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

Authors

Enrico Massignani[1,2], Roberto Giambruno[1,3,4], Marianna Maniaci[1,2], Luciano Nicosia[1+], Avinash Yadav[1#], Alessandro Cuomo[1], Francesco Raimondi[1,5] and Tiziana Bonaldi[1,6]

Affiliations

[1] Department of Experimental Oncology, European Institute of Oncology IRCCS, 20139 Milan, Italy

[2] European School of Molecular Medicine (SEMM), Milan, Italy

[3] Center for Genomic Science of Istituto Italiano di Tecnologia at European School of Molecular Medicine, Istituto Italiano di Tecnologia, Milan, Italy

[4] Institute of Biomedical Technologies, National Research Council, 20054 Segrate, Milan, Italy

[5] Bio@SNS, Scuola Normale Superiore, 56126 Pisa, Italy

[6] Department of Oncology and Haematology-Oncology, University of Milan, 20122 Milan, Italy

[+] Current address: Leukaemia Biology Laboratory, Cancer Research UK Manchester Institute, The University of Manchester, Oglesby Cancer Research Centre Building, Manchester, M20 4GJ, United Kingdom (UK)

[#] Current address: GSK Vaccines Srl, Siena, Italy

Corresponding author information:

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

SUPPLEMETARY 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:

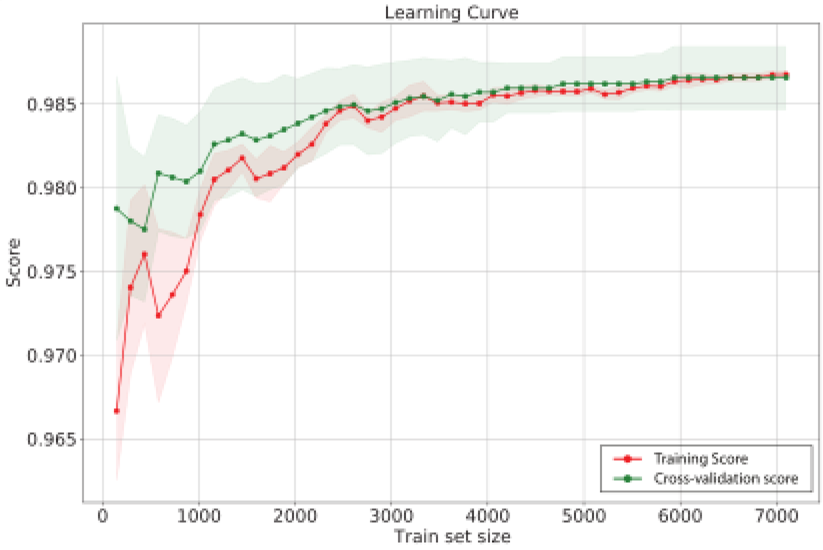
- "R/K-methyl-sites": A list of R/K-methyl-sites identified in the initial analysis of all the MS raw data available. For each site, the modification state, the sequence window and the related methyl-peptides are indicated.
- "R/K-methyl-peptides": A list of R/K-methyl-peptides identified in the initial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

Figure S1

A



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

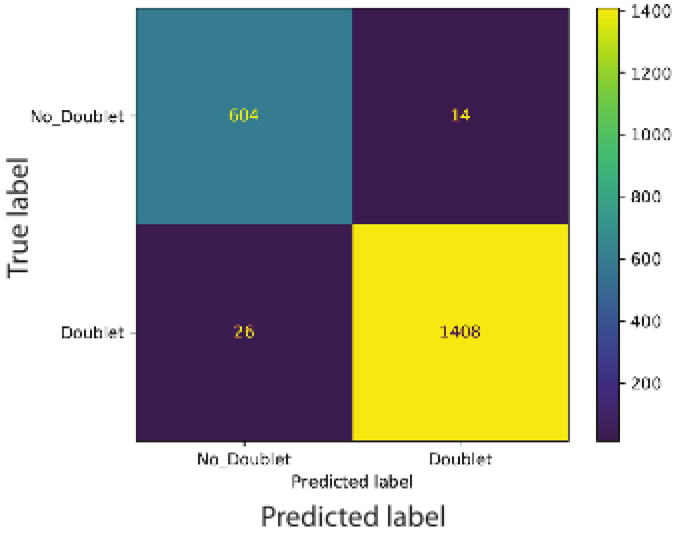
Training Set = 6000 doublets
(3000 True + 3000 False)

5-fold CV

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			
Positive	Positive_old	Positive_ML	Counts	%	
False	False	False	3511	43.6%	True Negative
		True	100	1.2%	False Positive we introduce with the ML
	True	False	0	0.0%	False Positive we remove with the ML
		True	7	0.1%	False Positives
True	False	False	77	1.0%	False negative
		True	768	9.5%	True Positive we recover with ML
	True	False	0	0.0%	False negatives we introduce with ML
		True	3589	44.6%	True Positive
TOT			8052	100%	

