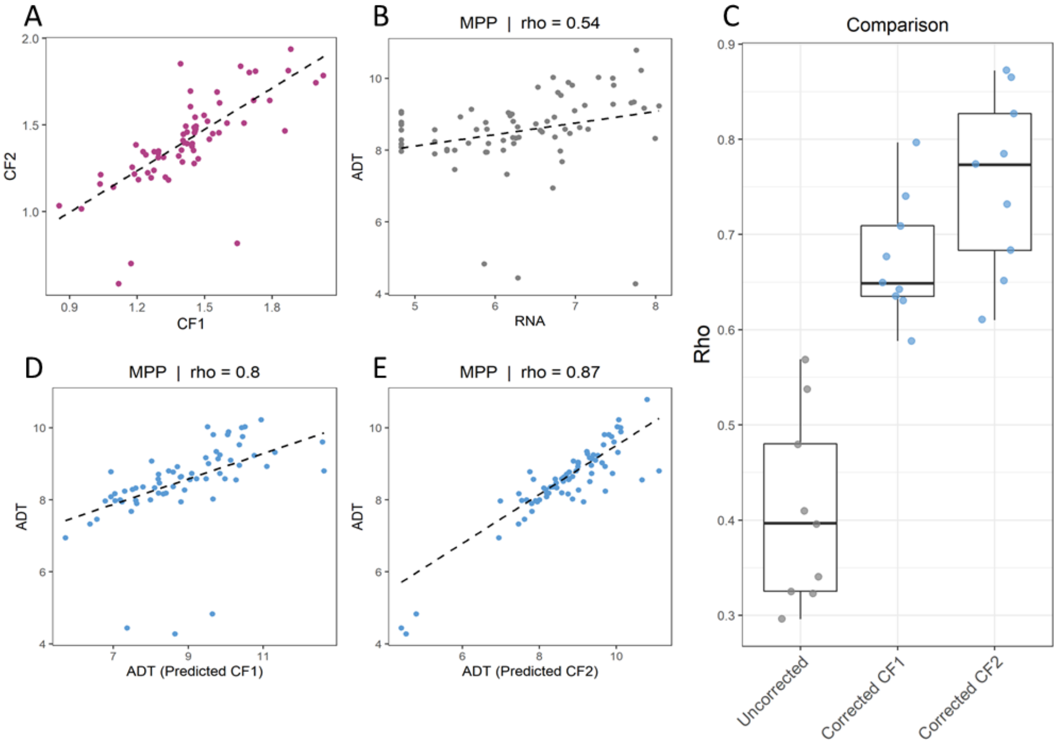


Fig. 11 A, comparison between CF1 and CF2. B, RNA-ADT scatterplot of MPPs before correction. C, comparison of the Spearman's correlation value before and after correction for all pseudo-bulk profiles. D, E, RNA-ADT scatterplots of MPPs after correction using CF1 (D) and CF2 (E)

6 Discussion

In the present thesis we have explored the concept of using translation rates (also referred to as correction factor) to estimate the protein abundance in a sample from the transcriptome.

As discussed in the State of the art section, previous studies have shown the potential of this measurement to improve RNA-Protein correlations. However, these reports were somewhat limited, and results needed to be validated using a more comprehensive dataset. In this sense, in the first part of our analysis we explored this approach using a proteogenomic atlas of 29 different human tissues, more than doubling the number of tissues included in previous efforts.

Our results showed good agreement with earlier studies. RNA and proteins showed modest correlations when compared directly in all 29 tissues (average

Spearman correlation 0.52). Moreover, we showed that Protein/RNA rations were roughly constant across tissues, validating previous observations of this phenomenon in a pan-tissue context.

More importantly, we showed that protein abundances from one tissue could be effectively predicted from the transcriptomic levels using the Protein/RNA ratios (correction factor) of the other 28 tissues. Interestingly, we could see this increase in tissues such as the brain, that have a distinct cellular composition than the rest of the tissues used to compute the CF. This result, suggest that the CF is independent of the cellular configuration and can be applied in any human sample regardless of the origin or biological architecture.

When we focused on cell surface proteins, we could also observe an increase in RNA-Protein correlations after correction in all tissues. Some tissues however showed moderate correlations such as the Ovary (Spearman correlation 0.45 after correction). We hypothesize that this modest correlation is due to technical artifacts in this sample since the correlation before correction was considerably lower compared to the rest of the tissues (Spearman correlation 0.25 for Ovary versus 0.47 average correlation for the other 28 tissues).

An interesting finding is that cell surface proteins display low Protein/RNA ratios. This result is in line with previous studies that revealed that cell-surface proteins have stable RNAs but high protein turnover rates [20] making protein and RNA levels more comparable, thus reducing their Protein/RNA ratios.

In the second part of our analysis, we set out to explore the application of this approach in single cell data. To do so, we used a recently published proteogenomic atlas of the bone marrow that surveyed almost 200 cell surface proteins alongside transcriptional profiles using CITE-sequencing.

Cell clustering and labelling showed that the dataset utilised had a good representation of the bone marrow cellular composition. To avoid unwanted artifacts in the RNA-Protein correlations that arise from cell-to-cell variation we computed the pseudo-bulk RNA and Protein profiles from the clusters. Interestingly, principal component analysis of the pseudo-bulk profiles showed similar clustering results as the ones observed using UMAP in single cell data.

These results showed that pseudo-bulk profiling not only captures the biological information of the clusters but also removes sources of variation that are challenging to model using linear dimensionality reduction methods.

When we applied the CF estimated from bulk data to the pseudo-bulk profiles, we observed a moderate increase in the correlation. We hypothesize there are two main reasons for the limited performance the correction factor in this case:

- On the one hand, the correction factor was estimated using proteomic data measured by mass spectrometry whereas CITE-seq measures protein quantities by sequencing antibody derived tags. Although both techniques aim to measure protein abundances, the approaches are radically different and are affected by their own particular biases that can in turn affect the final correction factor [9].
- On the other hand, proteomic data measures whole cell protein abundance as opposed to CITE-seq that only surveys proteins on the cell surface. Indeed, this can have a prominent effect on the RNA-protein ratio since proteins present in one cell compartment (the cell surface in this case) do not represent total protein abundance in the cell [14]. In the case of cell surface proteins, this bias can be particularly important since some membrane proteins can be downregulated in the cell surface without affecting the cellular total protein content [47].

To overcome these limitations, we explored the application of a CF computed from CITE-seq data. Our results showed that this second CF can be used to effectively predict cell surface protein levels from the transcriptomic profiles with average ADT-RNA correlations improving from 0.43 to 0.76. More importantly, we demonstrated that the CF estimated from different datasets are roughly similar, and that the CF from one dataset can be used to predict cell surface protein levels across studies.

