Proteomics
www.proteomics-journal.com

hmSEEKER: Identification of hmSILAC Doublets in MaxQuant Output Data

Enrico Massignani, Alessandro Cuomo, Daniele Musiani, SriGanesh Jammula, Giulio Pavesi, and Tiziana Bonaldi**

Heavy methyl Stable Isotope Labeling with Amino acids in Cell culture (hmSILAC) is a metabolic labeling strategy employed in proteomics to increase the confidence of global identification of methylated peptides by MS. However, to this day, the automatic and robust identification of heavy and light peak doublets from MS-raw data of hmSILAC experiments is a challenging task, for which the choice of computational methods is very limited. Here, hmSEEKER, a software designed to work downstream of a MaxQuant analysis for in-depth search of MS peak pairs that correspond to light and heavy methyl-peptide within MaxQuant-generated tables is described with good sensitivity and specificity. The software is written in Perl, and its code and user manual are freely available at Bitbucket (https://bit.ly/2scCT9u).

Protein methylation is a posttranslational modification (PTM) consisting in the addition of one or more methyl-groups to arginine and lysine residues of a protein. Lysine can be mono-, di-, or tri-methylated, while arginine can be mono- or di-methylated. Arginine di-methylation can, in turn, be either symmetrical or asymmetrical. While the functional implication of histone lysine and arginine methylation in the regulation of gene expression is well established, recent evidence revealed that this PTM is widespread at the proteome level, a notion that expands its regulatory potential well beyond chromatin structure and transcriptional regulation.[1] These discoveries are the results of several technical improvements in the field of MS-based proteomics, an analytical strategy widely adopted to investigate PTMs at the global proteome level.

Nevertheless, identifying in vivo protein methylations by MS is still challenging and error-prone, mainly because various amino acid substitutions (e.g., glycine to alanine) and chemical methyl-esterification of aspartic and glutamic acid (occurring when protein samples are exposed to methanol or ethanol) are isobaric to methylation. Thus, the use of orthogonal validation methods to reduce False Discovery Rates (FDRs) has been recommended.[2]

A reliable strategy for the identification of in vivo methylated peptides is called heavy methyl Stable Isotope Labeling with Amino acids in Cell culture (hmSILAC),[3] a metabolic labeling strategy achieved by growing cells in the presence of natural methionine (Light, L) or $^{13}CD_3$-methionine (Heavy, H). Cells convert methionine into *S*-adenosyl-methionine, the sole donor of methyl-groups used by protein methyltransferases to modify their substrates. Hence, methyl-groups added to the protein backbone through an enzymatic reaction will exist in either the light or the heavy form. Given a 1:1 mix of light- and heavy- labeled cells, followed by protein extraction, digestion, and peptide separation, the analysis by LC–MS/MS permits the identification of peptides bearing in vivo enzymatically driven methylations that in the MS readout will form pairs of H-L peaks, differing for a specific mass value. Conversely, false positive methylations do not produce H-L doublets, since they exist only in the light form. The hmSILAC strategy has been employed in numerous large-scale MS-based analyses of protein-methylation,[4,5] leading to a significant reduction of FDRs when compared to approaches not relying on hmSILAC or similar strategies (e.g., iMethyl-SILAC).[5]

At the moment, there is however a lack of bioinformatic tools specifically designed for the identification of methyl-peptide pairs from hmSILAC data. Among the few methods available, PyQuant is a MS data quantification package that supports several labeling strategies, both chemical and metabolic, such as SILAC, iTRAQ, and Neucode.[6] The package is highly flexible and is capable of processing hmSILAC data, although this specific labeling needs to be explicitly configured by the user.[7] Another suitable software package is MethylQuant, which is specifically designed to reconstruct and quantify hmSILAC doublets directly from raw MS1 spectra.[8]

E. Massignani, Dr. A. Cuomo, Dr. D. Musiani, Dr. S. Jammula[+]
Dr. T. Bonaldi
Department of Experimental Oncology
IEO
European Institute of Oncology IRCCS
Milan, Italy
E-mail: tiziana.bonaldi@ieo.it
Prof. G. Pavesi
Department of Biosciences
Università degli Studi di Milano
Milano, 20133, Italy
E-mail: giulio.pavesi@unimi.it

[+]Present address: Cancer Research UK Cambridge Institute, Cambridge, CB2 0RE, UK.

