On peptide selection for targeted protein quantitation

Cristina Chiva[1,2], Eduard Sabidó[1,2,*]

[1]Proteomics Unit, Centre de Regulació Genòmica (CRG), Barcelona Institute of Science and Technology (BIST), Dr. Aiguader 88, 08003 Barcelona, Spain

[2]Universitat Pompeu Fabra (UPF), Dr. Aiguader 88, 08003 Barcelona, Spain

Keywords: SRM, SWATH, targeted proteomics, quantitation

ABSTRACT

Targeted proteomics methods in their different flavors rely on the use of a few peptides as proxies for protein quantitation, which need to be specified either prior or after data acquisition. However, in contrast to discovery methods that use all identified peptides for a given protein to estimate its abundance, targeted proteomics methods are limited in the number of peptides that are used for protein quantitation. As only few peptides per protein are acquired or extracted in targeted experiments, the selection of peptides that are used for targeted protein quantitation becomes crucial. Several rules have been proposed to guide peptide selection for targeted proteomics studies, which have generally been based on the amino acidic composition of the peptide sequences. However, the compliance of these rules do not imply that not-conform peptides are not reproducibly generated nor they guarantee that the selected peptides correctly represent the behaviour of the protein abundance in different conditions.

Targeted proteomics has been the method of choice for reproducible protein quantitation in multiple experimental conditions[1] and large sample cohorts[2,3] as it overcomes the undersampling effects associated to discovery proteomics. Due to its analytical capabilities, targeted proteomics has gained significant popularity in the recent years, and multiple methods have been developed that combine different approaches of targeted acquisition and targeted data analysis.[4-7]

Targeted proteomics methods in their different flavors rely on the use of a few peptides as proxies for protein quantitation, which need to be specified either prior—SRM and PRM—or after—MSX and SWATH—data acquisition. However, in contrast to discovery methods that use all identified peptides for a given protein to estimate its abundance, targeted proteomics methods are limited in the number of peptides that are used for protein quantitation. As only few peptides per protein are acquired or extracted in targeted experiments, the selection of peptides that are used for targeted protein quantitation becomes crucial, and a wrong peptide selection (i.e. a peptide that does not represent the true fold-change of the protein of interest) might lead to a biased protein quantitation.

Some studies have experimentally defined peptides with good quantitative response for certain subsets of proteins,[8] while others have described heuristics to predict the detectability by mass spectrometry of unique peptides based on their physicochemical properties and previous experimental data.[9,10] Rules have been proposed to guide peptide selection for targeted proteomics studies, which have generally been based on the amino acidic composition of the peptide sequences. The avoidance of peptides prone to chemical reactions such as spontaneous deamidation, uncontrolled oxidation or water loss,[11] or bearing missed cleavages has traditionally been described as highly desirable for targeted protein quantitation. However, often it is difficult to find a peptide that satisfies all the described rules for a given protein, and even if a fully compliant peptide is selected, there is no guarantee that another version of the same peptide is present bearing a missed cleavage, a chemical or a post-translational modification, or a single point mutation, as true complexity of proteomes is generally difficult to predict. Therefore, the current procedure of peptide selection might be of limited use for the community,

as the common rules do not imply that not-conform peptides are not reproducibly generated nor they guarantee that the selected peptides correctly represent the behaviour of the protein abundance in different conditions.

To illustrate this situation, in this work we evaluated the impact of non-compliant peptides as surrogates for targeted protein quantitation and revisited the way peptides are selected in targeted proteomics studies. For this purpose we used an already published dataset from our group consisting of a mixture of thirty commercial proteins spiked in an *E. coli* background.[12] Briefly, five mixes were prepared in triplicate containing different ratios of the spiked-in proteins in the *E. coli* background, and the samples were subjected either to an in-solution or a filter-aided digestion with trypsin prior shotgun mass spectrometry acquisition.[12] Data were analyzed with an *E. coli* Uniprot database containing the spiked-in proteins using cysteine carbamidomethylation as fixed modification, and the modification of interest as variable modification in the Mascot search engine (v2.4). Areas for the identified peptides were extracted from MS1 chromatograms with Proteome Discoverer v1.4 (*Precursor Ions Area Detector* node). The use of a dataset with spiked-in proteins in a complex background allowed us to know beforehand the protein true fold-changes among the different mixes, and thus evaluate the accuracy of several types of peptides in the estimation of protein quantities, while maintaining the sample complexity.