Figure S2

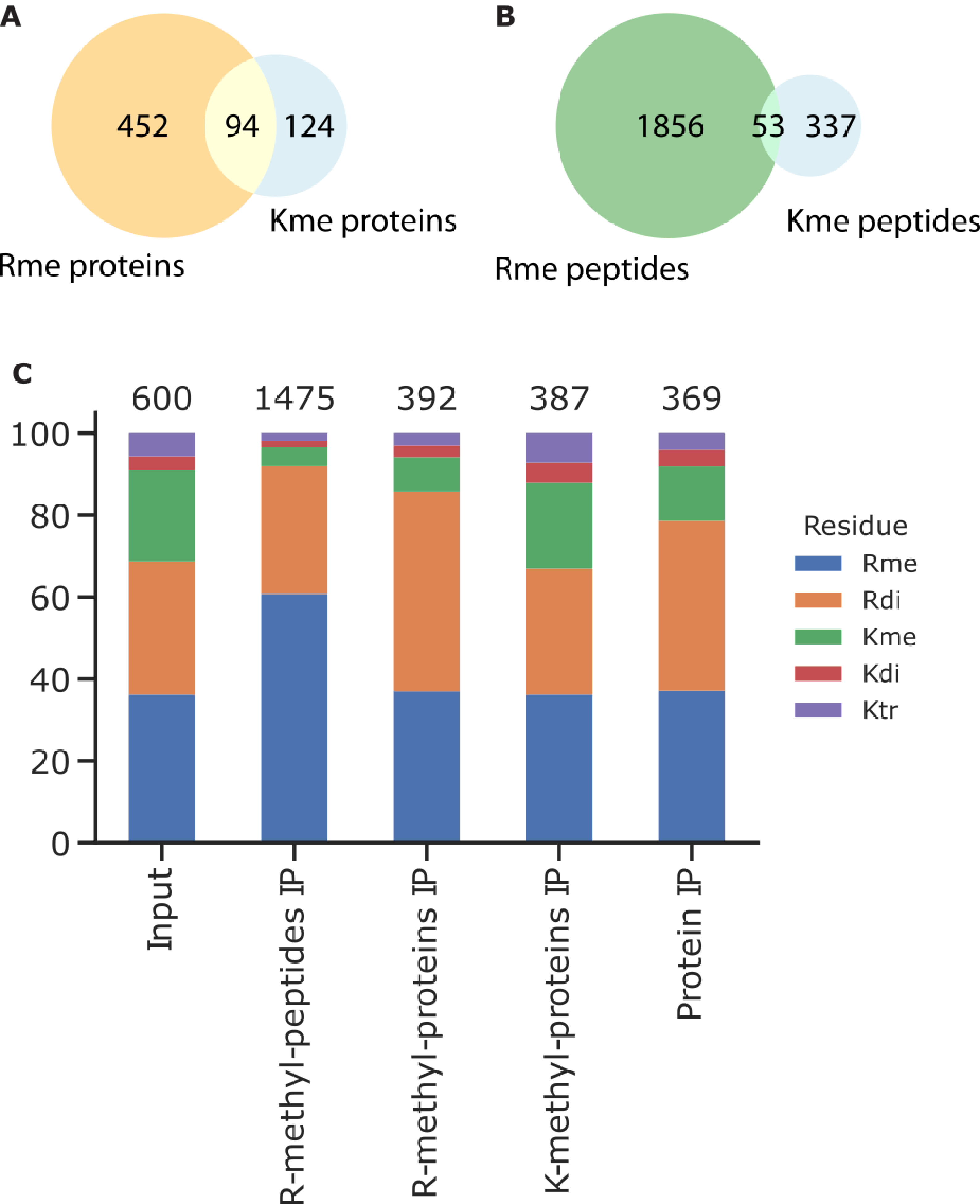

A
452
94
124
Rme proteins
Kme proteins
B
1856
53
337
Rme peptides
Kme peptides
C
600
1475
392
387
369
100
80
60
40
20
0
Input
R-methyl-peptides IP
R-methyl-proteins IP
K-methyl-proteins IP
Protein IP
Residue
Rme
Rdi
Kme
Kdi
Ktr

