Visualisation tools for dependent peptide searches to support the exploration of *in vitro* protein modifications

George W. Preston[1,2,#a*], Liping Yang[2], David H. Phillips[1], Claudia S. Maier[2*]

[1] MRC-PHE Centre for Environment & Health, Department of Analytical, Environmental & Forensic Sciences, School of Population Health & Environmental Sciences, Faculty of Life Sciences & Medicine, King's College London, London, United Kingdom

[2] Department of Chemistry, Oregon State University, Corvallis, OR, United States of America

[#a] Current address: Stoller Biomarker Discovery Centre, University of Manchester, Manchester, United Kingdom

[*] Corresponding authors

E-mail: george.preston@kcl.ac.uk (GWP), claudia.maier@oregonstate.edu (CSM)

Abstract

Dependent peptide searching is a method for discovering covalently-modified peptides – and therefore proteins – in mass-spectrometry-based proteomics experiments. Being more permissive than standard search methods, it has the potential to discover novel modifications (e.g., post-translational modifications occurring *in vivo*, or modifications introduced *in vitro*). However, few studies have explored dependent peptide search results in an untargeted way. In the present study, we sought to evaluate dependent peptide searching as a means of characterising proteins that have been modified *in vitro*. We generated a model data set by analysing *N*-ethylmaleimide-treated bovine serum albumin, and performed dependent peptide searches using the popular MaxQuant software. To facilitate interpretation of the search results (hundreds of dependent peptides), we developed a series of visualisation tools (R scripts). We used the tools to assess the diversity of putative modifications in the albumin, and to pinpoint hypothesised modifications. We went on to explore the tools' generality via analyses of public data from studies of rat and human proteomes. Of 19 expected sites of modification (one in rat cofilin-1 and 18 across six different human plasma proteins), eight were found and correctly localised. Apparently, some sites went undetected because chemical enrichment had depleted necessary analytes (potential 'base' peptides). Our results demonstrate (i) the ability of the tools to provide accurate and informative visualisations, and (ii) the usefulness of dependent peptide searching for characterising *in vitro* protein modifications. Our model data are available via PRIDE/ProteomeXchange (accession number PXD013040).

Introduction

By the time a protein is subjected to analysis, it can have acquired one or more covalent modifications. These could include modifications of biological origin, modifications introduced

deliberately (e.g., to probe protein structure and function), and modifications occurring during sample preparation and storage. In bottom-up mass-spectrometry-based proteomics, where proteins are digested and analysed as peptides, prior knowledge of modifications can enable more of the acquired spectra to be identified [1]. Known or suspected modifications are specified as parameters of a database search, enabling more of the protein sequence to be mapped, and also allowing the modifications themselves to be localised and quantified. For partially characterised or unknown modifications, however, this approach is not practical: specifying a long list of variable modifications (e.g., as a way of capturing unknown modifications) would expand the database dramatically, lengthening the search duration and reducing the number of confidently identified spectra [2]. New types of search have been developed to address this problem [2-9]. They include 'open' database searches, which permit precursor ions with shifted masses [6]; and 'spectral pair' searches, in which unidentified spectra are matched to spectral libraries [5]. An example of the latter approach is dependent peptide (DP) searching [3]. In a typical case, a DP is a chromatographic feature that is not identified by a database search, but whose fragment-ion spectrum partially matches that of one of the search hits (the 'base' peptide). The DP is typically a modified form of the base peptide, and the two features' masses differ. In theory, some of the DP's product ions will be the same as the corresponding product ions of the base peptide, while others will display the mass difference (Δm). Crucially, Δm does not need to be specified *a priori*, as it is calculated for every pair of unidentified feature and database-search hit. Identifying features in this way can take much less time than a database search [3] but does confer certain limitations: sites that are fully occupied by unknown modifications cannot be detected; and overall sequence coverage is unlikely to be extended.

Originally implemented as a stand-alone tool (ModifiComb [3, 10]), DP searching has recently been incorporated into the MaxQuant software [11, 12]. Within MaxQuant, the DP search can utilise hits (i.e., potential base peptides) generated by the Andromeda search engine [13]. Studies utilising MaxQuant's DP search function have confirmed its potential to discover modifications [12, 14-21]. Lassak et al. used the function to analyse a bacterial translation elongation factor, and discovered a novel type of glycosylation [14]. Mordret et al. used the function to detect single amino

acid substitutions, and also carried out a general test of its validity [12]. A large set of synthetic phosphopeptides and corresponding unmodified peptides was analysed, and the phosphoryl modifications were left for a DP search to find. The search identified over a thousand spectra as belonging to singly-phosphorylated peptides, and all of these were true positives [12]. Few studies, however, have explored DP search results in an untargeted way. In the present study, we sought to evaluate DP searching as a means of characterising *in vitro* modified proteins. First, we generated model data by analysing a model protein (bovine serum albumin, BSA) that had been treated with a protein-modifying reagent (*N*-ethylmaleimide, NEM). Then, we performed DP searches and attempted to rationalise the search results.

Visualisation tools can greatly facilitate the interpretation of proteomic mass spectrometry data and database search results [22-25]. We identified a need for tools that visualise DP search results, and to meet this need we wrote a set of five scripts in the R language [26]. Three of the scripts are for surveying distributions of DPs (i.e., are hypothesis-generating), and the other two are for pinpointing hypothesised modifications (i.e., are hypothesis-testing). Some of the scripts can enrich DPs for modifications that are unique to a test sample. Herein we report search results and visualisations for our own data, as well as for public data from two other studies [27, 28]. The results demonstrate how a combination of DP searching and visualisation can assist in the characterisation of *in vitro* modified proteins. The approach could be useful for characterising protein targets of enzyme activities and reactive small molecules.