Overall, the results of this research project shows that the use of translation rates to predict protein levels in single cell data is possible, although careful considerations must be taken to minimise potential biases affecting the Protein/RNA ratios

7 Conclusions

7.1 Conclusions

Based on the results obtained in this project we can conclude that:

- Protein/RNA ratios remain roughly constant across tissues, supporting their use for protein prediction from RNA levels.
- The translation rate for every gene expressed as the ratio Protein/RNA (also referred to as Correction Factor) can be used to impute protein abundances from transcriptomic data in a wide range of tissues regardless of their origin or cellular composition. Importantly, given the high number of tissues included we can say this is a robust conclusion of the study.
- Surface proteins prediction is also possible, which suggests that this approach can be used to improve surface markers detection.
- Translation rates estimated from bulk datasets are affected by specific biases that limit their implementation in single cell data. Nonetheless, additional research is needed to validate the exact sources of variation that leaded to this result.
- Protein/RNA ratios can be computed from single cell datasets and can be effectively applied to improve RNA-protein correlations in pseudo-bulk profiles of the single cell clusters.
- Correction factors estimated from single cell data can improve RNA-protein correlations when applied to external datasets, indicating that is possible to estimate a correction factor that can be applied across single cell datasets. Nevertheless, further research is needed to corroborate this hypothesis.

7.2 Future perspectives

Overall, this thesis serves as a starting point for the development of algorithms that predict protein levels in scRNA-seq. Building upon our results we propose different strategies to further explore unresolved questions in this topic.

We theorised that one of the reasons for the poor performance of the correction factor from bulk data in single cell is the intrinsic bias in protein detection by both techniques. Indeed, although bulk proteomics can detect far more surface proteins than CITE-seq (in our datasets, around 1300 in bulk and almost 200 in single cell), bulk proteomics measure whole cell protein levels, introducing an unwanted bias.

Plasma membrane profiling is a proteomic technique that employs aminooxy-biotinylation to label and extract cell surface proteins [48]. This technique retains the high coverage characteristic of bulk techniques while avoiding artifacts introduced by whole cell proteomics. For this reason, we hypothesize that correction factors computed using this data will perform better in CITE-seq datasets, increasing the number of proteins that can be corrected.

On the other hand, recent antibody based approaches were able to measure intracellular proteins in single cells [13]. It would be interesting to explore the possibility of applying the correction factor in this dataset and in future proteogenomic atlases developed with this technique.

Finally, our single cell analysis was limited to bone marrow samples. As more proteogenomic datasets are published it will be necessary to validate the conclusions obtained in this thesis in other human tissues, especially the ability of the correction factor to be applied across studies.

7.3 Planning Compliance

At the beginning of the project, we stablished two main goals, both of which were successfully completed according to the timeline established. The completion of the goals also means that all milestones were achieved on time, and the project progressed uninterruptedly. Furthermore, all the methods originally proposed were sufficient to address the main questions of this project. Overall, the planning was followed successfully, and the project completed satisfactorily.

7.3.1 Working plan deviations and justification.

For the second part of the project there was some deviations from the original timeline. The second goal was achieved 1.5 weeks ahead of schedule, which gave us more time to complete additional tasks. In this sense, we added an additional last task (i.e., apply the correction factor across datasets and estimate its performance) which enabled us to obtain more robust conclusions for this project.

8 Glossary

8.1 Definition of relevant terms used in the project

RNA, RiboNucleic Acid. Polimeryc molecule that contains genetic information essential for biological processes such as expression and regulation of genes. Although there are different types of RNA described (for instance ribosomal RNAs, transfer RNA or non-coding RNAs), in this thesis we focus on those RNAs that contain the genetic information used to produce proteins, that is, messenger RNA or mRNA. Hence, the term RNA in this project refers to the mRNA type.

Protein, biological molecule made of amino acids that exert most of the functions in the cell, from structural support to reaction catalysis and cell signalling.

CF, Correction Factor. Measure of the ratio Protein/RNA for each gene.

8.2 Other acronyms used in the project

- **Techniques**
  - **UMAP**, Uniform Manifold Approximation and Projection
  - **PCA**, Principal Components Analysis
  - **TSNE**, t-Distributed Stochastic Neighbour Embedding
- **Cell types**
  - **Neut**, neutrophil
  - **NK**, natural killer
  - **B**, B cell
  - **Mono**, monocytes
  - **T-CD8/CD4**, T-CD8/CD4+ lymphocytes
  - **EP/MkP**, Erythroid/Megakaryocyte Progenitor
  - **HSC/MPP**, haematopoietic stem cell/multipotent progenitor
  - **MP**, myeloid progenitor
  - **CD1c DC**, CD1c dendritic cell
  - **pDC**, plasmacytoid dendritic cell
  - **Myelo**, myelocyte.
  - **iB**, immature B cell

9 Bibliography

[1] C. Zhu, S. Preissl, and B. Ren, “Single-cell multimodal omics: the power of many,” *Nat. Methods*, vol. 17, no. 1, pp. 11–14, 2020.

[2] R. Stark, M. Grzelak, and J. Hadfield, “RNA sequencing: the teenage years,” *Nat. Rev. Genet.*, vol. 20, no. 11, pp. 631–656, 2019.

[3] T. Kalisky *et al.*, “A brief review of single-cell transcriptomic technologies,” *Brief. Funct. Genomics*, vol. 17, no. 1, pp. 64–76, 2018.

[4] T. M. Consortium, “Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris,” *Nature*, vol. 562, no. 7727, pp. 367–372, 2018.

[5] X. Han *et al.*, “Construction of a human cell landscape at single-cell level,” *Nature*, vol. 581, no. 7808, pp. 303–309, 2020.

[6] E. Papalexi and R. Satija, “Single-cell RNA sequencing to explore immune cell heterogeneity,” *Nat. Rev. Immunol.*, vol. 18, no. 1, p. 35, 2018.

[7] L. Li *et al.*, “What are the applications of single-cell RNA sequencing in cancer research: a systematic review,” *J. Exp. Clin. Cancer Res.*, vol. 40, no. 1, pp. 1–12, 2021.

[8] A. Doerr, “Single-cell proteomics,” *Nat. Methods*, vol. 16, no. 1, p. 20, 2019.

[9] J. R. Choi, K. W. Yong, J. Y. Choi, and A. C. Cowie, “Single-cell RNA sequencing and its combination with protein and DNA analyses,” *Cells*, vol. 9, no. 5, p. 1130, 2020.

[10] M. Stoeckius *et al.*, “Simultaneous epitope and transcriptome measurement in single cells,” *Nat. Methods*, vol. 14, no. 9, pp. 865–868, 2017.