Initially we assessed the frequency, extend and reproducibility of the most common peptide chemical modifications from the *E. coli* proteome background for both the in-solution and filter-aided original datasets (Table 1).[12] The results obtained from the re-analysis of this MS1 quantitative dataset suggest that the selection of tryptic peptides with potential sites of modifications for targeted proteomics studies have minimal incidence in relative protein quantitation as either these peptides are rarely modified or because when the modification occurs, it generates reproducible peptide areas that should not affect peptide relative quantitation (Table 1). Exceptions to these observations might be the spontaneously

thermodynamically favoured cyclization of N-terminal glutamine to pyro-glutamate, and in lower proportion asparagine deamidation and methionine oxidation.

However, to further assess the impact of selecting peptides with potential sites of modification in protein quantitation, we used the areas of the different identified peptides to estimate the known ratios of the spiked-in proteins within the original dataset (Table 2).[12] Noteworthy, only a low percentage of the quantified peptides within the analysed dataset (<10%) fulfilled all the requirements to be classified as fully-compliant peptides according to the commonly used guidelines i.e. without residues prone to cyclize, without residues prone to oxidation (Trp, His, Met), without residues prone to deamidation (Asn, Gln), and without residues that favour the presence of different *cis-trans* isomers (Pro) (Table 2). Nonetheless, peptides bearing potential sites of chemical modification did not exhibit a higher error in the protein fold-change estimation when compared to canonical proteotypic peptides. Similarly, no particular trend that affected protein relative quantitation was observed from peptides containing certain amino acidic residues, thus evidencing that the amino acid composition *per se* does not determine the quantitative properties of a peptide.

To ensure that the observations described were not dataset dependent, we took the MS1-based quantitation data as reported in another publicly available dataset and analysed the physicochemical properties of the peptides that better represent protein fold-changes between different human cell lines.[13] In most experimental designs, the true protein fold-change between biological samples is unknown. Therefore, we made the assumption that the protein fold-change calculated with all its unique peptides is the one that best represents the true protein fold-change of a given protein. Based on this assumption, we set the comparison HeLa *vs.* HepG2 cells, and GAMG *vs.* HEK293 cells, and estimated the protein abundance fold-changes using all available unique peptide areas with the MSstats R package.[14] Then we compared the protein log-fold change estimated with all peptides to the fold-change of each single peptide and ranked the peptides accordingly (Figure 1A, Table S-1). Thus, for each quantified protein we defined a *first peptide* that corresponded to the peptide with the lowest difference between the estimated

protein fold-change and the peptide fold-change; the *last peptide*, being the peptide with the highest difference between the calculated fold-changes; and the remaining peptides that were classified as *others*. Although *first* and *last* peptides differed substantially on the assessment of the protein fold-change (Figure 1B and 1C), the frequency of peptides bearing potential sites of chemical modification classified as *first* or *last peptide* were almost the same (Figure 1D and 1E).

