Review

Transcription factor proteomics—Tools, applications, and challenges

Jovan Simicevic[1,2]* and Bart Deplancke[1]

[1] Laboratory of Systems Biology and Genetics, Institute of Bioengineering, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), and Swiss Institute of Bioinformatics, Lausanne, Switzerland
[2] LimmaTech Biologics AG, Schlieren, Switzerland

Transcription factors (TFs) are a family of DNA-binding proteins whose gene regulatory capabilities are of vital importance in defining the molecular state of a cell. Despite their biological significance, our understanding of TF behavior and function is still limited. This is because we have so far mostly relied on gene expression data to approximate TF protein levels given that the latter information has been notoriously difficult to obtain due to the relatively low expression levels of many TFs. However, significant advances in mass spectrometry technologies combined with the development of sensitive methodologies aimed at detecting TFs are now allowing a transition from a predominantly qualitative to a quantitative protein landscape. Such a paradigm shift is expected to unravel dynamic aspects of TF function, potentially linking TF copy number fluctuations in cells with specific regulatory functions. This in turn may provide novel insights into the regulatory mechanisms underlying a wide range of fundamental and disease-related biological processes. In this review, we will present the latest advances in mass spectrometry-based TF proteomics and describe novel strategies tailored around the quantification of this important family of DNA-binding proteins.

Received: July 29, 2016
Revised: October 18, 2016
Accepted: November 11, 2016

Keywords:
Enrichment / Mass spectrometry / Quantification / Technology / Transcription factor

1 Introduction (transcription factors and their regulatory properties)

Understanding how the expression of genes is regulated is of fundamental importance in biology. This is because the vast majority of biological processes, from development to homeostasis maintenance, from cell cycle to cell differentiation, are tuned by differential gene expression. Due to significant advances in genomic, proteomic, and other molecular technologies, we are achieving an increasingly detailed view of the gene regulatory networks (GRNs) that control gene expression [1, 2]. GRNs capture the physical and functional interactions between DNA-binding regulatory proteins, transcription factors (TFs), and regulatory elements associated with their target genes (i.e. promoters, enhancers). The key function of these GRNs is to coordinate the establishment of distinct molecular states both in space and time. However, most GRNs are still vastly incomplete and their components far from being completely characterized. This data paucity significantly hampers our ability to model these networks and infer how the regulation of specific genes is orchestrated. To attain such a detailed level of understanding, novel approaches will need to be developed and data across a wide range of methods will need to be integrated.

2 Inferring transcription factor regulatory properties from their abundance in cells

Although qualitative information regarding the regulatory behavior of TFs is widely available in the literature [3–5], reliable TF protein measurements are much scarcer. This disparity can in part be explained by the fact that TFs tend not to

Correspondence: Dr. Bart Deplancke, Laboratory of Systems Biology and Genetics, Institute of Bioengineering, School of Life Sciences, Ecole Polytechnique Fédérale de Lausanne (EPFL), and Swiss Institute of Bioinformatics, CH-1015 Lausanne, Switzerland
E-mail: bart.deplancke@epfl.ch

Abbreviations: GRN, gene regulator network; **TF**, transcription factor

*Additional corresponding author: Dr. Jovan Simicevic
E-mail: jovan.simicevic@lmtbio.com

be highly abundant in cells [6–10], which makes quantitative analyses a substantial challenge considering the complexity of eukaryotic proteomes and the current state of protein analytical technologies. This is despite the importance of retrieving accurate, quantitative protein level data to generate in silico models of specific regulatory mechanisms. Indeed, the copy number of TFs is a key aspect of gene regulation given that the ability of a TF to bind to DNA and thus to exert its regulatory function is in part determined by its nuclear abundance [11–13]. Neglecting to incorporate such quantitative data in gene regulatory models or to simply infer it using gene expression data introduces therefore biases [14–17] and leads to incomplete understanding of regulatory mechanisms. Thus, TF levels need to be experimentally measured to determine how much of a binding site will be occupied by a TF to assess or model its regulatory input.

Moreover, interest in obtaining quantitative data on TFs derives not only from a fundamental interest in understanding how they regulate gene expression, but also from a biomedical perspective. This is because aberrations in TF levels and function are responsible for a variety of disorders e.g. [18,19]. It is therefore not surprising that there has always been a great interest in this particular family of proteins as potential pharmaceutical targets to treat a variety of diseases, including cancer [20–23], even though TFs are admittedly rather difficult to target [24].

Various techniques are available for the quantification of proteins nowadays, the majority of which rely on the use of antibodies (e.g. ELISA, protein microarrays [25]). Quantitative immunoassays are widely used because of their accuracy and the fact that they can be implemented even in a basic laboratory setting due to their low cost and simplicity. Nevertheless, these techniques suffer from a certain number of drawbacks, non-linearity in quantitation, and formation of unspecific reactions to name a few [26]. Moreover, quantitation of different proteins normally necessitates separate experiments, not to mention that only a limited number of TF-specific antibodies are commercially available, limiting thereby the applicability of these methodologies to small scale studies of several TFs at the time. In this regard, there is a strong need for robust methodologies that could bypass such limitations, possibly pushing the current limits in terms of sensitivity and specificity even further. In this review, we will describe the tools and methods that are currently available to study TFs from a quantitative standpoint utilizing mass spectrometry, giving much attention to strategies aimed at enriching these molecules from their complex native environment to ease their detectability. We will thereby mostly focus on studies introducing the approaches rather than on those applying them. As such, we will at first enumerate the most adopted techniques to detect and subsequently quantify DNA-binding proteins, and describe the most recent improvements in the isolation and enrichment of TFs. Subsequently, the different quantification approaches will be described in greater detail along with practical examples from the recent literature. Here, we will introduce the concepts of relative as well as absolute quantification, and describe in which circumstances the use of one over the other would be more appropriate. A separate section will be devoted to targeted mass spectrometry-based approaches, as the recent resurrection of such techniques has not only proven beneficial in the detection of medium-to-low abundant molecules due to their inherent high sensitivity and specificity, but also because targeted approaches are amenable to throughput increase and automation. Note that the concept and characteristics of each approach have been summarized in Table 1.