[11] V. M. Peterson *et al.*, “Multiplexed quantification of proteins and transcripts in single cells,” *Nat. Biotechnol.*, vol. 35, no. 10, pp. 936–939, 2017.

[12] Y. Hao *et al.*, “Integrated analysis of multimodal single-cell data.,” *Cell*, May 2021.

[13] J. Reimegård *et al.*, “A combined approach for single-cell mRNA and intracellular protein expression analysis,” *Commun. Biol.*, vol. 4, no. 1, pp. 1–11, 2021.

[14] E. A. O’donnell, D. N. Ernst, and R. Hingorani, “Multiparameter flow cytometry: advances in high resolution analysis,” *Immune Netw.*, vol. 13, no. 2, p. 43, 2013.

[15] X. Zhang *et al.*, “CellMarker: a manually curated resource of cell markers in human and mouse,” *Nucleic Acids Res.*, vol. 47, no. D1, pp. D721–D728, 2019.

[16] D.-M. Popescu *et al.*, “Decoding human fetal liver haematopoiesis,” *Nature*, vol. 574, no. 7778, pp. 365–371, 2019.

[17] C. Buccitelli and M. Selbach, “mRNAs, proteins and the emerging principles of gene expression control,” *Nat. Rev. Genet.*, vol. 21, no. 10, pp.

630–644, 2020.

[18] A. Regev *et al.*, "Science forum: the human cell atlas," *Elife*, vol. 6, p. e27041, 2017.

[19] Y. Liu, A. Beyer, and R. Aebersold, "On the dependency of cellular protein levels on mRNA abundance," *Cell*, vol. 165, no. 3, pp. 535–550, 2016.

[20] B. Schwanhäusser *et al.*, "Global quantification of mammalian gene expression control," *Nature*, vol. 473, no. 7347, pp. 337–342, 2011.

[21] F. Edfors *et al.*, "Gene-specific correlation of RNA and protein levels in human cells and tissues," *Mol. Syst. Biol.*, vol. 12, no. 10, p. 883, 2016.

[22] M. Wilhelm *et al.*, "Mass-spectrometry-based draft of the human proteome," *Nature*, vol. 509, no. 7502, pp. 582–587, 2014.

[23] Python Software Foundation, "Python Language Reference." [Online]. Available: http://www.python.org. [Accessed: 08-Jun-2021].

[24] R Core Team and R Foundation for Statistical Computing, "R: A Language and Environment for Statistical Computing." Vienna, Austria, 2021.

[25] P. RStudio, "RStudio: Integrated Development Environment for R." Boston, 2020.

[26] H. Wickham *et al.*, "Welcome to the Tidyverse," *J. Open Source Softw.*, vol. 4, no. 43, p. 1686, 2019.

[27] P. Hoffman and Satija Lab, "Seurat - Guided Clustering Tutorial," 2021. [Online]. Available: https://satijalab.org/seurat/articles/pbmc3k_tutorial.html. [Accessed: 08-Jun-2021].

[28] N. Nagaraj *et al.*, "Deep proteome and transcriptome mapping of a human cancer cell line," *Mol. Syst. Biol.*, vol. 7, no. 1, p. 548, 2011.

[29] D. Wang *et al.*, "A deep proteome and transcriptome abundance atlas of 29 healthy human tissues," *Mol. Syst. Biol.*, vol. 15, no. 2, p. e8503, 2019.

[30] C. T. Wohnhaas *et al.*, "DMSO cryopreservation is the method of choice to preserve cells for droplet-based single-cell RNA sequencing," *Sci. Rep.*, vol. 9, no. 1, pp. 1–14, 2019.

[31] S. Zaaijer, S. C. Groen, and N. E. Sanjana, "Tracking cell lineages to improve research reproducibility," *Nat. Biotechnol.*, pp. 1–5, 2021.

[32] S. H. Triana *et al.*, "Single-cell proteo-genomic reference maps of the hematopoietic system enable the purification and massive profiling of precisely defined cell states," *bioRxiv*, 2021.

[33] D. Bausch-Fluck *et al.*, "The in silico human surfaceome," *Proc. Natl. Acad. Sci.*, vol. 115, no. 46, pp. E10988–E10997, 2018.

[34] C. Hafemeister and R. Satija, "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression," *Genome Biol.*, vol. 20, no. 1, pp. 1–15, 2019.

[35] N. C. Chung and J. D. Storey, "Statistical significance of variables driving

systematic variation in high-dimensional data,” *Bioinformatics*, vol. 31, no. 4, pp. 545–554, 2015.

[36] C. Xu and Z. Su, “Identification of cell types from single-cell transcriptomes using a novel clustering method,” *Bioinformatics*, vol. 31, no. 12, pp. 1974–1980, 2015.

[37] T. Stuart *et al.*, “Comprehensive integration of single-cell data,” *Cell*, vol. 177, no. 7, pp. 1888–1902, 2019.

[38] A. Butler, P. Hoffman, P. Smibert, E. Papalexi, and R. Satija, “Integrating single-cell transcriptomic data across different conditions, technologies, and species,” *Nat. Biotechnol.*, vol. 36, no. 5, pp. 411–420, 2018.

[39] D. Kobak and P. Berens, “The art of using t-SNE for single-cell transcriptomics,” *Nat. Commun.*, vol. 10, no. 1, pp. 1–14, 2019.

[40] L. McInnes, J. Healy, and J. Melville, “Umap: Uniform manifold approximation and projection for dimension reduction,” *arXiv Prepr. arXiv1802.03426*, 2018.

[41] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, “edgeR: a Bioconductor package for differential expression analysis of digital gene expression data,” *Bioinformatics*, vol. 26, no. 1, pp. 139–140, 2010.

[42] B. Fabre *et al.*, “Comparison of label-free quantification methods for the determination of protein complexes subunits stoichiometry,” *EuPA Open Proteomics*, vol. 4, pp. 82–86, 2014.

[43] I. A. Kooij, S. Sahami, S. L. Meijer, C. J. Buskens, and A. A. Te Velde, “The immunology of the vermiform appendix: a review of the literature,” *Clin. Exp. Immunol.*, vol. 186, no. 1, pp. 1–9, 2016.

[44] E. Becht *et al.*, “Dimensionality reduction for visualizing single-cell data using UMAP,” *Nat. Biotechnol.*, vol. 37, no. 1, pp. 38–44, 2019.

[45] S. Haider, B. Ballester, D. Smedley, J. Zhang, P. Rice, and A. Kasprzyk, “BioMart Central Portal—unified access to biological data,” *Nucleic Acids Res.*, vol. 37, no. suppl_2, pp. W23–W27, 2009.

[46] S. Povey, R. Lovering, E. Bruford, M. Wright, M. Lush, and H. Wain, “The HUGO gene nomenclature committee (HGNC),” *Hum. Genet.*, vol. 109, no. 6, pp. 678–680, 2001.