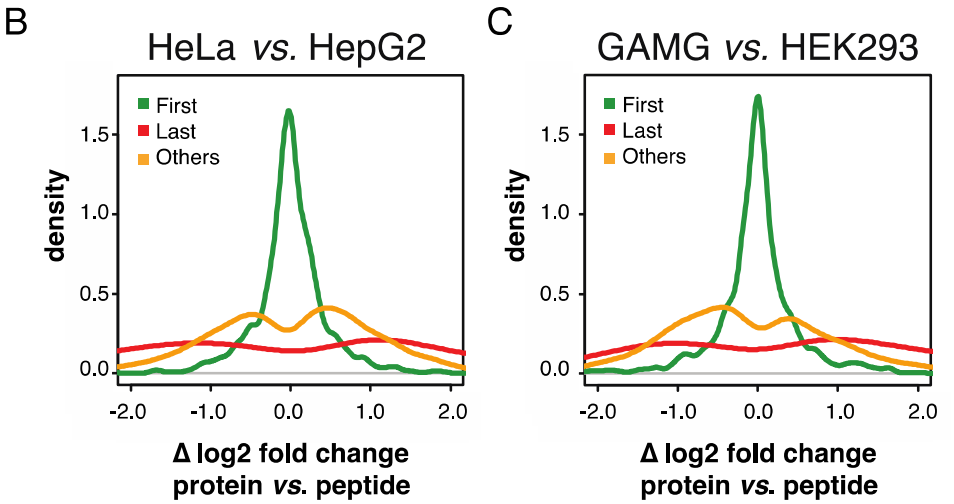
These observations show that aminoacid-based rules to select peptides for targeted studies such as SRM, PRM, MSX and SWATH do not guarantee good quantitative behaviour of the selected peptides, and that selection based on the avoidance of potential sites of modifications might be close to random peptide selection in terms of protein fold-change estimation. Indeed, several previous works had successfully used peptides with missed cleavages or containing several potential sites of chemical modification for targeted proteomics analyses, as these peptides were shown to be reproducibly generated from replicate samples.[15-17] Therefore, given these evidences it seems advisable to not rely on the aminoacidic composition of a peptide but rather on experimental data—i.e. pilot shotgun or data-independent experiments with the system under study—to assess the quantitative behaviour of the peptides to be targeted, and thus select peptides that correctly represent true protein fold-changes in targeted proteomics studies.

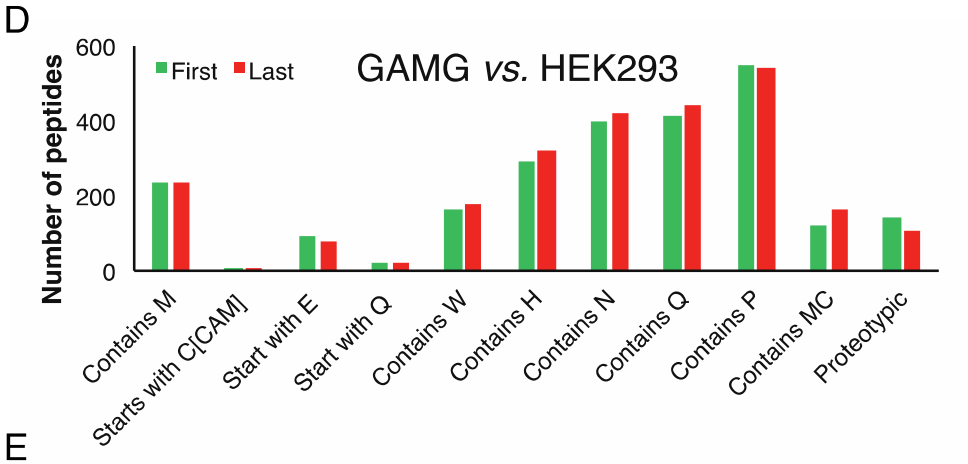
FIGURE 1

A

Comparison GAMG vs. HEK293 cells

Protein	Peptide	log2FC Peptide	log2FC Protein	|Δlog2FC|	Type
IPI00000105	DITPIQVVIPNTAIHIK	1.440	1.797	0.357	**First**
IPI00000105	VVAGDEWIFEGPGTYIPR	2.246	1.797	0.449	Other
IPI00000105	IAQDPFPIYPGEVIEK	0.936	1.797	0.861	Other
IPI00000105	AIIDFEDK	0.856	1.797	0.942	Other
IPI00000105	AIIDFEDKDGDK	3.164	1.797	1.367	**Last**





