Title

Sweet Google O' Mine - The Importance of Online Search Engines for MS-facilitated, Database-independent Identification of Peptide-encoded Book Prefaces

A EUPA YPIC challenge entry

Team

Axl Rose

Submission Date

8. September 2017

Authors

Alexander Hogrebe, Rosa R. Jersie-Christensen

Affiliation

Jesper V. Olsen's group at the Novo Nordisk Foundation Center for Protein Research, Proteomics Program, Faculty of Health and Medical Sciences, University of Copenhagen, Blegdamsvej 3b, DK-2200 Copenhagen

jesper.olsen@cpr.ku.dk

Correspondence

E-mail:

alexander.hogrebe@cpr.ku.dk

rosa.rjc@cpr.ku.dk

Abstract

In the recent year, we felt like we were not truly showing our full potential in our PhD projects, and so we were very happy and excited when YPIC announced the ultimate proteomics challenge. This gave us the opportunity of showing off and procrastinating at the same time :) The challenge was to identify the amino acid sequence of 19 synthetic peptides made up from an English text and then find the book that it came from. For this task we chose to run on an Orbitrap Fusion™ Lumos™ Tribrid™ Mass Spectrometer with two different sensitive MS^2 resolutions, each with both HCD and CID fragmentation consecutively. This strategy was chosen because we speculated that multiple MS^2 scans at high quality would be beneficial over lower resolution, speed and quantity in the relatively sparse sample. The resulting chromatogram did not reveal 19 sharp distinct peaks and it was not clear to us where to start a manual spectra interpretation. We instead used the de novo option in the MaxQuant software and the resulting output gave us two phrases with words that were specific enough to be searched in the magic Google search engine. Google gave us the name of a very famous physicist, namely Sir Joseph John Thomson, and a reference to his book “Rays of positive electricity” from 1913. We then converted the paragraph we believed to be the right one into a FASTA format and used it with MaxQuant to do a database search. This resulted in 16 perfectly FASTA search-identified peptide sequences, one with a missing PTM and one found as a truncated version. The remaining one was identified within the MaxQuant de novo sequencing results. We thus show in this study that our workflow combining de novo spectra analysis algorithms with an online search engine is ideally suited for all applications where users want to decipher peptide-encoded prefaces of 20th century science books.

Introduction

In this study, we faced the challenge of deciphering the amino acid sequence of 19 PTM-modified peptides. These peptides were synthesized specifically for this challenge and would yield a sentence taken from a book. The goal was to decipher both the peptide sequences and the book they came from.

We received the EuPA YPIC challenge peptides with an attached description stating that one peptide was still carrying a modification of m/z 89.97, probably as a protection group on a Ser, and that three PTMs were used to cipher the letters O (acetylated lysine), B (phosphorylated Ser) and U (methylated Arg). Each peptide should comprise one to five English words, which with an average English word length of five characters would indicate peptide lengths from five to 25 characters ("Wolfram|Alpha: Computational Knowledge Engine" 2017).

With the information given at hand, we developed an LC-MS/MS based workflow focused on high quality peptide spectra generation using an Orbitrap Fusion Lumos instrument. We reasoned that the physico-chemical properties of the peptides should not be substantially different from tryptic peptides in a standard bottom-up shotgun proteomics experiment. Thus a typical HPLC workflow strategy with standard C_{18} column and a separating gradient from low to high organic solvent should be applicable.

For MS analysis in a routine proteomics experiment, a fragmentation strategy optimized to the sample and instrument at hand with lower MS^2 scan resolutions and shorter injection times would be preferable to facilitate a higher scan speed (Kelstrup et al. 2012). However, we speculated that the challenge sample, even though lacking post-synthesis peptide purification, should contain less peptide variants than e.g. a standard tryptic HeLa proteome. Thus, a high spectrum quality for confident amino acid sequence determination would be desirable over a high peptide throughput with an as fast as possible MS scan speed.

