Immunology Letters 245 (2022) 8–17

Contents lists available at ScienceDirect

Immunology Letters

journal homepage: www.elsevier.com/locate/immlet

Statistical and machine learning methods to study human CD4$^+$ T cell proteome profiles

Tomi Suomi [a,*], Laura L. Elo [a,b,*]

[a] Turku Bioscience Centre, University of Turku and Åbo Akademi University, Turku, Finland
[b] Institute of Biomedicine, University of Turku, Turku, Finland

ARTICLE INFO

Keywords:
Bioinformatics
Computational systems biology
Data mining
Machine learning
Proteomics
Mass spectrometry, T cell

ABSTRACT

Mass spectrometry proteomics has become an important part of modern immunology, making major contributions to understanding protein expression levels, subcellular localizations, posttranslational modifications, and interactions in various immune cell populations. New developments in both experimental and computational techniques offer increasing opportunities for exploring the immune system and the molecular mechanisms involved in immune responses. Here, we focus on current computational approaches to infer relevant information from large mass spectrometry based protein profiling datasets, covering the different steps of the analysis from protein identification and quantification to further mining and modelling of the protein abundance data. Additionally, we provide a summary of the key proteome profiling studies on human CD4$^+$ T cells and their different subtypes in health and disease.

1. Introduction

Proteomics has become an important field in modern immunology, providing information about the identity, abundance, localization, modifications and interactions of proteins in different cell populations under different conditions [1]. Although system-wide studies of transcriptomes have provided important insights in immune cell type compositions and cellular signaling networks [2,3], correlation between mRNA and protein levels can vary [4,5]. Proteomics as a field has been driven by the fact that the final product of a gene is inherently more complex and closer to the functionality than the gene itself [6]. These functions are also greatly affected by post-translational modifications, which can be determined only through proteomics. Furthermore, most diseases manifest themselves at the level of protein activity [7] and, therefore, studying proteins holds potential to, for example, help identify new markers for disease diagnosis or new drug targets for disease treatment.

The rapid developments in high-resolution quantitative mass spectrometry have made it a powerful technology to directly study the proteomes at the system level, complementing the other molecular layers of information. Accordingly, mass spectrometry based high-throughput proteome profiling studies have made major contributions to understanding the complex molecular mechanisms in immune responses in both health and disease and identified candidate markers associated with them [8–11].

CD4$^+$ T cells perform important immunoregulatory roles, including activation of B cells, cytotoxic T cells, and macrophages. After activation, CD4$^+$ T cells differentiate into distinct subtypes, which play a key role in the immune response through secretion of cytokines. They also play a critical role in the pathogenesis of many diseases, including infectious, autoimmune and inflammatory diseases, and cancer. Comparisons of human and mouse CD4$^+$ T cells have suggested considerable differences in the protein expression profiles between the two species [12,13], underscoring the importance of human studies [14,15].

Interpretation of the high-throughput mass spectrometry proteome profiling data requires specialized computational tools. These include tools for protein identification and quantification, as well as tools for further mining and modelling of the protein abundances, such as finding significant differences between sample groups or modelling protein regulation and networks. Important steps also include quality control, normalization and possible imputation of missing values, which may have a significant impact on the final outcome of the analysis [16,17]. Here, we provide an overview of the recent developments in the field as outlined in **Fig. 1**. Additionally, we summarise recent applications of proteomics to study human CD4$^+$ T cell proteomes in healthy and disease states (**Table 1**).

* Corresponding authors.
E-mail addresses: tomi.suomi@utu.fi (T. Suomi), laura.elo@utu.fi (L.L. Elo).

https://doi.org/10.1016/j.imlet.2022.03.006
Received 30 November 2021; Received in revised form 11 March 2022; Accepted 15 March 2022
Available online 2 April 2022

2. Quantitative proteome profiling data

Quantitative mass spectrometry methods enable proteome-wide analysis of cellular states [18]. They are used, for instance, to reveal information about molecular composition, protein regulation, and pathways [19], to discover biomarkers as indicators of biological or pathogenic processes [20], and in drug discovery to design compounds that interfere with protein functions [21]. After acquiring the estimated protein abundances, further analysis typically includes additional quality control, normalization, imputation of missing values, differential expression analysis, and various pathway and network enrichment analyses. These steps are summarized in **Fig. 1**. Importantly, today, it has become a common practice to publish the mass spectrometry proteomics datasets in repositories, such as PeptideAtlas [22] or PRIDE [23]. This opens up the possibility of mining and reusing the data in new studies.

2.1. Quantitative analysis of protein abundances using mass spectrometry

The most common approach for mass spectrometry based proteomics is the so-called bottom-up paradigm, where the proteins are first cleaved into peptides before the analysis, for instance, enzymatically using trypsin [1]. The peptides are then separated and ionized for measuring the masses. Integrated liquid-chromatography (LC-MS) systems that are coupled to a mass spectrometer via electrospray ionization are often preferred [1].

A widely used method for obtaining an overall quantitative proteome profile of a sample is label-free *shotgun* proteomics, where the bottom-up proteomics approach is used for identifying proteins from a complex mixture [18]. Alternatively, there are various labeled techniques, where the samples are labeled using tandem mass tags (TMT) [24], chemical labeling (e.g. ICAT [25] or iTRAQ [26]), or metabolically in a cell culture (SILAC) [27]. While these approaches provide means for relative quantification of the proteins, absolute quantification can be achieved, for example, using synthetic peptides (e.g. AQUA [28]).

To identify peptides in complex mixtures, tandem mass spectrometry (MS/MS) involving two separate stages of mass analysis is used. In the first stage of mass spectrometry (MS1 or survey scan), mass-to-charge ratios and intensities of all peptide ions eluting over time are recorded. In the second stage of mass spectrometry (MS2), peptide ions of interest are further fragmented and analyzed to generate the fragment spectra for peptide identification. For this, the mass spectrometry instrument is often operated in the data-dependent acquisition (DDA) mode, where the machine selects and isolates the most intense precursor ions from the MS1 level and fragments them to produce the secondary spectra (i.e. tandem mass spectra, MS2) for peptide identification. This semi-stochastic nature of the selection procedure results in only a proportion of peptides being identified reliably in all samples [29]. To overcome this limitation, the data-independent acquisition (DIA) approach collects MS2 scans systematically over time [30]. For validation, targeted mass spectrometry can be used, including the selected reaction monitoring (SRM) [31], where only those molecular ions that match the mass of a targeted peptide are selected for fragmentation. This allows measuring the specified targets very accurately in all samples. For a more thorough introduction to the mass spectrometry technologies, the reader is referred to e.g. [1,18,32].

2.2. Protein identification and quantification

In mass spectrometry studies, individual peptides are identified by their masses after fragmenting them. A typical approach is to compare the acquired masses against theoretical masses produced computationally from a reference protein sequence database. The protein sequences in the reference database are digested *in silico* based on the expected cleavage sites (e.g. trypsin-specific cleavage sites) and the theoretical masses for the peptides and their fragments are calculated [33,34]. For performing the searches, there are many commercial, free, and open source tools available, such as SEQUEST [35], Mascot [36], OMSSA [37], TANDEM [38], Andromeda [39], Comet [40], and MS-GF+ [41]. Alternatively, peptide identifications can be made without relying on any database using *de novo* techniques that predict the peptide based on the spectral information, including PEAKS [42], pNovo [43], or Novor [44]. Recently, artificial neural networks (e.g. *deep learning*) have gained attraction and they have also been utilized for *de novo* peptide sequencing [45]. While these methods allow the identification of peptides that are not in any database, deriving a sequence solely from a fragment mass spectrum remains challenging and it is strongly

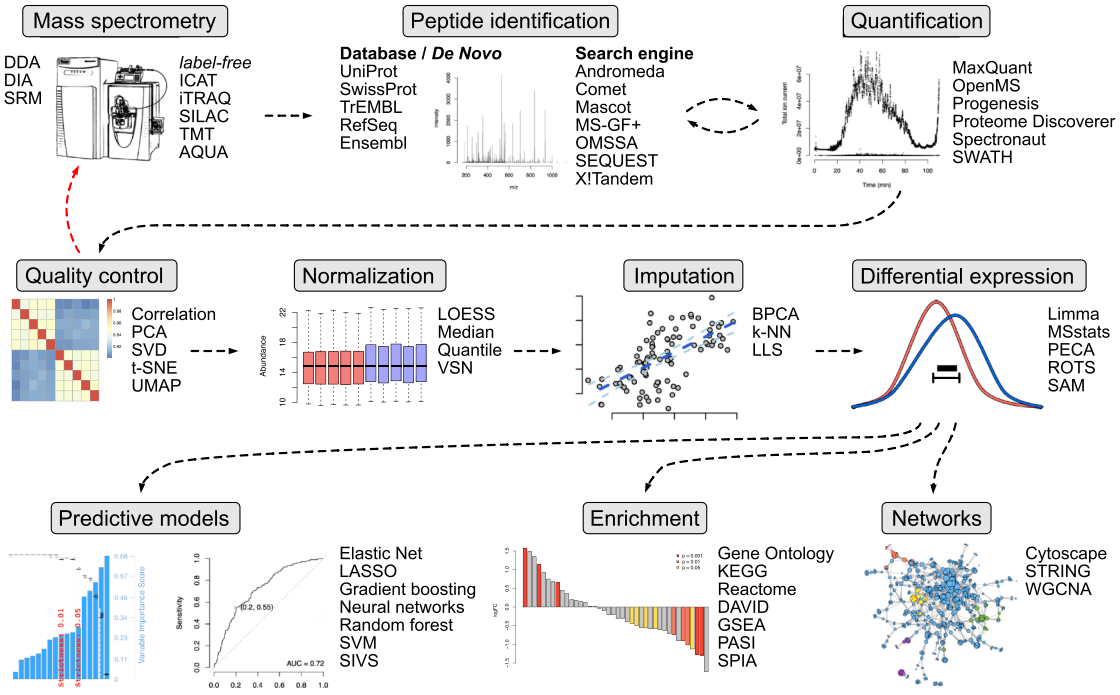

Fig. 1. High-throughput mass spectrometry proteomics workflow to investigate T cell proteome profiles

dependent on the overall spectral quality and accuracy of the instrument [46].