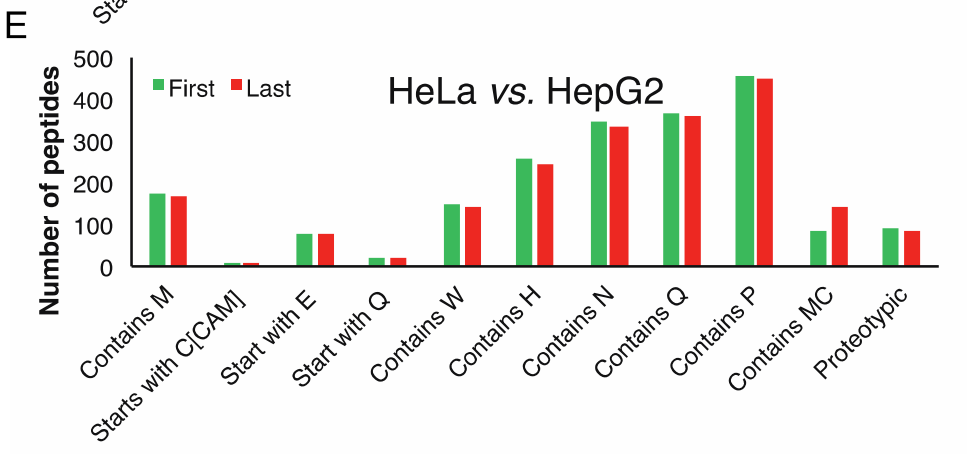


Figure 1 A: Example of peptide ranking for one protein (IPI00000105) as "*first*", "*other*" and "*last*" according to the difference between the observed peptide $\log_2$(fold-change) and the estimated protein $\log_2$(fold-change) using all unique peptides per protein. B-C: Density plots representing the difference between the protein $\log_2$(fold-change) calculated with all unique peptides identified (MS2 evidence) at least in 2 of the 3 replicates of each cell line and the protein $\log_2$(fold-change) estimated with each single peptide. The plot includes data for all proteins within the dataset[13] that comply with the aforementioned restrictions. Green: *first peptides*; red: *last peptides*; orange: *other peptides*. D-E: Distribution of the peptides between the *first* and *last* group according to their amino acid composition for all proteins within the dataset[13] that comply with the aforementioned restrictions.

TABLE 1

Frequency, extend and reproducibility of peptide chemical modifications within the identified peptides of the E. coli background proteome.

	% peptides with modification	Peptide Area Proportion *Area modified peptide / (Area modified peptide + Area unmodified peptide)*	
		Average *(n = 3 samples)*	Standard deviation *(n = 3 samples)*
	in-solution digestion		
M to M oxd	21%	0.21	0.05
H to Hoxd	5%	0.22	0.19
W to Woxd	5%	0.28	0.05
N-ter Q to pyroE	47%	0.53	0.05
N-ter E to pyroE	5%	0.10	0.02
N to D	27%	0.18	0.05
Q to E*	2%	0.08	0.03
	filter-aided digestion		
M to M oxd	43%	0.40	0.08
H to Hoxd	9%	0.27	0.22
W to Woxd	11%	0.23	0.05
N-ter Q to pyroE	40%	0.32	0.07
N-ter E to pyroE	7%	0.22	0.14
N to D	9%	0.18	0.05
Q to E*	1%	0.04	0.12

The frequency, extend and reproducibility of peptide chemical modifications were calculated from the identified peptides of the *E. coli* background proteome in the three replicates of *mix 3* from the original dataset.[12] The frequency of a modification was calculated as the number of peptides bearing a particular modification in respect to the total number of peptides containing the potentially modifiable residue (*% peptides*). In the case of peptides identified as both modified and unmodified, the extent of the modification was calculated as the relative area of the modified peptide form versus the sum of areas of the modified and unmodified peptide forms. Due to possible differences in the response factor between the modified and unmodified peptide versions, the calculated proportions are just an estimation of their relative abundances. Finally, the modification reproducibility among replicates was calculated based on the standard deviation of the extent of the modification among the three replicates of *mix 3* from the original dataset.[12] *Only peptides that contained Q but not N were considered.

TABLE 2

Average error on protein ratio estimation and average peptide $\log_2$(area) according to peptide amino acid composition.

	Average error on protein fold change estimation	Average peptide log2(*area*)	n
Starts with C	0.20	28.71	33
Starts with E	0.17	30.18	120
Starts with Q	0.21	28.40	21
Contains W	0.21	30.67	210
Contains H	0.19	30.41	399
Contains N	0.23	30.33	507
Contains Q	0.20	30.61	495
Contains P	0.21	30.70	654
Contains M	0.21	30.24	198
Proteotypic	0.18	30.90	88
All	0.20	30.49	1035

The areas of the different identified peptides were used to estimate the known ratios of the spiked-in proteins between mixes 3, 4 and 5 of the original dataset.[12] The error on the protein fold-change estimation was calculated for each spiked-in protein using the difference between the known protein fold-change and the fold-change estimated by each peptide, with the formula $|\log_2(\text{fold-change } \textit{peptide}) - \log_2(\text{fold-change } \textit{protein})| \,/\, |\log_2(\text{fold-change } \textit{protein})|$.