3 Mass spectrometry-based proteomics as a tool for transcription factor analysis

Until recently, mass spectrometry-based methodologies simply lacked the necessary sensitivity to be used for the identification of low abundant proteins. Noticeable improvements in mass spectrometer detection limits have enhanced our quantitative analysis capacities of proteins that are expressed at medium to low levels in cells, such as TFs. In essence, one can segregate MS-based quantification methodologies into two major classes: those requiring stable isotope labeling and those that do not necessitate labeling, so called "label-free".

(1) Label-free approaches: such methodologies are cost efficient, amicable to up-scaling and can be based on MS2 or MS1 level data [27,28]. Spectral counting label-free quantification uses the number of acquired MS2 spectra per protein as a proxy for protein abundance. Although easy to implement, spectra count based quantification is to be considered as semi-quantitative since it provides limited accuracy and precision. Better quantification properties in terms of dynamic range, accuracy and precision, can in this regard be achieved by using MS1-based label-free quantification. Here, dedicated software is used to extract peptide ion intensities that are subsequently summarized to protein level abundance values. Recently, considerable effort has been devoted to overcome such limitations (for an evaluation of label-free quantification methods, we refer to [29]), aiming to combine the benefits of label-free analyses with the sensitivity of targeted MS approaches.

(2) Labeling approaches: the majority of these techniques employ metabolic (SILAC or Stable Isotope Labeling of Amino Acids in Cell Culture) or variations of chemical stable isotope labeling to introduce a predictable mass shift between peptides from two or more experimental conditions (these approaches have been extensively reviewed in [30, 31]). Stable isotope-labeling quantification entails the use of "heavy"-labeled molecules, such as AQUA [32], in which selected chemically synthesized isotope-labeled peptides are carefully quantified and used as standards. "Heavy"-to-"light" peptide ratios define the amount of the endogenous protein present within the biological sample. Issues with the cost of these peptides and with its storage have pushed for an amelioration of the

Table 1. Overview of recent TF proteomics studies

MS approach	Enrichment/ separation	Quantification method	Type of measurement	Advantages	Disadvantages	Applications	Authors
Shotgun	Fluorophore-tagged protein immunoprecipitation	Label-free	Relative	The single-step affinity purification protocol is fast and relatively easy to perform. The data analysis by label-free quantification (LFQ) requires an inexpensive and relatively simple experimental setup.	*In-vitro* environment, measurement accuracy and reproducibility.	A sensitive, quantitative MS-based proteomics procedure to determine the composition of (plant) protein complexes.	[37]
	CatTFRE pull-down	Label-free (validation by SILAC)	Relative	High-throughput. The data analysis by label-free quantification (LFQ) requires an inexpensive and relatively simple experimental setup.	*In-vitro* environment. Creation of artificial TF binding sites due to DNA linkers and tandem copies of TFREs, measurement accuracy and reproducibility.	The catTFRE methodology enables high-throughput identification and quantification of DNA binding activity of TFs.	[39]
	TF pull-down with biotin-tagged oligonucleotides+ on bead digestion	Direct dimethyl labeling / label-free (validation)	Absolute quantification (by external standard)	Simple, high-throughput workflow. On bead digestion of proteins is low cost, reproducible, and does not require the integration of restriction sites or the use of expensive desthiobiotin tags. The direct dimethyl labeling implemented reduces sample handling, leading to higher reproducibility and time saving	*In-vitro* environment	A straightforward mass spectrometry-based DNA pulldown method that permits the study of protein-DNA interactions in a high-throughput manner (suitable for use with primary material as a protein source).	[40]
	TF pull-down with biotin-tagged oligonucleotides+ on bead digestion	Dimethyl labeling	Relative	Simple, high-throughput workflow. On bead digestion of proteins is low cost, reproducible, and does not require the integration of restriction sites or the use of expensive desthiobiotin tags. The direct dimethyl labeling implemented reduces sample handling, leading to higher reproducibility and time saving	*In-vitro* environment.	A description of proteome-wide, mutation-specific binding at recurrent, oncogenic TERT promoter mutations.	[41]
	Nascent Chromatin Capture (NCC)	SILAC	Relative	High enrichment of chromatin factors, easy to couple to other technologies.	*In-vitro* environment, limited to cell lines.	A methodology for profiling the dynamics of the chromatin proteome and histone marks during replication. The NCC strategy combined with quantitative mass spectrometry is a powerful tool to address how the epigenomic framework is maintained in dividing cells.	[43]
	Organelle purification (nuclei), SDS-PAGE gel	Label-free	Relative	Increased protein coverage due to a workflow based on a combination of bioinformatics tools.	Sub-optimal peptide recovery from gel. Measurement accuracy and reproducibility.	An approach with an improved analytical workflow for analyzing proteome datasets in conditions of low database representation.	[46]

