




Technical Note

Terminomics Methodologies and the Completeness of Reductive Dimethylation: A Meta-Analysis of Publicly Available Datasets

Mariella Hurtado Silva ¹, Iain J. Berry ^{1,2}, Natalie Strange ¹, Steven P. Djordjevic ² and Matthew P. Padula ^{1,*}

¹ Proteomics Core Facility and School of Life Sciences, Faculty of Science, University of Technology Sydney, Broadway NSW 2007, Australia; Mariella.I.Hurtado@alumni.uts.edu.au (M.H.S.); iain.berry@uts.edu.au (I.J.B.); natalie.strange@student.uts.edu.au (N.S.)

² The itthree Institute, Faculty of Science, University of Technology Sydney, Broadway NSW 2007, Australia; steven.djordjevic@uts.edu.au

* Correspondence: matthew.padula@uts.edu.au; Tel.: +61-403-838-981

Received: 16 January 2019; Accepted: 25 March 2019; Published: 29 March 2019



Abstract: Methods for analyzing the terminal sequences of proteins have been refined over the previous decade; however, few studies have evaluated the quality of the data that have been produced from those methodologies. While performing global N-terminal labelling on bacteria, we observed that the labelling was not complete and investigated whether this was a common occurrence. We assessed the completeness of labelling in a selection of existing, publicly available N-terminomics datasets and empirically determined that amine-based labelling chemistry does not achieve complete labelling and potentially has issues with labelling amine groups at sequence-specific residues. This finding led us to conduct a thorough review of the historical literature that showed that this is not an unexpected finding, with numerous publications reporting incomplete labelling. These findings have implications for the quantitation of N-terminal peptides and the biological interpretations of these data.

Keywords: terminomics; mass spectrometry; amine labelling

1. Introduction

Quantitative proteomics has become a widely used tool in the biological sciences, inferring biological significance from the changes in protein abundance between varying treatments or conditions. Protein quantification studies normally use an organism's reference genome to identify and quantify products of an organism's open reading frames (ORFs). This approach, while useful for quantitation of predicted protein products, does not take into consideration the various post-translational modifications (PTMs) that may occur on a protein, or the presence of biologically distinct proteoforms arising from a single ORF. Identifying specific proteoforms and quantifying their abundance levels is necessary for a complete assessment of proteomic variation and its biological significance.

The production of mature proteoforms is an intricate process, as outlined in Figure 1 [1–3]. ORFs by definition are inferred by bioinformatic prediction from genome sequencing projects, often without direct proteomic evidence to confirm their accuracy [4–6]. Variations resulting from chemical modifications to the nascent protein (PTMs), such as acetylation and phosphorylation, as well as primary structural modifications by proteolytic cleavage, introduces a level of proteome complexity that is often overlooked when using reference genomes to detect the presence of a particular ORF using mass spectrometry [7–9]. Considering proteolytic cleavage in particular, the most prevalent

PTM in biological systems, mature proteoforms that are products of proteolysis are often incorrectly represented in protein databases as a non-mature, direct translation of the ORF. Our extensive work examining *Mycoplasma spp.*, considered the ‘simplest’ self-replicating organism yet discovered, has shown that proteolytic cleavage is a critical process in the generation of mature proteoforms from large ORFs, producing a larger proteome than bioinformatically predicted for these genome-reduced bacteria [10]. In addition, proteolysis creates proteoforms that have different functionality to the parent proteoform, further extending proteome complexity [10–22]. This increase in proteome diversity through proteolysis also occurs in eukaryotes, and this presents a need to identify and characterize proteoforms produced through proteolysis in order to understand proteoform diversity and its effect on biological systems, rather than quantifying the abundance of aggregated ORF products. However, despite a range of methodologies being available to achieve this, the methods are often not able to definitively identify the diversity of proteoforms on a proteome-wide scale or in a high throughput manner.

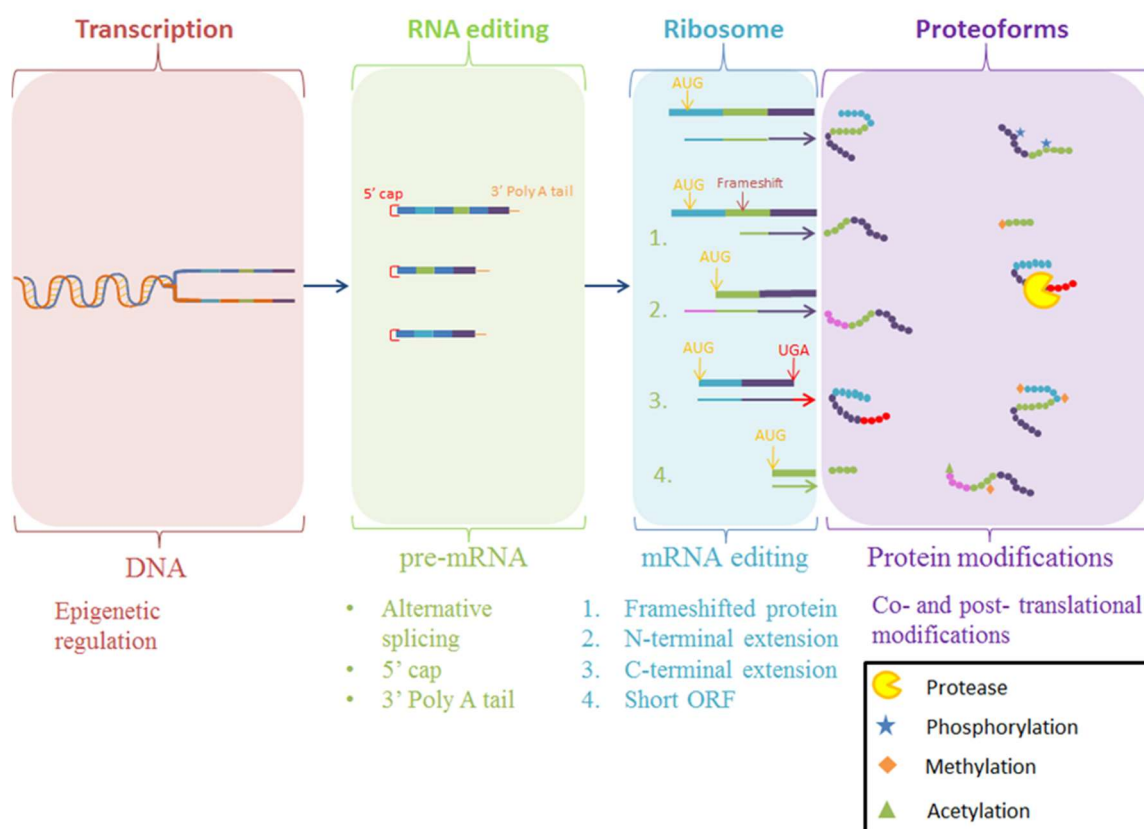


Figure 1. Schematic for the various methods for the production of proteoforms. Final protein products may be modified during transcription or during translation at the ribosome. Alternatively, nascent polypeptide chains may be modified after translation with a myriad of post-translational modifications. Once translated, the protein sequence can often require further modifications to perform a specialized function, generating proteoforms that vary from the original expressed protein [2,3,23]. Post-translational modifications such as phosphorylation, a reversible chemical addition to the protein, or proteolysis, a permanent hydrolysis event removing amino acid(s) from the polypeptide chain, cause important functional changes in proteins [24,25].