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/pmic.201800300

DOI: 10.1002/pmic.201800300

We present here hmSEEKER, a novel tool that allows a fast and reliable identification of hmSILAC H-L doublets by processing proteomic data outputs generated by MaxQuant.[9] The latter choice was motivated by the fact that MaxQuant is a widely used software suite designed for the accurate identification and quantitation of peptides, since it includes several functions like MS1 peak detection, integration of different mass measurements to increase mass accuracy, peptide identification based on their MS/MS spectra through the Andromeda search engine,[10] and quantitation of peptides and proteins. hmSEEKER thus permits a straightforward integration of hmSILAC with any other type of high-resolution quantitative proteomics data processed via MaxQuant (e.g., standard SILAC or Label Free), for orthogonal, high-confidence validation of methyl-sites.

The hmSEEKER software package is written in Perl (version 5.24.0) and is compatible with MaxQuant 1.5.8.3 and subsequent releases. The input consists of the "msms.txt" and "allPeptides.txt" files generated by MaxQuant, plus a protein sequence collection in FASTA format, to permit the mapping of the peptides to the respective sequence. The main output produced by hmSEEKER consists of the list of all the peptide doublets identified. It also returns a list of methyl-peptides for which a light/heavy counterpart was not retrieved (i.e., putative false positive methylations), for manual inspection by users. The software workflow is composed of three main steps (**Figure 1**A), detailed as follows:

Input Reading and Filtering

Information associated with each peptide and MS1 peak is extracted from the "msms" and "allPeptides" files. All methyl-peptides that 1) derive from contaminants, 2) have charge lower than 2, or 3) carry simultaneously heavy and light modifications are discarded. The tool also discards by default the methyl-peptides with 1) an Andromeda peptide score <25, 2) a delta score <12, or 3) a Localization Probability (LcPrb) of the methyl-site <0.75. These input filtering parameters can be however modified by users. Peptides are then assigned to the two H and L classes, and the expected delta mass is calculated as the mass shift between one light and one heavy methyl-group, multiplied by the number of methionines and methyl-groups carried by the peptide. This step leads to the generation of two tables: a "Peptide Index", containing information about all peptides that were fragmented by MS/MS and identified, and a "Peak Index", including information on all MS1 peaks that were detected, but not necessarily identified by MS/MS.

Doublets Search

To identify hmSILAC doublets, hmSEEKER first associates each peptide in the Peptide Index with the corresponding MS1 peak in the Peak Index, then searches for peak pairs comparing peak charges and m/z ratios, and by computing for each pair the retention time difference (ΔRT) and Mass Error (ME), defined as:

$$\text{Mass Error} = 10^6 \times \left(\frac{m/z_L + m/z_{\text{shift}}}{m/z_{\text{H}}} - 1 \right) \qquad (1)$$

where m/z_{H} and m/z_{L} are the m/z ratios of the H and L objects, respectively, and m/z_{shift} is the expected m/z difference between the H and L forms of the peptide. The underlying assumption is that the two isotopic counterparts of a peptide co-elute, undergo the same ionization process, and differ only for the presence of H or L methyl-groups. Thus, hmSEEKER can identify peptide pairs at the MS1 level, even if one of the two counterparts is not MS/MS fragmented, without requiring to directly parse the raw MS data. To be considered a true doublet, a pair of peaks must satisfy all the following conditions:

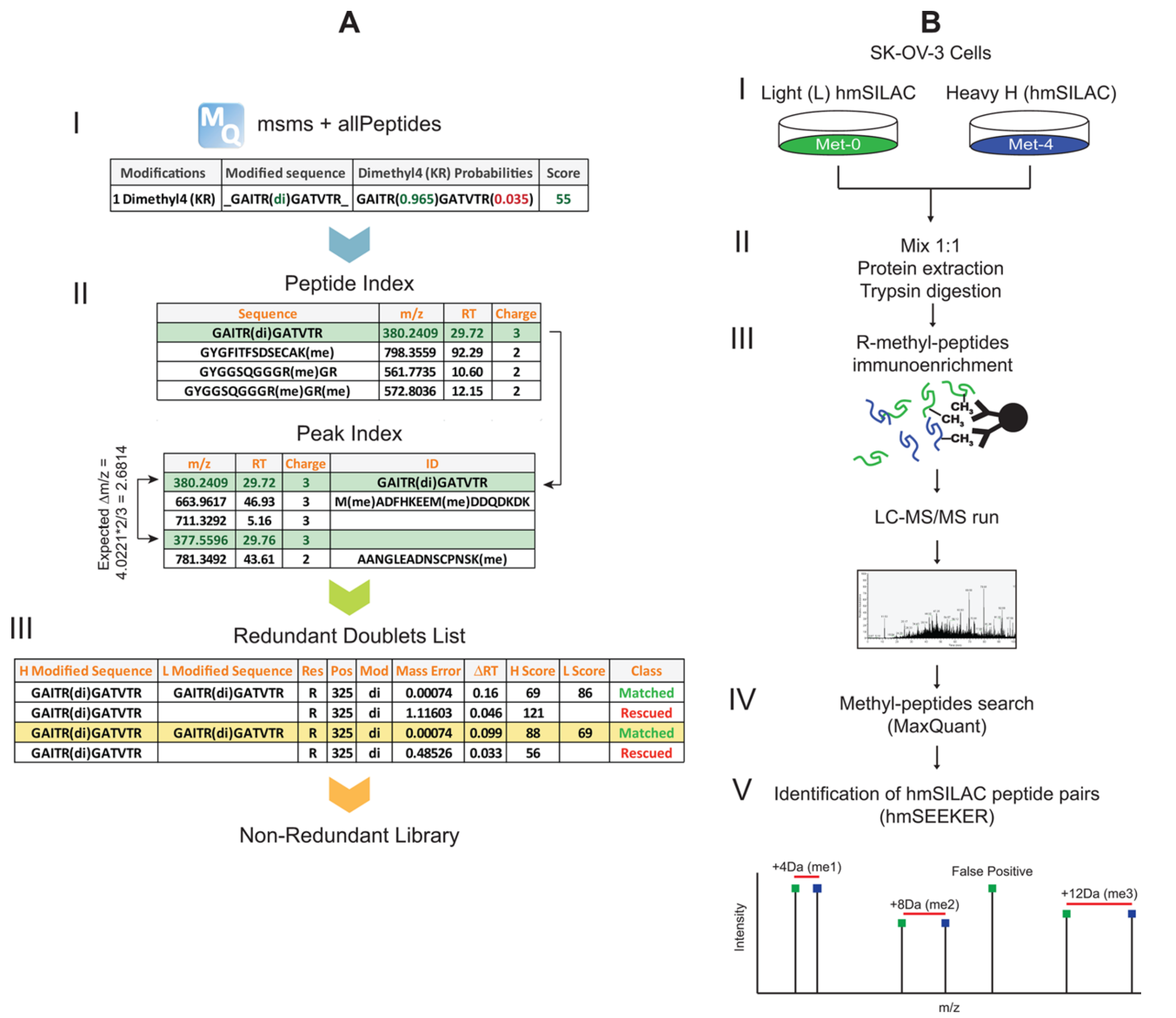
1) $m/z_{\text{H}} > m/z_{\text{L}}$,
2) $\text{Charge}_{\text{H}} = \text{Charge}_{\text{L}}$,
3) $|\,\text{RT}_{\text{H}} - \text{RT}_{\text{L}}\,| < \Delta\text{RT}$ threshold (default = 0.5 min),
4) $|\text{ME}| <$ ME threshold (default = 2 ppm),
5) $|\,\text{Log}_2(\text{H/L ratio})\,| < \text{Log}_2\text{Ratio}$ threshold (default = 1).

After this step, hmSEEKER produces a list containing all the identified doublets.

Output Refinement

For the doublets search step, hmSEEKER uses the values corresponding to charge, ME, RT, and H/L ratio of the peaks. However, the input also contains the peptide sequences derived from the Andromeda peptide search. This additional information is used by hmSEEKER after the search to define four classes of doublets:

- ”Matched”: This class includes doublets in which the two peptides are identical, allowing for site-specific calling of the modification.
- "Mismatched": This class consists of H-L pairs in which the peptides share the same amino acid sequence, but the modification is assigned to different sites. These doublets are reasonably true, yet MaxQuant could not pinpoint the modification on one of the two counterparts. The score and LcPrb values reported in the output can suggest which of the positional isoforms of the methyl-peptide is the most likely to be true. When both H and L peptides have similar scores, users may wish to directly check the MS/MS spectra via MaxQuant.
- "Putative False Positive (FP)": This class contains pairs in which two different amino acid sequences have been assigned to the H and L counterparts. These doublets are excluded from the final output.
- "Rescued": This class contains the doublets in which only one counterpart of the pair was identified. These doublets can be further classified as "L only" or "H only", depending on the channel that was identified by MS/MS. We generally consider "H only" pairs to be more reliable, because heavy methyl groups can only derive from in vivo enzymatic methylation. Rescued doublets expand the number of annotated methyl-peptides beyond those where both the H and L counterparts were fragmented by MS and identified during the database search; yet, such methyl-peptide pairs are in principle less reliable than the Matched and Mismatched ones. Hence, we also estimated an FDR associated to them (see below).



Modifications	Modified sequence	Dimethyl4 (KR) Probabilities	Score
1 Dimethyl4 (KR)	_GAITR(di)GATVTR_	GAITR(0.965)GATVTR(0.035)	55

Sequence	m/z	RT	Charge
GAITR(di)GATVTR	380.2409	29.72	3
GYGFITFSDSECAK(me)	798.3559	92.29	2
GYGGSQGGGR(me)GR	561.7735	10.60	2
GYGGSQGGGR(me)GR(me)	572.8036	12.15	2

m/z	RT	Charge	ID
380.2409	29.72	3	GAITR(di)GATVTR
663.9617	46.93	3	M(me)ADFHKEEM(me)DDQDKDK
711.3292	5.16	3	
377.5596	29.76	3	
781.3492	43.61	2	AANGLEADNSCPNSK(me)

H Modified Sequence	L Modified Sequence	Res	Pos	Mod	Mass Error	ΔRT	H Score	L Score	Class
GAITR(di)GATVTR	GAITR(di)GATVTR	R	325	di	0.00074	0.16	69	86	Matched
GAITR(di)GATVTR		R	325	di	1.11603	0.046	121		Rescued
GAITR(di)GATVTR	GAITR(di)GATVTR	R	325	di	0.00074	0.099	88	69	Matched
GAITR(di)GATVTR		R	325	di	0.48526	0.033	56		Rescued

Figure 1. A) The three-step workflow of hmSEEKER. I) Input Reading and filtering: The "msms" and "allpeptides" files from MaxQuant are filtered to remove low confidence methyl-peptides and contaminants. Data are then summarized by two indexes: one includes information on objects that were fragmented by MS/MS and thus were identified as peptides (Peptide Index), while the second contains information from all objects detected as MS1 peaks (Peak Index). II) Doublets Search: Each peptide in the Peptide Index is associated to its corresponding peak in the Peak Index, then its counterpart is searched. Since the doublet search is performed at the MS1 peak level, it is possible to find a counterpart even if the latter is not fragmented in MS2. In this case hmSEEKER retrieves a 'Rescued' doublet. If both counterparts have been identified, a 'Matched' doublet is reported. III) Output Refinement: All doublets identified are listed in a "Redundant" table. Starting from this table, hmSEEKER also generates a "Non-Redundant" table, which includes only the highest-scoring, non-redundant doublets. B) Workflow of the experiment carried out to generate the methyl-proteome on which hmSEEKER was tested. I) Ovarian cancer cells SK-OV-3 were grown in media containing either light (Met-0) or heavy (Met-4) methionine. II) Upon full incorporation of the isotopically labeled amino acids, cells were harvested and mixed in 1:1 proportion. Trypsin was used to digest in solution the proteins into peptides. III) Peptides were then subjected to an affinity enrichment step using pan-methyl-R-antibodies, LC–MS/MS aquisition was carried out on a Q Exactive Orbitrap mass spectrometer. IV) Raw files were processed using MaxQuant to identify methylated peptides. V) Finally, MaxQuant output was processed with hmSEEKER. Putative true positive methyl-peptides are detected as pairs of peaks separated by a specific delta mass, which reflects the methylation state of the peptide (each methyl group introduces a 4.0222 Da shift). See also Supporting Information.