Table 1. Continued

MS approach	Enrichment/ separation	Quantification method	Type of measurement	Advantages	Disadvantages	Applications	Authors
	2D-DIGE	2D-DIGE	Relative	The approach alllows for a large protein mass range. The large amount of proteins can be analyzed.	Relatively expensive. Image analysis can be labor-intensive. Large and hydrophobic proteins in the first dimension separation can be difficult to separate by electrophoresis. Extremely acidic and basic proteins are not well represented. Reproducibility is an issue; the technique requires a high level of laboratory skills, and is of low throughput.	A strategy based on 2D-DIGE coupled to MS to study early cold-regulated proteins in rice seedlings.	[47]
	-	SILAC-PrEST	Absolute/ relative	The approach is accurate (protein standards provide higher measurement accuracy when compared to peptide standards) and amicable to multiplexing. The PrEST standards do not need to be purified extensively. An inherent advantage of the ABP-PrEST fusion is that expression and purification of insoluble proteins are facilitated.	Labor-intensive, as it requires cloning procedures.	The SILAC-PrEST combination allows accurate and streamlined quantification of the absolute or relative amount of proteins of interest in a wide variety of applications.	[48]
SRM	SDS-PAGE	Isotope-labeled full-length protein standards	Absolute	Sensitive, reproducible and accurate (protein standards provide higher measurement accuracy when compared to peptide standards). Medium-throughput. Relatively inexpensive. Quick production of protein standards.	Labor-intensive, as it requires cloning procedures. Sub-optimal peptide recovery from gel. Inability to detect proteins/peptides that are not targeted.	A sensitive SRM-based proteomic assay that allows us to determine copy numbers per cell of up to ten proteins simultaneously.	[11]
	Isolation of nuclei / SCX	MS-QBiC	Absolute	Sensitive, reproducible and accurate. Multiplexed preparation of isotopically labeled peptides for use as internal standards.	Labor-intensive, as it requires cloning procedures. Peptide-construct digestion efficiency may not mirror the one from the native protein. Inability to detect proteins/peptides that are not targeted.	A systems-level analysis of dynamics of core circadian clock proteins by SRM-based targeted proteomics.	[56]
	-	Stable isotope dilution (SID)	Relative	Sensitive, reproducible and accurate.	Inability to detect proteins/peptides that are not targeted.	Application of stable isotope dilution as a quantitative technique to determine concentrations of low abundant TFs in the innate immune response.	[57]

Table 1. Continued

MS approach	Enrichment/ separation	Quantification method	Type of measurement	Advantages	Disadvantages	Applications	Authors
	-	Stable isotope-labeled standard peptides	Absolute	Sensitive, reproducible and accurate. Integrative approach.	Inability to detect proteins/peptides that are not targeted.	Methodology aimed at identifying lung cancer-related TFs by integrating previously reported genomic, transcriptomic, and proteomic data. Quantification was achieved by SRM in various cell lines.	[58]
	Isolation of nuclei	Stable isotope-labeled standard peptides	Relative	Sensitive, reproducible and accurate.	Inability to detect proteins/peptides that are not targeted.	A comprehensive approach based on computational modeling, quantitative mass spectrometry, and single-cell microscopy was used to show that cell-to-cell variability in protein abundance acts within a network of positive feedbacks that allow pre-adipocytes to differentiate at very low rates.	[59]
	Concatenated tandem array of TF response elements	SILAC	Relative	Sensitive, reproducible and accurate.	Inability to detect proteins/peptides that are not targeted.	A strategy based on SRM for the systematic measurement of proteins with known or suspected roles in transcriptional regulation at RNA polymerase II transcribed promoters in *Saccharomyces cerevisiae*.	[61]

technique, in which concatenated peptides (QconCAT) [33] are utilized. One of the criticisms though regarding QconCAT constructs revolves around the digestion of its tryptic peptides, which does not necessarily mirror the digestion of endogenous proteins. For this particular reason, many laboratories have oriented their methodologies towards the expression of full-length proteins, expressed either in vivo (e.g. Absolute SILAC) [34] or in vitro (PSAQ) [35]), which are spiked at some stage of sample preparation within the complex mixture. The main advantage is that all tryptic peptides generated from the protein, except the C-terminal one, can be readily monitored. This application tries to overcome issues related to peptide detection, because a large fraction of protein-specific peptides may not be identifiable due to sample complexity, solubility, and ionization issues. Furthermore, by selecting only a small subset of peptides, the methodology is more sensitive to post-translational modification (PTM). Methods based on full-length protein expression also allow for a more accurate quantification and a more robust statistical assessment, while precipitation issues related to peptide storage are systematically bypassed. It is generally agreed that the spiking of the labeled standard should be introduced as early as possible during the sample fractionation steps to secure that both the standard and its endogenous counterpart are subject to the same artifacts. Therefore, sample losses or differential proteolytic treatments that may affect downstream measurements are minimized.

4 Enrichment for TFs is crucial to permit their detection and subsequently their quantification

Within a eukaryotic cell, TFs tend to spend considerable time in the nucleus where they act in conjunction with other molecules to activate or repress the expression of their target genes. That is why the vast majority of proteomic TF studies focus on the nuclear fraction, even though several members of this family of regulatory proteins are also located on the nuclear envelope (e.g. nuclear receptors) or in the cytoplasm. The nuclear environment is deemed very complex, consisting of a large number of factors that are entangled within a chromatin meshwork, which may hamper the full recovery of relevant factors using conventional extraction procedures. Due to the fact that TFs tend to be lower expressed than other types of proteins, and considering the limited levels of sensitivity permitted by MS-based approaches in detecting low abundance molecules, several studies have opted for approaches aimed at isolating and enriching for TFs using specifically designed DNA motifs. Although the endogenous context in which these TFs operate in vivo is now disrupted, TFs can be retrieved in sufficient amounts by their affinity to their respective consensus DNA sequences to allow DNA-protein interaction studies and derive quantitative information in the process. Although we have previously reviewed techniques aimed at studying DNA-protein interactions in greater extent [36], a few interesting novel methodologies based on TF purification have recently been implemented (e.g. [37, 38]). In this context, improvements in DNA pull down-based affinity strategies have permitted to reliably identify and subsequently quantify consensus sequence-specific protein interactors. Ding and colleagues [39] developed an affinity reagent composed of synthetic DNA incorporating concatenated tandem array of consensus TFREs (TF response elements) termed catTFRe for a vast number of TF families. This procedure was shown effective in enriching TFs from cells in an in vitro context in a high throughput setting. The authors reported the identification of as high as 400 TFs from a single cell line and 878 TFs from 11 cell types in total. Their approach is also compatible with the quantification of proteome-wide changes in DNA-binding profiles of cellular TFs in response to specific stimuli or perturbations by utilizing a label-free strategy. Along the same lines, Hubner and co-workers [40] introduced a high throughput compatible DNA-based system that takes advantage of on-bead digestion for DNA pull-downs combined with label-free as well as direct dimethyl labeling protein quantification approaches. Using this approach, the researchers were able to isolate and quantify nearly 7000 nuclear proteins in K562 and PBMC cells utilizing an external universal protein standard mix. As such, their approach is intended to ease laboratory procedures and increase the throughput of DNA-protein interaction proteomics. A similar workflow based on on-bead protein digestion of DNA pulldowns coupled to dimethyl labeling was implemented by Makowski and colleagues [41] to explore variable TF binding at the human telomerase reverse transcriptase promoter region. In contrast with the previous study however, the variable nature of TF binding upon promoter mutations was elucidated by deriving relative measurements.