For obtaining the reference protein sequence databases, multiple resources are available, such as RefSeq [47], Ensembl [48], or the commonly used UniProt [49], which contains manually curated entries of SwissProt [50] and automatically added entries of TrEMBL [51]. There are also references for specific purposes, such as UniPept [48] containing unique tryptic peptides for metaproteomics purposes. Beyond the existing databases, proteogenomics can be used to generate a reference directly from DNA- or RNA-sequencing data of the same samples [52].

Quantification of peptide abundances can be done either at precursor (MS1) or fragment ion (MS2) level. At MS1 level, the intensities for precursor ions are measured over time, and the intensity at a particular time corresponds to the peptide abundance [53]. However, the peptides must simultaneously be identified at the MS2 level. At MS2 level, the number of identified peptide spectra can directly be used to estimate the abundances (spectral counting). It is easier to implement, but it typically has a poor signal-to-noise ratio [54,55]. Protein abundances are then estimated by aggregating the peptide abundance values using various roll-up methods, such as mean [56], sum [57], or linear models (e.g. [58, 59]). Since peptides originating from the same protein can behave differently, the use of the peptide-level quantifications have been suggested as a robust alternative for the protein-level values in the downstream analysis [60,61].

2.3. Importance of data pre-processing: normalization and missing values

Thanks to the rapid developments of mass spectrometry based proteomics, even thousands of proteins and their modifications can be quantified in a mass spectrometry experiment [62]. Despite the advances, however, the results are still prone to various biases [63], caused by, for example, instrument calibration and performance, differences in sample preparation or sample temperature [64]. As the underlying reasons for the biases are typically unknown, they cannot be compensated only by adjusting the experimental settings. Therefore, to counter the biases, there are a plethora of different normalization techniques that aim to computationally remove the unwanted technical variations from the data, such as median normalization, quantile normalization [65], linear or local regression normalization [66], and variance stabilization normalization [67]. Some of the methods originate from earlier transcriptomics technologies, while some are more tailored towards mass spectrometry proteomics data. There are also extensive reviews assessing the performance of the various normalization methods [17] and tools to help in decision making [63].

Another common challenge with mass spectrometry based proteomics data is that they traditionally suffer from missing values, which are essentially a result of unrecorded peptides. The missing values can be divided into two main categories; they are either abundance-dependent (i.e. instrument limitations), or the values are missing completely at random (e.g. stemming from the semi-stochastic nature of the acquisition). For imputation purposes, there are plenty of methods to choose from. Besides more simple approaches to replace the missing values with zero or the smallest value found from the data, they also include, for example, a *k*-nearest neighbour approach [68], singular value decomposition imputation [68], local least squares imputation [69], or Bayesian principal component analysis imputation [70]. There are several general reviews in this area [71,72], as well as practical evaluations of the performance of the different imputation approaches in the context of mass spectrometry proteomics data [16,73].

3. Computational approaches to analyze proteome profiling data

3.1. Dimensionality reduction and clustering

After the estimates of peptide or protein abundances are produced, various downstream data analysis can be done. This often starts with exploration of the overall patterns in the data. For this, principal component analysis (PCA) is a commonly used method for

Table 1
Recent proteome profiling studies of human CD4+ T cells.

Table 1A. Proteomic profiles of human CD4+ T cells upon activation

Dataset	Cellular components	Mass spec method	Disease state	Source of sample	Accession
[123]	whole cell	DDA, label-free	healthy	peripheral blood	Supplementary
[12]	whole cell	DDA, label-free	healthy	peripheral blood	PRIDE: PXD015872
[124]	whole cell	DIA	healthy	peripheral blood	PRIDE: PXD019446, PXD019542

Table 1B. Proteomic landscapes of human CD4+ T cell subsets

Dataset	**Cellular components**	**Mass spec method**	**Disease state**	**Source of sample**	**Accession**
[8]	whole cell	DDA, label-free	healthy	peripheral blood	PRIDE: PXD004352
[9]	whole cell	DDA, label-free	healthy	peripheral blood	PRIDE: PDX007745, PDX007744, PXD005477
[129]	whole cell	DDA, TMT	healthy	peripheral blood	PRIDE: PXD015315
[130]	whole cell	DDA, TMT	healthy	peripheral blood	PRIDE: PXD005703
[13]	whole cell	DDA, label-free	healthy	cord blood	PRIDE: PXD008973
[133]	whole cell	DDA, TMT	healthy	peripheral blood	PRIDE: PXD008563

Table 1C. Subcellular proteomes of human CD4+ T cells

Dataset	**Cellular components**	**Mass spec method**	**Disease state**	**Source of sample**	**Accession**
[135]	nucleus	DDA, iTRAQ	healthy	cord blood	Tranche
[136]	cytoplasm	DDA, label-free	healthy	Peripheral blood	
[137]	cytoplasm	DDA, iTRAQ	healthy, ageing	peripheral blood	PRIDE: PXD016039
[138]	membrane	DDA, label-free	healthy	peripheral blood	PRIDE: PXD001432
[139]	cytosol, membrane, nucleus	DDA, iTRAQ	healthy	peripheral blood	PRIDE: PXD000376
[134]	cytosol, membrane, nucleus	DDA, TMT	healthy	peripheral blood	PRIDE: PXD013284

Table 1D. Proteomic landscapes of CD4+ T cells in disease

Dataset	**Cellular components**	**Mass spec method**	**Disease state**	**Source of sample**	**Accession**
[141]	whole cell	DIA	HIV	peripheral blood	PRIDE: PXD005234
[10]	whole cell	DDA, TMT	HIV	peripheral blood	MassIVE: MSV000082229
[142]	whole cell	DDA, TMT	HIV	peripheral blood	PRIDE: PXD012263
[11]	whole cell	DDA, label-free and DIA	type 1 diabetes	peripheral blood	PRIDE: PXD006223, PXD007184
[143]	whole cell	DDA, label-free	multiple sclerosis	peripheral blood	PRIDE: PXD011785
[144]	whole cell	DDA, label-free	Crohn's disease	intestinal biopsies	Supplementary
[145]	whole cell	DDA, label-free	bladder cancer	lymph nodes, peripheral blood	PRIDE: PXD009569

dimensionality reduction. For visualization purposes, it is used for projecting high-dimensional data using only a few principal components to obtain lower-dimensional data. The principal components account as much of the variability in the data as possible, in descending order. Thus, the low-dimensional representation using the first components still preserves as much variability in the data as possible. Other popular alternatives to reduce the dimensionality of high-dimensional data are t-distributed stochastic neighbor embedding (t-SNE), or uniform manifold approximation and projection (UMAP), which allow to reveal grouping of the data points (e.g. samples or proteins) and their relative proximities [74].

Another widely used approach for exploration is clustering, which aims to group the data points into clusters so that the data points within a cluster are more similar than those in the other clusters. The clusters can then be projected on the lower dimensional space, such as those based on t-SNE or UMAP. In the context of proteomics data, perhaps the most common technique for clustering is the agglomerative hierarchical clustering, which allows grouping of the samples or proteins (or both) according to a similarity measure, which is often based on correlation or Euclidean distance. Agglomerative clustering is a bottom-up approach, which starts from singleton clusters and then recursively joins pairs of clusters until all data is in one large cluster. There are different linkage options that determine the order of connections, including complete linkage, average linkage, single linkage, and the Ward's method [75]. The similarities can then be visualized as a dendrogram; a graph where similar samples or proteins are closest to each other. Often, the hierarchical clustering is used together with heatmaps, illustrating the abundances of the proteins and samples of interest.

3.2. Marker discovery

Differential expression analysis is used to find statistically significant differences between sample groups. It is used in search of proteins that would act as markers of, for instance, different cell types, or different disease states. A traditional method to assess the differences has been the Student's t-test of the protein abundances, although it has been shown to be a sub-optimal solution for many high-throughput technologies [76,77], including proteomics [78], and many alternative methods have been proposed (e.g. [79–81]). Since the bottom-up proteomics produces the quantifications at the peptide-level, peptide-centric methods have also been proposed [60,61,82,83], which overcome the challenge of inferring the protein-level abundances. Especially in clinical studies, more complex statistical models may be required to accommodate various clinical or confounding factors, such as generalized linear models (GLM) or linear mixed effects (LME) models. GLM is a generalization of the linear regression to allow the response variable not to be normally distributed (e.g. discrete or categorical variables), whereas LMEs are extensions of the linear models to allow both fixed and random effects to account for dependencies in the data (e.g. longitudinal or hierarchical data).