Figure S3

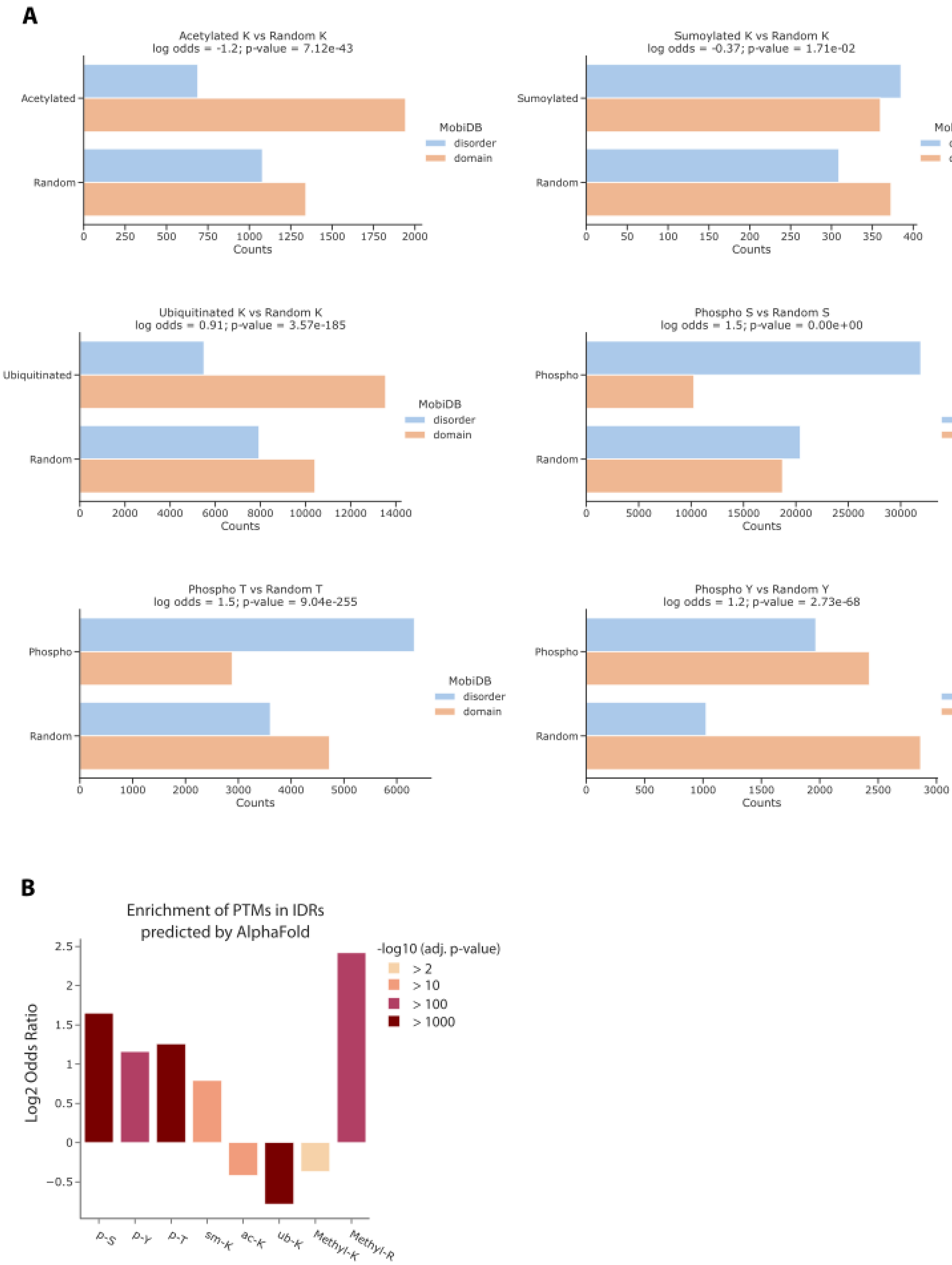

A
Acetylated K vs Random K
log odds = -1.2; p-value = 7.12e-43
Acetylated
Random
0
250
500
750
1000
1250
1500
1750
2000
Counts
MobiDB
disorder
domain
Sumoylated K vs Random K
log odds = -0.37; p-value = 1.71e-02
Sumoylated
Random
0
50
100
150
200
250
300
350
400
Counts
MobiDB
disorder
domain
Ubiquitinated K vs Random K
log odds = 0.91; p-value = 3.57e-185
Ubiquitinated
Random
0
2000
4000
6000
8000
10000
12000
14000
Counts
MobiDB
disorder
domain
Phospho S vs Random S
log odds = 1.5; p-value = 0.00e+00
Phospho
Random
0
5000
10000
15000
20000
25000
30000
Counts
MobiDB
disorder
domain
Phospho T vs Random T
log odds = 1.5; p-value = 9.04e-255
Phospho
Random
0
1000
2000
3000
4000
5000
6000
Counts
MobiDB
disorder
domain
Phospho Y vs Random Y
log odds = 1.2; p-value = 2.73e-68
Phospho
Random
0
500
1000
1500
2000
2500
3000
Counts
MobiDB
disorder
domain
B
Enrichment of PTMs in IDRs
predicted by AlphaFold
Log2 Odds Ratio
2.5
2
1.5
1
0.5
0
−0.5
-log10 (adj. p-value)
> 2
> 10
> 100
> 1000
p-S
p-Y
p-T
sm-K
ac-K
ub-K
Methyl-K
Methyl-R