Although affinity TF purification strategies have the advantage of allowing the enrichment of selected TFs through the use of their DNA consensus elements, such methods operate predominantly in an in vitro setting and are DNA-centric. However, isolating regulatory proteins directly from their native environment is, as already indicated, not a trivial task. Nevertheless, identifying and quantifying TFs in an endogenous context would alleviate biases that are introduced when employing synthetic molecular components. Cross-linking methodologies have in this context become an interesting option to enrich for transiently bound chromatin-associated proteins. This is because chromatin enrichment methodologies can be easily coupled to various strategies aimed at identifying and quantifying chromatin bound proteins in general. Recently, Kustatscher and colleagues [42] introduced such a Chromatin Enrichment for Proteomics (ChEP) strategy. This relatively straightforward biochemical approach based on interphase chromatin enrichment enabled a comprehensive investigation of global chromatin composition and its changes [43], as opposed to previously developed strategies that either focus on specific chromatin loci or have more limited specificity.

Gel-based methodologies are also an interesting alternative to selectively enrich for TFs of interest. We have previously reported the development of a multiplexed MS-based approach aimed at measuring selected TFs directly from cells which utilizes SDS-PAGE to isolate TFs of interest according to their molecular weight [11]. As only the fraction of the gel containing the bands of the proteins of interest is considered, a significant increase in sensitivity can be achieved. This in turn overcomes issues related to the high abundance of specific nuclear components (e.g. histones) which in a mass spectrometer can obscure signals of their less abundant counterparts. This methodology has allowed us to enrich for selected TFs and derive quantitative data along a process of cellular differentiation [11].

In sum, what is currently clear is that without a proper enrichment strategy, only the most abundant factors will be detected, thus scratching only the surface of the cellular TF proteome. Different methods are available to the researcher that can aid in enriching for DNA-binding proteins. The method that will ultimately be selected depends largely on the scope of the study. Nevertheless, it is important to emphasize that it remains challenging to efficiently enrich or extract chromatin-associated proteins under conditions that are compatible with mass spectrometry, independent of the chosen assay [44]. One important consequence is the elevated risk for underestimating protein amounts when dealing with TFs. Proteomic measurements are therefore best complemented with other technologies to corroborate the generated data [e.g. 11], and thus to possibly correct for the effects of a suboptimal nuclear protein extraction.

5 Relative versus absolute quantification in discovery-based studies

Relative quantification is achieved when comparing protein amounts between at least two different samples (e.g. healthy vs. disease, wild type vs. mutant) or conditions of the same samples (e.g. perturbation or time-course analyses). These applications therefore allow the capture of temporal changes and comparison among proteomes in a straightforward manner, and have become in the past decade a gold standard in proteomics. Although TFs do occasionally get caught in the fishing net of large-scale proteomics studies, their detectability with MS-based methods remains rather low. Even with the most sensitive instrumentation available, discovery-based comparative proteomics studies aimed at studying TF dynamics have to be carefully designed and the methodology needs to be tailored to its scope. For example, using a five-plexed SILAC-based MS approach, Molina and co-workers [45] identified 882 nuclear and secreted proteins (but only a few TFs) at five different time-points of adipogenesis. For about half of them, relative quantitative measurements were obtained. In addition, Pascual and colleagues [46] utilized a combination of mass spectrometry state-of-the-art methods based on the use of organelle purification followed by gel and label-free approaches that were coupled to novel bioinformatics methods. Using this approach, these researchers studied variations in the nuclear proteome resulting from UV time-dependent irradiation in *Pinus radiata* enabling them to identify 33 TFs among 388 nuclear proteins related to stress responsive mechanisms. This in turn allowed them to characterize these proteins as potential biomarker candidates for breeding programs destined to improve UV resistance, and thus productivity in forest species. Their effort is an elegant example of how a proteomics-centered integrative strategy can be used to gain a deeper understanding of nuclear protein dynamics. Remaining in plants, Huo and co-workers [47] utilized two dimensional difference gel electrophoresis (2D-DIGE) coupled to MS to study the early regulatory events that orchestrate the cold response in rice. The researchers identified gel spots that reflected altered protein expression levels upon cold treatment. Only five minutes after cold exposure, the abundance levels of 26 proteins were significantly altered. Additional evidence suggested that one of these 26 proteins, OsPLDα1, has a role in early cold-regulated cellular responses, consistent with its earlier established function in transducing cold signaling in rice. Thus, a combination of gel-based as well as LC-MS/MS approaches allowed the identification of a key cold stress-response protein.

Although the vast majority of studies aiming to quantitatively characterize TF amounts are comparative in nature, the necessity for researchers to obtain accurate protein measurements has spurred an increasing interest to develop methodologies aimed at quantifying fractions of the proteome in absolute amounts. Despite such efforts, absolute quantification remains rather challenging from a technical perspective compared to relative (comparative) quantification. Nevertheless, recent improvements in sensitivity and throughput have allowed for a more routine implementation of absolute quantification methodologies. An elegant implementation of the absolute SILAC methodology, SILAC-PrEST [48], used a solubilization tag to quantify in absolute terms the amount of recombinant PrESTs (Protein Epitope Signature Tags) produced in vivo to quantify 40 selected proteins in HeLa cells. One of the benefits of the utilized ABP (Albumin Binding Protein) solubilization tag is that most of its tryptic peptides can be used for quantification, increasing the overall robustness of the approach. Additionally, the two steps of quantification of the recombinant PrESTs and of their endogenous counterparts were collapsed in one single experiment, simplifying the workflow as a whole. Using this approach, Zeiler and co-workers [48] were able to accurately quantify TFs such as proto-oncogene c-Fos at approximately 5–6000 copies per cell, and Zfp828 at approximately 70–75 000 copies per cell.