Finally, the software performs a step of redundancy reduction of the output, where doublets associated to the same methyl-peptide are ranked according to their class and total Andromeda score, and for each modified peptide only the top-ranking doublet is retained.

hmSEEKER is controlled from the command line through a configuration file, in which the user can explicitly define any of the input filtering and the doublet search parameters. If any of these parameters is not defined the configuration file, its default value will be employed. For the doublet search step, the

ME, ΔRT, and $\mathrm{Log_2}$Ratio default values have been optimized empirically, by using a dataset generated in house from the MS-analysis of immuno-enriched methyl-peptides from an in-solution tryptic digestion of SK-OV-3 whole cell extract on a Q Exactive instrument, as detailed in Figure 1B and in the Supporting Information. The MS raw data were processed by MaxQuant and the output dataset was initially analyzed using permissive ME and ΔRT threshold values of 25 ppm and 5 min, respectively, and no $\mathrm{Log_2}$Ratio cut-offs. Once a first list of hmSILAC doublets was generated by the tool, we calculated mean (μ) and SD (σ) of their $\log_2$-transformed H/L ratio distribution. These values were then employed to distinguish reliable assignments from putative false positives. In fact, under the hypothesis that the two H and L samples were grown in the same condition and mixed in 1:1 proportion, the H/L ratio of genuine hmSILAC doublets is expected be close to 1, while false IDs will deviate from this value. Thus, we defined as High Confidence (HC) the doublets with a ratio within the $\mu \pm \sigma$ range and as Low Confidence (LC) those with ratios exceeding this range (**Figure 2**A).

We then progressively refined the cut-offs to retain the majority of the HC doublets and filter out LC doublets. This led us to define 2 ppm and 0.5 min as the ME and ΔRT default values (Figure 2B,C; see also Supporting Information).

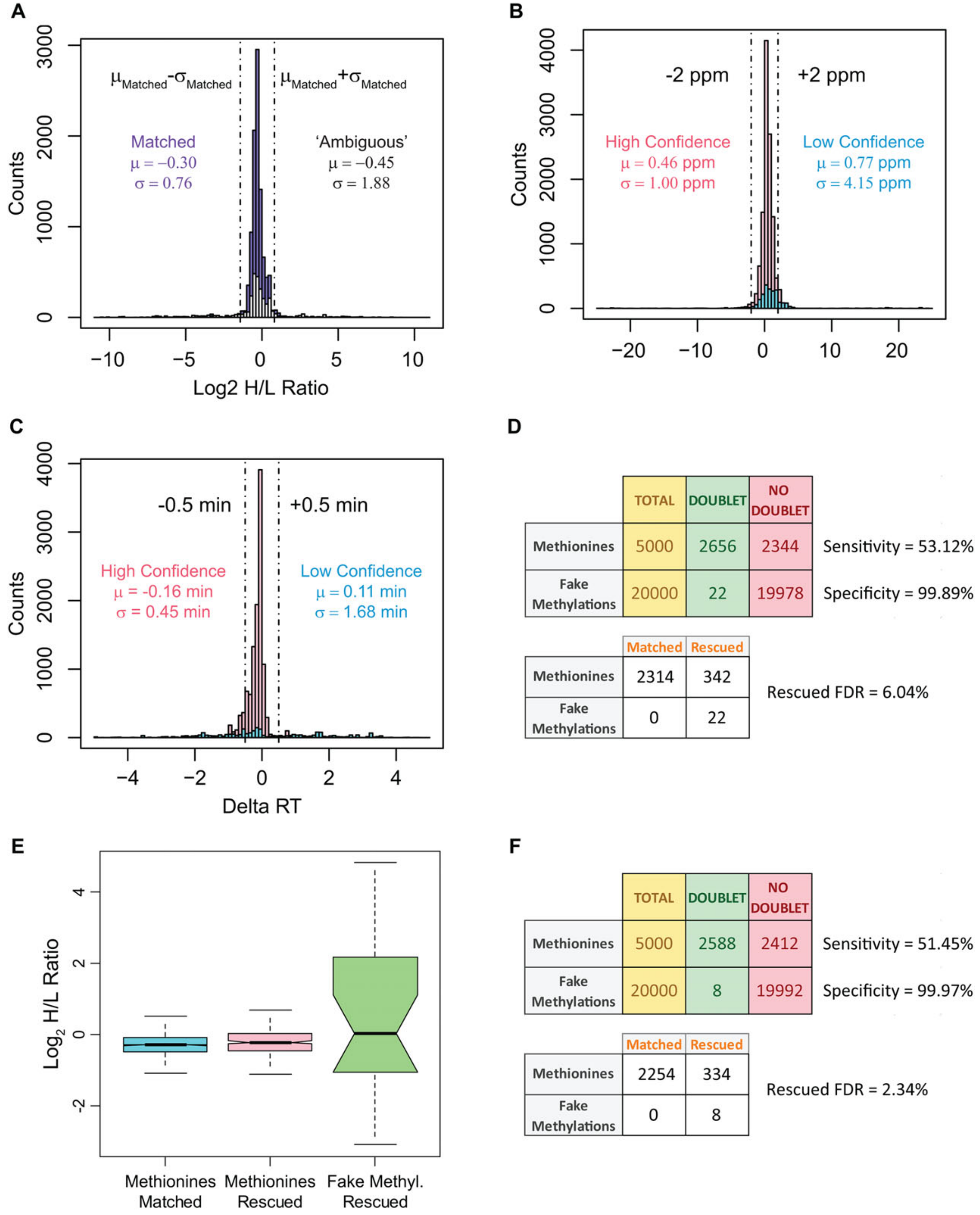
To further assess specificity and sensitivity of hmSEEKER, we focused on the methionine-containing peptides included in the test dataset. The rationale is that, since we are labeling cells with heavy methionine, peptides containing this amino acid will generate doublets similar to methylated peptides. In parallel, we built a dataset of "True Negatives", by considering peptides that do not contain methionine (hence, should not generate a doublet) and randomly assigning to them one type of methylation (i.e. mono-, di-, or tri-methylation, in either light or heavy form). From these two datasets, we sampled 5000 and 20 000 peptide-spectrum matches corresponding to 2136 and 13 767 non-redundant peptides, respectively. In this way, we reproduced the 80% FDR that has been observed for methyl-peptide-spectrum matches when the results of a target–decoy peptide search are filtered in order to obtain a global FDR of ≈1%.[2] We analyzed this dataset with hmSEEKER using the default parameters previously described. hmSEEKER identified a doublet for 2656 methionine-containing peptides out of 5000, corresponding to a sensitivity of 53.12%, and only 22 false methyl-peptides out of 20 000, resulting in a specificity of 99.89% (Figure 2D). Of the 2656 methionine-containing peptides, 2314 were identified as Matched doublets and 342 as Rescued, while all 22 "false positives" were in the Rescued class (Matched doublets FDR = 0%; Rescued doublets FDR = 22/364 = 6.04%, Figure 2D). We however observed that these false positives had a broader H/L peptide-pair ratio distribution with respect to methionine-containing ones (Figure 2E). This observation led us to include an additional cut-off threshold, set to 1, for the $\mathrm{Log_2}$Ratio parameter. This cut-off reduced by 64% the number of identifications in the "true negative" subset of peptides, but did not affect the methionine-containing ones (Figure 2F), reducing the FDR of Rescued doublets from 6.04% to 2.34% (Figure 2F). The downside of using the doublet Ratio as possible filtering criterion is that it is not applicable in a straightforward way to hmSILAC experiments where the mixing ratio is not 1:1.