We therefore decided to measure the sample twice with two different high resolution MS^2 settings including both CID and HCD fragmentation on the same precursor. Using two different fragmentation approaches should further increase the quality for peptides which have a preference for better fragmentation with either CID or HCD.

Taken all these consideration into account we ended with the workflow described in Fig. 1a.

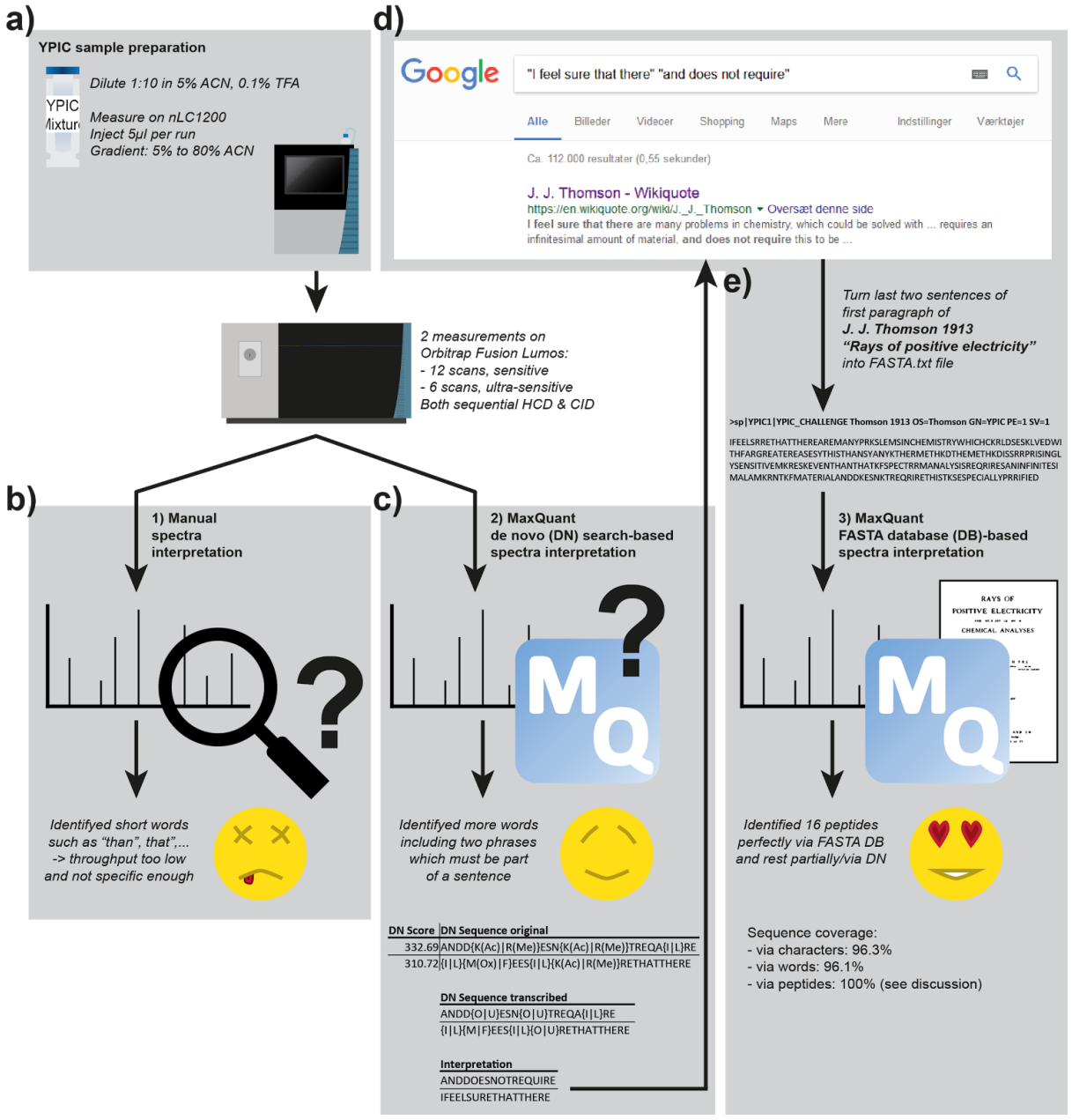


Figure 1: MS-based Workflow to Identify Peptide-encoded Book Prefaces. The scheme depicts the workflow and results for the identification of the correct text the 19 challenge peptides originated from. a) EuPA YPIC challenge peptides were diluted 1:10 in 5% ACN, 0.1% TFA and 5µl per injection measured on an nLC1200 + Orbitrap Fusion Lumos system. b) Manual spectra interpretation did not yield enough text-specific peptide sequences. c) MaxQuant de novo spectrum analysis helped in identifying initial peptide phrases. d) Two peptide phrases entered into the search engine Google with quotation marks were enough to identify the correct book from Sir J. J. Thomson. e) Transcribing two sentences from the preface of the book into a FASTA file and searching it with MaxQuant yielded a 100% peptide sequence coverage.

Results and Discussion

When first inspecting the resulting chromatogram (see Supp. Fig. 1), we saw highly abundant peaks and began manual interpretation of the spectra in the two raw files we recorded (see Fig. 1b). We were able to identify individual sequences forming words such as “than” and “that” (data not shown). While these findings indicated that we had indeed generated usable data, the throughput and specificity of this approach were both far too low for a meaningful analysis and identification of the book we were looking for. Instead, we needed an automated, high-throughput data analysis strategy.