While the differential expression analysis typically treats each peptide or protein separately, machine learning allows to discover a set of features (e.g. a panel of marker proteins) that together predict the state of interest, such as cell type or disease. A number of machine learning methods are available with different levels of complexity, including gradient boosting [84], support vector machines [85], random forests [86], and neural networks [87]. While the more complex methods hold potential to capture various underlying non-linear dependencies between the proteins, they often suffer from lack of interpretability and tend to be more prone to overfitting if the number of samples is not large enough. Among the simpler models, generalized linear modeling combined with shrinkage methods, such as LASSO [88] or Elastic Net [89], have become commonly used for high-dimensional data. Methods are also available that refine the feature selection to find as small as possible, yet meaningful sets of markers [90]. Common to all the marker detection approaches is that, in addition to computational cross-validation, it is crucial to evaluate the marker panels in large independent sample sets to ensure their generalizability.

3.3. Longitudinal modelling

With the rapid developments of quantitative mass spectrometry proteomics as an established tool, longitudinal and time course experiments have begun to emerge [91,92]. This allows considering individual variability over time. For instance, longitudinal studies are considered to have more statistical power than cross-sectional designs [93]. Several regression based approaches have been applied for longitudinal transcriptomics data, including both linear and non-linear approaches [94–99], whereas fewer examples are available for longitudinal proteomics data [92]. As proteomics data still remains noisy, has a lot of missing values, and often comes with a limited set of replicates, conventional differential expression methods for longitudinal data are sub-optimal, calling for specialized tools [100].

3.4. Pathways, networks and data integration

The results of a proteome profiling study are typically further interpreted using data available in various publicly available or commercial databases, such as Gene Ontology [101], KEGG [102], Reactome [103], ProteinAtlas [104], STRING [105], and MSigDB [106]. Gene set enrichment analysis allows researchers to gain further insights from filtered or ranked lists of proteins. These protein list-based methods can reveal, for instance, biological pathways that are overrepresented in a condition more than would be expected by chance. There are many tools for performing such enrichment analysis, such as GSEA [106] and g:Profiler [107], as well as their visualization, such as Cytoscape [108]. Although the methods based on protein lists have remained the most common approach, specialized tools are available for pathway enrichment analysis that take into account the pathway structures, such as CePa [109], NetGSA [110], and SPIA [111]), which may help determine the pathway activities over the protein list-based methods [112]. While the most common approach is to investigate the enrichment group-wise, techniques are available also for sample-wise enrichment analysis, which allow identification of deregulated pathways sample-by-sample [113,114].

Protein interactions play a central role in biology. Therefore, network analysis provides an interesting opportunity to study the proteome profiles [8,115]. There are several databases that contain known or predicted protein–protein interactions and associations, including one of the earliest and most widely used STRING database [105]. Additionally, networks can be constructed on the basis of protein co-expression levels using, for instance, the weighted correlation network analysis (WGCNA) method [116]. After constructing the networks, their topological properties can be investigated to highlight, for instance, key hub proteins.

The proteomic data can also be integrated with complementary omics datasets, including genomics, transcriptomics, epigenomics, and metabolomics. Excellent reviews exist that discuss the relationship between proteins and mRNAs [117], the use of genomic or transcriptomic data to generate customized protein sequence databases in proteogenomics [118], as well as integrative analysis of proteomics data with other omics data types [119]. While proteomics has started to become an integral part of multi-omics research and considerable progress has been made in the field, the field is still rapidly evolving [120]. Currently, a common approach is to focus on associations between proteomics and the other data types, instead of their quantitative modeling together in an integrative manner, which would hold great potential to provide more efficient utilization of the data to reveal the underlying multi-level relationships. Such integration techniques include, for instance, multi-omics factor analysis [121] and variational autoencoders [122]. In the context of clinical studies, other layers of information may also include, for instance, data from imaging, electronic medical records, or

clinical lab tests.

4. Human CD4$^+$ T cell proteomes

4.1. Proteomic profiles of human CD4$^+$ T cells upon activation

Multiple recent studies have characterized the proteome profiles of human CD4$^+$ T cells, with focus on protein expression changes upon activation of naive CD4$^+$ T cells (**Table 1A**).

[123] used quantitative label-free mass spectrometry to generate dynamic proteome profiles of human primary naive CD4$^+$ T cells from blood of healthy donors immediately after sorting and at multiple time points upon *in vitro* activation (12, 24, 48, 72 and 96 h). Additionally, dynamic metabolome profiles were investigated. The study revealed that intracellular L-arginine was a crucial regulator of the metabolic fitness of the T cells, as well as their survival capacity and anti-tumor activity.

Similarly, [12] used label-free mass spectrometry to investigate proteomic profiles of unactivated and *in vitro* activated (72 h) primary human CD4$^+$ T cells from peripheral blood of two healthy donors. They also compared the profiles to those of the human SUP-T1 and Jurkat T lymphoblast cell lines, suggesting a substantial overlap between the primary CD4+ T cell proteomic profiles and the human lymphoblastic cell lines.

[124] generated a DDA based spectral library of primary human T cells, including both *in vitro* activated and *ex vivo* CD3$^+$ T cell samples from human peripheral blood. The library was then confirmed by analyzing DIA and DDA data on three replicate series of CD4$^+$ T cell samples, including *ex vivo* (0 h) and *in vitro* activated cells (6, 12, 24 and 72 h, 7 d), supporting the utility of the library and the benefits of the DIA approach in protein quantification. The focus of the study was on the generation of the spectral library, while no detailed further analysis of the activation related signal was provided.

While all these three studies involved CD4+ T cells from peripheral blood of healthy donors before and after *in vitro* activation, there were considerable differences both in the mass spectrometry technologies applied as well as in the data analysis methods used, making a direct comparison between the studies difficult. The numbers of quantified proteins varied from 7815 proteins in [123] to 5237 proteins in [12], and 2850 proteins in [124]. Geiger et al. (2016) and Subbannayya et al. (2020) also studied differentially expressed proteins between the activated and unactivated cells and reported 2824 and 1119 proteins after 72 h of activation, respectively. Notably, however, Subbannayya et al. (2020) observed considerable differences in the proteomic profiles of the activated cells between the two donors investigated, with only ~20% of the reported differentially expressed proteins consistent in both donors.

4.2. Proteomic landscapes of human CD4$^+$ T cell subsets

After activation, the naive CD4$^+$ T cells differentiate into various subsets of effector and memory cells guided by the cytokine signals received. Over the years, multiple distinct subsets have been characterised by their cytokine secretion profiles and master transcriptional regulators, including conventional helper T cells Th1, Th2, Th17, regulatory T cells (Tregs), and follicular helper T cells (Tfh) [12]. The memory CD4$^+$ T cells include subsets of central memory, effector memory, and tissue-resident memory T cells [125]. It has also been shown that a substantial portion of the activated CD4$^+$ T cells remain plastic and may be later capable of acquiring other properties [126].

The first proteomic study of activated human primary T helper cells, published in 2001, identified 91 proteins using metabolic labeling, 2-dimensional gel electrophoresis, and MALDI-TOF mass spectrometry [127]. Since then the development of the technologies have enabled considerable increase in the depth of analysis with up to 10,000 proteins identified in a single study (**Table 1B**).

[8] applied label-free mass spectrometry to characterize the cellular proteomes of 28 primary human hematopoietic cell types from peripheral blood of healthy donors, including seven major lineages (granulocytes, monocytes, dendritic cells, natural killer, B cells, CD4 and CD8 lymphocytes), as well as erythrocytes, and platelets. The CD4$^+$ T cell profiles included naive, central memory, and effector memory cells, naive and memory regulatory T cells, and effector T cell subsets Th1, Th2, and Th17. Differential expression analysis confirmed known lineage specific marker proteins. Additionally, Lasso regression was used to reveal previously unknown combinations of cell surface receptors for the different cell types. To study cell type resolved functions, WGCNA was used to identify modules of co-expressed proteins, which were then studied for enrichment of functional properties. Intercellular communication networks were constructed by categorizing proteins as transcription factors, adaptor molecules, receptors, and secreted molecules, and defining intercellular connections using protein interaction data from the STRING database together with the protein expression values to determine pairwise intercellular connection scores.

[9] used label-free mass spectrometry to profile proteomes of five different human CD4$^+$ T cell subsets isolated from peripheral blood of healthy donors, including blood-derived naive and memory conventional CD4$^+$ T cells, naive and effector Treg cells, and a previously incompletely defined CD4$^+$ T cell population that produces effector cytokines despite expressing the Treg specific transcription factor FOXP3. To understand mechanisms that prevent production of effector cytokines by Treg cells and to identify markers that discriminate them, specific protein expression signatures were identified for all and effector Treg cells. The stability of the identified patterns was then confirmed in additional proteome profiling experiments of *in vitro* cultured samples with (24 h) and without activation. Simultaneous analysis of the transcriptomes of the five CD4$^+$ T cell subsets suggested general correlation of the expression levels but also revealed layer specific regulation, highlighting the importance of proteomic analysis for the functional characterization of cell types. Although the Treg signature could not be found from the earlier proteomic dataset of human Tregs with bulk populations [128], the authors could trace it back in the proteomic dataset of human CD4$^+$ T cell subsets [8].

