ProMetheusDB: an in-depth analysis of the high-quality human methyl-proteome

Authors
Enrico Massignani\textsuperscript{1,2}, Roberto Giambruno\textsuperscript{1,3,4}, Marianna Maniaci\textsuperscript{1,2}, Luciano Nicosia\textsuperscript{1+}, Avinash Yadav\textsuperscript{1#}, Alessandro Cuomo\textsuperscript{1}, Francesco Raimondi\textsuperscript{1,5} and Tiziana Bonaldi\textsuperscript{1,6}

Affiliations
\begin{itemize}
\item \textsuperscript{1}Department of Experimental Oncology, European Institute of Oncology IRCCS, 20139 Milan, Italy
\item \textsuperscript{2}European School of Molecular Medicine (SEMM), Milan, Italy
\item \textsuperscript{3}Center for Genomic Science of Istituto Italiano di Tecnologia at European School of Molecular Medicine, Istituto Italiano di Tecnologia, Milan, Italy
\item \textsuperscript{4}Institute of Biomedical Technologies, National Research Council, 20054 Segrate, Milan, Italy
\item \textsuperscript{5}Bio@SNS, Scuola Normale Superiore, 56126 Pisa, Italy
\item \textsuperscript{6}Department of Oncology and Haematology-Oncology, University of Milan, 20122 Milan, Italy
\item \textsuperscript{+}Current address: Leukaemia Biology Laboratory, Cancer Research UK Manchester Institute, The University of Manchester, Ogleby Cancer Research Centre Building, Manchester, M20 4GJ, United Kingdom (UK)
\item \textsuperscript{#}Current address: GSK Vaccines Srl, Siena, Italy
\end{itemize}

Corresponding author information:
Tiziana Bonaldi: Via Adamello 16, 20139 Milan, Italy; Phone: +390294375123; e-mail: tiziana.bonaldi@ieo.it; tiziana.bonaldi@unimi.it
SUPPLEMENTARY MATERIAL

Figure S1. Assessment of the performance of the Machine Learning model within hmSEEKER v2.0. A) We determined the optimal training set size by plotting the model learning curve, which reaches a plateau when the training set size is ~6000. Thus, 6000 doublets (3000 true and 3000 false) were extracted from a dataset of 8052 to train the model via five-fold cross-validation; the remaining 2052 were set aside to serve as a validation set. B) Confusion Matrix generated by applying the model to the validation set. C) Table panel showing how the doublets in the training set are classified differently by the original hmSEEKER cut-offs and by the machine learning model.

Figure S2. Comparison between the different protein methylation events detected in our experiments. A-B) Overlap between R-methylated and K-methylated peptides and proteins, respectively. C) Bar chart displaying the fraction of each methylation mark identified in each sub-category of the different hmSILAC experiments. Numbers on top of each bar indicate the total number of modifications annotated.

Figure S3. Structural analysis of protein regions bearing PTMs other than methylation. A) Counts of modified sites that occur in regions annotated as either domains or disordered regions in the MobiDB database, compared to randomly sampled sites. We observed that other PTMs beside K methylation, such as K acetylation and K ubiquitination, occur on domains more frequently than on disordered regions. B) Enrichment of PTMs in Intrinsically Disordered Regions (IDRs) predicted by AlphaFold, which confirms the results shown in panel A.

Figure S4. Cross-talk of R methylation with K acetylation, K ubiquitination and K sumoylation. A) Counts of R-methyl-sites that occur in proximity of a ubiquitination site. As a control, counts of randomly sampled R-sites from the human proteome were also assessed, indicating a significant anti-correlation between R methylation and K ubiquitination. Significance was calculated with a Fisher exact test. B-C) The same analysis performed on K acetylation and K sumoylation sites did not indicate significant correlation or anti-correlation of these PTM. Significance was calculated with a Fisher exact test.

Figure S5. Crystal structures of additional protein pairs emerged from the structural analysis with Mechismo. A) Crystal structure of SRSF1 (orange) and SRPK1 (green). The structure shows R154 of SRSF1 forming hydrogen bonds with E543 and Y549 of SRPK1; methylation of SRSF1 may disrupt the hydrogen bonds and thus reduce the interaction between the two proteins. B) Crystal structure of two MAT2A subunits: like the case in panel A, methylation on R264 could impair the hydrogen bond with E57 between two subunits of the dimeric SAM synthase enzyme.