Figure S4

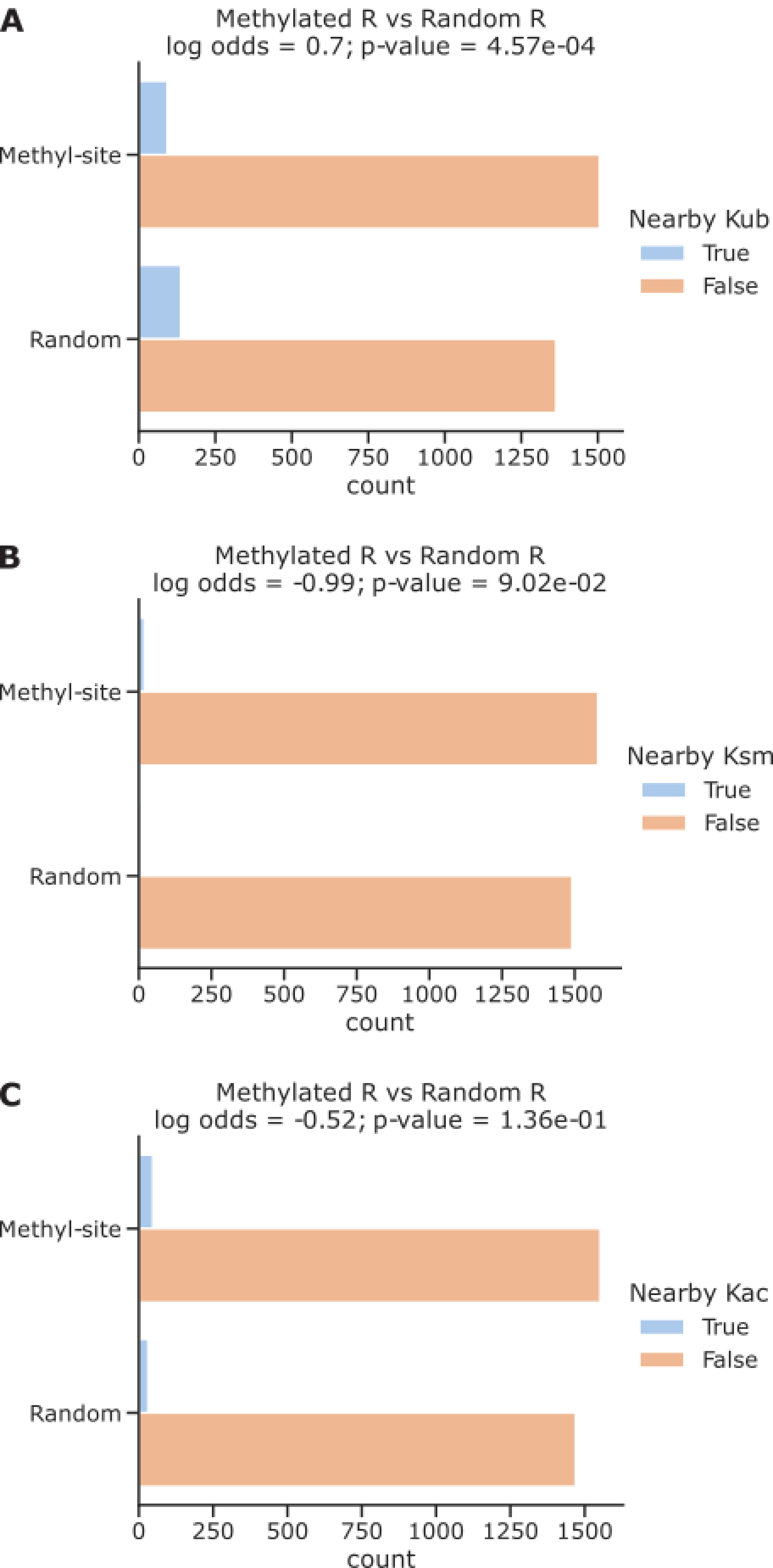

A
Methylated R vs Random R
log odds = 0.7; p-value = 4.57e-04
Methyl-site
Random
0
250
500
750
1000
1250
1500
count
Nearby Kub
True
False
B
Methylated R vs Random R
log odds = -0.99; p-value = 9.02e-02
Methyl-site
Random
0
250
500
750
1000
1250
1500
count
Nearby Ksm
True
False
C
Methylated R vs Random R
log odds = -0.52; p-value = 1.36e-01
Methyl-site
Random
0
250
500
750
1000
1250
1500
count
Nearby Kac
True
False