Since the peptides were not tryptic and the phrases were not derived from any organism’s genome, we could not start with a classic FASTA database search (DB) approach, which is why we tried out the de novo (DN) search feature of MaxQuant v1.5.8.4i (see Fig. 1c) (Cox and Mann 2008). This feature tries to de novo annotate amino acid sequences for all spectra from a raw file without resorting to a predefined FASTA database. It can also include PTMs in this process and even assigns a DN quality score to each annotation. After using this feature with our two raw files, we did indeed acquire lots of annotations, which we sorted by DN score in a highest to lowest order. After transcribing PTM-modified residues to their respective letters, the highest scoring peptide with a DN score of 332.69 yielded the sequence “ANDD{O|U}ESN{O|U}TREQA{I|L}RE”. Although wrongly spelled with an A instead of U, this had to indicate the phrase “and does not require” being part of the text which we were looking for. Since the synthetic peptides were not purified, there should be a high chance of finding wrongly or partly synthesized versions of the target sequences. Thus it could very well be possible for misspelled versions to score higher than the actual target sequences. The five next highest DN scoring peptides were variations of this sequence, but the seventh transcribed into a perfectly spelled “I feel sure that there”.

After identifying these two phrases, we next wanted to see if we could already now pinpoint which text they originated from. Since manually reading through literally decades of scientific research publications would be neither high-throughput nor specific, we instead relied on the power of a second automated database search tool - Google (see Fig. 1d). Since the identified sequences seemed to be part of sentences, we entered them with quotation marks into the search query. This would ensure that Google does not search for the words individually, but preserves their appearance as phrases. Indeed, the first hit of the more than 112,000 search pointed to the wikiquote entry of Sir Joseph John Thomson. To avoid a false-positive hit, we also tested the alternative search engines Yahoo and Bing. While the

former only identified two search hits, with the first correctly being the J. J. Thomson text, Bing only identified it as its seventh hit. Other articles with unhelpful, but at least for a PhD student relatable topics such as “Why do we sleep, anyway?” scored higher instead.