Materials and methods

Preparation and analysis of modified BSA

BSA (1 mg mL^{-1}) was reacted with NEM (1 mM) in potassium phosphate buffer (100 mM) at pH 7.4. Unreacted NEM was scavenged with 1,4-dithiothreitol (DTT). The protein was purified

(buffer exchange), reduced (DTT), alkylated (iodoacetamide), purified again (acetone precipitation), and digested (trypsin). The peptides were analysed in duplicate (analytical replicates 1 and 2) by reversed-phase nano liquid chromatography (nanoACQUITY liquid chromatograph; Waters, Milford, Massachusetts, USA) with online data-dependent tandem mass spectrometry (Orbitrap Fusion Lumos mass spectrometer; ThermoFisher Scientific, Waltham, Massachusetts, USA). A control sample (untreated BSA) was also prepared and analysed. Further details of materials and methods can be found in S1 Text. Data for NEM-treated and untreated BSA have been deposited in PRIDE/ProteomeXchange [29] (accession number PXD013040).

Public data

Further mass spectrometry data were obtained from PRIDE/ProteomeXchange [29]. Data were selected according to the following criteria: (i) experiment involving exposure of one or more proteins to a protein-modifying reagent; (ii) data collected using standard data-dependent acquisition mass spectrometry; (iii) control data and two replicates available (not an essential criterion); (iv) results of a variable-modification database search reported in the literature. The following files/groups of files met the criteria and were included in the present study (the selection was not exhaustive): DMF_Cofilin1A.raw from PRIDE project PXD008314 [27]; and 1362-cs774_0_a.raw, 1364-cs774_0_b.raw, 1380-cs774_5_a.raw and 1382-cs774_5_b.raw from PRIDE project PXD006663 [28]. Sequence data (*.fasta files) were obtained from the RCSB Protein Data Bank [30] (accession numbers 4F5S [31] and 1S81 [32]) and UniProt [33] (accession numbers P02042, P02647, P02766, P02768, P02787, P45592 and P68871, and the UniProtKB/Swiss-Prot human proteome, 4th June 2019). Where possible, sequences were obtained without extraneous elements such as signal peptides. The data from Protein Data Bank accession number 4F5S consisted of two identical sequences, and so one of these was removed. MaxQuant's database of contaminants [11] was used either as supplied or in an edited form (see S1 Text).

Database searches and dependent peptide searches

All searches were done in MaxQuant (Max Planck Institute of Biochemistry, version 1.6.0.1) [11]. Database searches were done using Andromeda [13]. Individual *.raw files were searched against databases consisting of either a protein of interest plus potential contaminants (1 + 244 or 1 + 245 sequences) or the UniProtKB/Swiss-Prot human proteome plus potential contaminants (20,406 + 82 sequences). The *in silico* digestion was done in 'specific' mode, using 'Trypsin/P' as the enzyme, and allowing for a maximum of two missed cleavages. The maximum peptide mass was adjusted so as to include ≥95% of relevant theoretical peptides (see S1 Text). The minimum peptide length was seven amino acid residues. Additional *in silico* digestions were done using PeptideMass [34] (see S1 Text).

When the purpose of the database search was to discover potential base peptides, a minimal set of variable modifications (methionine oxidation and protein N-terminal acetylation) and an appropriate fixed modification (cysteine *S*-carbamidomethylation [28] or *S*-pyridylethylation [27]) were specified. When the purpose was instead to localise a particular modification (a variable-modification search), the fixed and variable modifications were adjusted accordingly (see S1 Text). The maximum number of modifications per peptide was always five, and the 'second peptides' function was always enabled. DP searches were appended to their respective database searches by enabling the 'dependent peptides' function (false discovery rate of 1%, mass bin size of 0.0065 Da). Results of DP searches were obtained from allPeptides.txt files [11] and filtered as described below. For selected DPs, we investigated whether the same chromatographic feature could also be detected by a variable-modification search (see S1 Text).

Development of visualisation tools

All tools (Scripts I-V; S1-5 Scripts) were developed in R for Windows (R Core Team, version 3.4.0 or later) [26] using functions from the base packages, plus the additional function *read.fasta* from the 'seqinR' package (version 3.3-6 or later) [35]. Scripts I-IV were developed and tested on a

Dell desktop PC (Intel Core i5-7500 processor, 8 GB RAM) running Windows 10. Script V was developed and tested on a Toshiba laptop PC (AMD E1-2100 APU processor, 4 GB RAM) running Windows 8.1. Each script requires a set of search results (allPeptides.txt files), a protein sequence (*.fasta file), and the identifier of a protein of interest (e.g., a UniProt identifier). The search results are filtered (see Table B of S1 Text) and a table of DPs is prepared. DPs are localised to segments of the protein sequence using a 'sliding window' [36]. Different scripts require different numbers of allPeptides.txt files and process the data to different extents. Scripts I, II and III are for surveying DPs' Δ*m* values; they return DP localisation plots and Δ*m* frequency histograms. Scripts IV and V are for pinpointing particular modifications; they require an expected Δ*m* value, and they return plots of DPs' localisation probabilities [12]. Scripts II, III and IV have the ability to enrich the DPs for modifications that are unique to a test sample (see 'Results and discussion'). A set of notes explaining how the scripts work can be found in Table C of S1 Text. We will endeavour to maintain the scripts' compatibility with relevant software, and to address any limitations that come to light. Any future versions of the scripts will be made available from https://github.com/preston-gw/.

The accuracy of the visualisations was tested by manually mapping selected DPs onto graphics generated by scripts (20 DPs across five graphics, all confirmed to have been mapped correctly). After this, no significant changes were made to either the main data processing code or the mapping code (changes to graphical parameters, for example, were permitted). Certain graphics were prepared in batches by iterating an appropriate script. Figures were prepared from R output files using Inkscape (Free Software Foundation, version 0.91 or later) and GNU Image Manipulation Program (GIMP Development Team, version 2.10.8) (see S1 Text). Figures such as S4 Fig and S8 Fig are representative of the graphics generated by the scripts.