The original method for analyzing intact proteoforms and protein sequences was Edman degradation, which allows direct and unambiguous sequencing of the N-terminal amino acids of intact, purified proteins [3,26–28]. Nevertheless, Edman degradation is a time-consuming process, sequencing one amino acid residue per hour, and is limited by the efficiency of the chemical reagents to hydrolyze the peptide bonds of each subsequent residue [26–29]. The ideal solution for characterizing

intact proteoforms is direct sequencing in a complex proteome sample using intact mass spectrometry (MS), with high-resolution accurate mass measurements and MS/MS fragmentation. Intact protein MS/MS (or Top Down MS) analysis can determine the N- and C-terminal sequences of a proteoform by matching of spectral data to the predicted ORF, as well as detecting the presence and location of PTMs, where the difference in proteoform mass corresponds exactly to the PTM mass [30–32]. However, higher throughput and increased proteome depth (at the expense of proteoform identification) can be achieved by analyzing peptides from enzymatically digested proteins using traditional shotgun Liquid Chromatography-Tandem Mass Spectrometry (LC/MS/MS), with routine identification of ~11,000 ORF products [33–35] compared to <1500 proteoforms identified by Top Down MS [36]. In order to characterize the N- or C-terminus of the mature proteoforms, the terminal peptide must be unambiguously identified to avoid conjecture, which can be challenging if the true terminal peptide is not ‘MS friendly’ and thus undetectable due to poor ionization potential or unsuitable length after digestion. A solution for these issues which has been implemented by several groups [37–40] is the application of protein-level labelling and enrichment of terminal peptide sequences after proteolytic digestion, known collectively as “terminomics”. While these methods are able to reduce the aforementioned ambiguity, they are not free of problems.

Currently, the most popular terminomics techniques target the identification of the N-terminus of proteoforms by enriching and sequencing N-terminal peptides by MS, which are then mapped to ORFs in silico [41–43]. Enrichment aims to address the biggest obstacle encountered in terminomics and proteomics, which is the complexity of the sample being analyzed, as a sample containing tens of thousands of proteoforms will produce an exponentially larger number of peptides of varying abundance to be analyzed following enzymatic cleavage [44,45]. This is exemplified by the fact that the 15,721 human proteins or ORFs that have been detected and catalogued in ProteomicsDB, arguably the most comprehensive proteome resource available, are described by 455,289 unique peptides, with peptide evidence for 7977 N-termini and 6778 C-termini [46]. The competition of peptides for detection in MS can be addressed through the selected isolation of only N-terminal peptides via N-terminomics enrichment strategies, enabling the complexity of the sample to be reduced and minimizing signal competition during MS, thereby improving N-terminal sequence identification [41].

Bottom-up MS proteome analysis assigns peptide sequences to ORFs using bioinformatics; however, information about the original intact proteoform is often lost [3,47,48]. For example, proteoforms differing by a single amino substitution cannot be distinguished if the peptide containing the substitution is not detected. In the case of proteolytic processing, the main problem of bottom-up methodologies is the previously mentioned unambiguous assignment of the N-termini of mature proteoforms. This is best illustrated by considering a digest of Bovine Serum Albumin (BSA). In a majority of cases when analyzing peptides, the MS¹ scan range is usually set between 350–1500 m/z to optimize the transmission of ions through the quadrupole for high sensitivity and to avoid low mass, but high signal ‘background’ ions being detected. The lower mass of 350 m/z is also set to avoid selecting small peptides of less than 3–5 amino acids (AAs) for fragmentation and MS/MS, as these are scored against by search algorithms as they are more likely to be matched randomly. In the case of BSA, the N-terminal site is well known (UniProt P02796) and the mature proteoform starts at Aspartic Acid (D, position 19) after removal of the signal and pro-peptide. After digestion with trypsin, the peptide created at the N-terminal of the proteoform is DTHK, which would have a monoisotopic mass of 500.24 m/z, while the 2⁺ ion would be 250.62 m/z, which is below the normally utilized scan range. The first detectable and assigned peptide of the sequence in our experience is FKDLGEEHFK (AAs 34–44) which includes a missed tryptic cleavage site. If this was an uncharacterized protein with no knowledge of the mature proteoform, such shotgun LC/MS/MS data would erroneously assign the N-terminal amino acid.

The difficulty of identifying proteoforms using shotgun LC/MS/MS is further compounded when trying to characterize proteoforms generated by the cleavage of large precursor proteins, exemplified by the adhesin families of *Mycoplasma hyopneumoniae*. This organism overcomes a small

genome of less than 700 ORFs by performing extensive proteolytic cleavage of expressed proteins into mature proteoforms, such as the cleavage of mhp683 [12]. This ORF is processed into three main proteoforms of 45, 48, and 50 kDa, a fact not able to be determined by shotgun LC/MS/MS but revealed through prior fractionation by proteoform mass using one dimensional PolyAcrylamide Gel Electrophoresis (SDS-PAGE), where peptides mapping to the entire ORF are found at approximately 50 kDa. The individual proteoforms are resolvable by Two Dimensional PAGE (Isoelectric focusing followed by SDS-PAGE), and LC/MS/MS analysis of peptides from the 48 kDa proteoform (designated P48) revealed the true N-terminus due to the presence of pyroglutamate, formed by the cyclisation of the side chain of glutamate or glutamine with its free α -amine. This 'labelling' of the N-terminus, where the mass of the AA has been altered by the formation of pyroglutamate (which can be removed by enzymatic treatment with pyroglutaminase), led us to explore the literature for other methods of labelling to distinguish the true N-termini of mature proteoforms. Most strategies for N-terminomics, such as COmbined FRActional DIagonal Chromatography (COFRADIC) and Terminal Amine Isotopic Labeling of Substrates (TAILS), implement the use of a chemical label [42,49–52], where sample complexity is then simplified by either positive or negative enrichment of labelled N-terminal peptides for analysis by LC/MS/MS [3].