Thomson was a physicist and Nobel prize laureate from England, and is credited as one of the discoverers of the electron. More importantly with respect to this challenge, he also was the first person to measure the mass-to-charge ratio of positively charged molecules and the slowly dying unit Th (m/z) is actually named after him. These findings, which are often regarded as the birth of the field of mass spectrometry, were described in his 1913 publication “Rays of positive electricity and their application to chemical analyses” (Thomson et al. 2017). The preface of this book, written by Thomson himself, contains the two phrases we identified in our peptide mixture. Since the book in question played such an important role in the development of modern mass spectrometers, we were positive that we had found the correct text sequence. To confirm this, we next transformed the preface into a FASTA file and searched the two raw files again with MaxQuant. We transcribed the letters B, O and U into their PTM-modified amino acid counterparts, and searched the raw files applying unspecific digestion with a minimum amino acid sequence length of six. We did this two times, using the whole preface the first time (data not shown), and the second time only two sentences that due to their high number of peptide sequence hits had to be the correct text (see Fig. 1e). With this approach we were able to identify 16 out of 19 peptides without any spelling mistakes (See Fig. 2). Fun fact: the peptide corresponding to the sequence “specially”, as written in the preface, was also found as “especially”.

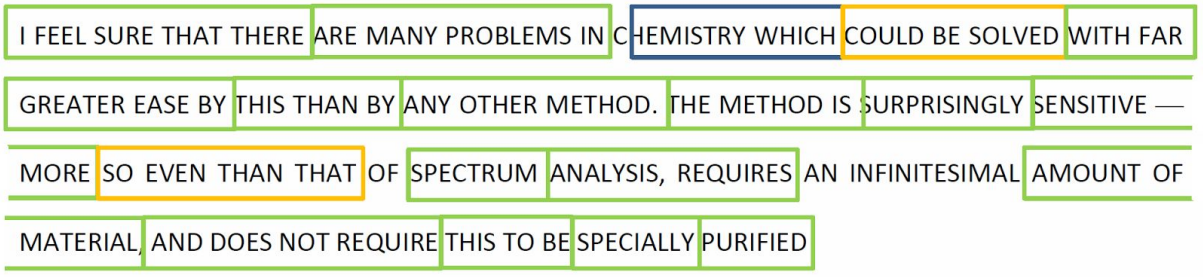


Figure 2: Peptide Sequence Coverage. Peptide sequences marked with green boxes were identified without spelling mistakes via the MaxQuant FASTA DB search. Peptides within yellow boxes showed slight spelling mistakes, and the one in blue was only identified by MaxQuant DN search. The phrase “an infinitesimal” was not found, but as indicated in the challenge by “two words still missing”, which would “explain its own absence if combined with ‘small amount’”, this could be the one/two words missing.

The remaining three peptides were still identified, although with spelling mistakes or truncations (see Table 1): 1) The peptide “could be solved” was misspelled with “s” instead of “b”. In the challenge it was mentioned that one peptide carries a possible protection group on a Ser. We argue that the phosphorylated Ser, which would have translated into the letter

B, got dephosphorylated, thus leading to the S in our sequence. 2) The peptide “so even than that of” was only identified partially as “so even tha”. In addition, MaxQuant DN search also found the variant “so even that”. The sequence “of” was only found in the correctly identified peptide “amount of material”. Since we do not find evidence of peptides with a “tha_tha_” motif, it is intriguing for us to speculate that the peptide was synthesized with only “that”, as indicated by the DN results. Since the sequential “of” was missing as well, this peptide could have been particularly difficult to synthesize. 3) The final missing peptide “chemistry which” was only identified without its N-terminal C via MaxQuant DN search. We speculate that the Cys might not have been carbamidomethylated, but oxidized instead, which is why our approach could not find it.

The phrase “an infinitesimal” could not be found. However, the challenge stated that two words were missing in the sample, with one of them being interchangeable with “small amount”. This fits our missing two words perfectly. We thus identified variants of all 19 peptides.

Table 1: Peptide Sequence Coverage. The table lists peptide sequences identified by MaxQuant FASTA DB (highlighted in white) or DN (highlighted in grey). The complete list of peptide sequences found by these two methods is provided in the supplementary. Peptide four was only identified with S instead of S(ph). Peptides 12.1 and 12.2 in FASTA DB and DN search, respectively, were only identified as partial constructs of the target sequences.