Figure S5

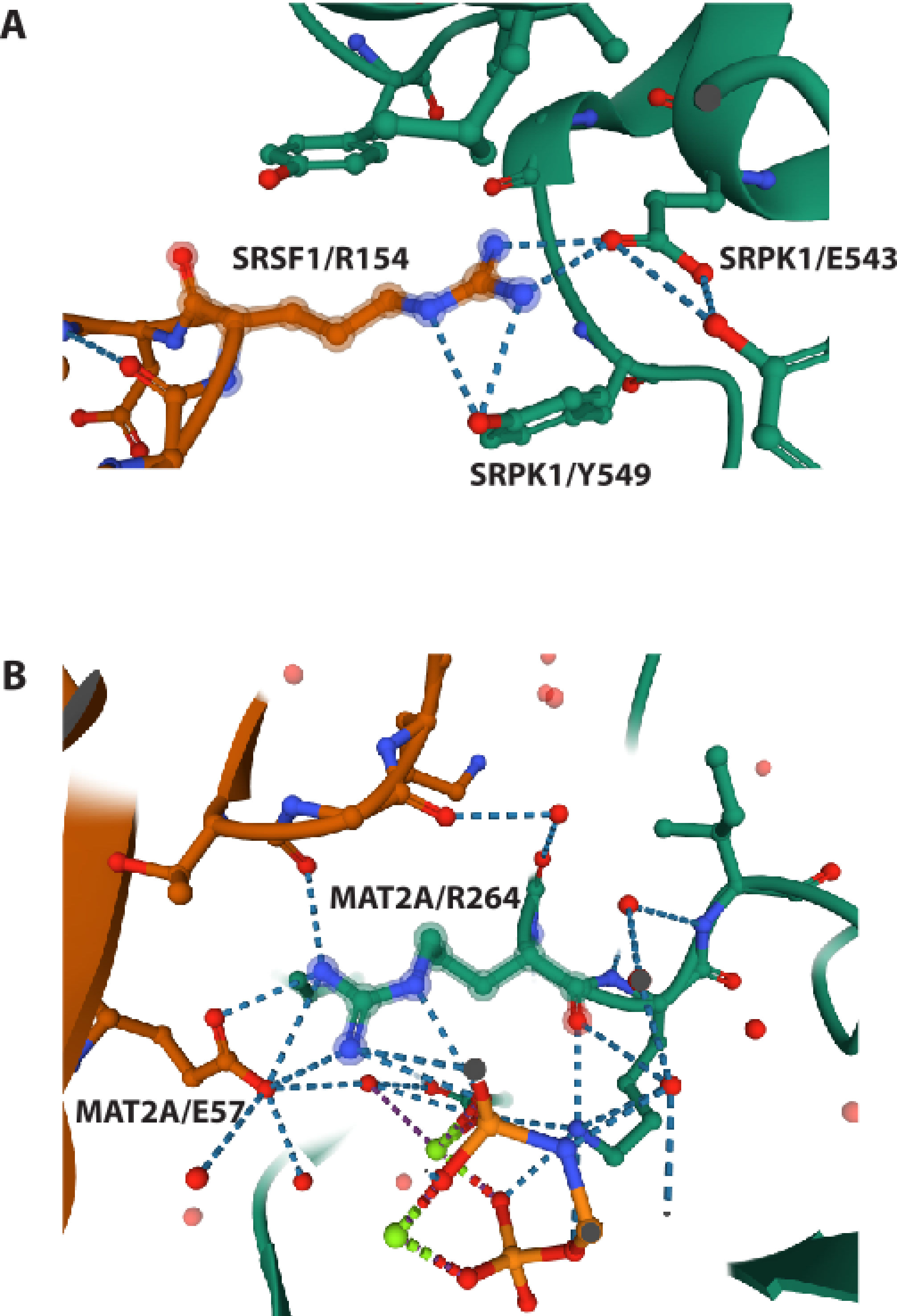
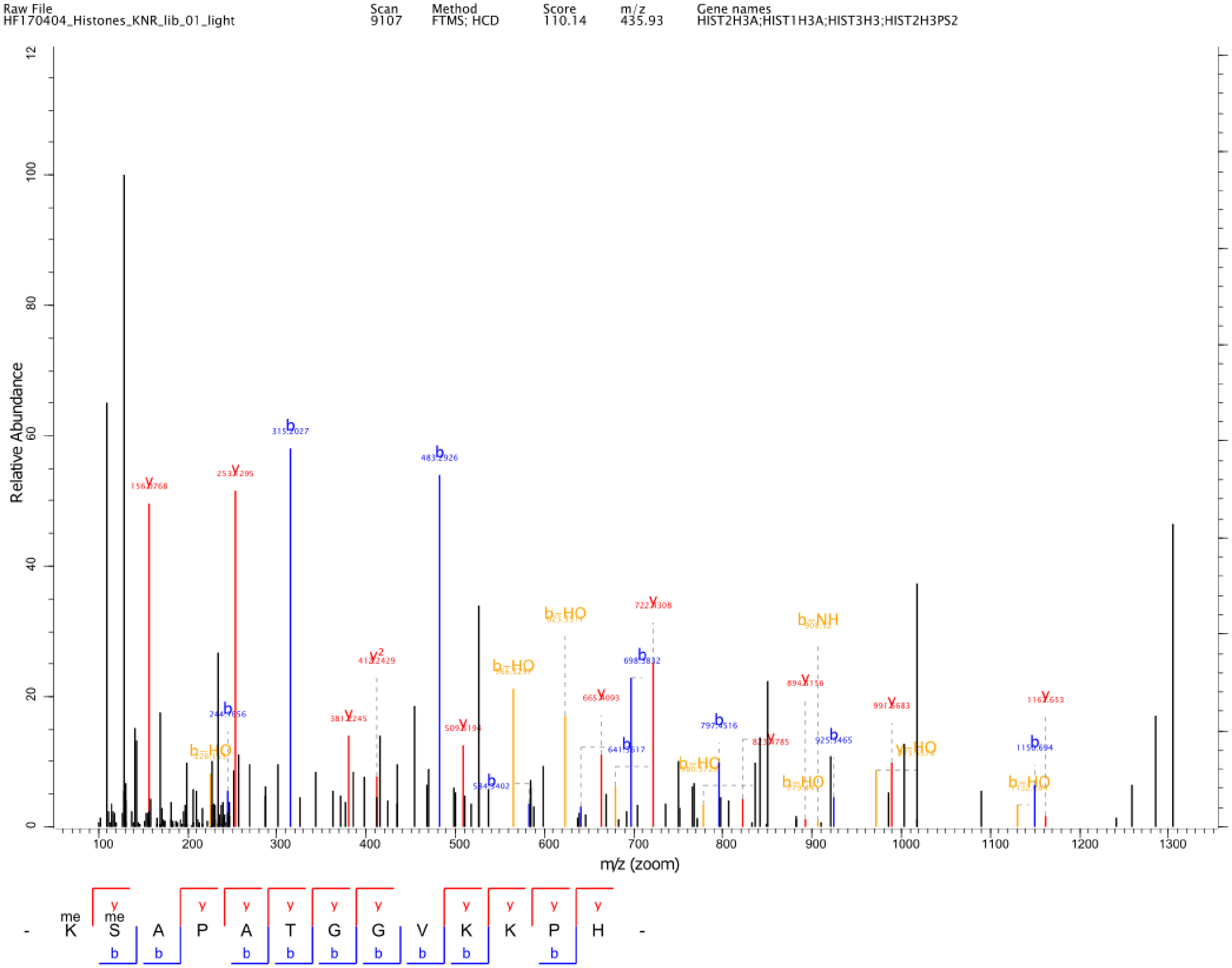