[129] used TMT-labelled mass spectrometry to characterize proteome-wide responses of naive and memory CD4$^+$ T cells to five different cytokine combinations (Th1, Th2, Th17, iTreg, and IFN-β) at 16 hours and 5 days after activation, and compared them to unactivated cells as well as activated cells cultured without cytokines (Th0). The overall profiles showed that the main source of variation was T cell activation, while the activated cells clustered by time and cell type (naive or memory). Similarly, the early changes in protein expression were dominated by T cell activation, compared to the effects of the cytokines. In general, the results suggested that the cytokines acted in a cell type specific manner to induce five cell states in naive CD4$^+$ T cells (Th1, Th2, Th17, iTreg, and IFN-β) and three in memory CD4$^+$ T cells (Th1, Th17/iTreg, and IFN-β, with no detectable Th2 response). In addition to proteomics, the responses were profiled using bulk and single-cell RNA-sequencing, which suggested high correlation between RNA and protein expression. Identification of cell state specific signatures using jointly the protein and RNA expression identified 105 signature genes/proteins across the five cell states in the naïve CD4$^+$ T cells, and 162 signature genes/proteins across the three cell states in the memory CD4$^+$ T cells.

[130] used TMT-labelled mass spectrometry to profile human induced Tregs (iTregs) over time. To enable a broad analysis of universal FOXP3-inducing pathways, two differentiation protocols were considered along with control cells activated without Treg-inducing factors at four time points of differentiation (6, 24 and 48 h, 6 d). Additional controls included unstimulated naive CD4$^+$ T cells (0 h) and naturally occurring Treg cells from the same three healthy donors. In general, the number of differentially expressed proteins related to both activation and iTreg-specific effects increased over time. RNA-sequencing analysis of the same samples suggested considerable concordance between the

genes and proteins detected as differentially expressed but also several differences. This was in line with previous studies showing that RNA levels are not always good predictors of the corresponding protein abundance [131,132]. Integration of the transcriptome and proteome data revealed enrichment of a newly defined iTreg subnetwork for immune disease-associated genes, including many known Treg regulators as well as novel candidates.

[13] used label-free mass spectrometry to study quantitative changes in the cellular proteome of naive human $CD4^+$ T cells derived from umbilical cord blood before and after activation (Th0), and after polarization towards Th17 cells at 24 and 72 hours. Th17 cell specific signature of proteins regulated during early Th17 cell differentiation was determined using reproducibility-optimized statistical testing. Although the majority of the proteome was shared between the activated and Th17 polarized cells at this early stage of differentiation and the temporal changes were dominated by T cell activation, several significant lineage specific changes were identified, involving both previously known and unknown proteins with Th17-related functions. Examination of protein-protein interaction networks indicated coordinated regulation of proteins related to distinct biological processes and cellular pathways during early Th17 cell differentiation. Comparison of the proteomic profiles with corresponding transcriptomic profiles revealed overall high concordance between the two molecular layers. A comparison with corresponding published mouse Th17 proteome data, on the other hand, showed only limited overlap between the two species, highlighting the importance of human studies for translational research.

[133] used TMT-labelled mass spectrometry to explore quantitative proteome profiles of human naive $CD4^+$ T cells derived from peripheral blood of healthy donors together with *in vitro* induced follicular helper T (iTfh) cells after 5 days of polarization. The results revealed biological processes and pathways related to both T cell activation and Tfh cell differentiation.

4.3. Subcellular proteomes of human $CD4^+$ T cells

In addition to analyses of the whole cell proteomes, multiple studies have used mass spectrometry to generate proteome-wide data on subcellular localizations of proteins, which is important for their biological function [134] (**Table 1C**). The studies of subcellular proteomes include, for instance, the study of T cell subproteomes in the nucleus of activated human cord blood $CD4^+$ T cells after activation and polarization towards Th2 cells (6 and 24 h) [135], the study of proteome alterations in the cytoplasmic fraction of human primary $CD4^+$ and $CD8^+$ T cells from peripheral blood of healthy donors unactivated and after activation (24 h) [136], and the study of protein profiles of cytoplasmic extracts of human $CD4^+$ T cells from peripheral blood between young (21-34 years) and older (68–83 years) participants [137]. Additionally, several studies have been conducted to study the surface proteome of human naive $CD4^+$ T cells and their changes upon activation [138], as well as more global mapping of proteins and their translocations between different cellular compartments [134].

Cell surface proteins are crucial in response to other cells or environmental changes. [138] used label-free mass spectrometry to characterize the expression of cell surface proteins of human naive and activated naive $CD4^+$ T cells during the first hours of activation (3, 6, 12, 24 and 48 h). Unsupervised clustering of the proteins grouped them into three clusters according to their dynamic expression profiles, characterized by specific Gene Ontology terms. To extend their *ex vivo* cell surface atlas, the non-targeted mass spectrometry data were complemented with flow cytometry based surface screen of known surface markers, as well as a transcriptomic approach with microarrays. Comparison to corresponding transcriptomic results suggested that around half of the proteins identified with the proteomic approaches could not be found in the transcriptional surface expression data, underscoring the need for proteomic approaches. [139] complemented the surface atlas by [138] with deeper proteomic profiling and by adding comparative analysis of naive and resting memory $CD4^+$ T cells, with the aim to separate transient protein changes during naive $CD4^+$ T cell activation from changes that persist to the resting memory state. In addition to proteomics, they also utilized multiple other high-throughput technologies to derive a comprehensive profile of the two cell types, including whole genome sequencing, methylation arrays, RNA-sequencing, miRNA-sequencing, and phosphoproteomics. In addition to investigating pairwise associations between the different layers of the data, an integrated T cell receptor signaling pathway was constructed by overlaying the measurements onto the pathway.

[134] generated an in-depth subcellular proteomic map of primary human $CD4^+$ T cells using high-resolution isoelectric focusing (HiRIEF) combined with TMT-labelled mass spectrometry. Conventional $CD4^+$ T cells (excluding recently activated and regulatory T cells) from three healthy donors were measured in resting state and upon 15 min and 1 h of activation, fractionated into cytosolic, membrane (including membranous organelles like mitochondria) and nuclear compartments. The study provided a global mapping of the subcellular location of proteins. The proteome-wide identification of translocations of proteins in response to stimulation revealed both known and novel T cell receptor induced translocations. Overall, the study provided wide coverage for the subcellular proteome of $CD4^+$ T cells, with the previous studies mainly limited to profiling a particular subcellular fraction.

4.4. Proteomic landscapes of $CD4^+$ T cells in disease

$CD4^+$ T cells play an important role in the pathogenesis of many diseases, including various infectious, autoimmune, and inflammatory diseases, as well as cancer. Accordingly, an increasing number of studies have been conducted to characterize the proteomes of the different $CD4^+$ T cell subsets in these diseases (**Table 1D**).

$CD4^+$ T cells have a crucial role in the development of HIV infection and the acquired immunodeficiency syndrome (AIDS) pathogenesis, where progressive depletion of $CD4^+$ T cell populations is one of the hallmarks of the disease [140]. [141] used DIA mass spectrometry proteomics to study modulations of the human host cell systems during HIV-1 infection both *in vitro* and *in vivo*. In the *in vitro* experiment, the proteome of $CD4^+$ T cells was quantified over time (0, 12, 24 and 48 h) following HIV-1 infection. In the *in vivo* experiment, paired samples of $CD4^+$ T cells were analyzed from viremic and subsequently successfully treated patients with no detectable viral load. The results revealed a range of changes in the proteome of HIV-infected human $CD4^+$ T cells. Although the overall overlap of the specific changes in the proteome between *in vivo* and *in vitro* infected $CD4^+$ T cells remained low, perturbations in the type 1 interferon signaling pathway were found at both levels. [10] employed TMT-labelled mass spectrometry to analyze *in vitro* activated primary $CD4^+$ T cells infected with GFP-encoding HIV-1 (96 h) from four HIV-1-negative donors. The results suggested that HIV-1 infection activated cellular survival and viability programs, while the *in silico* functional network linkage analysis suggested a central role of a molecular inhibitor of cell apoptosis BIRC5 and its upstream regulator OX40. Similarly, [142] used TMT-labelled mass spectrometry to study protein dynamics of HIV-infected and mock-infected primary human $CD4^+$ T cells (24 and 48 h) from peripheral blood to identify cellular proteins regulated by HIV in its natural target cell.

$CD4^+$ T cells also play an important role in various autoimmune diseases, such as type 1 diabetes. [11] used label-free mass spectrometry to study the proteome profiles of peripheral $CD4^+$ T cells in a pediatric cohort of newly diagnosed type 1 diabetes subjects and their age- and sex-matched healthy controls to identify cellular signatures associated with the onset of the disease. In total, samples from 114 individuals were analyzed using either the DDA or the DIA method. In line with previous studies, considerable heterogeneity was observed between the individuals. However, highly overlapping and statistically significant differences in protein abundances were observed between the type 1 diabetes and control children using both the DIA and the DDA

approaches. In particular, the results revealed an inflammatory signature in children with type 1 diabetes, suggesting an important role of the activation of the innate immune system in the disease onset. [143] used label-free mass spectrometry to analyse proteome profiles of $CD4^+$ and $CD8^+$ T cells purified from whole blood of genotyped 13 newly diagnosed, treatment-naive patients with relapsing-remitting multiple sclerosis (RRMS) and 14 age- and sex-matched healthy controls. The results suggested dysregulation in T cells at the protein level, including several proteomic differences in $CD4^+$ T cells from RRMS patients compared to healthy controls. Additionally, some associations were found between protein expression and genotypes of previously identified risk loci, suggesting potential novel protein expression quantitative trait loci (pQTL).