Peptide number	Modified sequence	Translated sequence	Score	Length	m/z	Mass	Charge	Fragmen-tation
1	_IFEELSR(me)RETHATTHERE_	_IFEELSURETHATTHERE_	332.96	18	752.377	2254.11	3	CID
2	_AREM(ox)ANYPRK(ac)S(ph)LEM(ox)SIN_	_AREMANYPROBLEMSIN_	148.75	17	721.990	2162.95	3	CID
3	_HE{M|F}{I|L}STRYWH{I|L}CH_ *	_HEMISTRYWHICH_	253.27	13	643.604	1927.79	3	HCD
4	_CK(ac)R(me)LDSESK(ac)LVED_	_COULDSESOLVED_	167.71	13	838.911	1675.81	2	CID
5	_WITHFAR_	_WITHFAR_	158.38	7	465.751	929.487	2	CID
6	_GREATEREASES(ph)Y_	_GREATEREASEBY_	230.50	13	782.818	NaN	0	CID
7	_THISTHANS(ph)Y_	_THISTHANBY_	250.13	10	605.748	1209.48	2	HCD
8	_ANYK(ac)THERM(ox)ETHK(ac)D_	_ANYOTHERMETHOD_	255.25	14	930.421	1858.83	2	HCD
9	_THEMETHK(ac)DIS_	_THEMETHODIS_	181.53	11	685.306	NaN	0	HCD
10	_SR(me)RPRISINGLY_	_SURPRISINGLY_	85.73	12	723.420	1444.83	2	CID
11	_SENSITIVEM(ox)K(ac)RE_	_SENSITIVEMORE_	319.81	13	797.393	1592.77	2	CID
12.1	_SK(ac)EVENTHA_	_SOEVENTHA_	191.17	9	528.751	1055.49	2	CID
12.2	_S{K(Ac)|R(Me)}EVENTHAT_ *	_SOEVENTHAT	169.81	10	579.276	1156.54	2	CID
13	_SPECTRR(me)M_	_SPECTRUM_	129.68	8	525.744	1049.47	2	HCD
14	_ANALYSISREQR(me)IRES_	_ANALYSISREQUIRES_	228.73	16	636.341	NaN	0	HCD
15	_AM(ox)K(ac)R(me)NTK(ac)FM(ox)ATERIAL_	_AMOUNTOFMATERIAL_	231.71	16	671.020	2010.04	3	CID
16	_ANDDK(ac)ESNK(ac)TREQR(me)IRE_	_ANDDOESNOTREQUIRE_	204.43	17	547.524	2186.07	4	HCD
17	_THISTK(ac)S(ph)E_	_THISTOBE_	213.03	8	512.721	NaN	0	HCD
18.1	_SPECIALLY_	_SPECIALLY_	158.53	9	533.268	1064.52	2	CID
18.2	_ESPECIALLY_	_ESPECIALLY_	174.21	10	597.789	1193.56	2	CID
19	_PR(me)RIFIED_	_PURIFIED_	89.05	8	530.301	1058.59	2	CID

Eleven out of 19 high-scoring peptide identifications resulted from fragmentation with CID, as compared to eight for HCD. This illustrated that for non-tryptic peptides, both fragmentation techniques can have complementary power in producing high-quality MS^2 spectra. Interestingly, twelve out of the 21 spectra considered in Table 1 were recorded with the more sensitive top6 method, indicating that the higher MS^2 resolution and longer fill times increased the overall spectrum quality as expected. To confirm if the DB search algorithm

identified all peptides correctly, we manually checked the annotated spectra (see Supp. Fig. 2). As an example, we present the annotated spectrum of our initially identified peptide “I feel sure that there”, which is the first one in the text, with a nearly perfect b- and y-ion series covering 100% of the amino acid sequence (see Fig. 3).