The most popular approach to N-terminal labelling of proteoforms involves exploiting the apparent chemical reactivity of the free primary amine group on the N-terminus and lysine side chain [3,43,53]. As is the case for numerous chemical reactions, including Edman degradation, labelling procedures are restrained by the efficiency of the chemical reaction to modify the targeted amine groups and attach the chemical label [40,43,54]. In our laboratory, there has been evidence to suggest incomplete dimethyl labelling of proteoforms isolated from bacteria, especially the lysine-rich Mycoplasmas, with cases of lysine being identified as the N-terminal amino acid but possessing only one dimethyl tag instead of the expected two (Figure 2). This observation was made because our peptide search parameters have dimethylation set as a variable modification, which is in contrast to the consensus in the literature to use fixed modifications. This led us to suspect that incomplete labelling was more widespread than being reported. Several groups have explored the efficiency of the dimethylation procedure (listed in Table 1), but each quote different sets of conditions that are "optimal" for protein sample labelling. Additionally, none of these groups have systematically reported the effect of protein concentration or the level of proteome complexity that is compatible with the dimethyl–amine chemistry or other labelling techniques [55–57].

Examination of data from experiments using reductive dimethylation in our laboratory and comparing it to other reports prompted us to question whether the dimethylation labelling was complete with all primary amines labelled when performed on different organisms with higher lysine content. This also raised the question of whether the data in current N-terminomics literature is underreporting the completeness of labelling, potentially leading to inaccurate quantitation. To address this, we have performed a meta-analysis on a selection of publically available, quantitative datasets which implemented a dimethyl labelling protocol without peptide depletion.

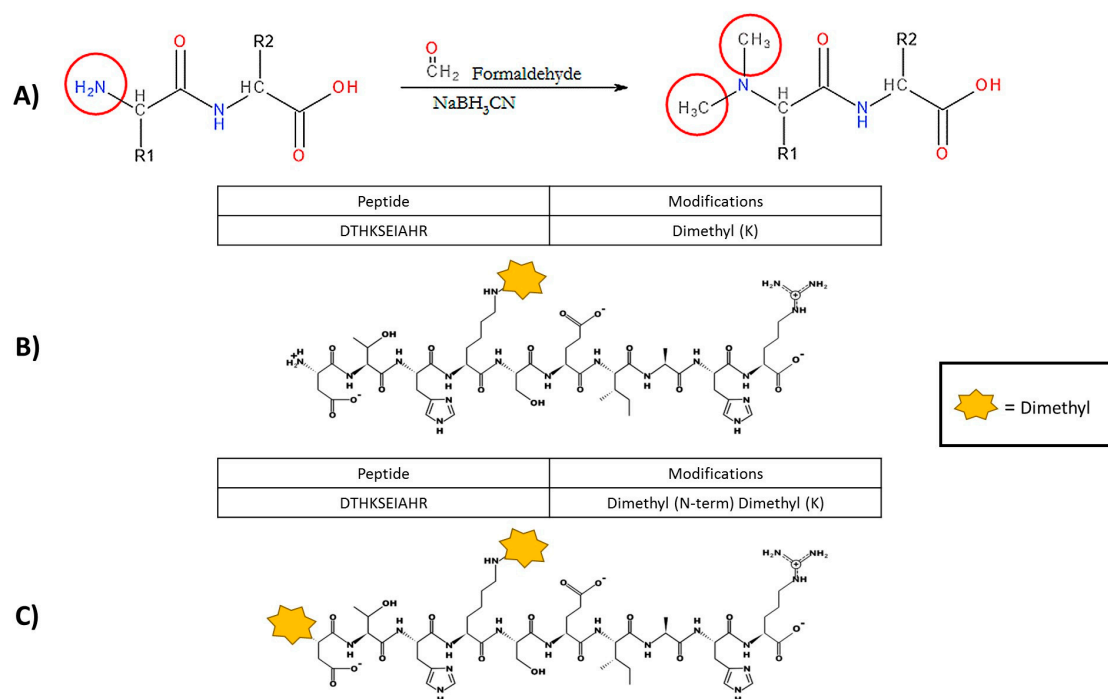


Figure 2. Dimethylation reaction implemented in the UTS Proteomics Core Facility with examples of incomplete labelling. **(A)** Theoretical reductive dimethylation reaction shown, which will attach two methyl groups to every primary amine in a protein sample (N-terminus and lysine residues). However, preliminary experimental results of the dimethylation process indicate that the current method is not modifying all primary amine groups in complex samples. **(B)** Unpublished data of incomplete labelling occurring in a model protein sample (bovine serum albumin). The N-terminal sequence that was obtained indicated that the protein amino acid sequence began with aspartic acid (Letter symbol, D), which was confirmed with bovine serum albumin data from UniProt (accession number: P02769). Identical peptide sequences containing lysine residues have been sequenced by mass spectrometry; however, the first peptide, indicated in the red rectangle, has been identified with only the lysine dimethylated. **(C)** The second sequence from the same mass spectrometry experiment has a dimethyl label on both the N-terminal amine (the aspartic acid residue) and lysine residue.

2. Materials and Methods

2.1. Data Selection

Raw MS/MS data files were obtained from the online proteomics data repository, PRoteomics IDentifications (PRIDE) Archive (Table 2). All analyzed datasets implemented dimethyl labelling protocols involving no enrichment strategy and labelling at the peptide level (after sample digestion); thus, all detected peptides should be dimethylated if all primary amines had reacted. Our meta-analysis omits datasets with ambiguously named data files, no raw MS/MS data, or labelling strategies which utilized peptide depletion protocols (e.g., N-TAILS). The reason for omitting datasets generated with depletion strategies is that the negative selection of amine-containing peptides, which should be the internal peptides generated by proteolytic digestion after dimethylation, will remove any incompletely labelled peptides, preventing their analysis by LC/MS/MS. We were therefore forced to omit studies that used dimethylation at the protein level, the technique that is of most value for the determination of the N-termini of mature proteoforms, because the relevant data was not being captured in the experiment. The datasets used for re-analysis were selected as randomly as possible to cover as many types of organisms as possible.

2.2. Data Search and Analysis

Data files obtained from PRIDE were subsequently searched using the PEAKS Studio software package (v8.5) with the relevant organism sequence database, while the specific instrument parameters were as per those reported in each study, but with the relevant modifications set to variable rather than fixed (see Supplemental Table S1 for full specifications). The PEAKS search results were filtered to remove sample contaminants and then sorted to identify spectra that match to the same sequence (duplicate sequences). All significant ($p \leq 0.05$) duplicate sequences were interrogated for frequency of complete, partial, or total lack of dimethyl labelling by assessing the proportion of duplicate sequences detected with a label at the N-terminus and at any lysine residues.

3. Results

In this meta-analysis, we sought to examine whether dimethylation was present on all primary amines in peptides in published data utilizing a non-depleted dimethyl labelling protocol. It is important to point out that many of the studies examined performed database searching with dimethylation as a fixed modification, thus assuming that all primary amines present were dimethylated. If dimethylation of all available reactive amines does not occur, as we suspect that it does not, the search parameters applied in these studies will either not assign all of the acquired spectra, resulting in a false negative, or assign a spectrum to the incorrect sequence, resulting in a false positive. Our analysis, presented in Table 2, found that between 6–18% of duplicate sequences detected in each dataset have amine groups that are not dimethylated, which was not reported by the authors. These peptide matches are therefore false negatives, a value that is normally not able to be calculated. As these are peptide-based, shotgun LC/MS/MS experiments, the need to calculate false discovery rates (FDR) is a mandatory requirement of many reviewers and journals, and it is generally accepted that the FDR be below 1% [58–60]. Our analysis assigned FDRs to each dataset of $\leq 3\%$. In the case of these studies, the false negative rate is far greater than the false positive rate reported in the individual studies, and this is of serious concern when quantitation is considered as it will lead to false values.