In addition to $CD4^+$ T cells derived from blood, studies have also been conducted to profile proteomes of $CD4^+$ T cells from other tissues. [144] employed label-free mass spectrometry to study proteomes of tissue-derived Th cell clones, with focus on human Th1 and Th1/Th17 clones derived from intestinal biopsies of patients with Crohn's disease. Major differences were observed between the two phenotypes especially in cytotoxic proteins, which were overrepresented in the Th1 clones. [145] used label-free mass spectrometry to study proteomes of T regulatory cells and $CD4^+$ T effector cells derived from sentinel lymph nodes (SN), non-SNs, and peripheral blood from two patients with muscle-invasive urothelial bladder cancer collected at cystectomy. The results revealed upregulation of growth and immune signalling pathways in the SN-resident Tregs. Furthermore, centrality analysis of the constructed SN-Treg interaction network identified a central role for the cytokine IL-16, suggesting altered IL-16 signalling as a candidate tumour immune escape mechanism.

5. Discussion and future perspectives

Mass spectrometry based proteomics has greatly advanced our understanding of complex biological systems, including various healthy and disease states. Currently, increasingly automated pipelines have been developed for the preprocessing of the mass spectrometry data, including protein identification and quantification. For example, there is an initiative for gathering community-curated bioinformatics pipelines for Nextflow, a popular workflow manager, to promote standardization and reproducibility [146]. Overall, it currently has the largest curated collection of ready-to-use bioinformatics workflows [147]. These include many automated analysis workflows for quantitative mass spectrometry based proteomics. Furthermore, there are for example automated tools for downstream analysis of proteomics data processed by the popular MaxQuant software [148,149]. Similarly, there are automated tools for more special purposes, such as metaproteomics [150,151] or phosphoproteome profiling [152]. However, several issues remain challenging, such as quantification of low-abundant proteins and missing values. Another major challenge is the ability to analyze an adequate number of samples, which is related to the cost and the throughput of the mass spectrometry experiments. To this end, the current developments hold great potential to considerably speed up the analyses and reduce the costs, including sample multiplexing with isobaric tags and additional separation techniques, such as ion mobility spectrometry [153,154].

Statistical and machine learning tools are needed to extract useful information from the proteomics datasets, which are complex and noisy by nature despite the technological developments, including biological noise from individual variation. With the continuous development of the mass spectrometry technology, the associated bioinformatics tools also need continuous development. An important direction is increasing use of artificial intelligence for different tasks from peptide identification to biomarker discovery and data integration [155]. The studies on transcriptomics data support the utility of artificial intelligence in predicting cellular outcomes [156]. Since proteins are more directly linked to the biological functions, they hold great potential to enhance the predictions, as time-resolved data and data from systematic perturbations accumulate. A benefit of machine learning techniques over standard statistical models is the ability to discover unknown patterns from the high-dimensional data, which holds great potential, for instance, to improve prediction of disease or treatment risks. However, the full deployment of machine learning techniques for precision medicine requires further efforts. In terms of the models themselves, interpretability is a key issue. In terms of developing reliable models, the major challenges are the typically limited numbers of individuals considered and lack of independent datasets for validation. This increases the risk of overfitting and poor generalizability of the results to new datasets.

Another emerging future direction is integration of the proteome data with other omics data, such as genomics, transcriptomics, epigenomics and metabolomics. For instance, comparison of proteomics with transcriptomics may enable identification of discordant trends due to involvement of post-transcriptional regulation mechanisms [157]. In addition to identifying associations between the different omics layers, multi-omics integration of the data holds potential to identify more complex multi-omics patterns and also improve the construction of protein signalling networks. For instance, application of the virtual inference of protein activity by enriched regulon (VIPER) algorithm suggested that protein activity can be predicted from gene expression using relationships between transcription factors and their potential targets [158]. Moreover, spatial approaches will preserve the information about the *in vivo* context of the cells. Combining spatial and temporal analysis allows elucidation of global reorganization of proteins in time and space and the involved molecular processes [159].

International and interdisciplinary efforts have facilitated generation of diverse proteomics datasets. Mining such resources in comprehensive meta-analyses hold potential to extract new biologically relevant information from the already existing datasets and suggest new hypotheses for further experimental studies. To this end, standardized approaches and modern data repositories, such as PRIDE, play an important role, including appropriate metadata following controlled vocabulary and minimal standards. Another important area of development are open source software ecosystems for the developed computational tools, such as the R/Bioconductor, supporting reproducibility of research and allowing the research community to utilize the tools widely.

In immunological research, a key question is detailed understanding of the immune cells and their molecular networks, and ability to predict immune reactions. Here, proteomics provides one important layer of information towards precision medicine and new candidate therapies.

Funding

Prof. Elo reports grants from the European Research Council ERC (677943), European Union's Horizon 2020 research and innovation programme (955321), Academy of Finland (296801, 310561, 314443, 329278, 335434, 335611 and 341342), and Sigrid Juselius Foundation during the conduct of the study. Our research is also supported by University of Turku Graduate School (UTUGS), Biocenter Finland, and ELIXIR Finland.

Declaration of Competing Interest

None of the authors have any conflicts of interest to declare.

References

1 R Aebersold, M. Mann, Mass spectrometry-based proteomics, Nature 422 (2003) 198–207.
2 N Novershtern, A Subramanian, LN Lawton, RH Mak, WN Haining, ME McConkey, et al., Densely interconnected transcriptional circuits control cell states in human hematopoiesis, Cell 144 (2011) 296–309.