Visualisation tools' instructions for use

1. Open R (version 3.6.0 or later)
2. Make sure that package 'seqinR' [35] is installed. Installation is achieved by entering *install.packages("seqinr")* in the R console.

3. Open the script within R (File > Open script)
4. Review the script and complete file paths as appropriate. Further instructions are included in the header and body of the script. Explanatory notes can be found in Tables B and C of S1 Text.
5. Save a copy of the script if desired (File > Save as)
6. Run the script (Edit > Run all)
7. A graphic should appear on-screen. The times taken to visualise our model data using one of the aforementioned PCs were as follows: Script I, 8 s (desktop); Script II, 27 s (desktop), Script III, 62 s (desktop); Script IV, 16 s (desktop); Script V, 26 s (laptop).

Mass calculations

Expected Δ*m* values were calculated in R (version 3.4.0 or later) [26] using monoisotopic masses from ChemDraw (various versions, PerkinElmer) or Unimod [37] (http://www.unimod.org/masses.html). Masses used for calculations were accurate to at least four decimal places. Δ*m* values mentioned in the text have been rounded to two decimal places.

Statistical methods

Pairwise comparisons of Δ*m* frequency histograms were done using the *cor.test* function (Spearman method) in R (version 3.4.0) [26].

Results and discussion

Exploration of model data

BSA was selected as a model protein because it is well-characterised and contains a number of nucleophilic (i.e., potentially modifiable) amino acid residues [38, 39]. NEM was selected as the protein-modifying reagent because it is reactive towards a variety of amino acid side chains (those of cysteine, lysine and histidine) [40]. We predicted that NEM would modify BSA's only reduced cysteine residue, Cys-34 [31], as well as one or more lysine and/or histidine residues. The primary products of the reaction of BSA with NEM were expected to be Michael adducts, in which a hydrogen atom of the protein has effectively been replaced by an *N*-ethylsuccinimidyl (NESyl) group (Δm = +125.05 Da). NESyl groups attached to cysteine residues are susceptible to hydrolysis (additional Δm = +18.01 Da) [41, 42], and we assumed that this would also be the case for NESyl groups attached to lysine or histidine residues. Additionally, sulfur atoms to which NESyl groups are attached may oxidise [41].

Large numbers of chromatographic features were detected in analyses of NEM-treated BSA ($N \geq 34{,}812$), and also in analyses of untreated BSA ($N \geq 44{,}430$). Five to six percent of the features ($1958 \leq N \leq 2328$) were identified by MaxQuant as either potential base peptides (Andromeda search hits, 23-25% of identified features) or DPs (75-77% of identified features) (S1 Fig). Peptides of BSA (95% of identified features) were 9-10 times as numerous as predicted by *in silico* digestion and oxidation ($N = 218$). The high ratio of observed to expected features implies that large numbers of modifications had occurred independently of NEM treatment (e.g., artefacts of sample preparation or modifications pre-existing in the BSA). The detection of so many 'background' modifications, although difficult to account for, is consistent with Nielsen and coauthors' estimate of 8-12 modified peptides per unmodified tryptic peptide [10]. Filters were employed to isolate the DPs, to limit the number of 'background' modifications (see Table B of S1 Text), and to limit Δm to ±500 Da (for clarity of visualisation). Filtering removed 70-74% of the identified features (S1 Fig).

Visualisation of Δ*m* distributions

The filtered Δ*m* values were visualised in two ways: firstly by mapping DPs to segments of the protein sequence, and secondly using a frequency histogram [3, 25]. These modes of visualisation, both achieved using Script I, revealed a diversity of putative modifications in NEM-treated BSA (Fig 1A, B). Visualisations of putative modifications to porcine trypsin, also detected in analyses of NEM-treated BSA, demonstrated the flexibility of Script I (S2 Fig).

Fig 1. Localisation plots and mass-shift (Δ*m*) frequency histograms for dependent peptides (DPs) of *N*-ethylmaleimide-treated bovine serum albumin. In the localisation plots (left-hand panels), the protein sequence is represented as a dashed line that becomes solid in regions for which peptides were observed. X-axis values refer to positions in the protein sequence (position 1 = N-terminal amino acid residue). Each DP is represented as a rectangle whose height is proportional to Δ*m*, and whose grey border is partially transparent. The Δ*m* values are summarised in frequency histograms (right-hand panels). The DPs were unenriched (A, B), enriched using Script II (C, D) or enriched using Script III (E, F).

Δ*m* frequency histograms were used to investigate whether modifications in NEM-treated and untreated BSA were the same. Histograms for the two samples were similar (nearly as similar, or more similar, than histograms for analytical replicates; see S3 Fig). This suggested that the samples had many modifications in common. In order to selectively visualise the NEM-derived modifications, we investigated ways of enriching DPs. We started with a method (Script II) that subtracts the DPs observed in an analysis of untreated protein from those observed in an analysis of treated protein (Fig 1C, D). For this purpose, DPs were regarded as simple combinations of sequence and Δ*m* (in principle, retention time could also be used, but this was not attempted here). We then took the enrichment idea a stage further (Script III) by looking for DPs that were ‘constantly conjoined’ with NEM treatment (i.e., observed in both analyses of NEM-treated BSA, but in neither analysis of untreated BSA). Enrichment was quantified as an increase in the percentage of DPs having either of two putative NEM-derived groups: intact NESyl (Δ*m* ± tolerance = +125.05 ± 0.01 Da) or hydrolysed

NESyl (Δm ± tolerance = +143.06 ± 0.01 Da). Both modifications were observed for NEM-treated BSA (8.7% of the DPs from analysis 1, Fig 1B) and neither was observed for untreated BSA. Script II effected 2.3-fold enrichment of DPs from analysis 1 of NEM-treated BSA (Fig 1D). Script III effected 6.0-fold enrichment of these same DPs, but its greater stringency led to the exclusion of eleven relevant DPs (Fig 1F). The results suggest that our scripts should be able to enrich DPs even when the modifications of interest are unknown. In other words, the scripts might be able to discover novel modifications and attribute them to a given reagent or condition.