The meta-analysis also revealed that a portion of the acquired MS/MS spectra are unassigned in the original publications, indicating that potentially important protein/peptide information is being overlooked in the final quantification. The number of peptides displaying a dimethyl label varied between each sample, with only 81–94% of all peptides demonstrating a dimethylated -amine and, in the case of lysine terminating peptides, a dimethylated -amine. It is possible that this is an underestimation. Prior published evidence [61,62] indicates that the ability of a primary amine to be dimethyl-labelled may be relative to the characteristics of the protein sample (reviewed in greater detail by Feeney and Blankenhorn [63]). This data may suggest that there are variations in labelling between organisms; however, this view is not supported by the meta-analysis conducted here. Jentoft and Dearborn [61] provided evidence that extreme concentrations of formaldehyde and NaBH_3CN will not result in complete modification of primary amines in proteins, but concluded that reductive methylation may be implemented quantitatively. In contrast, Gidley and Sanders [64] found that the yields are never quantitative, and there always appears to be unchanged starting material present at the completion of the reaction, which is in contrast to the current understanding and implementation of the technique as reported in the literature.

One study that does report a completeness of labelling is Rowland et al. [65], where the labelling efficiency of lysine residues in the chloroplast proteome of *A. thaliana* was reported to be $\geq 99\%$ for detected peptides. This may be attributed to protein preparation and reaction conditions, or the inherent characteristics of proteoforms present in the investigated proteome [61,62,66]. While these data [65] are available on the online proteomics data repository PRIDE (dataset identifier PXD002476 and 10.6019/PXD002476), it was difficult for us to determine with complete certainty which files corresponded to the dimethyl efficiency testing, so the data was unable to be included in our meta-analysis. This is not an uncommon issue, and there needs to be a dedicated effort made to properly label raw data files so that the results can be independently validated.

Inconsistency in labelling of raw data files available from online data repositories was only one issue encountered that hindered data acquisition. A large number of studies performing reductive dimethylation experiments failed to upload raw data files onto online data repositories, which restricted the number of datasets available for analysis. Results from our meta-analysis indicate that in the datasets analyzed, dimethylation was not complete; however, more data is required to understand the more widespread implications of this observed inefficiency. As such, we acknowledge and echo the recommendations of Lange et al. [67] and strongly encourage others to upload raw data files with clear, understandable filenames onto online data repositories.

In our meta-analysis, we were unable to include datasets from studies utilizing TAILS or negative selection of peptides because incompletely labelled peptides would be captured by the polyaldehyde polymers used to capture all molecules with free primary amines. These studies implement dimethylation at the protein level, which is of most relevance to our need to identify mature proteoforms. In our experience, protein level labelling is not complete, and we have no reason to suspect that protein level labelling is complete in TAILS-based analysis or other systems. Studies implementing chemical labelling strategies for quantitative analyses need to be aware that labelling efficiency can be variable and incomplete with important implications for data analysis, so we suggest that the extent of this should be empirically analyzed by searching the data with demethylation as a variable modification to determine the completeness of labelling. Once determined, the researcher can decide whether to proceed with the quantitative experiment and report the completeness of amine labelling. In SILAC experiments, it is generally accepted that the heavy amino acid be incorporated into >95% of the peptides detected before the quantitative experiment is performed, and this should be the case for chemical labelling strategies.

Table 1. Published reaction conditions of reductive methylation protocols.

Reference (Year)	Reaction Conditions	Reactant Concentrations	Significant Observations
Friedman et al. [66] (1974)	4–16 h (room temperature) Alcohol/lithium acetate buffer pH 5.2	~11 mM NaBH ₃ CN ~11 mM aldehyde (various)	Modification of lysine residues ranged from 40–90% using different aldehyde reagents, between protein molecules and different amino acid residues
Jentoft et al. [61] (1979)	2–24 h (22 °C) HEPES buffer pH 7.5	20 mM NaBH ₃ CN Concentration formaldehyde ~ concentration of lysyl residues in sample	80–90% dimethyl conversion of lysine residues with a 6 fold excess of formaldehyde Lower concentrations of NaBH ₃ CN (5 mM to 20 mM) yielded in the highest modifications of lysyl residues Maximal rates of labelling observed at pH 8
Hsu et al. [68] (2005)	Sodium acetate buffer pH 5–85 min	~22 mM NaBH ₃ CN ~52 mM formaldehyde	Observation of immonium ion signal with dimethyl labelling
Krusemark et al. [69] (2008)	2 h, room temperature 300 mM triethanolamine and 6 mM Guanidine-HCL buffer pH 7.5 20% MeOH 1 mg/mL protein	30 mM Pyridine-BH ₃ (reducing agent) 20 mM formaldehyde	4 model proteins containing various abundance of amine groups, dimethyl labelled to completeness NaBH ₃ CN and NaBH ₄ found to produce side reactions resulting in reduced purity of products
Boersema et al. [70] (2009)	1 h, room temperature 100 mM Triethylammonium bicarbonate buffer pH 5–8.5	~22 mM NaBH ₃ CN ~52 mM formaldehyde	(protocol paper)
Kleinfeld et al. [43] (2011)	4 h—overnight incubation at 37 °C 100 mM HEPES pH 7.0	20 mM NaBH ₃ CN 40 mM formaldehyde	(protocol paper)
Jhan et al. [71] (2017)	30 s–2h, room temperature 100 mM sodium acetate pH 5–6	1.4–85 mM NaBH ₃ CN 156 mM formaldehyde	Accessibility of primary amines on the protein greatly affects dimethylation efficiency At 30 s 80% of amines were dimethylated

Table 2. Meta-analysis results of dimethyl labelling studies [72–75].