[47] B. D. Berkovits and C. Mayr, “Alternative 3′ UTRs act as scaffolds to regulate membrane protein localization,” *Nature*, vol. 522, no. 7556, pp. 363–367, 2015.

[48] M. P. Weekes *et al.*, “Proteomic plasma membrane profiling reveals an essential role for gp96 in the cell surface expression of LDLR family members, including the LDL receptor and LRP6,” *J. Proteome Res.*, vol. 11, no. 3, pp. 1475–1484, 2012.

10 Supplementary Materials

10.1 Supplementary methods

Supplementary method 1

To evaluate the number of human genes annotated with the term GO:0009897 we followed the next process:

1. Using BioMart, we filtered all Human Ensembl Gene IDs that were annotated with the GO term GO:0009897.
2. Since the list can have some duplicates, we filtered and counted the dataset using the R software with the following command:

   ```
   IDs %>% unique %>% length
   ```

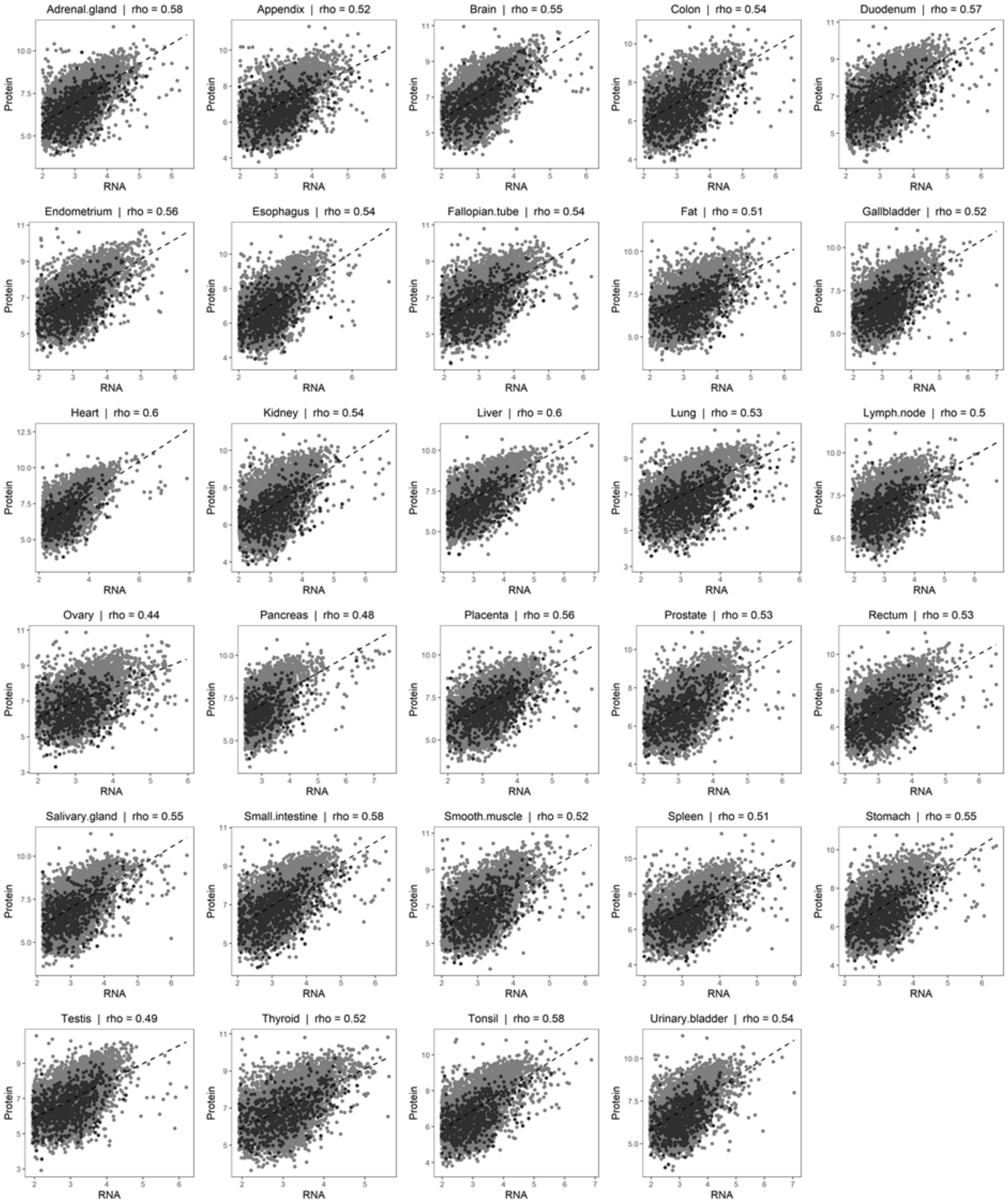
   Where “IDs” is a vector object with the Ensembl gene IDs.