Visualisation of expected modifications

As well as surveying the diversity of modifications, we were also interested in visualising the distributions of specific hypothesised modifications. For this purpose, we developed a method that maps 'constantly conjoined' DPs to the protein sequence and highlights their probable sites of modification (Script IV). A sliding window is used as before, but in Script IV its role is to direct the entry of values into matrices. Each DP is 'etched' (as a line of ones) into a blank matrix, and the localisation probabilities for that DP's modification are deposited in a corresponding zero matrix (localisation probability is a computed measure of the likelihood of a modification occurring at a given site [11, 12]). The two matrices are subsequently converted to images (R function *image* [26]) and merged. Fig 2 is a formatted version of one of the resulting graphics, showing the distribution of putative hydrolysed NESyl groups. S4 Fig is an example of an unformatted graphic, showing the distribution of putative intact NESyl groups. Note how some of the DPs in Fig 2 occur as pairs of putative diastereoisomers [43] (i.e., modified peptides with identical *m*/*z* values and fragmentation patterns but different retention times).

Fig 2. Localisation plot for putative hydrolysed *N*-ethylsuccinimidyl groups in *N*-ethylmaleimide-treated bovine serum albumin (Δm ± tolerance = +143.06 ± 0.01 Da). The protein sequence is represented as a dashed line that becomes solid in regions for which peptides were observed. X-axis values refer to positions in the protein sequence (position 1 = N-terminal amino acid residue). Dependent peptides are represented as coloured strips with shading to indicate the localisation

probability (darker = more probable). Any site with a non-zero probability is annotated. One dependent peptide (amino acid residues 336-347, probably modified at Arg-336 or His-337) does not appear because the relevant matrices were full.

Using the hydrolysed NESyl group as an example, we examined whether the putative modifications had been localised to plausible sites. In 23 of 34 DPs with putative hydrolysed NESyl groups (68%), the highest localisation probability had been assigned to a cysteine, histidine or lysine residue (S1 Table). In cases where the same probability had been assigned to multiple sites, we used a system of prioritisation to narrow down the possibilities (see S1 Text for details). Six of the 23 plausible localisations were confirmed by a variable-modification search (S1 Table). In all six cases, the modification could be localised to a histidine or lysine residue. Modifications to cysteine residues, which represent a special case, are discussed below. A seventh, less plausible localisation (to Asp-13) was also confirmed by the variable-modification search (S1 Table).

In order to discover the DPs with modified cysteine residues, we had to account for the fact that the corresponding residues in the base peptides would also be modified (*S*-carbamidomethylated). Subtracting the Δm for carbamidomethylation gave new values for NESyl (+68.03 Da) and hydrolysed NESyl (+86.04 Da), both of which we recognised from the Δm frequency histograms (Fig 1). Surprisingly, neither modification was localised to Cys-34 (S5 and S6 Figs), and no modified cysteine residues were confirmed by the variable-modification search. Seeking to understand the apparent absence of modifications to Cys-34, we turned to a group of unexplained DPs ($\Delta m = -25.03$ Da; Fig 1F), which we speculated might contain oxidised cysteine residues (cysteinesulfinic acid). The modification was found to have been localised to Cys-34 in some DPs, but none that were 'constantly conjoined' with NEM treatment (S7 Fig). The ambiguous results for Cys-34 are possibly a consequence of modifications at this site having decomposed prior to or during analysis. We did see some evidence of modification to cysteine residues other than Cys-34 (S1 Table, S5-7 Figs), and this was unexpected because these residues are normally disulfide-bonded to other cysteine residues [31]. It is possible that modifications to cysteine residues other than Cys-34 occurred when DTT was added

to scavenge unreacted NEM. It is also possible that some of the other modifications observed in the study occurred following this addition of DTT.

The above results highlight the fact that observed Δ*m* values do not always correspond to real chemical transformations, and cannot always be interpreted directly. Direct interpretation is permitted if the DP is singly-modified and the base peptide is truly unmodified, and likewise if the DP and base peptide contain modifications that 'cancel out' (e.g., cysteine *S*-carbamidomethylation). However, if the base peptide contains modifications not found in the DP, or if the base peptide and DP contain different modifications at the same site, then interpretation will be less straightforward. Problems of this nature could be avoided by omitting modifications from the database search, but this would of course restrict the overall number of identifications.

Validation of tools using public data

We explored the scripts' generality by applying them to analyses of public data. We hypothesised that a combination of DP searching and visualisation would reveal the same adducts as other authors had found by variable-modification database searching. We expected to observe these adducts via targeted visualisation (e.g., using Script IV). We also predicted that they would be evident from an untargeted survey (e.g., using Script III).

First, we analysed data from a study by Piroli et al. [27]. In this work, cultured cells (rat astrocytes) were exposed to the drug dimethyl fumarate. The authors collected proteins from the exposed cells, resolved them on gels, and then subjected individual protein bands to reduction, alkylation (4-vinylpyridine), digestion (trypsin) and analysis. In one of the protein bands, the authors detected a monomethyl fumarate adduct of cofilin-1 and localised the modification to a cysteine residue (Cys-139). Using the methods developed for the analysis of NEM-treated BSA, we performed an independent analysis of Piroli and coauthors' cofilin-1 data. Script I was used to survey Δ*m* values, and an additional script was developed for mapping the expected Δ*m* (Script V, an analogue of Script IV that works with single allPeptides.txt files). Script I did not highlight the expected nominal Δ*m*

(+25 Da; S8 Fig), which is perhaps understandable given this script's inability to enrich DPs. The overall sparsity of S8 Fig could reflect a real lack of modifications in the rat cofilin-1, or alternatively it could reflect qualities of the sample and/or data. Script V revealed that the expected Δ*m* (+24.97 ± 0.01 Da) was present in one DP, and that it had been localised to the correct cysteine residue (S9 Fig). The Δ*m* itself is also evidence of correct localisation, since this is the difference in mass with respect to a pyridylethylated base peptide. The discrepancy in the site numbers (138 versus 139) arose probably because we used the sequence of mature rat cofilin-1 (no N-terminal methionine residue) whereas Piroli et al. used the sequence of the full-length protein (we used mature sequences where possible to ensure that N-terminal peptides could be found by Andromeda).