6 Towards the use of targeted proteomic approaches: SRM/MRM

Novel applications in quantitative proteomics and advances in MS technology development are now permitting to accurately measure protein amounts in a given mixture in absolute terms. One technique in particular, named Selected Reaction Monitoring (SRM), also known as Multiple Reaction Monitoring (MRM), has become a benchmark in targeted proteomics approaches, since it allows the detection and quantification of predetermined sets of proteins, based on selected peptide fragmentation reactions, in complex samples with previously never achieved sensitivity and specificity. This in turn allows for a more in-depth analysis of the proteome, particularly those proteins that are expressed at levels that tend to obscure detection with canonical MS approaches. The most important aspect of SRM, when it comes to quantification, is the consistency and the uniqueness of the selected peptides. Only peptides that uniquely identify a protein of interest, and that are consistently detected in different MS runs should be utilized; such peptides are termed "proteotypic." Moreover, when selecting such peptides, one has to be careful in selecting the highest responding peptides for each protein of interest. There is no *gold standard* for the identification of such peptides. Nevertheless, several bioinformatic tools that guide the user in the selection of proteotypic peptide candidates based on a set of physicochemical properties are currently available [49–52]. The best responding peptides are usually selected for validation. In recent years, SRM coupled to stable-isotope labeling techniques has been adopted for estimating cellular protein levels in large-scale proteomic analyses. Most of such efforts were aimed at quantifying a large fraction of the proteome, covering the largest possible dynamic range [53]. As the complexity of the model system studied increases, mainly due to technical limitations, venturing in the lower levels of the expression range where many TFs reside quickly becomes very challenging. In this regard, few large scale proteomics studies have attempted to quantify TFs identified in a discovery-based mode utilizing a targeted approach. An example of such a strategy is the study of Beck and colleagues [54], who estimated that the least abundant TFs detected in their study using a human cell line were present at less than 500 copies per cell. TF-focused efforts would therefore be greatly aided by the availability of TF peptide-specific information since such data would not only enable improvements in protein identification speed and accuracy, but also ameliorate cross-comparisons of quantitative proteomics data and allow for a more efficient development of targeted proteomics assays. However, to date, no comprehensive TF proteotypic peptide database has been developed. Nevertheless, an interesting advance was introduced by Stergachis and colleagues [9]. These researchers developed a high-throughput, cost-effective methodology for the discovery of optimal precursor- and fragment- ions that can be utilized in targeted proteomics assays based on the use of in vitro-synthesized full-length proteins. Absolute quantification of in vitro-expressed TFs is accomplished via two GST (Glutathione-S-transferase) signature peptides. Using their approach, optimal transitions for 96 human TFs were experimentally derived, after which the expression and enrichment of 44 of these TFs was empirically tested and verified. The utility of the derived ion transitions to quantify endogenous TFs was tested by measuring the relative abundance of six candidates between four human cell lines. To also address this evident lack of TF peptide data in public repositories, we generated a relatively large, experimentally derived TF proteotypic peptide spectral library dataset based on in vitro-expressed TFs using a high-yield Gateway-compatible protein expression system [55]. Our library currently contains peptide information for 89 TFs and this number is set to increase in the near future. Such an effort enabled us to utilize TF-specific peptide information to develop a sensitive SRM-based mass spectrometry assay to quantify TFs in absolute terms. Using this assay, we were able to simultaneously determine the copy numbers of ten predetermined TFs. We subsequently applied the methodology to profile the absolute levels of pro-adipogenic TFs, including the master regulators PPARγ and RXRα, over the course of terminal differentiation of mouse 3T3-L1 pre-adipocytes. Our analyses revealed that the individual abundance of TFs differs dramatically (from ~250 to >300 000 copies per nucleus) and that their dynamic range during differentiation can vary up to fivefold. Using these data, we were able to formulate a highly predictive DNA binding model for PPARγ. This model was not only based on TF copy number, but also on binding affinity data and local chromatin state, demonstrating the feasibility of studying and even predicting the DNA binding behavior of TFs using these parameters across a wide range of biological processes. A comparable approach was used by Narumi and co-workers, who devised a methodology termed MS-QBiC (MS-based Quantification By isotope-labeled Cell-free products), that makes use of appositely designed internal standards for protein quantification by using a reconstituted cell-free protein system [56]. Such an SRM-based workflow allowed the researchers to monitor the abundance fluctuations of core circadian rhythm proteins (including TFs).

As targeted methodologies seemed to have hit the mark with respect to the detection and quantification of DNA-binding proteins, much effort is being devoted to further optimize these SRM-based protocols. For example, Zhao and colleagues [57] introduced an interesting proof of concept study in which they described the important steps in building a comprehensive SRM-based workflow to quantify relatively low abundant molecules using the innate immune response as a model system and the TF interferon response factor (IRF)-3 as target. Kim and colleagues [58] implemented a comparable SRM-based methodology that was optimized for quantifying TFs in lung cancer related cell lines without depletion or fractionation of the cell lysates (28 TFs in eight cell lines). These researchers managed to for example quantify the TF STAT3 at less than 20 amol/μg of proteins, demonstrating high detection sensitivity. To better understand the regulated mechanisms underlying mammalian cellular differentiation, Ahrends and colleagues [59] utilized an integrated approach based on a combination of methods, including SRM. Specifically, a panel of selected nuclear proteins (including TFs) was selected for quantitative analyses upon chemical and genetic perturbations to obtain comparative, quantitative data that enabled the development of models to study feedback loops in adipogenic cell fate decisions. In follow-up work, the abundance of 42 nuclear proteins in subcutaneous and visceral white adipose tissue was compared to provide novel insights into functional differences between different types of adipocytes in healthy and disease tissues [60].