These parameters were further confirmed by processing two additional hmSILAC methyl-proteomes acquired in leukemic NB4 cells (Supporting Information). In fact, we observed that the HC doublets were reproducibly found inside the defined cut-offs, in spite of small differences in the LC–MS/MS run.

By applying these cut-off parameters to the same dataset used for their estimation, we retrieved 646 non-redundant methyl-peptide doublets of which 440 (68%) are identified as Matched doublets, 44 (7%) as Mismatched doublets, and 162 (25%) as Rescued doublets.

To benchmark the performance of hmSEEKER, we compared our results with those obtained by re-processing the same dataset with MethylQuant (version 1.0). To ensure that the two analyses were truly comparable, we applied the 2-ppm and 0.5-min cut-off values to the search and also manually removed low-scoring peptides and peptides with mixed labeling from the input for MethylQuant, which mimics the initial filtering operated by hmSEEKER. MethylQuant returned 609 non-redundant doublets, of which 69 (11%) were classified as 'Very High Confidence', 154 (25%) as 'High Confidence' and 386 (63%) as 'Low Confidence'. We found 469 non-redundant doublets (70%) in common between the methods (**Figure 3**A; Figure S3, Supporting Information). Of the 140 doublets uniquely found by MethylQuant, 110 (79%) are Low Confidence, whereas 98 doublets out of the 177 (55%) uniquely identified by our tool are Matched (Figure 3B). Through further analysis, we also found that hmSEEKER shows a mild bias toward peptides with two positive charges, 1 or 2 missed cleavages and a higher Andromeda score, which might explain the differences in the results (Supporting Information). All analyses were performed on a Dell Precision T7600 workstation with 32 GB of RAM and two 2.40 GHz processors. The run time of MaxQuant, which generated the input for both tools was ≈ 3 days, whereas the run times for hmSEEKER and MethylQuant were 4 min and 15 h, respectively. In fact, hmSEEKER takes advantage of MaxQuant tabular output and does not need to parse the raw data on its own. Therefore, once the database search is completed, a list of putatively true methyl-peptides can be generated almost immediately and data can be quickly reprocessed. Overall, this comparison suggests that our strategy of parsing MaxQuant output tables performs comparably to the more computationally intensive algorithm MethylQuant, and confirms the high quality of the doublets found by our tool.