3. The output obtained is the number of human genes annotated with the GO:0009897 term.

10.2 Supplementary figures

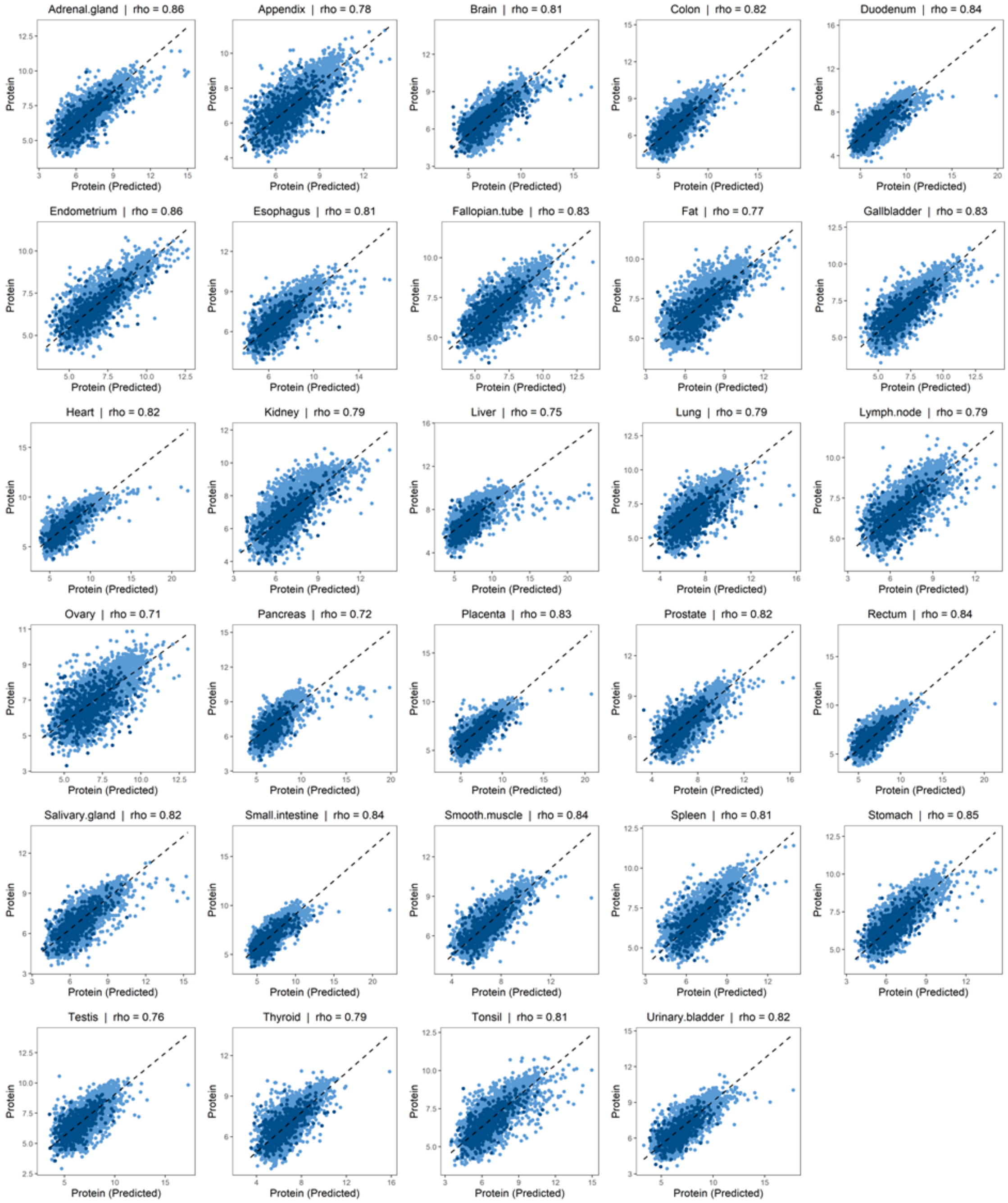
Supplementary figure 1

Protein/RNA Scatterplots for all tissues before correction. Darker points represent cell surface genes.



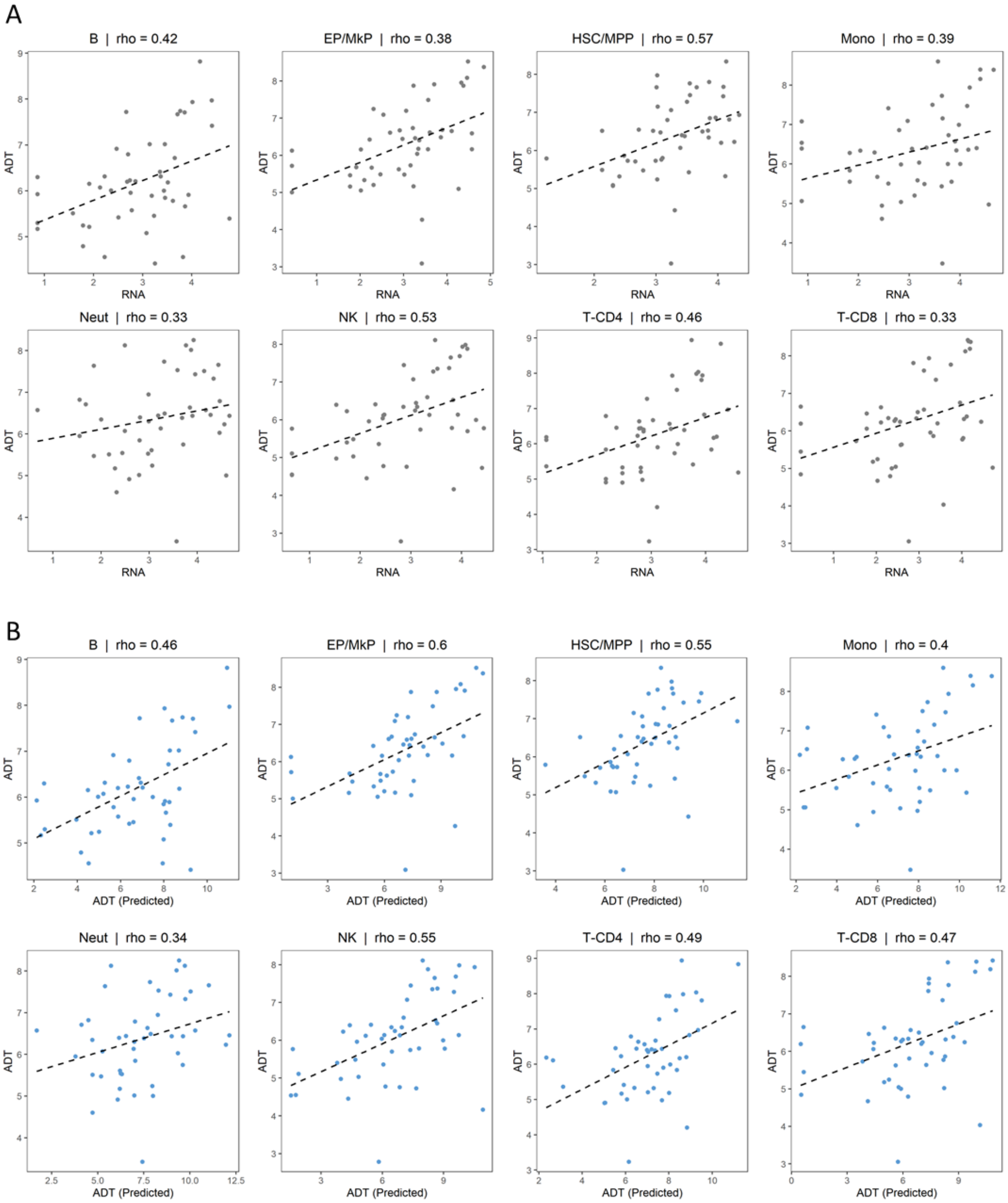
Supplementary figure 2

Protein/RNA Scatterplots for all tissues after correction. Darker points represent cell surface genes.