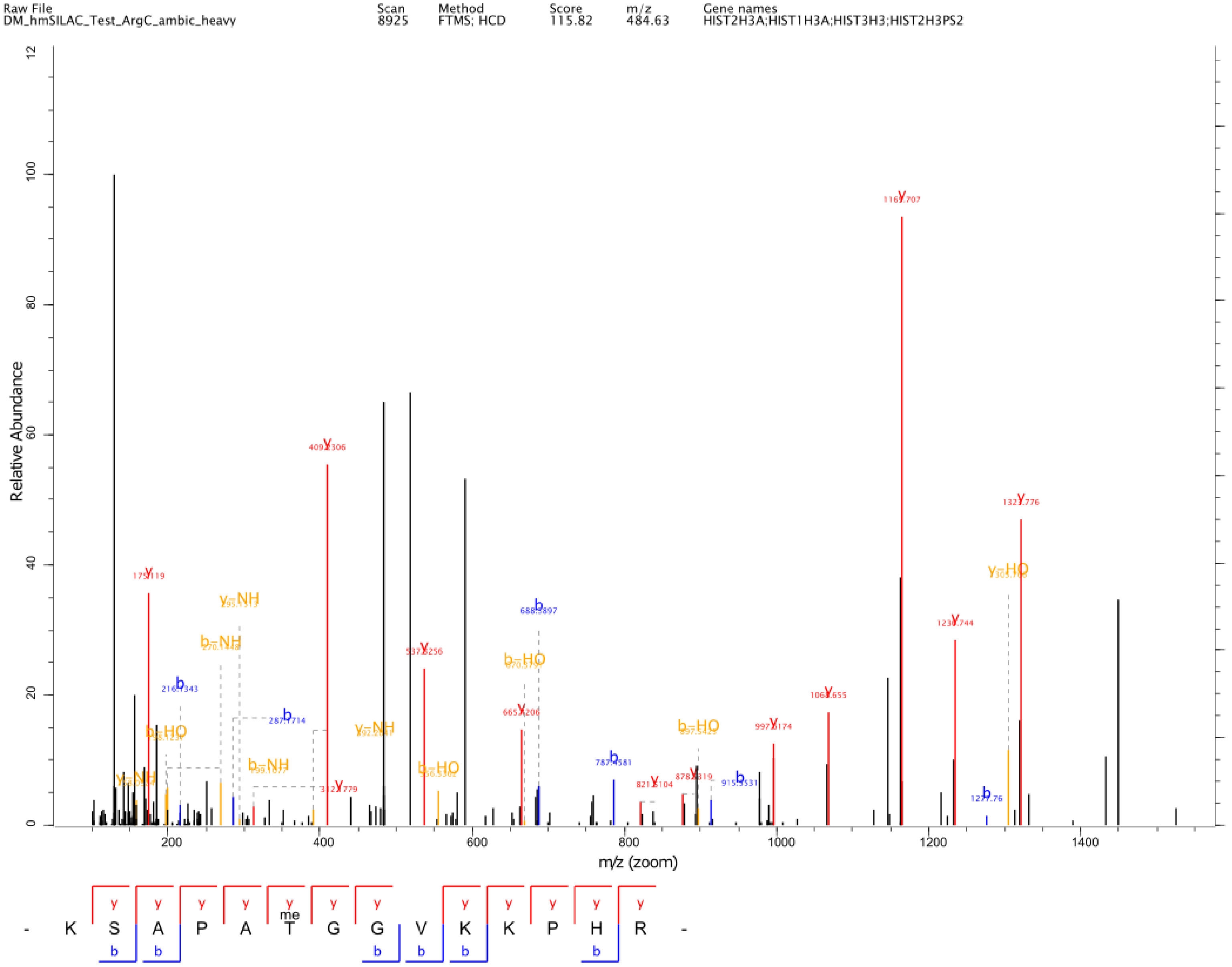
Figure S6

A



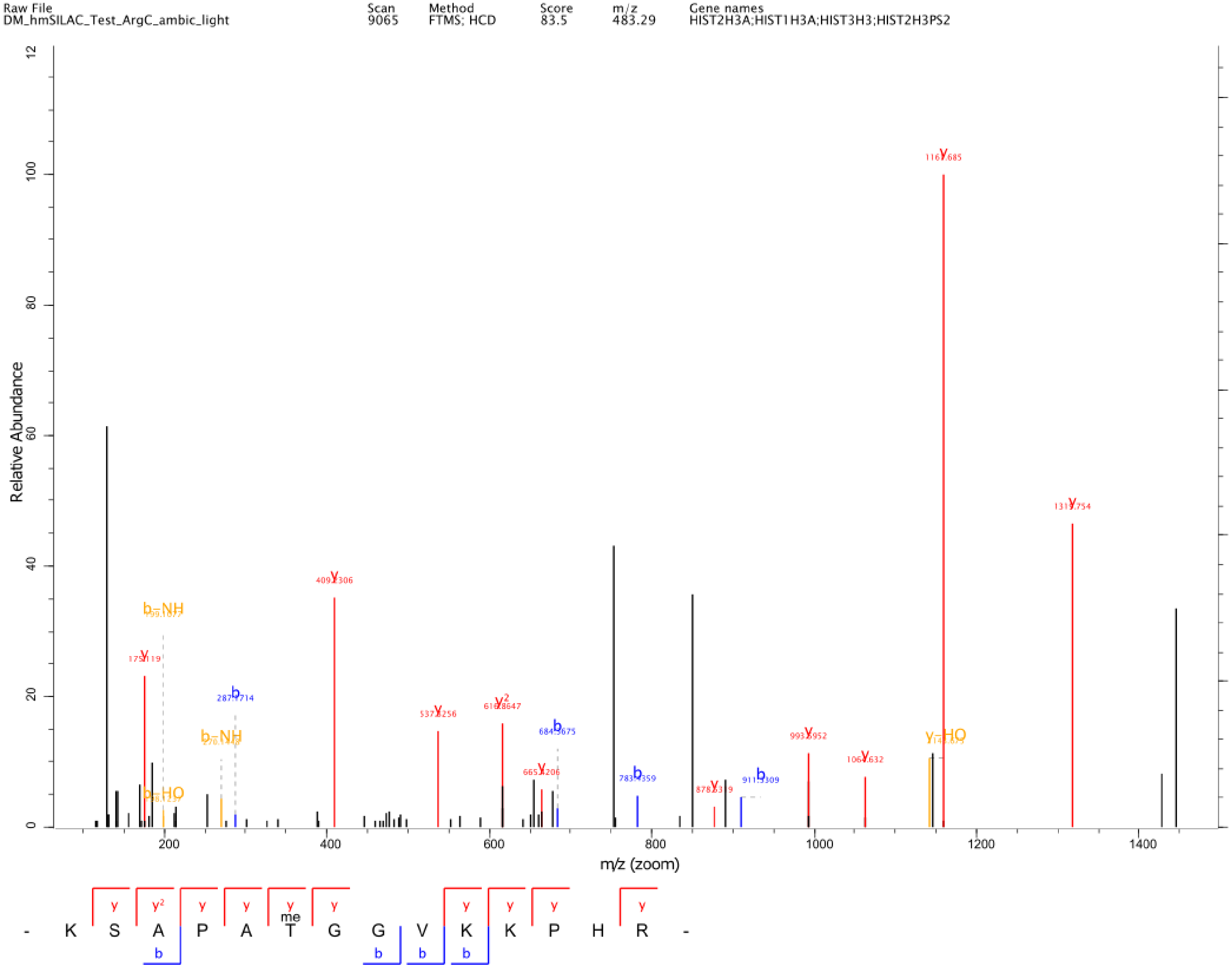
Raw File	Scan	Method	Score	m/z	Gene names
HF170404_Histones_KNR_lib_01_light	9107	FTMS; HCD	110.14	435.93	HIST2H3A;HIST1H3A;HIST3H3;HIST2H3PS2

B



Raw File	Scan	Method	Score	m/z	Gene names
DM_hmSILAC_Test_ArgC_ambic_heavy	8925	FTMS; HCD	115.82	484.63	HIST2H3A;HIST1H3A;HIST3H3;HIST2H3PS2