PRIDE Dataset Identifier	FDR PEAKS Generated (%)	Duplicate Peptide Sequences Detected	Duplicate Sequences with Complete Labelling	Complete Labelling (%)	Duplicate Sequences with Partial Labelling	Partial Labelling (%)	Duplicate Sequences with No Dimethyl Label	Unlabeled (%)	Total Partial and Unlabeled Duplicate Sequences	Total Partial and Unlabeled (%)
PXD002785 PXD003833 (125)	1.7	6658	5454	81.92	1161	17.44	43	0.65	1204	18.08
PRD000055 (115)	0.6	5395	5062	93.83	315	5.84	18	0.33	333	6.17
PXD005920 (126)	1.5	3269	2847	87.09	404	12.36	18	0.55	422	12.91
PXD003298 (127)	1.6	3531	2893	81.93	584	16.54	54	1.53	638	18.07
PXD004654 (128)	3.0	6293	5498	87.37	715	11.36	80	1.27	795	12.63

4. Discussion

Over a number of years, our laboratory has been interested in the characterization of proteolytic processing that occurs in prokaryotes to generate proteome diversity from a relatively small genome. Through the use of protein-centric techniques, especially 2D-PAGE, we showed the extent of processing in the model bacteria *Mycoplasma hyopneumoniae*, but we were unable to definitively identify the point of cleavage because we could not be sure that a peptide closer to the N-terminal was not being detected because it was not 'MS-friendly'. In an attempt to resolve this, we turned to reductive dimethylation as a method of labelling the N-terminal amine of mature proteoforms, but we found that the 'completeness' of the labelling was less than that reported in the wider literature. In an attempt to understand why this was the case here, we have performed a meta-analysis on a random selection of N-terminomics datasets with surprising results. The significant issue is that the search parameters used to identify dimethylated peptides assume that all amines are modified, which is a flawed assumption as very few chemical reactions proceed to absolute completion and some potential reactants are always left over. During reanalysis, we found a significant number of unlabeled peptides in the datasets, which are false negatives in the original published analysis. This brings into question some of the conclusions of any publications seeking to use reductive dimethylation in a quantitative manner. It is clear that more work needs to be performed to further characterize the reductive dimethylation chemistry and workflows. During our analysis, we had difficulty identifying datasets which met the selection criteria for the meta-analysis, due to poor annotation of datasets in PRIDE or the use of enrichment tools which disguise the presence of unlabeled peptides. It is clear that we need to investigate other chemistries, such as succinimide-based chemistries utilized in Isobaric Tags for Relative and Absolute Quantitation (iTRAQ) or Tandem Mass Tags (TMT) protocols, which may provide a more complete labelling or may suffer similar shortfalls as the dimethylation reaction.

Supplementary Materials: The following are available online at <http://www.mdpi.com/2227-7382/7/2/11/s1>, Table S1: Parameters used to re-analyze the datasets listed in Table 1.

Author Contributions: Conceptualization, M.H.S., I.J.B., and M.P.P.; methodology, M.H.S., I.J.B., N.S., and M.P.P.; software, N.S.; validation, M.H.S., I.J.B., N.S., and M.P.P.; formal analysis, M.H.S., I.J.B., N.S., and M.P.P.; writing—original draft preparation, M.H.S. and M.P.P.; writing—review and editing, M.H.S., I.J.B., N.S., S.P.D., and M.P.P.; supervision, S.P.D. and M.P.P.; project administration, S.P.D. and M.P.P.; funding acquisition, S.P.D. and M.P.P.

Funding: This research is supported by Australian Government Research Training Program Scholarships awarded to I.J.B. and N.S.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Lenasi, T.; Barboric, M. Mutual relationships between transcription and pre-mRNA processing in the synthesis of mRNA. *Wiley Interdiscip. Rev. RNA* **2013**, *4*, 139–154. [[CrossRef](#)] [[PubMed](#)]
2. Marshall, N.C.; Finlay, B.B.; Overall, C.M. Sharpening Host Defenses during Infection: Proteases Cut to the Chase. *Mol. Cell. Proteom.* **2017**, *16* (Suppl. 1), S161–S171. [[CrossRef](#)]
3. Hartmann, E.M.; Armengaud, J. N-terminomics and proteogenomics, getting off to a good start. *Proteomics* **2014**, *14*, 2637–2646. [[CrossRef](#)] [[PubMed](#)]
4. Muller, S.A.; Findeiss, S.; Pernitzsch, S.R.; Wissenbach, D.K.; Stadler, P.F.; Hofacker, I.L.; von Bergen, M.; Kalkhof, S. Identification of new protein coding sequences and signal peptidase cleavage sites of *Helicobacter pylori* strain 26695 by proteogenomics. *J. Proteom.* **2013**, *86*, 27–42. [[CrossRef](#)] [[PubMed](#)]
5. Castellana, N.E.; Payne, S.H.; Shen, Z.; Stanke, M.; Bafna, V.; Briggs, S.P. Discovery and revision of Arabidopsis genes by proteogenomics. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 21034–21038. [[CrossRef](#)] [[PubMed](#)]
6. Castellana, N.; Bafna, V. Proteogenomics to discover the full coding content of genomes: A computational perspective. *J. Proteom.* **2010**, *73*, 2124–2135. [[CrossRef](#)]
7. Boguski, M.S.; McIntosh, M.W. Biomedical informatics for proteomics. *Nature* **2003**, *422*, 233–237. [[CrossRef](#)]