These examples illustrate how targeted proteomics approaches are proving to be rather versatile and robust, as they can be easily coupled to various TF enrichment procedures that afford the necessary sensitivity to detect low abundant molecules while providing an unprecedented level of reproducibility. Mirzaei and colleagues [61] further demonstrated the utility of targeted proteomics methodologies by implementing SRM-based protein profiling aimed at systematically detecting a large fraction of the TF proteome with high reproducibility and measurement accuracy in unfractionated nuclear extracts. Their strategy based on the enrichment of specific TFs using a concatenated tandem array of TF response elements, has permitted the measurement of 464 proteins with known or suspected roles in transcriptional regulation at RNA polymerase II transcribed promoters in *Saccharomyces cerevisiae*. The researchers further validated the utility of their strategy by demonstrating that two of the SRM identified TFs, Mot3 and Azf1, were required for proper flocculation protein FLO11 expression.

In conclusion, SRM appears to live up to its expectations. Its high degree of sensitivity and specificity, combined with the reproducibility and ease of implementation of the methodology even in novice proteomics laboratories has propelled this technique to the forefront of MS-based quantitative proteomics technologies.

7 Concluding remarks: digging deeper in the TF proteome—The future quest for TF detection and quantification

In this review, we presented a broad picture of methodologies and recent studies aimed at deriving measurements of this particular family of DNA-binding proteins for comparative purposes, or at obtaining absolute amounts (Table 1). TF studies are also becoming rather comprehensive, for example, by including other regulatory molecules of importance, and by moving beyond a static picture, consistent with the ever-changing and adapting nature of biological processes. However, while becoming more frequent, there is still a general lack of quantitative TF information, which stands in stark contrast to their well-established biological importance. Pioneering work was first accomplished in bacteria and lower eukaryotes before TF copy numbers per cell were measured in mouse and human. However, although TFs are being fished out in the context of large-scale efforts aimed at determining copy numbers for a large fraction of an organism's entire proteome, still too few comprehensive studies have so far focused on quantitatively monitoring the dynamic behavior of TFs in specific biological processes. One underlying reason is the relatively low expression of TFs, at least compared to other protein types. This presents a great challenge in terms of obtaining the necessary sensitivity to identify and quantify TFs. Moreover, since the copy number of TFs tends to span several orders of magnitude in cells, it is of primary importance that high quantitative accuracy is maintained across this dynamic expression range. However, stochastic, systematic, and scaling errors can contribute to biased measurements, which can introduce important errors as outlined in Li *et al.* [10]. Although there is no clear consensus on what technology should be utilized to validate quantitative mass spectrometry data (and to what extent mass spectrometry data should be validated, if at all (e.g. [62]), we urge researchers to critically approach quantitative mass spec data, and if possible, to confirm these data with orthogonal technologies (at least for a subset of protein candidates). Several recent studies have implemented this strategy involving for example immuno-based methods to increase data robustness and accuracy (e.g. [11,63–65]). Thus, using such a validation strategy, along with selecting good quantification standards [66], it will be easier to control for different sources of errors.

To conclude, it is well understood that it will be impossible to fully describe biological processes without unraveling the regulatory mechanisms that orchestrate these same processes. Such a description will require the type of accurate quantitative protein data that MS-based approaches are increasingly able to provide. This explains the significant efforts that are being invested in advancing MS-based technologies, now enabling analytical capabilities that only a decade ago seemed unreachable. Exciting research lies therefore ahead.

The authors would like to thank Dr. Marc Moniatte from the Proteomics Core Facility at the Ecole Polytechnique Fédérale de Lausanne (EPFL) for careful reading of the manuscript and for his help with literature search, and Dr. Alexander Schmidt and Dr. Erik Ahrné for careful reading of the manuscript and for their constructive comments. Funding for this work was provided by the Swiss National Science Foundation (#31003A_162735 and #CRSII3_147684), by SystemsX.ch Special Opportunity Project 2015/323, and by Institutional support from the Ecole Polytechnique Fédérale de Lausanne (EPFL).

The authors have no conflict of interest to declare.

8 References

[1] Gerstein, M., Kundaje, A., Hariharan, M., Architecture of the human regulatory network derived from ENCODE data. *Nature* 2012, *489*, 91–100.

[2] Marbach, D., Lamparter, D., Quon, G., Kellis, M. et al., Tissue-specific regulatory circuits reveal variable modular perturbations across complex diseases. *Nat. Methods* 2016, (January), 1–44.

[3] Kalra, I. S., Alam, M. M., Choudhary, P. K., Pace, B. S., Kruppel-like Factor 4 activates HBG gene expression in primary erythroid cells. *Br. J. Haematol.* 2011, *154*, 248–259.

[4] Akinyeke, T. O., Stewart, L. V., Troglitazone suppresses c-Myc levels in human prostate cancer cells via a PPARgamma-independent mechanism. *Cancer Biol. Ther.* 2011, *11*, 1046–1058.

[5] Bintu, L., Buchler, N. E., Garcia, H. G., Gerland, U. et al., Transcriptional regulation by the numbers: applications. *Curr. Opin. Genet. Dev.* 2005, *15*, 125–135.

[6] Biggin, M. D., Animal transcription networks as highly connected, quantitative continua. *Dev. Cell*. 2011, *21*, 611–626.

[7] Deshmukh, A. S., Murgia, M., Nagaraj, N., Treebak, J. T. et al., Deep proteomics of mouse skeletal muscle enables quantitation of protein isoforms, metabolic pathways, and transcription factors. *Mol. Cell. Proteomics* 2015, *14*, 841–853.

[8] Tacheny, A., Dieu, M., Arnould, T., Renard, P., Mass spectrometry-based identification of proteins interacting with nucleic acids. *J. Proteomics* 2013, *94*, 89–109.

[9] Stergachis, A. B., MacLean, B., Lee, K., Stamatoyannopoulos, J. A., MacCoss, M. J., Rapid empirical discovery of optimal peptides for targeted proteomics. *Nat. Methods* 2011, *8*, 1041–1043.

[10] Li, J. J., Bickel, P. J., Biggin, M. D., System wide analyses have underestimated protein abundances and the importance of transcription in mammals. *PeerJ* 2014, *2*, e270.

[11] Simicevic, J., Schmid, A. W., Gilardoni, P. A., Zoller, B. et al., Absolute quantification of transcription factors during cellular differentiation using multiplexed targeted proteomics. *Nat. Methods* 2013, *10*, 570–576.