3 F Paul, Arkin Y 'ara, A Giladi, DA Jaitin, E Kenigsberg, H Keren-Shaul, et al., Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors, Cell 163 (2015) 1663–1677.
4 B Schwanhäusser, D Busse, N Li, G Dittmar, J Schuchhardt, J Wolf, et al., Global quantification of mammalian gene expression control, Nature 473 (2011) 337–342.
5 M Jovanovic, MS Rooney, P Mertins, D Przybylski, N Chevrier, R Satija, et al., Immunogenetics. Dynamic profiling of the protein life cycle in response to pathogens, Science 347 (2015), 1259038.
6 PR Graves, TAJ. Haystead, Molecular biologist's guide to proteomics, Microbiol Mol Biol Rev 66 (2002) 39–63, table of contents.
7 MW Gonzalez, MG. Kann, Chapter 4: Protein interactions and disease, PLoS Comput Biol 8 (2012), e1002819.
8 JC Rieckmann, R Geiger, D Hornburg, T Wolf, K Kveler, D Jarrossay, et al., Social network architecture of human immune cells unveiled by quantitative proteomics, Nat Immunol 18 (2017) 583–593.
9 E Cuadrado, M van den Biggelaar, S de Kivit, Y-Y Chen, M Slot, I Doubal, et al., Proteomic Analyses of Human Regulatory T Cells Reveal Adaptations in Signaling Pathways that, Protect Cellular Identity. Immunity. 48 (2018) 1046–1059, e6.
10 H-H Kuo, R Ahmad, GQ Lee, C Gao, H-R Chen, Z Ouyang, et al., Anti-apoptotic Protein BIRC5 Maintains Survival of HIV-1-Infected CD4 T Cells, Immunity 48 (2018) 1183–1194, e5.
11 MF Lepper, U Ohmayer, C von Toerne, N Maison, A-G Ziegler, SM Hauck, Proteomic Landscape of Patient-Derived CD4+ T Cells in Recent-Onset Type 1 Diabetes, J Proteome Res 17 (2018) 618–634.
12 Y Subbannayya, M Haug, SM Pinto, V Mohanty, HZ Meås, TH Flo, et al., The Proteomic Landscape of Resting and Activated CD4+ T Cells Reveal Insights into Cell Differentiation and Function, Int J Mol Sci 22 (2020), https://doi.org/10.3390/ijms22010275.
13 SK Tripathi, T Välikangas, A Shetty, MM Khan, R Moulder, SD Bhosale, et al., Quantitative Proteomics Reveals the Dynamic Protein Landscape during Initiation of Human Th17 Cell Polarization, iScience 11 (2019) 334–355.
14 J Mestas, CCW. Hughes, Of mice and not men: differences between mouse and human immunology, J Immunol 172 (2004) 2731–2738.
15 IW Mak, N Evaniew, M. Ghert, Lost in translation: animal models and clinical trials in cancer treatment, Am J Transl Res 6 (2014) 114–118.
16 T Välikangas, T Suomi, LL. Elo, A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation, Brief Bioinform 19 (2018) 1344–1355.
17 T Välikangas, T Suomi, LL. Elo, A systematic evaluation of normalization methods in quantitative label-free proteomics, Brief Bioinform 19 (2018) 1–11.
18 H Steen, M. Mann, The ABC's (and XYZ's) of peptide sequencing, Nat Rev Mol Cell Biol 5 (2004) 699–711.
19 BF Cravatt, GM Simon, JR 3rd Yates, The biological impact of mass-spectrometry-based proteomics, Nature 450 (2007) 991–1000.
20 DA Megger, T Bracht, HE Meyer, B. Sitek, Label-free quantification in clinical proteomics, Biochim Biophys Acta 1834 (2013) 1581–1590.
21 M Schirle, M Bantscheff, B. Kuster, Mass spectrometry-based proteomics in preclinical drug discovery, Chem Biol 19 (2012) 72–84.
22 Desiere F. The PeptideAtlas project. Nucleic Acids Research. 2006. pp. D655–D658. doi:10.1093/nar/gkj040.
23 Y Perez-Riverol, A Csordas, J Bai, M Bernal-Llinares, S Hewapathirana, DJ Kundu, et al., The PRIDE database and related tools and resources in 2019: improving support for quantification data, Nucleic Acids Res 47 (2019) D442–D450.
24 A Thompson, J Schäfer, K Kuhn, S Kienle, J Schwarz, G Schmidt, et al., Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS, Anal Chem 75 (2003) 1895–1904.
25 SP Gygi, B Rist, SA Gerber, F Turecek, MH Gelb, R. Aebersold, Quantitative analysis of complex protein mixtures using isotope-coded affinity tags, Nat Biotechnol 17 (1999) 994–999.
26 PL Ross, YN Huang, JN Marchese, B Williamson, K Parker, S Hattan, et al., Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents, Mol Cell Proteomics 3 (2004) 1154–1169.
27 S-E Ong, B Blagoev, I Kratchmarova, DB Kristensen, H Steen, A Pandey, et al., Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics, Mol Cell Proteomics 1 (2002) 376–386.
28 SA Gerber, J Rush, O Stemman, MW Kirschner, SP. Gygi, Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS, Proc Natl Acad Sci U S A 100 (2003) 6940–6945.
29 R Bruderer, OM Bernhardt, T Gandhi, SM Miladinović, L-Y Cheng, S Messner, et al., Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues, Mol Cell Proteomics 14 (2015) 1400–1410.
30 Y Liu, R Hüttenhain, S Surinova, LCJ Gillet, J Mouritsen, R Brunner, et al., Quantitative measurements of N-linked glycoproteins in human plasma by SWATH-MS, Proteomics 13 (2013) 1247–1256.
31 V Lange, P Picotti, B Domon, R. Aebersold, Selected reaction monitoring for quantitative proteomics: a tutorial, Mol Syst Biol 4 (2008) 222.
32 B Domon, R. Aebersold, Mass spectrometry and protein analysis, Science 312 (2006) 212–217.
33 J Zhang, E Gonzalez, T Hestilow, W Haskins, Y. Huang, Review of peak detection algorithms in liquid-chromatography-mass spectrometry, Curr Genomics 10 (2009) 388–401.
34 H. Lam, Building and searching tandem mass spectral libraries for peptide identification, Mol Cell Proteomics 10 (2011). R111.008565.
35 JK Eng, AL McCormack, JR. Yates, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database, J Am Soc Mass Spectrom 5 (1994) 976–989.
36 DN Perkins, DJ Pappin, DM Creasy, JS. Cottrell, Probability-based protein identification by searching sequence databases using mass spectrometry data, Electrophoresis 20 (1999) 3551–3567.
37 LY Geer, SP Markey, JA Kowalak, L Wagner, M Xu, DM Maynard, et al., Open mass spectrometry search algorithm, J Proteome Res 3 (2004) 958–964.
38 R Craig, RC. Beavis, TANDEM: matching proteins with tandem mass spectra, Bioinformatics 20 (2004) 1466–1467.
39 J Cox, N Neuhauser, A Michalski, RA Scheltema, JV Olsen, M. Mann, Andromeda: a peptide search engine integrated into the MaxQuant environment, J Proteome Res 10 (2011) 1794–1805.
40 JK Eng, TA Jahan, MR. Hoopmann, Comet: an open-source MS/MS sequence database search tool, Proteomics 13 (2013) 22–24.
41 S Kim, PA. Pevzner, MS-GF+ makes progress towards a universal database search tool for proteomics, Nat Commun 5 (2014) 5277.
42 B Ma, K Zhang, C Hendrie, C Liang, M Li, A Doherty-Kirby, et al., PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry, Rapid Commun Mass Spectrom 17 (2003) 2337–2342.
43 H Chi, R-X Sun, B Yang, C-Q Song, L-H Wang, C Liu, et al., pNovo: de novo peptide sequencing and identification using HCD spectra, J Proteome Res 9 (2010) 2713–2724.
44 B. Ma, Novor: real-time peptide de novo sequencing software, J Am Soc Mass Spectrom 26 (2015) 1885–1894.
45 NH Tran, R Qiao, L Xin, X Chen, C Liu, X Zhang, et al., Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry, Nat Methods 16 (2019) 63–66.
46 C Hughes, B Ma, GA. Lajoie, De novo sequencing methods in proteomics, Methods Mol Biol 604 (2010) 105–121.
47 NA O'Leary, MW Wright, JR Brister, S Ciufo, D Haddad, R McVeigh, et al., Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation, Nucleic Acids Res 44 (2016) D733–D745.
48 B Mesuere, B Devreese, G Debyser, M Aerts, P Vandamme, P. Dawyndt, Unipept: tryptic peptide-based biodiversity analysis of metaproteome samples, J Proteome Res 11 (2012) 5773–5780.
49 The UniProt Consortium, UniProt: the universal protein knowledgebase, Nucleic Acids Res. 45 (2017) D158–D169.
50 E Gasteiger, E Jung, A. Bairoch, SWISS-PROT: connecting biomolecular knowledge via a protein database, Curr Issues Mol Biol 3 (2001) 47–55.
51 A Bairoch, R. Apweiler, The SWISS-PROT protein sequence data bank and its supplement TrEMBL in 1998, Nucleic Acids Res 26 (1998) 38–42.
52 AI. Nesvizhskii, Proteogenomics: concepts, applications and computational strategies, Nat Methods 11 (2014) 1114–1125.
53 X Wang, W Zhu, K Pradhan, C Ji, Y Ma, OJ Semmes, et al., Feature extraction in the analysis of proteomic mass spectra, Proteomics 6 (2006) 2095–2100.
54 DH Lundgren, S-I Hwang, L Wu, DK. Han, Role of spectral counting in quantitative proteomics, Expert Rev Proteomics 7 (2010) 39–53.
55 H Choi, D Fermin, AI Nesvizhskii, Significance analysis of spectral count data in label-free shotgun proteomics, Mol Cell Proteomics 7 (2008) 2373–2385.
56 F-Y Cheng, K Blackburn, Y-M Lin, MB Goshe, JD. Williamson, Absolute protein quantification by LC/MS(E) for global analysis of salicylic acid-induced plant protein secretion responses, J Proteome Res 8 (2009) 82–93.
57 K Ning, D Fermin, AI Nesvizhskii, Comparative analysis of different label-free mass spectrometry based protein abundance estimates and their correlation with RNA-Seq gene expression data, J Proteome Res 11 (2012) 2261–2271.
58 Y Karpievitch, J Stanley, T Taverner, J Huang, JN Adkins, C Ansong, et al., A statistical framework for protein quantitation in bottom-up MS-based proteomics, Bioinformatics 25 (2009) 2028–2034.
59 T Clough, M Key, I Ott, S Ragg, G Schadow, O. Vitek, Protein quantification in label-free LC-MS experiments, J Proteome Res 8 (2009) 5275–5284.
60 T Suomi, GL Corthals, OS Nevalainen, LL. Elo, Using Peptide-Level Proteomics Data for Detecting Differentially Expressed Proteins, J Proteome Res 14 (2015) 4564–4570.
61 T Suomi, LL. Elo, Enhanced differential expression statistics for data-independent acquisition proteomics, Sci Rep 7 (2017) 5869.
62 F Meissner, M. Mann, Quantitative shotgun proteomics: considerations for a high-quality workflow in immunology, Nat Immunol 15 (2014) 112–117.
63 A Chawade, E Alexandersson, F. Levander, Normalyzer: a tool for rapid evaluation of normalization methods for omics data sets, J Proteome Res 13 (2014) 3114–3120.
64 YV Karpievitch, AR Dabney, RD. Smith, Normalization and missing value imputation for label-free LC-MS analysis, BMC Bioinformatics 13 (Suppl 16) (2012) S5.
65 Amaratunga D, Cabrera J. Analysis of Data From Viral DNA Microchips. Journal of the American Statistical Association. 2001. pp. 1161–1170. doi:10.1198/016214501753381814.
66 SJ Callister, RC Barry, JN Adkins, ET Johnson, W-J Qian, B-JM Webb-Robertson, et al., Normalization approaches for removing systematic biases associated with mass spectrometry and label-free proteomics, J Proteome Res 5 (2006) 277–286.
67 W Huber, A von Heydebreck, H Sültmann, A Poustka, M. Vingron, Variance stabilization applied to microarray data calibration and to the quantification of differential expression, Bioinformatics 18 (Suppl 1) (2002) S96–S104.
68 O Troyanskaya, M Cantor, G Sherlock, P Brown, T Hastie, R Tibshirani, et al., Missing value estimation methods for DNA microarrays, Bioinformatics 17 (2001) 520–525.