8. Johnson, R.S.; Davis, M.T.; Taylor, J.A.; Patterson, S.D. Informatics for protein identification by mass spectrometry. *Methods* **2005**, *35*, 223–236. [[CrossRef](#)]
9. Cottrell, J.S. Protein identification using MS/MS data. *J. Proteom.* **2011**, *74*, 1842–1851. [[CrossRef](#)]
10. Tacchi, J.L.; Raymond, B.B.; Haynes, P.A.; Berry, I.J.; Widjaja, M.; Bogema, D.R.; Woolley, L.K.; Jenkins, C.; Minion, F.C.; Padula, M.P.; et al. Post-translational processing targets functionally diverse proteins in *Mycoplasma hyopneumoniae*. *Open Biol.* **2016**, *6*, 150210. [[CrossRef](#)] [[PubMed](#)]
11. Deutscher, A.T.; Jenkins, C.; Minion, F.C.; Seymour, L.M.; Padula, M.P.; Dixon, N.E.; Walker, M.J.; Djordjevic, S.P. Repeat regions R1 and R2 in the P97 paralogue Mhp271 of *Mycoplasma hyopneumoniae* bind heparin, fibronectin and porcine cilia. *Mol. Microbiol.* **2010**, *78*, 444–458. [[CrossRef](#)]
12. Bogema, D.R.; Scott, N.E.; Padula, M.P.; Tacchi, J.L.; Raymond, B.B.; Jenkins, C.; Cordwell, S.J.; Minion, F.C.; Walker, M.J.; Djordjevic, S.P. Sequence TTKF ↓ QE defines the site of proteolytic cleavage in Mhp683 protein, a novel glycosaminoglycan and cilium adhesin of *Mycoplasma hyopneumoniae*. *J. Biol. Chem.* **2011**, *286*, 41217–41229. [[CrossRef](#)] [[PubMed](#)]
13. Deutscher, A.T.; Tacchi, J.L.; Minion, F.C.; Padula, M.P.; Crossett, B.; Bogema, D.R.; Jenkins, C.; Kuit, T.A.; Walker, M.J.; Djordjevic, S.P. *Mycoplasma hyopneumoniae* Surface proteins Mhp385 and Mhp384 bind host cilia and glycosaminoglycans and are endoproteolytically processed by proteases that recognize different cleavage motifs. *J. Proteome Res.* **2012**, *11*, 1924–1936. [[CrossRef](#)] [[PubMed](#)]
14. Jarocki, V.M.; Santos, J.; Tacchi, J.L.; Raymond, B.B.; Deutscher, A.T.; Jenkins, C.; Padula, M.P.; Djordjevic, S.P. MHJ_0461 is a multifunctional leucine aminopeptidase on the surface of *Mycoplasma hyopneumoniae*. *Open Biol.* **2015**, *5*, 140175. [[CrossRef](#)]
15. Raymond, B.B.; Jenkins, C.; Seymour, L.M.; Tacchi, J.L.; Widjaja, M.; Jarocki, V.M.; Deutscher, A.T.; Turnbull, L.; Whitchurch, C.B.; Padula, M.P.; et al. Proteolytic processing of the cilium adhesin MHJ_0194 (P123J) in *Mycoplasma hyopneumoniae* generates a functionally diverse array of cleavage fragments that bind multiple host molecules. *Cell. Microbiol.* **2015**, *17*, 425–444. [[CrossRef](#)] [[PubMed](#)]
16. Seymour, L.M.; Deutscher, A.T.; Jenkins, C.; Kuit, T.A.; Falconer, L.; Minion, F.C.; Crossett, B.; Padula, M.; Dixon, N.E.; Djordjevic, S.P.; et al. A processed multidomain mycoplasma hyopneumoniae adhesin binds fibronectin, plasminogen, and swine respiratory cilia. *J. Biol. Chem.* **2010**, *285*, 33971–33978. [[CrossRef](#)] [[PubMed](#)]
17. Seymour, L.M.; Jenkins, C.; Deutscher, A.T.; Raymond, B.B.; Padula, M.P.; Tacchi, J.L.; Bogema, D.R.; Eamens, G.J.; Woolley, L.K.; Dixon, N.E.; et al. Mhp182 (P102) binds fibronectin and contributes to the recruitment of plasmin(ogen) to the *Mycoplasma hyopneumoniae* cell surface. *Cell. Microbiol.* **2012**, *14*, 81–94. [[CrossRef](#)]
18. Robinson, M.W.; Buchtman, K.A.; Jenkins, C.; Tacchi, J.L.; Raymond, B.B.; To, J.; Roy Chowdhury, P.; Woolley, L.K.; Labbate, M.; Turnbull, L.; et al. MHJ_0125 is an M42 glutamyl aminopeptidase that moonlights as a multifunctional adhesin on the surface of *Mycoplasma hyopneumoniae*. *Open Biol.* **2013**, *3*, 130017. [[CrossRef](#)] [[PubMed](#)]
19. Wilton, J.; Jenkins, C.; Cordwell, S.J.; Falconer, L.; Minion, F.C.; Oneal, D.C.; Djordjevic, M.A.; Connolly, A.; Barchia, I.; Walker, M.J.; et al. Mhp493 (P216) is a proteolytically processed, cilium and heparin binding protein of *Mycoplasma hyopneumoniae*. *Mol. Microbiol.* **2009**, *71*, 566–582. [[CrossRef](#)]
20. Djordjevic, S.P.; Cordwell, S.J.; Djordjevic, M.A.; Wilton, J.; Minion, F.C. Proteolytic processing of the *Mycoplasma hyopneumoniae* cilium adhesin. *Infect. Immun.* **2004**, *72*, 2791–2802. [[CrossRef](#)] [[PubMed](#)]
21. Jarocki, V.M.; Tacchi, J.L.; Djordjevic, S.P. Non-proteolytic functions of microbial proteases increase pathological complexity. *Proteomics* **2015**, *15*, 1075–1088. [[CrossRef](#)] [[PubMed](#)]
22. Tacchi, J.L.; Raymond, B.B.; Jarocki, V.M.; Berry, I.J.; Padula, M.P.; Djordjevic, S.P. Cilium adhesin P216 (MHJ_0493) is a target of ectodomain shedding and aminopeptidase activity on the surface of *Mycoplasma hyopneumoniae*. *J. Proteome Res.* **2014**, *13*, 2920–2930. [[CrossRef](#)] [[PubMed](#)]
23. Fortelny, N.; Pavlidis, P.; Overall, C.M. The path of no return—Truncated protein N-termini and current ignorance of their genesis. *Proteomics* **2015**, *15*, 2547–2552. [[CrossRef](#)] [[PubMed](#)]
24. Bastos, P.A.; da Costa, J.P.; Vitorino, R. A glimpse into the modulation of post-translational modifications of human-colonizing bacteria. *J. Proteom.* **2017**, *152*, 254–275. [[CrossRef](#)]
25. Cain, J.A.; Solis, N.; Cordwell, S.J. Beyond gene expression: The impact of protein post-translational modifications in bacteria. *J. Proteom.* **2014**, *97*, 265–286. [[CrossRef](#)] [[PubMed](#)]