Further data were from a study by Salomón et al. [28]. In this work, the reactivity of human blood proteins towards a metabolite, 3-hydroxy-2,5-hexanedione, was explored using an alkyne-functionalised probe ('alk-3-HHD'). The authors prepared plasma from probe-treated blood and collected the plasma proteins. The proteins were reduced, alkylated (iodoacetamide) and digested (trypsin), and the resulting peptides were chemically enriched for alk-3-HHD adducts. The authors detected two different types of modification ('HTO' and 'HDMP', both specific to lysine residues) among six polypeptides (apolipoprotein A-I, haemoglobin β- and δ-chains, serotransferrin, serum albumin, and transthyretin). In total, the authors detected 18 unique sites of modification among 22 analytes. Again, we applied methods that had been developed for the BSA adducts. Salomón and coauthors' dataset included replicates and a control, permitting the use of Scripts III and IV. To maximise contrast, we used data for the highest concentration of alk-3-HHD. Script III revealed multiple 'constantly conjoined' DPs of each of the six aforementioned polypeptides, but did not highlight any of the expected Δ*m* values (Fig 3A, B; S10-14 Figs). Six DPs mapped to both of the haemoglobin chains, and therefore could not be localised unambiguously. Script IV was used to map HTO- and HDMP-type modifications (with or without sulfonation [28]) to the sequences of the polypeptides. In total, 14 DPs with putative alk-3-HHD-derived modifications were detected (11 unique combinations of sequence and modification) (S2 Table). In each of the DPs, a lysine residue was either the site with the highest localisation probability, or was one of multiple such sites. Most of

the DPs (93%) were of either haemoglobin β-chain (Fig 3C) or serum albumin (S15 Fig). Of the 22 analytes reported by Salomón et al., we detected six (27%) as DPs. Of the 18 expected sites of modification, we observed seven (39%). It is perhaps unsurprising that some of the expected DPs were not detected, since the chemical enrichment performed by Salomón et al. had the potential to remove their corresponding base peptides (an effect alluded to by Tyanova et al. in their protocol [11]). Indeed, for seven of the 22 expected analytes (32%), the absence of a required base peptide was sufficient to explain the absence of the DP.

Fig 3. Visualisation of dependent peptides of human haemoglobin β-chain using public data from the study by Salomón et al. [28]. (A) Dependent peptide localisation plot. (B) Mass-shift (Δ*m*) frequency histogram. (C) Localisation plot for putative sulfonated 'HDMP'-type modifications (Δ*m* ± tolerance = +390.10 ± 0.01 Da) [28]. Six of 106 dependent peptides also mapped to the sequence of haemoglobin δ-chain (S11 Fig). Base peptide VLGAFSDGLAHLDNLKGTFATLSELHCDK went undetected in analyses of untreated proteins and therefore does not appear (see Table C of S1 Text).

Scope for extending the present study

There is scope beyond the present study for developing and integrating the visualisation tools. As R scripts, they are highly amenable to modification, and could be adapted for more specialised purposes. The filters and the Δ*m* tolerance could made more stringent or permissive as required. The plots could be customised by changing the colour scheme or narrowing the limits of the Δ*m* axes. Another idea would be to rotate the histogram so as to align its Δ*m* axis with that of the DP localisation plot.

One area in which there is significant scope for development is annotation. Currently, the scripts highlight mass shifts but do not attempt to identify them. Some identifications are already made by MaxQuant, and these could be easily transferred onto the plots. Further identifications could be made via public protein-modification databases such as Unimod [37] and RESID [44]. These databases contain calculated Δ*m* values via which observed Δ*m* values could be linked to

modifications' identities. Identities could be added to the plots or visualised separately (e.g., a word-cloud of modifications' names). Another idea would be to highlight particular features of the protein sequence, such as nucleophilic amino acid residues. This could be done using lines, symbols or text.

Currently, each script visualises results for a single protein (Scripts I-III) or combination of protein and Δ*m* (Scripts IV and V). One way of extending the approach would be to iterate scripts so that they cycle through lists of proteins and/or Δ*m* values (in fact, we prepared certain groups of figures in this way). In theory, this could be done in a 'data-dependent' fashion by extracting the lists directly from allPeptides.txt. If this were attempted, each graphic (e.g., *.svg file) would have to be stamped with the protein identifier and/or Δ*m* value.

Scripts II, III and IV are able to enrich DPs for modifications that are unique to a test sample. The modes of enrichment employed by the scripts are simple but should work well for many *in vitro* modifications (especially modifications not found *in vivo*). The visualisation of *in vivo* modifications would be an impactful next step, but one that might require a more quantitative approach: it would be helpful to visualise differences in abundance in addition to the simple difference between presence and absence.

Finally, there is scope for combining the scripts in an R package [45]. This would promote their usability beyond the present study.