[12] Biggin, M. D., Animal transcription networks as highly connected, quantitative continua. *Dev. Cell*. 2011, *21*, 611–626.

[13] Brewster, R. C., Weinert, F. M., Garcia, H. G., Song, D. et al., The transcription factor titration effect dictates level of gene expression. *Cell* 2014, *156*, 1312–1323.

[14] Greenbaum, D., Colangelo, C., Williams, K., Gernstein, M., Comparing protein abundance and mRNA expression levels on a genomic scale. *Genome Biol* 2003, *4*, 117.

[15] de Sousa Abreu, R., Penalva, L. O., Marcotte, E. M., Vogel, C., Global signatures of protein and mRNA expression levels. *Mol. Biosyst.* 2009 *5*, 1512–1526.

[16] Karlebach, G., Shamir, R., Modelling and analysis of gene regulatory networks. *Nat. Rev. Mol. Cell Biol.* 2008, *9*, 770–780.

[17] Kim, H. D., Shay, T., O'Shea, E. K., Regev, A., Transcriptional regulatory circuits: predicting numbers from alphabets. *Science* 2009, *325*, 429–432.

[18] Deplancke, B., Alpern, D., Gardeux, V., The genetics of transcription factor DNA binding variation. *Cell* 2016, *166*, 538–544.

[19] Lee, T. I., Young, R. A., Transcriptional regulation and its misregulation in disease. *Cell* 2013, *152*, 1237–1251.

[20] Choi, J. H., Banks, A. S., Kamenecka, T. M., Busby, S. A. et al., Antidiabetic actions of a non-agonist PPARγ ligand blocking Cdk5-mediated phosphorylation. *Nature* 2011, *477*, 477–481.

[21] Karamouzis, M. V., Gorgoulis, V. G., Papavassiliou, A. G., Transcription factors and neoplasia: vistas in novel drug design. *Clin. Cancer. Res.* 2002.

[22] Bhagwat, A. S., Vakoc, C. R., Targeting transcription factors in cancer. *Trends Cancer* 2015, *1*, 53–65.

[23] Johnston, S. J., Carroll, J. S., Transcription factors and chromatin proteins as therapeutic targets in cancer. *Biochim. Biophys. Acta* 2015, *1855*, 183–192.

[24] Darnell, J. E., Transcription factors as targets for cancer therapy. *Nat. Rev. Cancer* 2002, *2*, 740–749.

[25] Pratsch, K., Wellhausen, R., Seitz, H., Advances in the quantification of protein microarrays. *Curr. Opin. Chem. Biol.* 2014, *18*, 16–20.

[26] Dodig, S., Interferences in quantitative immunochemical methods. *Biochemia Medica* 2009, *19*, 50–62.

[27] Schmidt, A., Kochanowski, K., Vedelaar, S., Ahrne, E. et al., The quantitative and condition-dependent Escherichia coli proteome. *Nat. Biotechnol.* 2015, *34*, 104–110.

[28] Schwanhäusser, B., Busse, D., Li, N., Global quantification of mammalian gene expression control. *Nature* 2011, *473*, 337–342.

[29] Ahrné, E., Molzahn, L., Glatter, T., Schmidt, A., Critical assessment of proteome-wide label-free absolute abundance estimation strategies. *Proteomics* 2013, *13*, 2567–2578.

[30] Ong, S. E., Foster, L. J., Mann, M., Mass spectrometric-based approaches in quantitative proteomics. *Methods* 2003, *29*, 124–130.

[31] Treumann, A., Thiede, B., Isobaric protein and peptide quantification: perspectives and issues. *Expert Rev. Proteomics* 2010, *7*, 647–653.

[32] Gerber, S. a, Rush, J., Stemman, O., Kirschner, M. W., Gygi, S. P., Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proc. Natl. Acad. Sci. U. S. A.* 2003, *100*, 6940–6945.

[33] Simpson, D. M., Beynon, R. J., QconCATs: design and expression of concatenated protein standards for multiplexed protein quantification. *Anal. Bioanal. Chem.* 2012, *404*, 977–989.

[34] Hanke, S., Besir, H., Oesterhelt, D., Mann, M., Absolute SILAC for accurate quantitation of proteins in complex mixtures down to the attomole level. *J. Proteome Res.* 2008, *7*, 1118–1130.

[35] Brun, V., Dupuis, A., Adrait, A., Marcellin, M. et al., Isotope-labeled protein standards: toward absolute quantitative proteomics. *Mol. Cell. Proteomics* 2007, *6*, 2139–2149.

[36] Simicevic, J., Deplancke, B., DNA-centered approaches to characterize regulatory protein-DNA interaction complexes. *Mol. Biosyst.* 2010, *6*, 462–468.

[37] Smaczniak, C., Li, N., Boeren, S., America, T. et al., Proteomics-based identification of low-abundance signaling and regulatory protein complexes in native plant tissues. *Nat. Protoc.* 2012, *7*, 2144–2158.

[38] Smaczniak, C., Immink, R. G. H., Muiño, J. M., Blanvillain, R. et al., Characterization of MADS-domain transcription factor complexes in Arabidopsis flower development. *Proc. Natl. Acad. Sci. U. S. A.* 2012, *109*, 1560–1565.

[39] Ding, C., Chan, D. W., Liu, W., Liu, M. et al., Proteome-wide profiling of activated transcription factors with a concatenated tandem array of transcription factor response

elements. *Proc. Natl. Acad. Sci. U. S. A.* 2013, *110*, 6771–6776.

[40] Hubner, N. C., Nguyen, L. N., Hornig, N. C., Stunnenberg, H. G., A quantitative proteomics tool to identify DNA-protein interactions in primary cells or blood. *J. Proteome Res.* 2015, *14*, 1315–1329.

[41] Makowski, M. M., Willems, E., Fang, J., Choi, J. et al., An interaction proteomics survey of transcription factor binding at recurrent TERT promoter mutations. *Proteomics* 2016, *16*, 417–426.