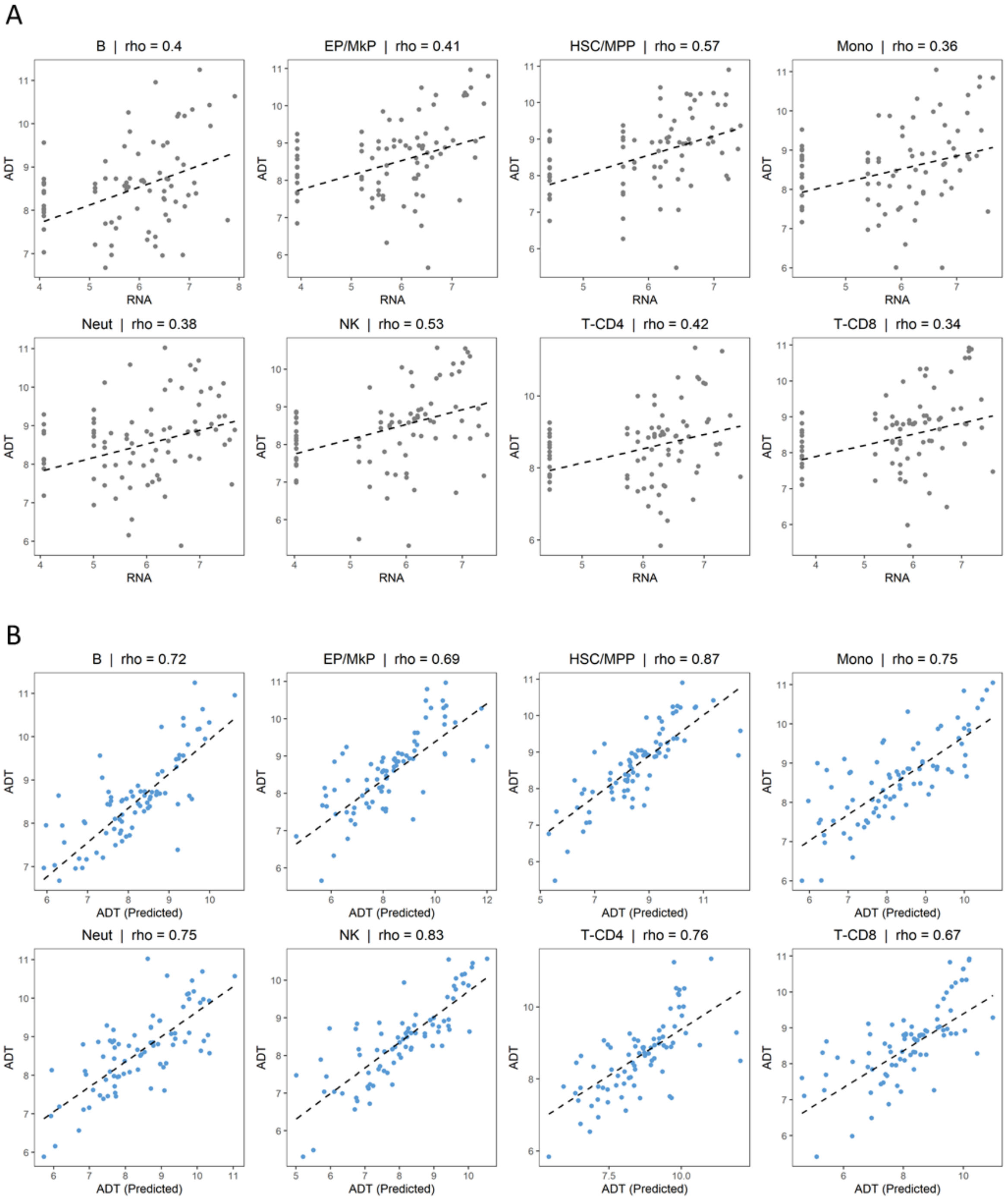
Supplementary figure 3

RNA/ADT scatterplots for all cell types before (A) and after (B) correction with the CF estimated from bulk data.



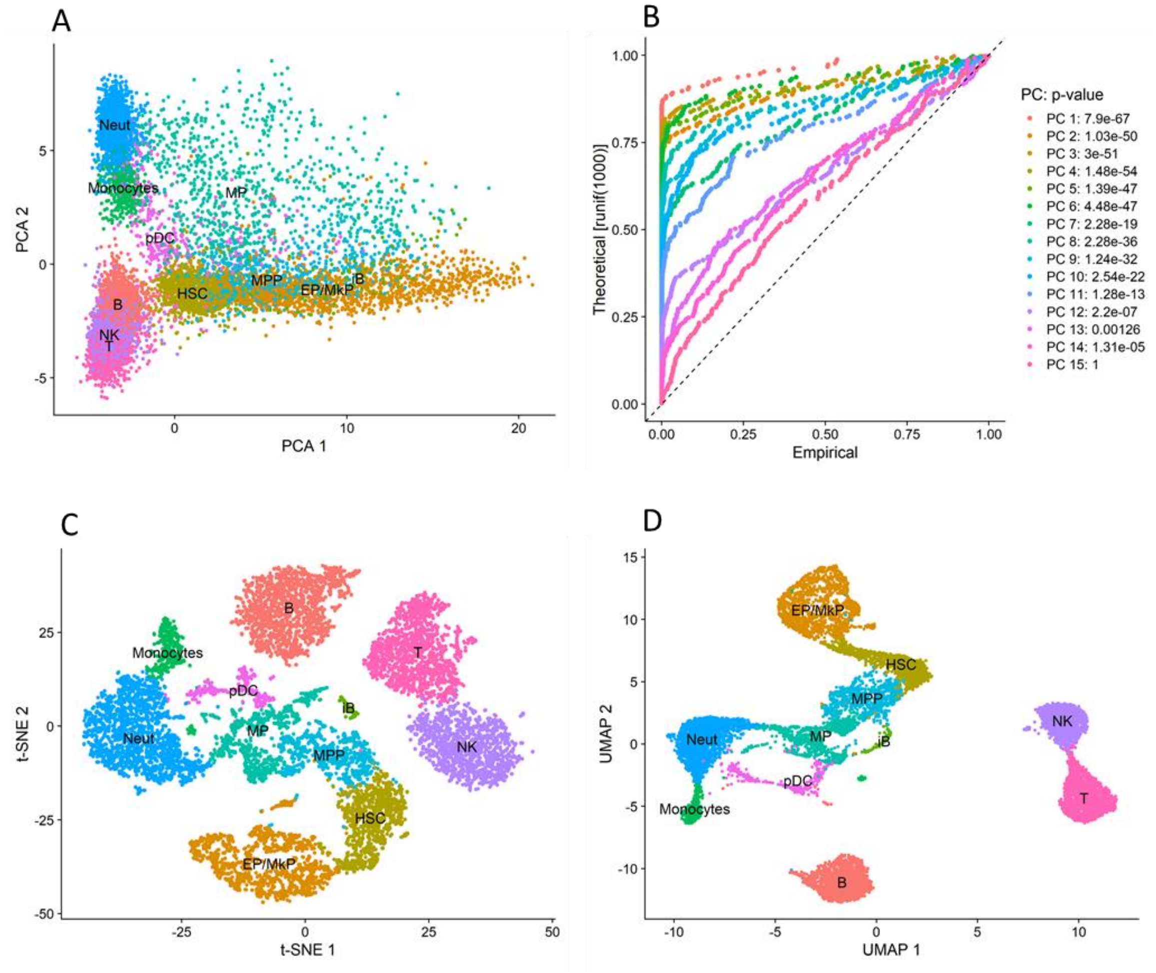
Supplementary figure 4

RNA/ADT scatterplots (testing dataset) for all cell types before (A) and after (B) correction with the CF estimated from single cell data.