Summary

We have developed and tested a set of analytical tools with which to interpret the results of a DP search. The tools visualise putative modifications (Δ*m* values) for a protein of interest (either an isolated protein or a component of a proteome). Some of the tools are able to enrich DPs for modifications that are unique to a test sample. We envisaged that an untargeted survey of DPs (using Scripts I-III) might generate hypotheses that could then be tested via targeted visualisation (using Scripts IV and V). This approach helped us to achieve our aim of rationalising DP search results for NEM-treated BSA. Expected modifications were found, and the majority of these were localised to

chemically plausible sites. In formal tests involving public data, a number of expected modifications were detected and correctly localised (although here the methods for surveying Δ*m* values proved less helpful than they had for the BSA study). Analyses of cysteine-specific modifications led us to consider the effect of fixed and variable modifications on Δ*m*; and analyses of data for chemically enriched peptides led us to consider the potential of chemical enrichment to limit DP searches. We conclude (i) that the tools can summarise DP search results accurately and informatively, and (ii) that DP searching can be useful for characterising *in vitro* modified proteins.

References

1. Cottrell JS. Protein identification using MS/MS data. J Proteomics. 2011;74(10):1842-51. doi: 10.1016/j.jprot.2011.05.014.

2. Ahrné E, Müller M, Lisacek F. Unrestricted identification of modified proteins using MS/MS. Proteomics. 2010;10(4):671-86. doi: 10.1002/pmic.200900502.

3. Savitski MM, Nielsen ML, Zubarev RA. ModifiComb, a New Proteomic Tool for Mapping Substoichiometric Post-translational Modifications, Finding Novel Types of Modifications, and Fingerprinting Complex Protein Mixtures. Mol Cell Proteomics. 2006;5(5):935-48. doi: 10.1074/mcp.T500034-MCP200.

4. Ahrné E, Nikitin F, Lisacek F, Müller M. QuickMod: A Tool for Open Modification Spectrum Library Searches. J Proteome Res. 2011;10(7):2913-21. doi: 10.1021/pr200152g.

5. Na S, Paek E. Software eyes for protein post-translational modifications. Mass Spectrom Rev. 2015;34(2):133-47. doi: 10.1002/mas.21425.

6. Chick JM, Kolippakkam D, Nusinow DP, Zhai B, Rad R, Huttlin EL, et al. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. Nat Biotechnol. 2015;33:743-9. doi: 10.1038/nbt.3267.

7. Bagwan N, Bonzon-Kulichenko E, Calvo E, Lechuga-Vieco AV, Michalakopoulos S, Trevisan-Herraz M, et al. Comprehensive Quantification of the Modified Proteome Reveals Oxidative Heart Damage in Mitochondrial Heteroplasmy. Cell Rep. 2018;23(12):3685-97. doi: 10.1016/j.celrep.2018.05.080.

8. Devabhaktuni A, Lin S, Zhang L, Swaminathan K, Gonzalez CG, Olsson N, et al. TagGraph reveals vast protein modification landscapes from large tandem mass spectrometry datasets. Nat Biotechnol. 2019;37(4):469-79. doi: 10.1038/s41587-019-0067-5.

9. Na S, Kim J, Paek E. MODplus: Robust and Unrestrictive Identification of Post-Translational Modifications Using Mass Spectrometry. Anal Chem. 2019;91(17):11324-33. doi: 10.1021/acs.analchem.9b02445.

10. Nielsen ML, Savitski MM, Zubarev RA. Extent of Modifications in Human Proteome Samples and Their Effect on Dynamic Range of Analysis in Shotgun Proteomics. Mol Cell Proteomics. 2006;5(12):2384-91. doi: 10.1074/mcp.M600248-MCP200.

11. Tyanova S, Temu T, Cox J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nat Protoc. 2016;11:2301-19. doi: 10.1038/nprot.2016.136.

12. Mordret E, Dahan O, Asraf O, Rak R, Yehonadav A, Barnabas GD, et al. Systematic Detection of Amino Acid Substitutions in Proteomes Reveals Mechanistic Basis of Ribosome Errors and Selection for Translation Fidelity. Mol Cell. 2019;75(3):427-41. doi: 10.1016/j.molcel.2019.06.041.

13. Cox J, Neuhauser N, Michalski A, Scheltema RA, Olsen JV, Mann M. Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. J Proteome Res. 2011;10(4):1794-805. doi: 10.1021/pr101065j.

14. Lassak J, Keilhauer EC, Fürst M, Wuichet K, Gödeke J, Starosta AL, et al. Arginine-rhamnosylation as new strategy to activate translation elongation factor P. Nat Chem Biol. 2015;11:266-70. doi: 10.1038/nchembio.1751.

15. Soufi B, Krug K, Harst A, Macek B. Characterization of the E. coli proteome and its modifications during growth and ethanol stress. Front Microbiol. 2015;6:103. doi: 10.3389/fmicb.2015.00103.

16. Bogdanow B, Zauber H, Selbach M. Systematic Errors in Peptide and Protein Identification and Quantification by Modified Peptides. Mol Cell Proteomics. 2016;15(8):2791-801. doi: 10.1074/mcp.M115.055103.

17. Cvetesic N, Semanjski M, Soufi B, Krug K, Gruic-Sovulj I, Macek B. Proteome-wide measurement of non-canonical bacterial mistranslation by quantitative mass spectrometry of protein modifications. Sci Rep. 2016;6:28631. doi: 10.1038/srep28631.

18. Rykær M, Svensson B, Davies MJ, Hägglund P. Unrestricted Mass Spectrometric Data Analysis for Identification, Localization, and Quantification of Oxidative Protein Modifications. J Proteome Res. 2017;16(11):3978-88. doi: 10.1021/acs.jproteome.7b00330.

19. van Mierlo G, Wester RA, Marks H. Quantitative subcellular proteomics using SILAC reveals enhanced metabolic buffering in the pluripotent ground state. Stem Cell Res. 2018;33:135-45. doi: 10.1016/j.scr.2018.09.017.