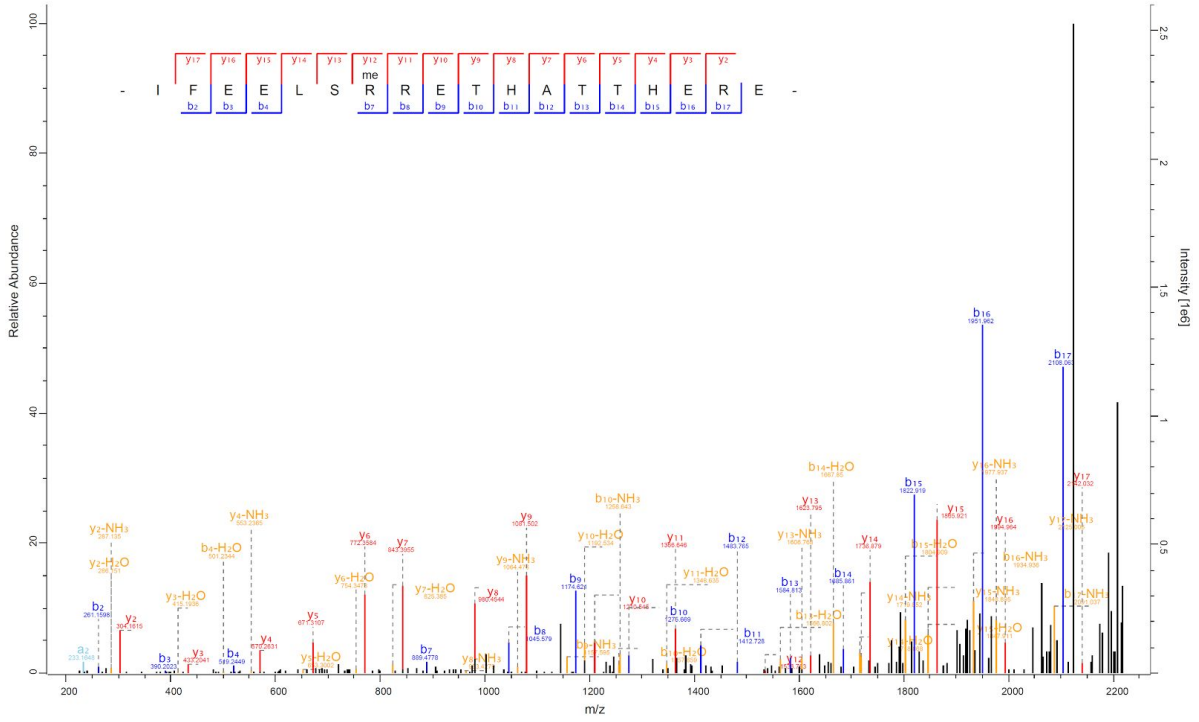


Figure 3: Example Peptide Sequence “I feel sure that there”. The spectrum shows annotated a-, b- and y-ions for the peptide sequence IFEELSR(me)RETHATTHERE, with R(me) translating to U. The spectrum was recorded following CID fragmentation.

Conclusion

We conclude that our combined MaxQuant DN and DB + Google search strategy proved extremely successful for the initial annotation of peptide sequences, identification of the correct book and text paragraph, and the final confirmation of all 19 peptide sequences in the sample. The only limitation we encountered, was that it was not possible for us to identify the peptide reported to carry a protection group on Ser. This is due to the fact that MaxQuant requires precise information on chemical composition for modifications it has to search, as not even an arbitrary protection group on Ser of the same mass yielded any true hits. Importantly, our approach highlights that due to the sheer number of impure peptide variants (even including ambiguities such as specially/especially), putting together the whole preface

text without errors just from the peptide sequences would be a very tedious process. The importance of Google for solving comparable tasks can thus not be stressed enough.