69 W Stacklies, H Redestig, M Scholz, D Walther, J Selbig, pcaMethods—a bioconductor package providing PCA methods for incomplete data, Bioinformatics 23 (2007) 1164–1167.
70 S Oba, M-A Sato, I Takemasa, M Monden, K-I Matsubara, S. Ishii, A Bayesian missing value estimation method for gene expression profile data, Bioinformatics 19 (2003) 2088–2096.
71 Lin W-C, Tsai C-F. Missing value imputation: a review and analysis of the literature (2006–2017). Artificial Intelligence Review. 2020. pp. 1487–1509. doi:10.1007/s10462-019-09709-4.
72 M Song, J Greenbaum, J 4th Luttrell, W Zhou, C Wu, H Shen, et al., A Review of Integrative Imputation for Multi-Omics Datasets, Front Genet 11 (2020), 570255.
73 B-JM Webb-Robertson, HK Wiberg, MM Matzke, JN Brown, J Wang, JE McDermott, et al., Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics, J Proteome Res 14 (2015) 1993–2001.
74 McInnes L, Healy J, Saul N, Großberger L. UMAP: Uniform Manifold Approximation and Projection. Journal of Open Source Software. 2018. p. 861. doi:10.21105/joss.00861.
75 Ward Jr. JH Hierarchical grouping to optimize an objective function. J Am Stat Assoc. 1963;58: 236–244.
76 Mukherjee S, Roberts SJ. A theoretical analysis of gene selection. Proc IEEE Comput Syst Bioinform Conf. 2004; 131–141.
77 L-X Qin, KF Kerr, Contributing Members of the Toxicogenomics Research Consortium. Empirical evaluation of data transformations and ranking statistics for microarray analysis, Nucleic Acids Res 32 (2004) 5471–5479.
78 A Pursiheimo, AP Vehmas, S Afzal, T Suomi, T Chand, L Strauss, et al., Optimization of statistical methods impact on quantitative proteomics data, J Proteome Res 14 (2015) 4118–4126.
79 ME Ritchie, B Phipson, D Wu, Y Hu, CW Law, W Shi, et al., limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Res 43 (2015) e47.
80 T Suomi, F Seyednasrollah, MK Jaakkola, T Faux, LL. Elo, ROTS: An R package for reproducibility-optimized statistical testing, PLoS Comput Biol 13 (2017), e1005562.
81 VG Tusher, R Tibshirani, G. Chu, Significance analysis of microarrays applied to the ionizing radiation response, Proc Natl Acad Sci U S A 98 (2001) 5116–5121.
82 M Choi, C-Y Chang, T Clough, D Broudy, T Killeen, B MacLean, et al., MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments, Bioinformatics 30 (2014) 2524–2526.
83 LJE Goeminne, A Argentini, L Martens, L. Clement, Summarization vs Peptide-Based Models in Label-Free Quantitative Proteomics: Performance, Pitfalls, and Data Analysis Guidelines, J Proteome Res 14 (2015) 2457–2465.
84 P Bühlmann, T. Hothorn, Boosting algorithms: Regularization, prediction and model fitting, Stat Sci 22 (2007) 477–505.
85 CJC. Burges, Data Min Knowl Discov 2 (1998) 121–167.
86 L. Breiman, Mach Learn 45 (2001) 5–32.
87 J. Schmidhuber, Deep learning in neural networks: an overview, Neural Netw 61 (2015) 85–117.
88 Tibshirani R. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological). 1996. pp. 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x.
89 Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005. pp. 301–320. doi:10.1111/j.1467-9868.2005.00503.x.
90 Mahmoudian M, Venäläinen MS, Klén R, Elo LL. Stable Iterative Variable Selection. Bioinformatics. 2021. doi:10.1093/bioinformatics/btab501.
91 N Lietzén, L Cheng, R Moulder, H Siljander, E Laajala, T Härkönen, et al., Characterization and non-parametric modeling of the developing serum proteome during infancy and early childhood, Sci Rep 8 (2018) 5883.
92 C-W Liu, L Bramer, B-J Webb-Robertson, K Waugh, MJ Rewers, Q. Zhang, Temporal expression profiling of plasma proteins reveals oxidative stress in early stages of Type 1 Diabetes progression, J Proteomics 172 (2018) 100–110.
93 Z Xu, X Shen, W Pan, Alzheimer's Disease Neuroimaging Initiative. Longitudinal analysis is more powerful than cross-sectional analysis in detecting genetic association with neuroimaging phenotypes, PLoS One 9 (2014), e102312.
94 MJ Aryee, JA Gutiérrez-Pabello, I Kramnik, T Maiti, J. Quackenbush, An improved empirical bayes approach to estimating differential gene expression in microarray time-course data: BETR (Bayesian Estimation of Temporal Regulation), BMC Bioinformatics 10 (2009) 409.
95 YC Tai, TP. Speed, On gene ranking using replicated microarray time course data, Biometrics 65 (2009) 40–51.
96 A Conesa, MJ Nueda, A Ferrer, M. Talón, maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments, Bioinformatics 22 (2006) 1096–1102.
97 GK. Smyth, Linear models and empirical bayes methods for assessing differential expression in microarray experiments, Stat Appl Genet Mol Biol 3 (2004) Article3.
98 MK Kerr, GA. Churchill, Statistical design and the analysis of gene expression microarray data, Genet Res 77 (2001) 123–128.
99 L Cheng, S Ramchandran, T Vatanen, N Lietzén, R Lahesmaa, A Vehtari, et al., An additive Gaussian process regression model for interpretable non-parametric analysis of longitudinal data, Nat Commun 10 (2019) 1798.
100 Valikangas T, Suomi T, Chandler CE, Scott AJ, Tran BQ, Ernst RK, et al. Enhanced longitudinal differential expression detection in proteomics with robust reproducibility optimization regression. bioRxiv. bioRxiv; 2021. doi:10.1101/2021.04.19.440388.
101 M Ashburner, CA Ball, JA Blake, D Botstein, H Butler, JM Cherry, et al., Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, Nat Genet 25 (2000) 25–29.
102 M Kanehisa, S. Goto, KEGG: kyoto encyclopedia of genes and genomes, Nucleic Acids Res 28 (2000) 27–30.
103 B Jassal, L Matthews, G Viteri, C Gong, P Lorente, A Fabregat, et al., The reactome pathway knowledgebase, Nucleic Acids Res 48 (2020) D498–D503.
104 M Uhlén, L Fagerberg, BM Hallström, C Lindskog, P Oksvold, A Mardinoglu, et al., Proteomics. Tissue-based map of the human proteome, Science 347 (2015), 1260419.
105 D Szklarczyk, AL Gable, KC Nastou, D Lyon, R Kirsch, S Pyysalo, et al., The STRING database in 2021: customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets, Nucleic Acids Res 49 (2021) D605–D612.
106 A Subramanian, P Tamayo, VK Mootha, S Mukherjee, BL Ebert, MA Gillette, et al., Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles, Proc Natl Acad Sci U S A 102 (2005) 15545–15550.
107 U Raudvere, L Kolberg, I Kuzmin, T Arak, P Adler, H Peterson, et al., g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update), Nucleic Acids Res 47 (2019) W191–W198.
108 P Shannon, A Markiel, O Ozier, NS Baliga, JT Wang, D Ramage, et al., Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res 13 (2003) 2498–2504.
109 Z Gu, J Liu, K Cao, J Zhang, J. Wang, Centrality-based pathway enrichment: a systematic approach for finding significant pathways dominated by key genes, BMC Syst Biol 6 (2012) 56.
110 A Shojaie, G. Michailidis, Network enrichment analysis in complex experiments, Stat Appl Genet Mol Biol 9 (2010). Article22.
111 AL Tarca, S Draghici, P Khatri, SS Hassan, P Mittal, J-S Kim, et al., A novel signaling pathway impact analysis, Bioinformatics 25 (2009) 75–82.
112 MK Jaakkola, LL. Elo, Empirical comparison of structure-based pathway methods, Brief Bioinform 17 (2016) 336–345.
113 Y Drier, M Sheffer, E. Domany, Pathway-based personalized analysis of cancer, Proc Natl Acad Sci U S A 110 (2013) 6388–6393.
114 MK Jaakkola, AJ McGlinchey, R Klén, LL. Elo, PASI: A novel pathway method to identify delicate group effects, PLoS One 13 (2018), e0199991.
115 JK Huang, DE Carlin, MK Yu, W Zhang, JF Kreisberg, P Tamayo, et al., Systematic Evaluation of Molecular Networks for Discovery of Disease Genes, Cell Syst 6 (2018) 484–495, e5.
116 B Zhang, S. Horvath, A general framework for weighted gene co-expression network analysis, Stat Appl Genet Mol Biol 4 (2005). Article17.
117 Y Liu, A Beyer, R. Aebersold, On the Dependency of Cellular Protein Levels on mRNA Abundance, Cell 165 (2016) 535–550.
118 Ruggles KV, Krug K, Wang X, Clauser KR, Wang J, Payne SH, et al. Methods, Tools and Current Perspectives in Proteogenomics. Molecular & Cellular Proteomics. 2017. pp. 959–981. doi:10.1074/mcp.mr117.000024.
119 B Vitrinel, HWL Koh, F Mujgan Kar, S Maity, J Rendleman, H Choi, et al., Exploiting Interdata Relationships in Next-generation Proteomics Analysis, Mol Cell Proteomics 18 (2019) S5–S14.
120 B Zhang, B. Kuster, Proteomics Is Not an Island: Multi-omics Integration Is the Key to Understanding Biological Systems, Mol Cell Proteomics 18 (2019) S1–S4.
121 R Argelaguet, B Velten, D Arnol, S Dietrich, T Zenz, JC Marioni, et al., Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets, Mol Syst Biol 14 (2018) e8124.
122 N Simidjievski, C Bodnar, I Tariq, P Scherer, H Andres Terre, Z Shams, et al., Variational Autoencoders for Cancer Data Integration: Design Principles and Computational Practice, Front Genet 10 (2019) 1205.
123 R Geiger, JC Rieckmann, T Wolf, C Basso, Y Feng, T Fuhrer, et al., L-Arginine Modulates T Cell Metabolism and Enhances Survival and Anti-tumor Activity, Cell. 167 (2016) 829–842, e13.
124 H Weerakoon, J Potriquet, AK Shah, S Reed, B Jayakody, C Kapil, et al., A primary human T-cell spectral library to facilitate large scale quantitative T-cell proteomics, Sci Data 7 (2020) 412.
125 BV Kumar, TJ Connors, DL. Farber, Human T Cell Development, Localization, and Function throughout Life, Immunity 48 (2018) 202–213.
126 M DuPage, JA. Bluestone, Harnessing the plasticity of CD4(+) T cells to treat immune-mediated disease, Nat Rev Immunol 16 (2016) 149–163.
127 TA Nyman, A Rosengren, S Syyrakki, TP Pellinen, K Rautajoki, R Lahesmaa, A proteome database of human primary T helper cells, Electrophoresis 22 (2001) 4375–4382.
128 C Procaccini, F Carbone, D Di Silvestre, F Brambilla, V De Rosa, M Galgani, et al., The Proteomic Landscape of Human Ex Vivo Regulatory and Conventional T Cells Reveals Specific Metabolic Requirements, Immunity 44 (2016) 406–421.
129 E Cano-Gamez, B Soskic, TI Roumeliotis, E So, DJ Smyth, M Baldrighi, et al., Single-cell transcriptomics identifies an effectorness gradient shaping the response of CD4 T cells to cytokines, Nat Commun 11 (2020) 1801.
130 A Schmidt, F Marabita, NA Kiani, CC Gross, HJ Johansson, S Éliás, et al., Time-resolved transcriptome and proteome landscape of human regulatory T cell (Treg) differentiation reveals novel regulators of FOXP3, BMC Biol 16 (2018) 47.
131 M Yang, F Petralia, Z Li, H Li, W Ma, X Song, et al., Community Assessment of the Predictability of Cancer Protein and Phosphoprotein Levels from Genomics and Transcriptomics, Cell Syst. 11 (2020) 186–195, e9.
132 N Fortelny, CM Overall, P Pavlidis, GVC. Freue, Can we predict protein from mRNA levels? Nature 547 (2017) E19–E20.