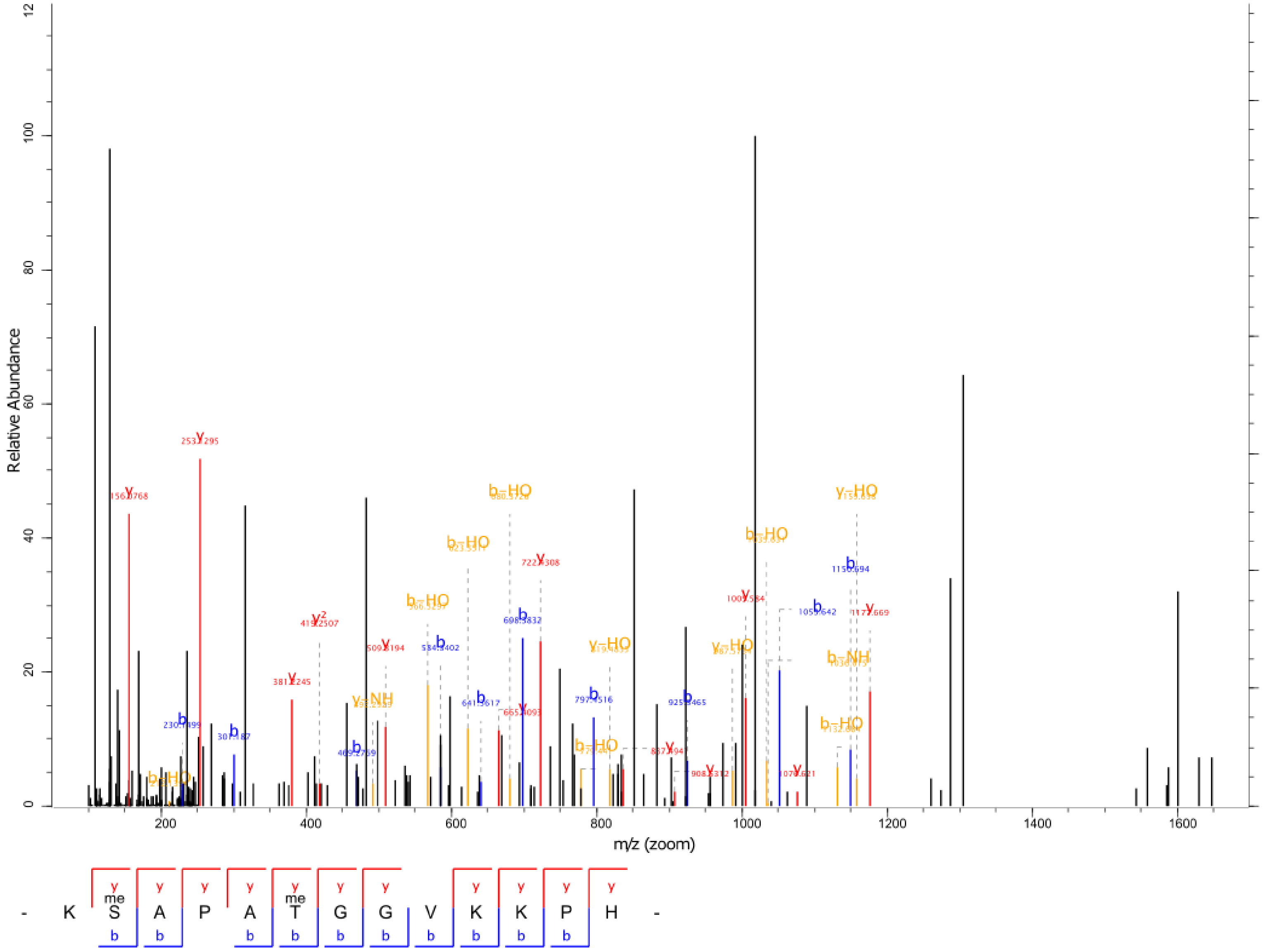
C



Raw File	Scan	Method	Score	m/z	Gene names
DM_hmSILAC_Test_ArgC_ambic_light	9065	FTMS; HCD	83.5	483.29	HIST2H3A;HIST1H3A;HIST3H3;HIST2H3PS2

D

Raw File	Scan	Method	Score	m/z	Gene names
HF170404_Histones_KNR_lib_03_light	9181	FTMS; HCD	128.15	435.93	HIST2H3A;HIST1H3A;HIST3H3;HIST2H3PS2


Relative Abundance
m/z (zoom)
y 156.0768
y 253.1295
b 230.1199
b 307.187
b-HO 213.134
y 381.2245
y2 419.2507
b 469.2739
y-NH 492.2525
y 509.3194
b-HO 566.5297
b 584.3402
b-HO 623.5517
b 641.3617
y 665.4093
b-HO 680.3728
b 698.3832
y 722.4308
y-HO 819.4833
b 797.4516
b-HO 779.441
y 837.4947
b 925.5465
y 908.5312
y-HO 987.5779
y 1005.584
b-HO 1037.631
b 1055.642
y 1076.621
b 1150.694
y-HO 1155.658
b-NH 1036.675
y 1173.669
b-HO 1132.684
- K S A P A T G G V K K P H -
me
me
y
b