Supplementary figure 5

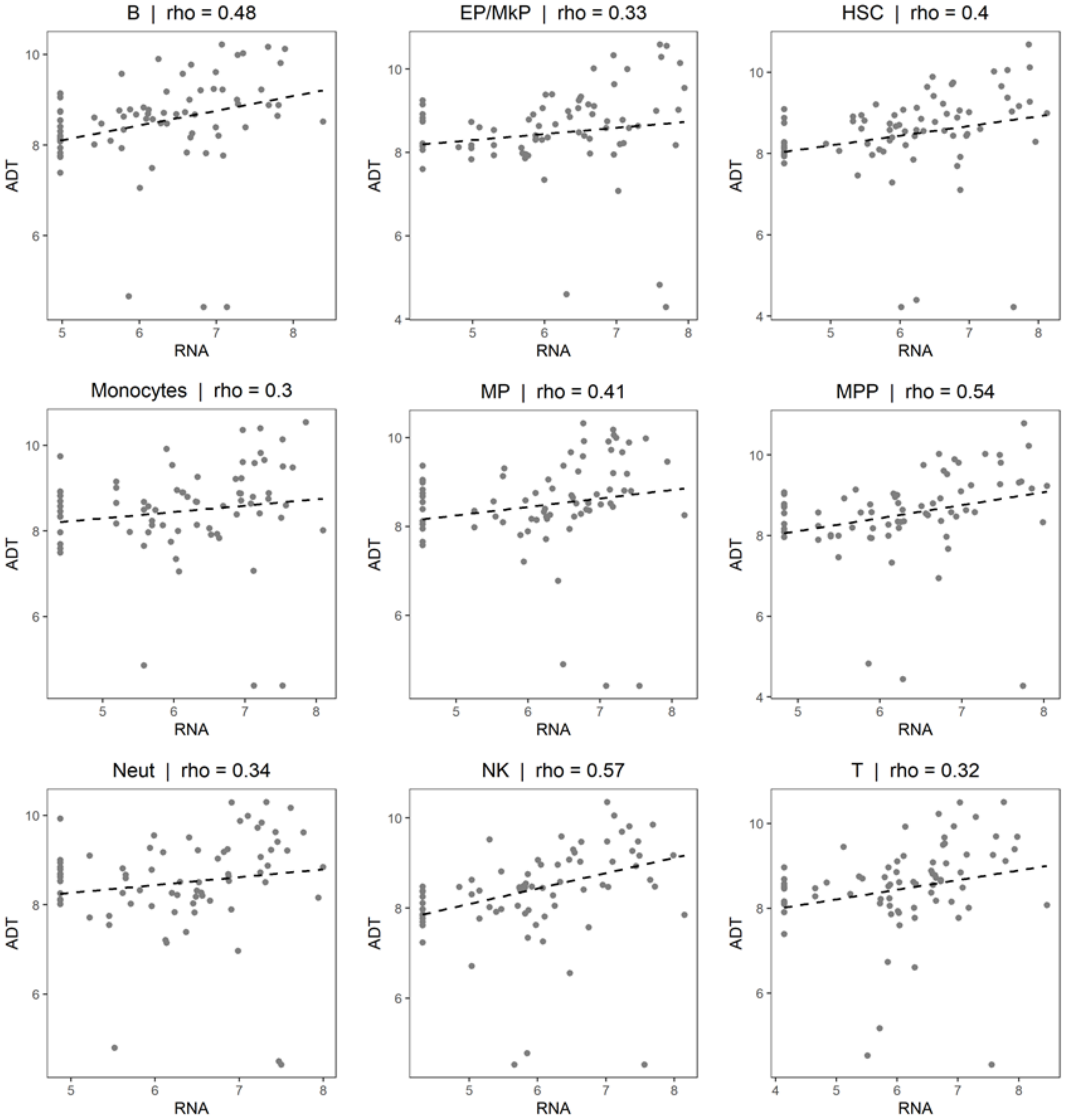
Dimensionality reduction and clustering results of the second scRNA-seq dataset. A, PCA; B, JackStraw permutation results; C, tSNE; D, UMAP.



Neut, Neutrophil, B, B cell; T, T cell; NK, Natural Killer cell; EP/MkP, Erythroid Progenitor/Megakaryocyte Progenitor; HSC, Haematopoietic Stem Cell; MP, Myeloid Progenitor; MPP, Multipotent Progenitor; pDC, plasmacytoid Dendritic Cell; iB, immature B cell.