133 M Zhao, S Jia, X Gao, H Qiu, R Wu, H Wu, et al., Comparative Analysis of Global Proteome and Lysine Acetylome Between Naive CD4 T Cells and CD4 T Follicular Helper Cells, Front Immunol 12 (2021), 643441.

134 RN Joshi, C Stadler, R Lehmann, J Lehtiö, J Tegnér, A Schmidt, et al., TcellSubC: An Atlas of the Subcellular Proteome of Human T Cells, Front Immunol. 10 (2019) 2708.

135 R Moulder, T Lönnberg, LL Elo, J-J Filén, E Rainio, G Corthals, et al., Quantitative proteomics analysis of the nuclear fraction of human CD4+ cells in the early phases of IL-4-induced Th2 differentiation, Mol Cell Proteomics 9 (2010) 1937–1953.

136 MC Gerner, L Niederstaetter, L Ziegler, A Bileck, A Slany, L Janker, et al., Proteome Analysis Reveals Distinct Mitochondrial Functions Linked to Interferon Response Patterns in Activated CD4+ and CD8+ T Cells, Front Pharmacol 10 (2019) 727.

137 A Bektas, SH Schurman, M Gonzalez-Freire, CA Dunn, AK Singh, F Macian, et al., Age-associated changes in human CD4 T cells point to mitochondrial dysfunction consequent to impaired autophagy, Aging 11 (2019) 9234–9263.

138 A Graessel, SM Hauck, C von Toerne, E Kloppmann, T Goldberg, H Koppensteiner, et al., A Combined Omics Approach to Generate the Surface Atlas of Human Naive CD4+ T Cells during Early T-Cell Receptor Activation, Mol Cell Proteomics 14 (2015) 2085–2102.

139 CJ Mitchell, D Getnet, M-S Kim, SS Manda, P Kumar, T-C Huang, et al., A multi-omic analysis of human naïve CD4+ T cells, BMC Syst Biol 9 (2015) 75.

140 AA Okoye, LJ. Picker, CD4(+) T-cell depletion in HIV infection: mechanisms of immunological failure, Immunol Rev 254 (2013) 54–64.

141 J Nemeth, V Vongrad, KJ Metzner, VP Strouvelle, R Weber, P Pedrioli, et al., and Proteome Analysis of Human Immunodeficiency Virus (HIV)-1-infected, Human CD4 T Cells, Mol Cell Proteomics 16 (2017) S108–S123.

142 A Naamati, JC Williamson, EJ Greenwood, S Marelli, PJ Lehner, NJ. Matheson, Functional proteomic atlas of HIV infection in primary human CD4+ T cells, Elife 8 (2019), https://doi.org/10.7554/eLife.41431.

143 T Berge, A Eriksson, IS Brorson, EA Høgestøl, P Berg-Hansen, A Døskeland, et al., Quantitative proteomic analyses of CD4 and CD8 T cells reveal differentially expressed proteins in multiple sclerosis patients and healthy controls, Clin Proteomics 16 (2019) 19.

144 T Riaz, LM Sollid, I Olsen, GA. de Souza, Quantitative Proteomics of Gut-Derived Th1 and Th1/Th17 Clones Reveal the Presence of CD28+ NKG2D- Th1 Cytotoxic CD4+ T cells, Mol Cell Proteomics 15 (2016) 1007–1016.

145 D Krantz, M Mints, M Winerdal, K Riklund, D Rutishauser, R Zubarev, et al., IL-16 processing in sentinel node regulatory T cells is a factor in bladder cancer immunity, Scand J Immunol 92 (2020) e12926.

146 PA Ewels, A Peltzer, S Fillinger, H Patel, J Alneberg, A Wilm, et al., The nf-core framework for community-curated bioinformatics pipelines, Nat Biotechnol 38 (2020) 276–278.

147 L Wratten, A Wilm, J. Göke, Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers, Nat Methods 18 (2021) 1161–1168.

148 JL Gallant, T Heunis, SL Sampson, W. Bitter, ProVision: a web-based platform for rapid analysis of proteomics data processed by MaxQuant, Bioinformatics 36 (2020) 4965–4967.

149 AD Shah, RJA Goode, C Huang, DR Powell, RB. Schittenhelm, LFQ-Analyst: An Easy-To-Use Interactive Web Platform To Analyze and Visualize Label-Free Proteomics Data Preprocessed with MaxQuant, J Proteome Res 19 (2020) 204–211.

150 K Cheng, Z Ning, X Zhang, L Li, B Liao, J Mayne, et al., MetaLab: an automated pipeline for metaproteomic data analysis, Microbiome 5 (2017) 157.

151 J Aakko, S Pietilä, T Suomi, M Mahmoudian, R Toivonen, P Kouvonen, et al., Data-Independent Acquisition Mass Spectrometry in Metaproteomics of Gut Microbiota-Implementation and Computational Analysis, J Proteome Res 19 (2020) 432–436.

152 Y Hong, D Flinkman, T Suomi, S Pietilä, P James, E Coffey, et al., PhosPiR: an automated phosphoproteomic pipeline in R, Brief Bioinform 23 (2022), https://doi.org/10.1093/bib/bbab510.

153 SP Couvillion, Y Zhu, G Nagy, JN Adkins, C Ansong, RS Renslow, et al., New mass spectrometry technologies contributing towards comprehensive and high throughput omics analyses of single cells, Analyst 144 (2019) 794–807.

154 J Li, JG Van Vranken, L Pontano Vaites, DK Schweppe, EL Huttlin, C Etienne, et al., TMTpro reagents: a set of isobaric labeling mass tags enables simultaneous proteome-wide measurements across 16 samples, Nat Methods 17 (2020) 399–404.

155 M Mann, C Kumar, W-F Zeng, MT. Strauss, Artificial intelligence for proteomics and biomarker discovery, Cell Syst 12 (2021) 759–770.

156 M Lotfollahi, FA Wolf, FJ. Theis, scGen predicts single-cell perturbation responses, Nat Methods 16 (2019) 715–721.

157 MS Robles, J Cox, M. Mann, In-vivo quantitative proteomics reveals a key contribution of post-transcriptional mechanisms to the circadian regulation of liver metabolism, PLoS Genet 10 (2014), e1004047.

158 MJ Alvarez, Y Shen, FM Giorgi, A Lachmann, BB Ding, BH Ye, et al., Functional characterization of somatic mutations in cancer using network-based inference of protein activity, Nat Genet 48 (2016) 838–847.

159 PM Jean Beltran, RA Mathias, IM Cristea, A Portrait of the Human Organelle Proteome In Space and Time during Cytomegalovirus Infection, Cell Syst 3 (2016) 361–373, e6.