Figure S6. MS/MS spectra for peptide 27-40 of histone H3. Complete list of fragmentation spectra that were used to annotate the newly identified histone marks H3S28me and H3T32me (linked to Fig 7D).
Table S1 (individual Excel file). The table contains the current version of ProMetheusDB and is organized in 6 sheets, as follows:

- **“R/K-methyl-sites”:** A list of R/K-methyl-sites identified in the initial analysis of all the MS raw data available. For each site, the modification state, the sequence window and the related methyl-peptides are indicated.

- **“R/K-methyl-peptides”:** A list of R/K-methyl-peptides identified in the initial analysis and unequivocally mapping on one protein. If a peptide was detected in a SILAC experiment, its regulation state is indicated.

- **“R/K-methylpeps- ambiguous”:** A list of R/K-methyl-peptides identified in the initial analysis and mapping on two or more different proteins. If a peptide was detected in a SILAC experiment, its regulation state is indicated.

- **“R/K/D/E/N/Q/S/T/H-methyl-sites”:** A list of R/K/D/E/N/Q/S/T/H-methyl-sites identified in the re-analysis of the non-enriched (Input) data. For each site, the modification state, the sequence window and the related methyl-peptides are indicated.

- **“R/K/D/E/N/Q/S/T/H-methyl-peptides”:** A list of R/K/D/E/N/Q/S/T/H-methyl-peptides identified in the re-analysis of the non-enriched (Input) data and unequivocally mapping on one protein.

- **“R/K/D/E/N/Q/S/T/H-methylpeps-ambiguous”:** A list of R/K/D/E/N/Q/S/T/H-methyl-peptides identified in the re-analysis of the non-enriched (Input) data and mapping on two or more different proteins.

- **“hmSILAC exps summary”:** Summary of the hmSILAC experiments analysed to generate the orthogonally validated methylation dataset.

- **“SILAC exps summary”:** Summary of the SILAC experiments that were combined with the hmSILAC data to generate the comprehensive ProMetheusDB.

Table S2 (individual Excel file). Complete results of the functional enrichment analysis performed with the “gprofiler2” R package, as described in the Materials and Methods section. Terms with FDR < 0.01 are highlighted. The file is organized in 4 sheets, as follows:

- **“Clusters”:** Functional enrichment of the proteins in the eight clusters shown in Fig 3A (linked to Fig 3B).

- **“Regulated-Unregulated Rme”:** Functional enrichment of proteins bearing one or more regulated R-methyl-sites (linked to Fig 4B).

- **“Cross-talk”:** Functional enrichment of proteins on which we observed a significant correlation of R-methyl-sites and phospho-S/T-Y sites (linked to Fig 6B).

- **“Non-canonical methylations”:** Functional enrichment of proteins bearing D/E/N/Q/S/T/H-methyl-sites (linked to Fig 7B).
A

Learning Curve

Score

Train set size

Total Doublets = 8052 doublets
(4434 True + 3618 False)

Training Set = 6000 doublets
(3000 True + 3000 False)

Validation Set = 2052 doublets
(1434 True + 618 False)

B

Results on Validation Set
(Absolute Features)

C

| True Label | Prediction based on initial cutoffs | ML Prediction | Counts | % |
|------------|-----------------------------------|---------------|--------|---|
| Positive   | False                             | False         | 3511   | 43.6% |
| False      | False                             | True          | 100    | 1.2%  | False Positive we introduce with the ML |
| False      | True                              | False         | 0      | 0.0%  | False Positive we remove with the ML |
| False      | True                              | True          | 7      | 0.1%  | False Positives |
| True       | False                             | False         | 77     | 1.0%  | False negative |
| True       | True                              | True          | 768    | 9.5%  | True Positive we recover with ML |
| True       | False                             | False         | 0      | 0.0%  | False negatives we introduce with ML |
| True       | True                              | True          | 3589   | 44.6% | True Positive |
| TOT        | 8052                              |               | 100%   |       |