26. Han, K.-K.; Belaiche, D.; Moreau, O.; Briand, G. Current developments in stepwise edman degradation of peptides and proteins. *Int. J. Biochem.* **1985**, *17*, 429–445. [[CrossRef](#)]
27. Lobas, A.A.; Verenchikov, A.N.; Goloborodko, A.A.; Levitsky, L.I.; Gorshkov, M.V. Combination of Edman degradation of peptides with liquid chromatography/mass spectrometry workflow for peptide identification in bottom-up proteomics. *Rapid Commun. Mass Spectrom.* **2013**, *27*, 391–400. [[CrossRef](#)] [[PubMed](#)]
28. Edman, P. A method for the determination of amino acid sequence in peptides. *Arch. Biochem.* **1949**, *22*, 475. [[CrossRef](#)] [[PubMed](#)]
29. Berry, I.J.; Steele, J.R.; Padula, M.P.; Djordjevic, S.P. The application of terminomics for the identification of protein start sites and proteoforms in bacteria. *Proteomics* **2016**, *16*, 257–272. [[CrossRef](#)] [[PubMed](#)]
30. Lorenzatto, K.R.; Kim, K.; Ntai, I.; Paludo, G.P.; Camargo de Lima, J.; Thomas, P.M.; Kelleher, N.L.; Ferreira, H.B. Top Down Proteomics Reveals Mature Proteoforms Expressed in Subcellular Fractions of the *Echinococcus granulosus* Preadult Stage. *J. Proteome Res.* **2015**, *14*, 4805–4814. [[CrossRef](#)] [[PubMed](#)]
31. Zheng, Y.; Huang, X.; Kelleher, N.L. Epiproteomics: Quantitative analysis of histone marks and codes by mass spectrometry. *Curr. Opin. Chem. Biol.* **2016**, *33*, 142–150. [[CrossRef](#)]
32. Roth, M.J.; Parks, B.A.; Ferguson, J.T.; Boyne, M.T., 2nd; Kelleher, N.L. “Proteotyping”: Population proteomics of human leukocytes using top down mass spectrometry. *Anal. Chem.* **2008**, *80*, 2857–2866. [[CrossRef](#)]
33. Kulak, N.A.; Geyer, P.E.; Mann, M. Loss-less nano-fractionator for high sensitivity, high coverage proteomics. *Mol. Cell. Proteom.* **2017**, *16*, 694–705. [[CrossRef](#)]
34. Kulak, N.A.; Pichler, G.; Paron, I.; Nagaraj, N.; Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **2014**, *11*, 319–324. [[CrossRef](#)] [[PubMed](#)]
35. Meier, F.; Geyer, P.E.; Virreira Winter, S.; Cox, J.; Mann, M. BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat. Methods* **2018**, *15*, 440–448. [[CrossRef](#)] [[PubMed](#)]
36. Durbin, K.R.; Fornelli, L.; Fellers, R.T.; Doubleday, P.F.; Narita, M.; Kelleher, N.L. Quantitation and Identification of Thousands of Human Proteoforms below 30 kDa. *J. Proteome Res.* **2016**, *15*, 976–982. [[CrossRef](#)] [[PubMed](#)]
37. Doucet, A.; Overall, C.M. Protease proteomics: Revealing protease in vivo functions using systems biology approaches. *Mol. Asp. Med.* **2008**, *29*, 339–358. [[CrossRef](#)] [[PubMed](#)]
38. Jagdeo, J.M.; Dufour, A.; Klein, T.; Solis, N.; Kleifeld, O.; Kizhakkedathu, J.; Luo, H.; Overall, C.M.; Jan, E. N-Terminomics TAILS Identifies Host Cell Substrates of Poliovirus and Coxsackievirus B3 3C Proteinases That Modulate Virus Infection. *J. Virol.* **2018**, *92*, e02211-17. [[CrossRef](#)]
39. Biniossek, M.L.; Niemer, M.; Maksimchuk, K.; Mayer, B.; Fuchs, J.; Huesgen, P.F.; McCafferty, D.G.; Turk, B.; Fritz, G.; Mayer, J.; et al. Identification of Protease Specificity by Combining Proteome-Derived Peptide Libraries and Quantitative Proteomics. *Mol. Cell. Proteom.* **2016**, *15*, 2515–2524. [[CrossRef](#)]
40. Impens, F.; Rolhion, N.; Radoshevich, L.; Becavin, C.; Duval, M.; Mellin, J.; Garcia Del Portillo, F.; Pucciarelli, M.G.; Williams, A.H.; Cossart, P. N-terminomics identifies Prli42 as a membrane miniprotein conserved in Firmicutes and critical for stressosome activation in *Listeria monocytogenes*. *Nat. Microbiol.* **2017**, *2*, 17005. [[CrossRef](#)]
41. Stroh, J.G.; Loulakis, P.; Lanzetti, A.J.; Xie, J. LC-mass spectrometry analysis of N- and C-terminal boundary sequences of polypeptide fragments by limited proteolysis. *J. Am. Soc Mass Spectrom.* **2005**, *16*, 38–45. [[CrossRef](#)] [[PubMed](#)]
42. Staes, A.; Impens, F.; Van Damme, P.; Ruttens, B.; Goethals, M.; Demol, H.; Timmerman, E.; Vandekerckhove, J.; Gevaert, K. Selecting protein N-terminal peptides by combined fractional diagonal chromatography. *Nat. Protoc.* **2011**, *6*, 1130–1141. [[CrossRef](#)] [[PubMed](#)]
43. Kleifeld, O.; Doucet, A.; Prudova, A.; auf dem Keller, U.; Gioia, M.; Kizhakkedathu, J.N.; Overall, C.M. Identifying and quantifying proteolytic events and the natural N terminome by terminal amine isotopic labeling of substrates. *Nat. Protoc.* **2011**, *6*, 1578–1611. [[CrossRef](#)] [[PubMed](#)]
44. Ahram, M.; Springer, D.L. Large-scale proteomic analysis of membrane proteins. *Expert Rev. Proteom.* **2004**, *1*, 293–302. [[CrossRef](#)] [[PubMed](#)]
45. Yagoub, D.; Tay, A.P.; Chen, Z.; Hamey, J.J.; Cai, C.; Chia, S.Z.; Hart-Smith, G.; Wilkins, M.R. Proteogenomic Discovery of a Small, Novel Protein in Yeast Reveals a Strategy for the Detection of Unannotated Short Open Reading Frames. *J. Proteome Res.* **2015**, *14*, 5038–5047. [[CrossRef](#)] [[PubMed](#)]