This challenge has been a true pleasure to work on and has resulted in many good and quite different talks over the coffee machine. We thank the organizers for preparing the challenge including peptide synthesis and required logistics to ship them out, and are looking forward to use our MS-skills again on the (hopefully coming) next EuPA YPIC challenge.

Acknowledgments

We would like to thank Jesper V. Olsen for encouraging us to participate in this challenge and permitting us to use chemicals and MS instrument time to solve it. We would further like to thank our colleagues for being nice and there, and Guns n' Roses for bearing us to plagiarize their lead singer's name.

Last but not least, we would like to thank our HGZ s200 disco coffee machine for being with us in times of dire need, Tuborg and Carlsberg for enabling creative brainstorming events, cake club for being delicious, our meeting room screens for being able to play Game of Thrones, and of course the Dalai Lama for always believing in us. Tusind tak!

References

Cox, J., and M. Mann. 2008. "MaxQuant Enables High Peptide Identification Rates, Individualized P.p.b.-Range Mass Accuracies and Proteome-Wide Protein Quantification." *Nature Biotechnology* 26 (12): 1367–72.

Kelstrup, C. D., C. Young, R. Lavallee, M. L. Nielsen, and J. V. Olsen. 2012. "Optimized Fast and Sensitive Acquisition Methods for Shotgun Proteomics on a Quadrupole Orbitrap Mass Spectrometer." *Journal of Proteome Research* 11 (6): 3487–97.

Thomson, J. J. (Joseph John), Sir, and 1856-. 2017. "Rays of Positive Electricity and Their Application to Chemical Analyses : Thomson, J. J. (Joseph John), Sir, 1856-1940 : Free Download & Streaming : Internet Archive." *Internet Archive*. Accessed September 8. https://archive.org/details/rayspositiveele01thomgoog.

Vizcaino, J. A., E. W. Deutsch, R. Wang, A. Csordas, F. Reisinger, D. Rios, J. A. Dianes, et al. 2014. "ProteomeXchange Provides Globally Coordinated Proteomics Data Submission and Dissemination." *Nature Biotechnology* 32 (3): 223–26.

Vizcaíno, Juan Antonio, Attila Csordas, Noemi del-Toro, José A. Dianes, Johannes Griss, Ilias Lavidas, Gerhard Mayer, et al. 2016. "2016 Update of the PRIDE Database and Its Related Tools." *Nucleic Acids Research* 44 (D1): D447–56.

"Wolfram|Alpha: Computational Knowledge Engine." 2017. Accessed September 4. http://www.wolframalpha.com/input/?i=average+english+word+length.

Materials and Methods

Nanoflow LC-MS/MS

EuPA YPIC challenge peptides supplied at a solution at ~0.5nmol/peptide in ~40µl 30% acetonitrile (ACN) were diluted 1:10 in 5% ACN, 0.01% trifluoroacetic acid (TFA). 5µl per injection were analyzed on an Easy-nLC 1200 coupled to an Orbitrap Fusion Lumos instrument (Thermo Fisher Scientific) equipped with a nanoelectrospray source. Peptides were separated on a 15cm analytical column (75µm inner diameter) in-house packed with 1.9µm C18 beads (Dr. Maisch, r119.b9). The column temperature was maintained at 40°C using an integrated column oven (PRSO-V1, Sonation). We used a 77min gradient at a flow rate of 250nl/min ramping from 5% buffer B (80% ACN and 0.1% formic acid) to 25% B in 50min, to 40% B in 10min, to 80% B in 2min, kept 5min, to 5% B in 5min and kept 5min.
MS analysis was performed with two different instrument methods. Both measured MS^1 scans at 120,000 resolution from 300-1500 m/z at 30% RF Lens with an AGC target of 4e5 and max. 50ms injection time. Dynamic exclusion was set to 30s. Both methods triggered two MS^2 scans on each selected precursor, first with HCF at 30% NCE, and then with CID at 35% NCE and activated MSA with neutral loss mass 97.9673. The two different methods termed sensitive and ultra-sensitive used top12/top6 analysis at 60,000/120,000 resolution with 0.8/0.4 m/z isolation window, an AGC target 1e5/2e5 and max. 118/256ms injection time, respectively. Coffee was infused at three cups per day on median into both authors.