ASSOCIATED CONTENT SECTION

Table S-1

List of unique peptides per protein classified as "*first*", "*other*" and "*last*" according how close the peptide fold-change is to the estimated protein fold change. Protein and peptide fold changes have been calculated with the MSstats R package from the peptide areas reported in the original manuscript (Geiger et al. Mol Cell Proteomics 11, M111.014050; PMID:22278370).

AUTHOR INFORMATION

Corresponding author

Eduard Sabidó

Tel. 0034933160834

eduard.sabido@crg.cat

Proteomics Unit, Centre de Regulació Genòmica (CRG), Dr. Aiguader 88, 08003 Barcelona, Spain

ACKNOWNLEDGEMENTS

The CRG/UPF Proteomics Unit is part of the “Plataforma de Recursos Biomoleculares y Bioinformáticos (ProteoRed)” supported by grant PT13/0001 of Instituto de Salud Carlos III (ISCIII). We acknowledge support of the Spanish Ministry of Economy and Competitiveness, “Centro de Excelencia Severo Ochoa 2013-2017”, SEV-2012-0208, and from “Secretaria d’Universitats i Recerca del Departament d’Economia i Coneixement de la Generalitat de Catalunya” (Project 2014 SGR 678).

REFERENCES

1. Picotti, P., Bodenmiller, B., Mueller, L. N., Domon, B., & Aebersold, R. Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell* **2009**, 138, 795-806.

2. Surinova, S., Radová, L., Choi, M., Srovnal, J., Brenner, H., Vitek, O., Hajdúch, M., & Aebersold, R. Non-invasive prognostic protein biomarker signatures associated with colorectal cancer. *EMBO Mol Med* **2015**, 7(9):1153-65. doi: 10.15252/emmm.201404874.

3. Surinova, S., Choi, M., Tao, S., Schüffler, P. J., Chang, C.-Y., Clough, T., Vysloužil, K., Khoylou, M., Srovnal, J., Liu, Y., Matondo, M., Hüttenhain, R., Weisser, H., Buhmann, J. M., Hajdúch, M., Brenner, H., Vitek, O., & Aebersold, R. (2015) Prediction of colorectal cancer diagnosis based on circulating plasma proteins. *EMBO Mol Med* , **2015**, 7(9):1166-78. doi: 10.15252/emmm.201404873.

4. Egertson, J. D., Kuehn, A., Merrihew, G. E., Bateman, N. W., MacLean, B. X., Ting, Y. S., Canterbury, J. D., Marsh, D. M., Kellmann, M., Zabrouskov, V., Wu, C. C., & MacCoss, M. J. Multiplexed MS/MS for improved data-independent acquisition. *Nat Methods* **2013**, 10, 744-6.

5. Gallien, S., Duriez, E., Crone, C., Kellmann, M., Moehring, T., & Domon, B. Targeted proteomic quantification on quadrupole-orbitrap mass spectrometer. *Mol Cell Proteomics* **2012**, 11, 1709-23.

6. Gillet, L. C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., Bonner, R., & Aebersold, R. Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Mol Cell Proteomics* **2012**, 11, O111.016717.

7. Peterson, A. C., Russell, J. D., Bailey, D. J., Westphall, M. S., & Coon, J. J. Parallel reaction monitoring for high resolution and high mass accuracy quantitative, targeted proteomics. *Mol Cell Proteomics* **2012**, 11, 1475-88.

8. Worboys, J. D., Sinclair, J., Yuan, Y., & Jørgensen, C. Systematic evaluation of quantotypic peptides for targeted analysis of the human kinome. *Nat Methods* **2014**, 11, 1041-4.