46. Wilhelm, M.; Schlegl, J.; Hahne, H.; Moghaddas Gholami, A.; Lieberenz, M.; Savitski, M.M.; Ziegler, E.; Butzmann, L.; Gessulat, S.; Marx, H.; et al. Mass-spectrometry-based draft of the human proteome. *Nature* **2014**, *509*, 582–587. [[CrossRef](#)] [[PubMed](#)]
47. Coorsen, J.; Yergey, A. Proteomics Is Analytical Chemistry: Fitness-for-Purpose in the Application of Top-Down and Bottom-Up Analyses. *Proteomes* **2015**, *3*, 440. [[CrossRef](#)]
48. Oliveira, B.M.; Coorsen, J.R.; Martins-de-Souza, D. 2DE: The phoenix of proteomics. *J. Proteom.* **2014**, *104*, 140–150. [[CrossRef](#)]
49. Lange, P.F.; Overall, C.M. Protein TAILS: When termini tell tales of proteolysis and function. *Curr. Opin. Chem. Biol.* **2013**, *17*, 73–82. [[CrossRef](#)]
50. Eckhard, U.; Marino, G.; Butler, G.S.; Overall, C.M. Positional proteomics in the era of the human proteome project on the doorstep of precision medicine. *Biochimie* **2016**, *122*, 110–118. [[CrossRef](#)]
51. Gevaert, K.; Vandekerckhove, J. COFRADIC™: The Hubble telescope of proteomics. *Drug Discov. Today TARGETS* **2004**, *3*, 16–22. [[CrossRef](#)]
52. Venne, A.S.; Solari, F.A.; Faden, F.; Paretto, T.; Dissmeyer, N.; Zahedi, R.P. An improved workflow for quantitative N-terminal charge-based fractional diagonal chromatography (ChaFRADIC) to study proteolytic events in *Arabidopsis thaliana*. *Proteomics* **2015**, *15*, 2458–2469. [[CrossRef](#)] [[PubMed](#)]
53. Lai, Z.W.; Gomez-Auli, A.; Keller, E.; Mayer, B.; Biniousek, M.; Schilling, O. Enrichment of protein N-termini by charge reversal of internal peptides. *Proteomics* **2015**, *15*, 2470–2478. [[CrossRef](#)]
54. Thingholm, T.E.; Jørgensen, T.J.D.; Jensen, O.N.; Larsen, M.R. Highly selective enrichment of phosphorylated peptides using titanium dioxide. *Nat. Protoc.* **2006**, *1*, 1929–1935. [[CrossRef](#)] [[PubMed](#)]
55. Boutilier, J.M.; Warden, H.; Doucette, A.A.; Wentzell, P.D. Chromatographic behaviour of peptides following dimethylation with H₂/D₂-formaldehyde: Implications for comparative proteomics. *J. Chromatogr. B Anal. Technol. Biomed. Life Sci.* **2012**, *908*, 59–66. [[CrossRef](#)]
56. Mommen, G.P.M.; van de Waterbeemd, B.; Meiring, H.D.; Kersten, G.; Heck, A.J.R.; de Jong, A.P.J.M. Unbiased selective isolation of protein N-terminal peptides from complex proteome samples using phospho tagging (PTAG) and TiO₂-based depletion. *Mol. Cell. Proteom.* **2012**, *11*, 832–842. [[CrossRef](#)]
57. Guryca, V.; Lamerz, J.; Ducret, A.; Cutler, P. Qualitative improvement and quantitative assessment of N-terminomics. *Proteomics* **2012**, *12*, 1207–1216. [[CrossRef](#)] [[PubMed](#)]
58. Li, Q.; Roxas, B.A. An assessment of false discovery rates and statistical significance in label-free quantitative proteomics with combined filters. *BMC Bioinform.* **2009**, *10*, 43. [[CrossRef](#)] [[PubMed](#)]
59. Gupta, N.; Bandeira, N.; Keich, U.; Pevzner, P.A. Target-decoy approach and false discovery rate: When things may go wrong. *J. Am. Soc. Mass Spectrom.* **2011**, *22*, 1111–1120. [[CrossRef](#)]
60. Barboza, R.; Cociorva, D.; Xu, T.; Barbosa, V.C.; Perales, J.; Valente, R.H.; Franca, F.M.G.; Yates, J.R.; Carvalho, P.C. Can the false-discovery rate be misleading? *Proteomics* **2011**, *11*, 4105–4108. [[CrossRef](#)] [[PubMed](#)]
61. Jentoft, N.; Dearborn, D.G. Labeling of proteins by reductive methylation using sodium cyanoborohydride. *J. Biol. Chem.* **1979**, *254*, 4359–4365. [[PubMed](#)]
62. Means, G.E.; Feeney, R.E. Reductive alkylation of amino groups in proteins. *Biochemistry* **1968**, *7*, 2192–2201. [[CrossRef](#)] [[PubMed](#)]
63. Feeney, R.E.; Blankenhorn, G.; Dixon, H.B. Carbonyl-amine reactions in protein chemistry. *Adv. Protein Chem.* **1975**, *29*, 135–203. [[PubMed](#)]
64. Gidley, M.J.; Sanders, J.K. Reductive methylation of proteins with sodium cyanoborohydride. Identification, suppression and possible uses of N-cyanomethyl by-products. *Biochem. J.* **1982**, *203*, 331–334. [[CrossRef](#)] [[PubMed](#)]
65. Rowland, E.; Kim, J.; Bhuiyan, N.H.; van Wijk, K.J. The Arabidopsis Chloroplast Stroma N-Terminome: Complexities of Amino-Terminal Protein Maturation and Stability. *Plant Physiol.* **2015**, *169*, 1881–1896. [[CrossRef](#)]
66. Friedman, M.; Williams, L.D.; Masri, M.S. Reductive alkylation of proteins with aromatic aldehydes and sodium cyanoborohydride. *Int. J. Pept. Protein Res.* **1974**, *6*, 183–185. [[CrossRef](#)]
67. Lange, P.F.; Huesgen, P.F.; Overall, C.M. TopFIND 2.0—linking protein termini with proteolytic processing and modifications altering protein function. *Nucleic Acids Res.* **2012**, *40*, D351–D361. [[CrossRef](#)]

68. Hsu, J.L.; Huang, S.Y.; Shiea, J.T.; Huang, W.Y.; Chen, S.H. Beyond quantitative proteomics: Signal enhancement of the a1 ion as a mass tag for peptide sequencing using dimethyl labeling. *J. Proteome Res.* **2005**, *4*, 101–108. [[CrossRef](#)] [[PubMed](#)]
69. Krusemark, C.J.; Ferguson, J.T.; Wenger, C.D.; Kelleher, N.L.; Belshaw, P.J. Global amine and acid functional group modification of proteins. *Anal. Chem.* **2008**, *80*, 713–720. [[CrossRef](#)]
70. Boersema, P.J.; Raijmakers, R.; Lemeer, S.; Mohammed, S.; Heck, A.J.R. Multiplex peptide stable isotope dimethyl labeling for quantitative proteomics. *Nat. Protoc.* **2009**, *4*, 484–494. [[CrossRef](#)] [[PubMed](#)]
71. Jhan, S.Y.; Huang, L.J.; Wang, T.F.; Chou, H.H.; Chen, S.H. Dimethyl Labeling Coupled with Mass Spectrometry for Topographical Characterization of Primary Amines on Monoclonal Antibodies. *Anal. Chem.* **2017**, *89*, 4255–4263. [[CrossRef](#)] [[PubMed](#)]
72. Boersema, P.J.; Aye, T.T.; van Veen, T.A.; Heck, A.J.; Mohammed, S. Triplex protein quantification based on stable isotope labeling by peptide dimethylation applied to cell and tissue lysates. *Proteomics* **2008**, *8*, 4624–4632. [[CrossRef](#)] [[PubMed](#)]
73. Roperto, S.; Varano, M.; Russo, V.; Lucà, R.; Cagiola, M.; Gaspari, M.; Ceccarelli, D.M.; Cuda, G.; Roperto, F. Proteomic analysis of protein purified derivative of Mycobacterium bovis. *J. Transl. Med.* **2017**, *15*, 68. [[CrossRef](#)] [[PubMed](#)]
74. Salih, M.; Demmers, J.A.; Bezstarosti, K.; Leonhard, W.N.; Losekoot, M.; van Kooten, C.; Gansevoort, R.T.; Peters, D.J.; Zietse, R.; Hoorn, E.J.; et al. Proteomics of Urinary Vesicles Links Plakins and Complement to Polycystic Kidney Disease. *J. Am. Soc. Nephrol.* **2016**, *27*, 3079–3092. [[CrossRef](#)] [[PubMed](#)]
75. Varano, M.; Gaspari, M.; Quirino, A.; Cuda, G.; Liberto, M.C.; Foca, A. Temperature-dependent regulation of the Ochrobactrum anthropi proteome. *Proteomics* **2016**, *16*, 3019–3024. [[CrossRef](#)] [[PubMed](#)]



© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).