In conclusion, hmSEEKER can identify in vivo methyl-peptides by analyzing hmSILAC data with a low rate of false positive doublets. However, since false positives fall in the Rescued doublets category, it is advisable to perform a two-step analysis, in which two different set of parameters, a broad one (i.e., RT < 1 min, ME < 8 ppm, $\mathrm{Log_2}$Ratio < 1.5) and a stringent one (i.e., the default parameters), are applied consecutively to the same dataset. The final output could thus include the Matched and Mismatched doublets from the broad analysis (which has higher sensitivity) and the Rescued doublets only from the stringent one. While the dependency on MaxQuant can be seen as a constraint, it can also represent an advantage, since MaxQuant is widely used for the analysis of quantitative proteomics data, which can be more easily integrated with the hmSEEKER results for methyl-peptides orthogonal validation. Hence, we deem hmSEEKER



D

	TOTAL	DOUBLET	NO DOUBLET	
Methionines	5000	2656	2344	Sensitivity = 53.12%
Fake Methylations	20000	22	19978	Specificity = 99.89%

	Matched	Rescued	
Methionines	2314	342	Rescued FDR = 6.04%
Fake Methylations	0	22	

F

	TOTAL	DOUBLET	NO DOUBLET	
Methionines	5000	2588	2412	Sensitivity = 51.45%
Fake Methylations	20000	8	19992	Specificity = 99.97%

	Matched	Rescued	
Methionines	2254	334	Rescued FDR = 2.34%
Fake Methylations	0	8	

Figure 2. Optimization of hmSEEKER parameters. A) Distribution of H/L peptide ratios for 'Matched' doublets (purple) and the rest of the population (white). We define as 'High confidence' (HC) peptides in the μ ± σ range calculated for the Matched doublets, and as 'Low confidence' the rest of the population. B) Distribution of mass errors in the HC (pink) and LC (cyan) doublets sub-populations. C) Distribution of ΔRT in the HC (pink) and LC (cyan) doublets sub-populations. D) Results of the analysis performed on 5000 methionine-containing peptides and 20 000 false methyl-peptides. E) Boxplot representation of the H/L doublet ratio distributions of methionine-containing peptides (cyan and pink) and false positives (green). F) Results of the second analysis, performed on the same dataset, in which we filtered doublets based on their H/L ratio, in addition to ME and ΔRT.