9. Mallick, P., Schirle, M., Chen, S. S., Flory, M. R., Lee, H., Martin, D., Ranish, J., Raught, B., Schmitt, R., Werner, T., Kuster, B., & Aebersold, R. Computational prediction of proteotypic peptides for quantitative proteomics. *Nat Biotechnol* **2007**, 25, 125-31.

10. Kuster, B., Schirle, M., Mallick, P., & Aebersold, R. Scoring proteomes with proteotypic peptide probes. *Nat Rev Mol Cell Biol* **2005**, 6, 577-83.

11. Lange, V., Picotti, P., Domon, B., & Aebersold, R. Selected reaction monitoring for quantitative proteomics: a tutorial. *Mol Syst Biol* **2008**, 4, 222.

12. Chiva, C., Ortega, M., & Sabidó, E. Influence of the digestion technique, protease, and missed cleavage peptides in protein quantitation. *J Proteome Res* **2014**, 13, 3979-86.

13. Geiger, T., Wehner, A., Schaab, C., Cox, J., & Mann, M. Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. *Mol Cell Proteomics* **2012**, 11, M111.014050.

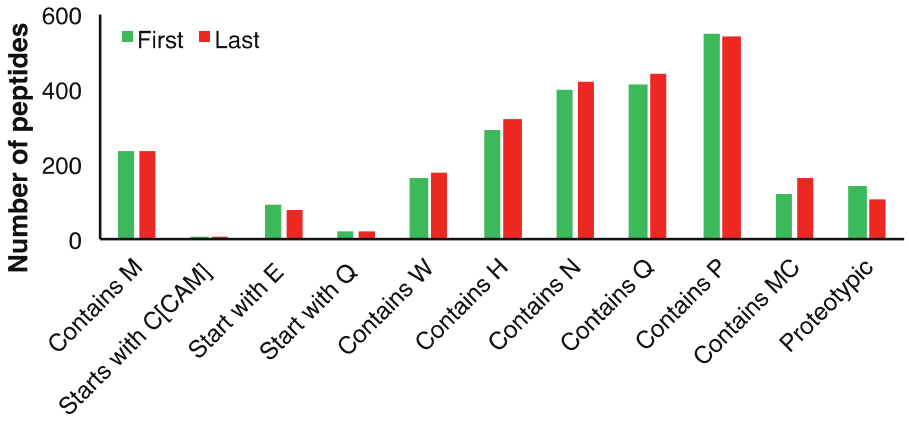
14. Choi M, Chang CY, Clough T, Broudy D, Killeen T, MacLean B, Vitek O. MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. Bioinformatics. 2014 Sep 1;30(17):2524-6. doi: 10.1093/bioinformatics/btu305.

15. Blankley, R. T., Fisher, C., Westwood, M., North, R., Baker, P. N., Walker, M. J., Williamson, A., Whetton, A. D., Lin, W., McCowan, L., Roberts, C. T., Cooper, G. J. S., Unwin, R. D., & Myers, J. E. A label-free selected reaction monitoring workflow identifies a subset of pregnancy specific glycoproteins as potential predictive markers of early-onset pre-eclampsia. *Mol Cell Proteomics* **2013**, 12, 3148-59.

16. Demeure, K., Fack, F., Duriez, E., Tiemann, K., Bernard, A., Golebiewska, A., Bougnaud, S., Bjerkvig, R., Domon, B., & Niclou, S. P. Targeted Proteomics to Assess the Response to Anti-Angiogenic Treatment in Human Glioblastoma. *Mol Cell Proteomics* **2016**, Feb;15(2):481-92. doi: 10.1074/mcp.M115.052423.

17. Steiner, C., Tille, J.-C., Lamerz, J., Kux van Geijtenbeek, S., McKee, T. A., Venturi, M., Rubbia-Brandt, L., Hochstrasser, D., Cutler, P., Lescuyer, P., & Ducret, A. Quantification of HER2 by Targeted Mass Spectrometry in Formalin-Fixed Paraffin-Embedded (FFPE) Breast Cancer Tissues. *Mol Cell Proteomics* **2015**, 14, 2786-99.

TOC GRAPHIC


Number of peptides
600
400
200
0
First
Last
Contains M
Starts with C[CAM]
Start with E
Start with Q
Contains W
Contains H
Contains N
Contains Q
Contains P
Contains MC
Proteotypic