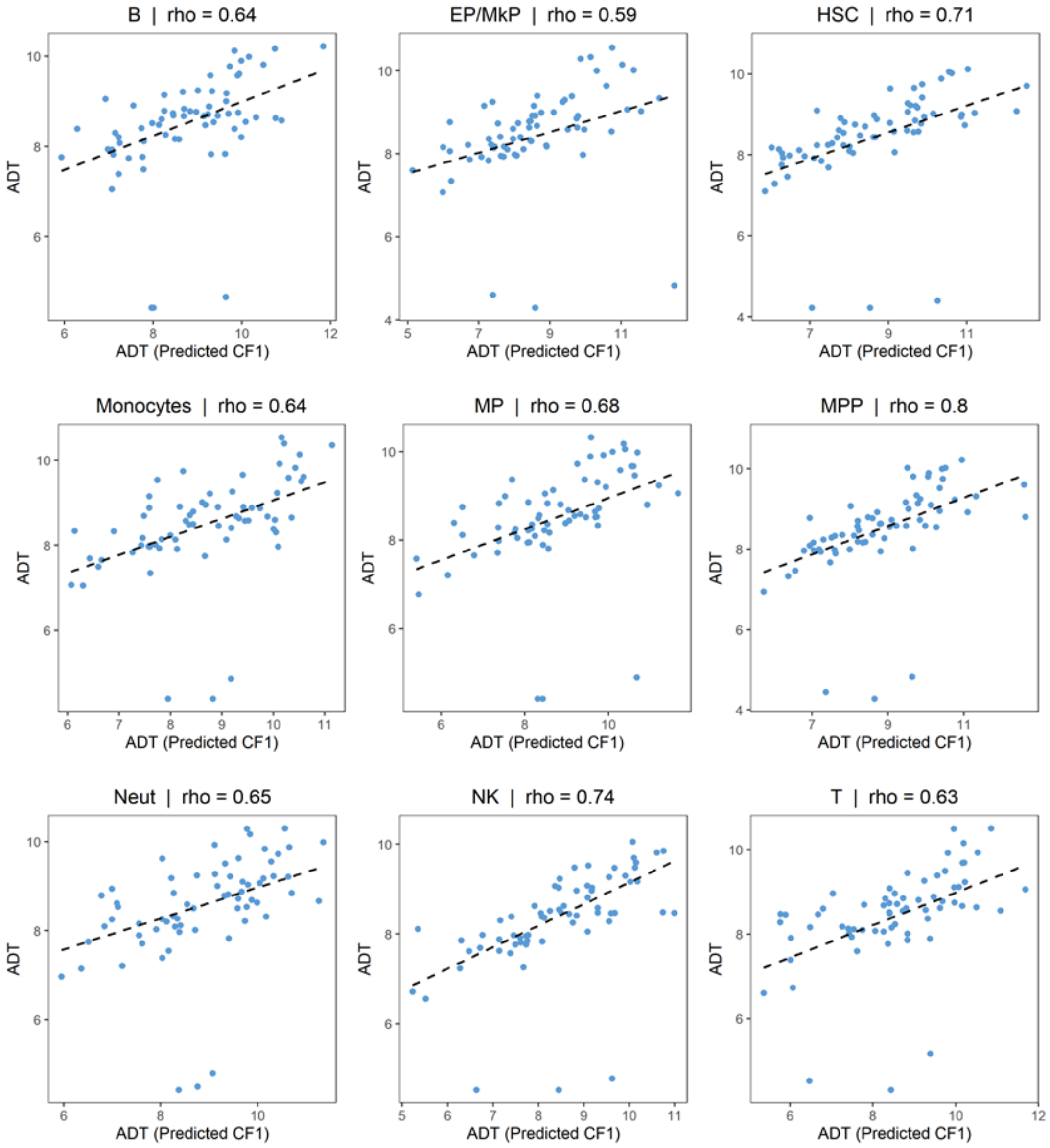
Supplementary figure 6

Uncorrected RNA-ADT scatterplots from pseudo-bulk profiles calculated from the second CITE-seq dataset.



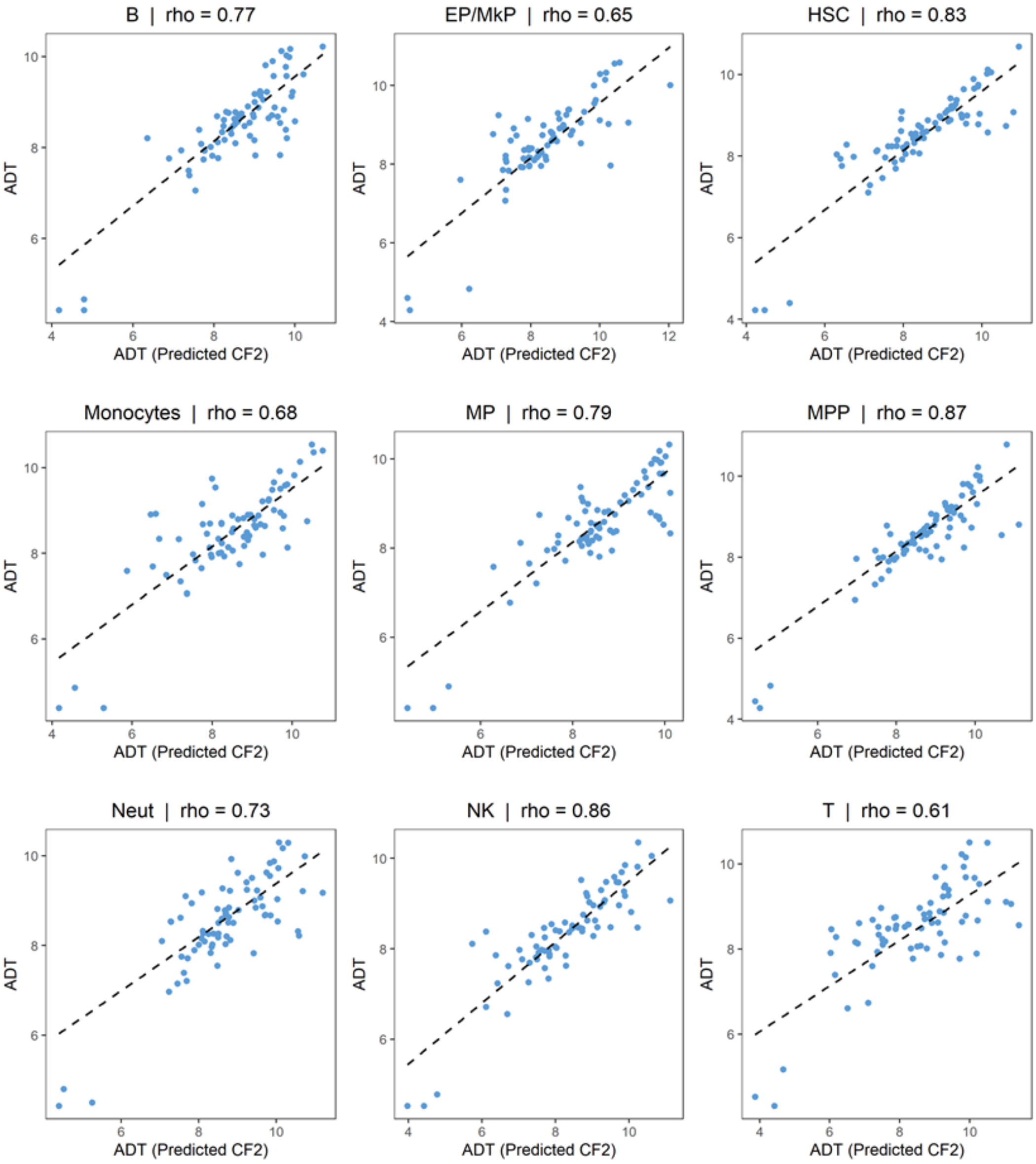
Supplementary figure 7

RNA-ADT scatterplots from pseudo-bulk profiles calculated from the second CITE-seq dataset and corrected with CF1.



Supplementary figure 8

RNA-ADT scatterplots from pseudo-bulk profiles calculated from the second CITE-seq dataset and corrected with CF2.


B | rho = 0.77
EP/MkP | rho = 0.65
HSC | rho = 0.83
Monocytes | rho = 0.68
MP | rho = 0.79
MPP | rho = 0.87
Neut | rho = 0.73
NK | rho = 0.86
T | rho = 0.61
ADT
ADT (Predicted CF2)
4
6
8
10
12