Raw data processing

Raw LC-MS/MS files were processed with MaxQuant v1.5.8.4i (Cox and Mann 2008) with activated de-novo sequencing and a custom FASTA database. The database was generated from two sentences from the preface of Sir J. J. Thomson's 1913 book "Rays of positive electricity and their application to chemical analyses", with letters B, O and U transformed to S, K and R, respectively. Variable modifications were set to Oxidation (M), Acetyl (K), Methyl (R), Phospho (S), and a custom defined protection group (S) at total m/z 89.97. Carbamidomethyl (C) was set as fixed modification. Both raw files were set to the same experiment name. Digestion was set to unspecific, with a minimum amino acid length of six. The false discovery rate (FDR) was set to 1% on PSM, PTM site and protein level.

Bioinformatics analysis

Bioinformatics analysis was performed using Microsoft Excel and Word v14.0.6112.5000 (32bit). Figures were created using Adobe Illustrator CS6 v16.0.3 (64bit).

Data availability

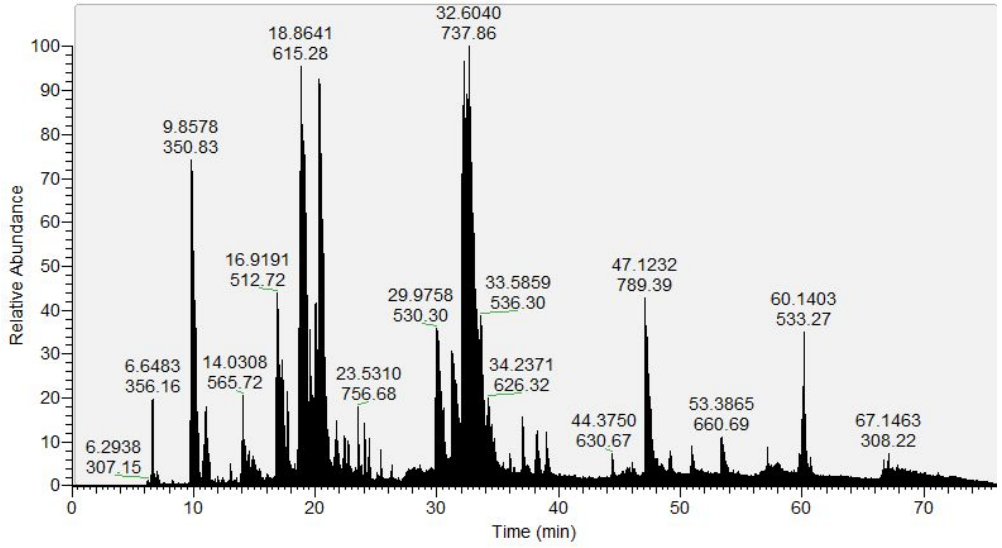
The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository (Vizcaíno et al. 2016; Vizcaino et al. 2014) with the dataset identifier PXD007693.

Reviewer account details:

Username: reviewer80489@ebi.ac.uk

Password: E8yDyXLr

Supplementary



Supplementary Fig. 1: Chromatogram of YPIC peptide top12 measurement. The peptides were measured with the above described sensitive top12 method. Even though more than 19 peaks can be identified, the chromatogram visually appears less complex than a standard tryptic HeLa digest.

Supplementary Fig. 2: Annotated MS^2 spectra of the identified EuPA YPIC challenge peptides. The document contains annotated spectra of all FASTA DB identified peptides in Table 1:
https://drive.google.com/open?id=0BwV0OAZgBdBzVjQ1ZkY4dUhqcUU