[42] Kustatscher, G., Wills, K. L. H., Furlan, C., Rappsilber, J., Chromatin enrichment for proteomics. *Nat. Protoc.* 2014, *9*, 2090–2099.

[43] Alabert, C., Bukowski-Wills, J.-C., Lee, S.-B., Kustatscher, G. et al., Nascent chromatin capture proteomics determines chromatin dynamics during DNA replication and identifies unknown fork components. *Nat. Cell Biol.* 2014, *16*, 281–293.

[44] Lelong, C., Chevallet, M., Diemer, H., Luche, S. et al., Improved proteomic analysis of nuclear proteins, as exemplified by the comparison of two myeloid cell lines nuclear proteomes. *J. Proteomics* 2012, *77*, 577–602.

[45] Molina, H., Yang, Y., Ruch, T., Kim, J. W. et al., Temporal profiling of the adipocyte proteome during differentiation using a five-plex SILAC based strategy. *J. Proteome Res.* 2009, *8*, 48–58.

[46] Pascual, J., Alegre, S., Nagler, M., Escandón, M. et al., The variations in the nuclear proteome reveal new transcription factors and mechanisms involved in UV stress response in Pinus radiata. *J. Proteomics* 2016, *143*, 390–400.

[47] Huo, C., Zhang, B., Wang, H., Wang, F. et al., Comparative study of early cold-regulated proteins by two dimensional difference gel electrophoresis reveals a key role for phospholipase Dα1 in mediating cold acclimation signaling pathway in rice. *Mol. Cell. Proteomics* 2016, *15*, 1397–1411.

[48] Zeiler, M., Straube, W. L., Lundberg, E., Uhlen, M., Mann, M., A Protein Epitope Signature Tag (PrEST) library allows SILAC-based absolute quantification and multiplexed determination of protein copy numbers in cell lines. *Mol. Cell. Proteomics* 2012, *11*, O111.009613.

[49] Mallick, P., Schirle, M., Chen, S. S., Flory, M. R. et al., Computational prediction of proteotypic peptides for quantitative proteomics. *Nat. Biotechnol.* 2007, *25*, 125–131.

[50] MacLean, B., Tomazela, D. M., Shulman, N., Chambers, M. et al., Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* 2010, *26*, 966–968.

[51] Eyers, C. E., Lawless, C., Wedge, D. C., Lau, K. W. et al., CONSeQuence: prediction of reference peptides for absolute quantitative proteomics using consensus machine learning approaches. *Mol. Cell. Proteomics* 2011, *10*, M110.003384–M110.003384.

[52] Fusaro, V. a, Mani, D. R., Mesirov, J. P., Carr, S. A., Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nat. Biotechnol.* 2009, *27*, 190–198.

[53] Lawless, C., Holman, S. W., Brownridge, P., Lanthaler, K. et al., Direct and absolute quantification of over 1800 yeast proteins via selected reaction monitoring. *Mol. Cell. Proteomics* 2016, *15*, 1309–1322.

[54] Beck, M., Schmidt, A., Malmstroem, J., Claassen, M. et al., The quantitative proteome of a human cell line. *Mol. Syst. Biol.* 2014, *7*, 549–549.

[55] Simicevic, J., Moniatte, M., Hamelin, R., Ahrné, E., Deplancke, B., A mammalian transcription factor-specific peptide repository for targeted proteomics. *Proteomics* 2015, *15*, 752–756.

[56] Narumi, R., Shimizu, Y., Ukai-Tadenuma, M., Kanda, G.N. et al., Mass-spectrometry-based quantification reveals rhythmic variation of mouse circadian clock proteins. *Proc. Natl. Acad. Sci. U. S. A.* 2016, *113*, E3461–E3467.

[57] Zhao, Y., Brasier, A. R., Applications of selected reaction monitoring (SRM)-mass spectrometry (MS) for quantitative measurement of signaling pathways. *Methods* 2013.

[58] Kim, J. S., Lee, Y., Lee, M. Y., Shin, J. et al., Multiple reaction monitoring of multiple low-abundance transcription factors in whole lung cancer cell lysates. *J. Proteome Res.* 2013, *12*, 2582–2596.

[59] Ahrends, R., Ota, A., Kovary, K. M., Kudo, T. et al., Controlling low rates of cell differentiation through noise and ultrahigh feedback. *Science (New York, N.Y.)* 2014, *344*, 1384–1389.

[60] Ota, A., Kovary, K. M., Wu, O. H., Ahrends, R. et al., Using SRM-MS to quantify nuclear protein abundance differences between adipose tissue depots of insulin-resistant mice. *J. Lipid Res.* 2015, *56*, 1068–1078.

[61] Mirzaei, H., Knijnenburg, T. A, Kim, B., Robinson, M. et al., Systematic measurement of transcription factor-DNA interactions by targeted mass spectrometry identifies candidate gene regulatory proteins. *Proc. Natl. Acad. Sci.* 2013, *110*, 1–6.

[62] Aebersold R., Burlingame A.L., Bradshaw R.A., Western blots *versus* selected reaction monitoring assays: time to turn the tables? *Mol. Cell. Proteomics* 2013, *12*, 2381–2382.

[63] Tomechko, S. E., Liu, G., Tao, M., Schlatzer, D. et al., Tissue specific dysregulated protein subnetworks in type 2 diabetic bladder urothelium and detrusor muscle. *Mol. Cell. Proteomics* 2015, *14*, 635–645.

[64] Zhou, H., Yan, H., Yan, W., Wang, X. et al., Quantitative proteomics analysis with iTRAQ in human lenses with nuclear cataracts of different axial lengths. *Mol. Vis.* 2016, *22*, 933–943.

[65] Hoover, H., Li, J., Marchese, J., Rothwell, C. et al., Quantitative proteomic verification of membrane proteins as potential therapeutic targets located in the 11q13 amplicon in cancers. *J. Proteome Res.* 2015, *14*, 3670–3679.

[66] Wisniewski, J.R., Mann, M., A proteomics approach to the protein normalization problem: selection of unvarying proteins for MS-based proteomics and Western Blotting. *J. Proteome Res.* 2016, *15*, 2321–2326.