20. Tiwari MK, Hägglund PM, Møller IM, Davies MJ, Bjerrum MJ. Copper ion / H_2O_2 oxidation of Cu/Zn-Superoxide dismutase: Implications for enzymatic activity and antioxidant action. Redox Biol. 2019;26:101262. doi: 10.1016/j.redox.2019.101262.

21. Sinitcyn P, Rudolph JD, Cox J. Computational Methods for Understanding Mass Spectrometry-Based Shotgun Proteomics Data. Annu Rev Biomed Data Sci. 2018;1(1):207-34. doi: 10.1146/annurev-biodatasci-080917-013516.

22. Martín-Campos T, Mylonas R, Masselot A, Waridel P, Petricevic T, Xenarios I, et al. MsViz: A Graphical Software Tool for In-Depth Manual Validation and Quantitation of Post-translational Modifications. J Proteome Res. 2017;16(8):3092-101. doi: 10.1021/acs.jproteome.7b00194.

23. Avtonomov DM, Kong A, Nesvizhskii AI. DeltaMass: Automated Detection and Visualization of Mass Shifts in Proteomic Open-Search Results. J Proteome Res. 2019;18(2):715-20. doi: 10.1021/acs.jproteome.8b00728.

24. Tyanova S, Temu T, Carlson A, Sinitcyn P, Mann M, Cox J. Visualization of LC-MS/MS proteomics data in MaxQuant. Proteomics. 2015;15(8):1453-6. doi: 10.1002/pmic.201400449.

25. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods. 2016;13(9):731-40. doi: 10.1038/nmeth.3901.

26. R Core Team. A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2017.

27. Piroli GG, Manuel AM, Patel T, Walla MD, Shi L, Lanci SA, et al. Identification of Novel Protein Targets of Dimethyl Fumarate Modification in Neurons and Astrocytes Reveals Actions Independent of Nrf2 Stabilization. Mol Cell Proteomics. 2019;18(3):504-19. doi: 10.1074/mcp.RA118.000922.

28. Salomón T, Sibbersen C, Hansen J, Britz D, Svart MV, Voss TS, et al. Ketone Body Acetoacetate Buffers Methylglyoxal via a Non-enzymatic Conversion during Diabetic and Dietary Ketosis. Cell Chem Biol. 2017;24(8):935-43. doi: 10.1016/j.chembiol.2017.07.012.

29. Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res. 2019;47(D1):D442-50. doi: 10.1093/nar/gky1106.

30. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000;28(1):235-42. doi: 10.1093/nar/28.1.235.

31. Bujacz A. Structures of bovine, equine and leporine serum albumin. Acta Crystallogr D. 2012;68(10):1278-89. doi: 10.1107/S0907444912027047.

32. Transue TR, Krahn JM, Gabel SA, DeRose EF, London RE. X-ray and NMR characterization of covalent complexes of trypsin, borate, and alcohols. Biochemistry. 2004;43(10):2829-39. doi: 10.1021/bi035782y.

33. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2018;47(D1):D506-15. doi: 10.1093/nar/gky1049.

34. Wilkins MR, Lindskog I, Gasteiger E, Bairoch A, Sanchez J-C, Hochstrasser DF, et al. Detailed peptide characterization using PEPTIDEMASS – a World-Wide-Web-accessible tool. Electrophoresis. 1997;18(3-4):403-8. doi: 10.1002/elps.1150180314.

506 35. Charif D, Lobry JR. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical
507 Computing Devoted to Biological Sequences Retrieval. In: Bastolla U, Porto M, Roman E,
508 Vendruscolo M, editors. Structural Approaches to Sequence Evolution: Molecules, Networks,
509 Populations. New York: Springer Verlag; 2007. pp. 207-32.
510 36. States DJ, Boguski MS. Dot Matrix Methods. In: Gribskov M, Devereux J, editors. Sequence
511 Analysis Primer. New York: Oxford University Press; 1992. pp. 92-124.
512 37. Creasy DM, Cottrell JS. Unimod: Protein modifications for mass spectrometry. Proteomics.
513 2004;4(6):1534-6. doi: 10.1002/pmic.200300744.
514 38. Diez MJF, Osuga DT, Feeney RE. The sulfhydryls of avian ovalbumins, bovine β-
515 lactoglobulin, and bovine serum albumin. Arch Biochem Biophys. 1964;107(3):449-58. doi:
516 10.1016/0003-9861(64)90301-7.
517 39. Alaiz M, Giron J. Modification of Histidine Residues in Bovine Serum Albumin by Reaction
518 with (E)-2-Octenal. J Agric Food Chem. 1994;42(10):2094-8. doi: 10.1021/jf00046a005.
519 40. Brewer CF, Riehm JP. Evidence for possible nonspecific reactions between N-
520 ethylmaleimide and proteins. Anal Biochem. 1967;18(2):248-55. doi: 10.1016/0003-2697(67)90007-
521 3.
522 41. Boyatzis AE, Bringans SD, Piggott MJ, Duong MN, Lipscombe RJ, Arthur PG. Limiting the
523 Hydrolysis and Oxidation of Maleimide-Peptide Adducts Improves Detection of Protein Thiol
524 Oxidation. J Proteome Res. 2017;16(5):2004-15. doi: 10.1021/acs.jproteome.6b01060.
525 42. Fontaine SD, Reid R, Robinson L, Ashley GW, Santi DV. Long-Term Stabilization of
526 Maleimide-Thiol Conjugates. Bioconjugate Chem. 2015;26(1):145-52. doi: 10.1021/bc5005262.
527 43. Kuninori T, Nishiyama J. Some Properties of Diastereomers formed in the Reactions of N-
528 Ethylmaleimide with Biological Thiols. Agric Biol Chem. 1985;49(8):2453-4. doi:
529 10.1080/00021369.1985.10867100.
530 44. Garavelli JS. The RESID Database of Protein Modifications as a resource and annotation tool.
531 Proteomics. 2004;4(6):1527-33. doi: 10.1002/pmic.200300777.

45. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, et al. Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 2004;5(10):R80. doi: 10.1186/gb-2004-5-10-r80.

Supporting information

S1 Fig. Numbers of features identified in analyses of NEM-treated and untreated BSA. After filtering, the numbers of dependent peptides (DPs) were all similar (no two counts differed by more than 5%). 'Conjoined' DPs were those detected in analysis 1 of NEM-treated BSA and not detected in analysis 1 of untreated BSA. 'Constant' DPs were those detected in both analyses of NEM-treated BSA. 'Constantly conjoined' DPs were those detected in both analyses of NEM-treated BSA and not detected in either analysis of untreated BSA.

S2 Fig. Localisation plot and mass-shift frequency histogram for dependent peptides of porcine trypsin. MaxQuant identified three putative deamidations and a putative methylation. Two of the deamidations were localised to asparagine residues.

S3 Fig. Similarity of mass-shift frequency histograms. Three pairwise comparisons are shown: treated/treated, untreated/untreated and treated/untreated. The treated/untreated pair shown is the least similar of four possible combinations. ρ = Spearman correlation coefficient.

S4 Fig. Probability localisation plot for putative intact NESyl groups in NEM-treated BSA (Δm ± tolerance = +125.05 ± 0.01 Da). X-axis values refer to positions in the protein sequence.

S5 Fig. Probability localisation plot for additional intact NESyl groups in NEM-treated BSA (Δm ± tolerance = +68.03 ± 0.01 Da). X-axis values refer to positions in the protein sequence.

S6 Fig. Probability localisation plot for additional putative hydrolysed NESyl groups in NEM-treated BSA (Δ*m* ± tolerance = +86.04 ± 0.01 Da). X-axis values refer to positions in the protein sequence.

S7 Fig. Probability localisation plot for putative oxidations in NEM-treated BSA (Δ*m* ± tolerance = −25.03 ± 0.01 Da). X-axis values refer to positions in the protein sequence.

S8 Fig. Localisation plot and mass-shift frequency histogram for dependent peptides of cofilin-1 from dimethyl-fumarate-treated rat astrocytes [27]. The nominal Δ*m* of +105 Da is consistent with pyridylethylation of non-cysteine residues.

S9 Fig. Probability localisation plot for 1-carboxy-2-methylcarboxyethyl groups in cofilin-1 from dimethyl-fumarate-treated rat astrocytes [27] (Δ*m* ± tolerance = +24.97 ± 0.01 Da). X-axis values refer to positions in the protein sequence.

S10 Fig. Localisation plot and mass-shift frequency histogram for dependent peptides of apolipoprotein A-I from alk-3-HHD-treated human blood [28].

S11 Fig. Localisation plot and mass-shift frequency histogram for dependent peptides of haemoglobin δ-chain from alk-3-HHD-treated human blood [28]. All six dependent peptides also mapped to the sequence of haemoglobin β-chain (Fig 3).

S12 Fig. Localisation plot and mass-shift frequency histogram for dependent peptides of serotransferrin from alk-3-HHD-treated human blood [28].

S13 Fig. Localisation plot and mass-shift frequency histogram for dependent peptides of serum albumin from alk-3-HHD-treated human blood [28].

S14 Fig. Localisation plot and mass-shift frequency histogram for dependent peptides of transthyretin from alk-3-HHD-treated human blood [28].

S15 Fig. Probability localisation plot for putative sulfonated HDMP-type modifications in serum albumin from alk-3-HHD-treated human blood [28] (Δ*m* ± tolerance = +390.10 ± 0.01 Da). X-axis values refer to positions in the protein sequence.

S1 Script. Script I. An R script that filters dependent peptides and generates a dependent-peptide localisation plot and a mass-shift frequency histogram.

S2 Script. Script II. An R script that filters dependent peptides, enriches them on the basis of 'conjunction' and generates a dependent-peptide localisation plot and a mass-shift frequency histogram.

S3 Script. Script III. An R script that filters dependent peptides, enriches them on the basis of 'constant conjunction' and generates a dependent-peptide localisation plot and a mass-shift frequency histogram.

S4 Script. Script IV. An R script that isolates dependent peptides with a specified mass shift, enriches them on the basis of 'constant conjunction' and generates a probability localisation plot.

S5 Script. Script V. An R script that isolates dependent peptides with a specified mass shift and generates a probability localisation plot.

S1 Table. Dependent peptides and matching variable-modification search results. Localisation probabilities for hydrolysed NESyl groups are given in parentheses after the respective amino acid symbols. Potential sites of modification are underlined, with the most plausible sites in boldface (see S1 Text). Dependent peptides were allowed to have either of two mass shifts: +86.04 ± 0.01 Da or +143.06 ± 0.01 Da (see 'Results and discussion').

S2 Table. Dependent peptides with putative alk-3-HHD-derived modifications. Dependent peptides containing HDMP-/HTO-type modifications were identified using Script IV. All modifications were detected as sulfonyl derivatives, and all could be localised to lysine residues. DPs were matched to Salomón and coauthors' search results [28] ('+' = match) by sequence, site of modification and modification type ('Expected analyte'), or by protein site only ('Expected site'). Two site numbers are given: the number used for the matching (first number); and the equivalent number for the mature protein (second number, in parentheses).

S1 Text. Supplementary methods. Chemicals; Preparation of BSA adducts; Sample preparation for mass spectrometry; Nano liquid chromatography and mass spectrometry; Enumeration of tryptic peptides; Calculation of maximum peptide mass; Contaminant databases; Comparison of dependent-peptide and variable-modification search results; Figure preparation; References; Table A (Gradient elution timetable); Table B (Criteria used to filter dependent-peptide search results); Table C (Explanatory notes to accompany scripts).