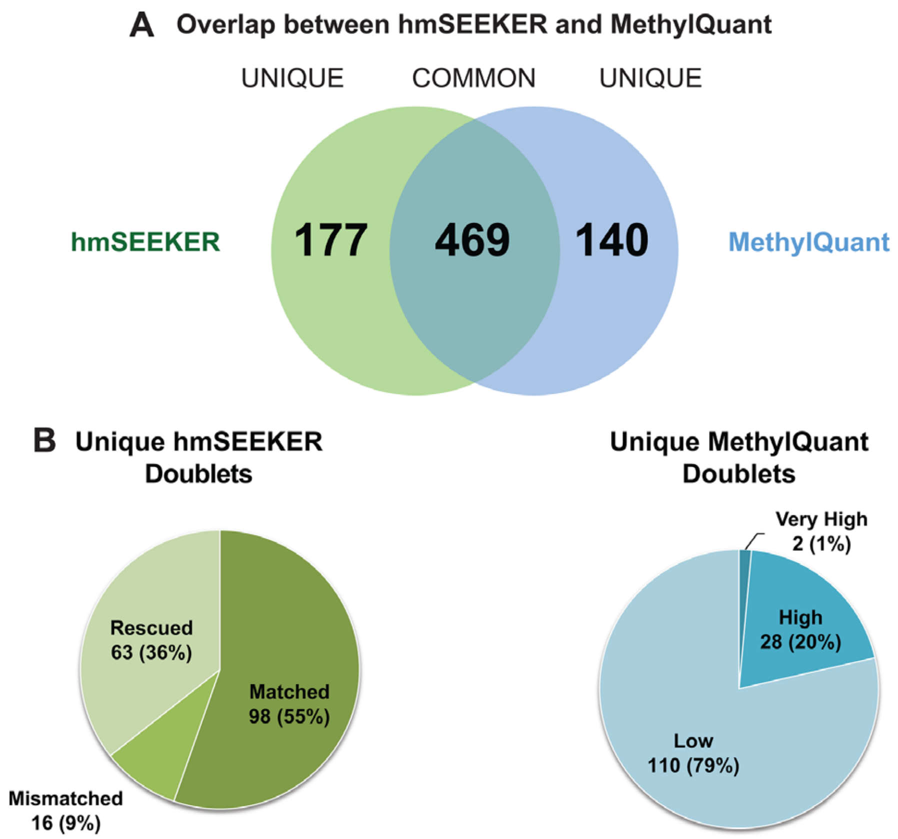


Figure 3. Comparison between the output of hmSEEKER and MethylQuant. A) Overlap between the methyl-peptide doublets identified as putative true. B) Left: Pie-chart showing the percentages of Matched, Mismatched, and Rescued doublets among the 177 peptides that are unique for hmSEEKER. Right: Pie-chart showing the percentages of Very High, High, and Low confidence doublets among the 140 peptides that are unique for MethylQuant.

a highly useful tool for the annotation of high-confidence methyl-proteomes.

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

The authors would like to thank Valeria Spadotto and Roberto Giambruno for feedback on the software and on its user manual. Research in TB group was supported by grants from the Italian Ministry of Health (GR-2011-02347880), the Epigen Flagship Project (CNR), and the Italian Association for Cancer Research (no. 15741).

Conflict of Interest

The authors declare no conflict of interest.

Keywords

data processing and analysis, heavy methyl SILAC, methylation, MS–LC–MS/MS, posttranslational modifications analysis, stable isotope labeling

Received: July 23, 2018
Revised: January 14, 2019
Published online:

[1] M. A. Erce, C. N. Pang, G. Hart-Smith, M. R. Wilkins, *Proteomics* **2012**, *12*, 564.
[2] G. Hart-Smith, D. Yagoub, A. P. Tay, R. Pickford, M. R. Wilkins, *Mol. Cell. Proteomics* **2016**, *15*, 989.
[3] S. E. Ong, G. Mittler, M. Mann, Nat Methods **2004**, *1*, 119.
[4] a) M. Bremang, A. Cuomo, A. M. Agresta, M. Stugiewicz, V. Spadotto, T. Bonaldi, *Mol. Biosyst.* **2013**, *9*, 2231; b) X. J. Cao, A. M. Arnaudo, B. A. Garcia, *Epigenetics* **2013**, *8*, 477.
[5] V. Geoghegan, A. Guo, D. Trudgian, B. Thomas, O. Acuto, *Nat. Commun.* **2015**.
[6] a) G. K. Potts, E. A. Voigt, D. J. Bailey, C. M. Rose, M. S. Westphall, A. S. Hebert, J. Yin, J. J. Coon, *Anal. Chem.* **2016**, *88*, 3295; b) P. L. Ross, Y. N. Huang, J. N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson, D. J. Pappin, *Mol. Cell. Proteomics : MCP* **2004**, *3*, 1154.
[7] C. J. Mitchell, M. S. Kim, C. H. Na, A. Pandey, *Mol. Cell. Proteomics* **2016**, *15*, 2829.
[8] A. P. Tay, V. Geoghegan, D. Yagoub, M. R. Wilkins, G. Hart-Smith, *J. Proteome Res.* **2017**.
[9] J. Cox, M. Mann, *Nat. Biotechnol.* **2008**, *26*, 1367.
[10] J. Cox, N. Neuhauser, A. Michalski, R. A. Scheltema, J. V. Olsen, M. Mann, *J. Proteome Res.* **2011**, *10*, 1794.