



**HAL**  
open science

# A Computational Tool for Analysis of Mass Spectrometry Data of Ubiquitin-Enriched Samples

Rune Matthiesen, Manuel S Rodriguez, Ana Sofia Carvalho

► **To cite this version:**

Rune Matthiesen, Manuel S Rodriguez, Ana Sofia Carvalho. A Computational Tool for Analysis of Mass Spectrometry Data of Ubiquitin-Enriched Samples. Manuel S. Rodriguez; Rosa Barrio. The Ubiquitin Code, 2602, Springer US, pp.205-214, 2023, Methods in Molecular Biology, 978-1-0716-2858-4. 10.1007/978-1-0716-2859-1\_15 . hal-04269366

**HAL Id: hal-04269366**

**<https://hal.science/hal-04269366>**

Submitted on 6 Nov 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1 **A computational tool for analysis of mass spectrometry data of ubiquitin enriched samples**

2 Rune Matthiesen<sup>1\*</sup>, Manuel S. Rodriguez<sup>2</sup> and Ana Sofia Carvalho<sup>1</sup>

3

4 <sup>1</sup>iNOVA4Health, NOVA Medical School, Faculdade de Ciências Médicas, NMS, FCM,

5 Universidade Nova de Lisboa; Lisboa, Portugal.

6 <sup>2</sup>Laboratoire de Chimie de Coordination (LCC)-CNRS, UPS, Toulouse, France.

7

8 \*) Corresponding author: [rune.matthiesen@nms.unl.pt](mailto:rune.matthiesen@nms.unl.pt); <https://orcid.org/0000-0002-6353-2616>

9

10 **Running title:** Ubiquitin proteasome analysis in R

11

12 **Key words:** Ubiquitin proteasome system, ubiquitin enrichment, mass spectrometry, R

13 programming language, ubiquitin branch statistics, proximity drugs, proteolysis targeting chimera

14

15

16 **Abstract**

17 Mass spectrometry data on ubiquitin and ubiquitin-like modifiers are becoming increasingly more

18 accessible and the coverage progressively deepen as methodologies mature. This type of mass

19 spectrometry data is linked to specific data analysis pipelines for ubiquitin. This chapter describes a

20 computational tool to facilitate analysis of mass spectrometry data obtained on ubiquitin enriched

21 samples. For example, the analysis of ubiquitin branch site statistics and functional enrichment

22 analysis against ubiquitin proteasome system protein sets are completed with a few functional calls.

23 We foresee that the proposed computational methodology can aid in proximity drug design by, for

24 example, elucidating the expression of E3 ligases and other factors related to the ubiquitin proteasome

25 system.

26

## 27 **1. Introduction**

28

29 Mass spectrometry (MS) methodologies for studying post translational modifications (PTMs)  
30 continue to develop. We might have elucidated the genetic code, but there are a lot of unexplored  
31 combinations of PTMs that are poorly understood in terms of their effect on protein function. The  
32 number of possible combinations of PTMs on a specific protein grows exponentially with the number  
33 of modifications and modification sites we consider in the analysis. For example, given hundreds of  
34 possible modifications, each with 5-50 possible modification sites, leads to an astronomically large  
35 number of combinations **(1)**. Furthermore, a number of modifications can form heterogeneous branch  
36 structures such as for example glycosylation **(2)**, ubiquitination and ubiquitin-like modifiers.  
37 Ubiquitination and ubiquitin-like modifiers **(3)** mainly function intracellularly. However, the system  
38 is complex and incompletely understood, which constitute one justification for the proposed  
39 computational tool presented in this chapter.

40 The design of proximity drugs is based on understanding the ubiquitin proteasome system (UPS),  
41 which constitute a vast number of protein families with main players such as ubiquitin and ubiquitin-  
42 like modifiers, E1, E2 and E3 ligases, deubiquitinases and the proteasome. For example, the  
43 expression level of E3 ligases can vary depending on the cellular state as demonstrated in the analysis  
44 in this chapter. Given that proximity drugs are dependent on the presence of a specific targeted ligase,  
45 this type of analysis should prove useful in the design process. Furthermore, the majority of the  
46 literature on UPS is not systematized. For example, frequently specific proteins are referred to by  
47 nonstandard names, making systematic computational analysis complex. One example is FAT10  
48 which is often used in the literature but the official gene name is UBD **(4)**. Another example is  
49 ubiquitin which is coded by four genes in the human genome, namely RPS27A, UBB, UBC and  
50 UBA52. These genes contain one, three, nine and one ubiquitin modifying protein, respectively, all  
51 with exactly the same sequence. Therefore, it can be concluded that identification of N terminal di-  
52 glycine can either arise from one of the polyubiquitin coding genes or from N-terminal modification

53 with ubiquitin by ubiquitin. The proposed R package contains annotation of the official gene names  
54 of protein in the ubiquitin system, which facilitates automated analysis of MS data.

55 Ubiquitin and ubiquitin-like modifiers are covalently bound to amino groups of lysine residues or to  
56 the N-terminal of the proteins that they modify. Ubiquitin and ubiquitin-like modifiers also contain  
57 lysine residues that, in turn, can be modified. Specific branch linkages of ubiquitin have been linked  
58 to specific molecular functions. Although, the system is far more complex, having heterogeneous  
59 linkage structures that mix ubiquitin with ubiquitin-like modifiers. Furthermore, ubiquitin can also  
60 undergo other modifications such as acetylation and phosphorylation (3). The proposed R package  
61 contains functions that automatically estimate the overall average level of linkages in each sample  
62 for cross comparisons. This tool was developed based on our past research on tandem ubiquitin  
63 binding entities (TUBEs) combined with MS (5-7). However, it can be applied to basically all the  
64 current ubiquitin enrichment strategies that currently are being combined with MS identification (*see*  
65 **Figure 1**). These technologies TUBEs (8), UbiSite (9), His<sub>6</sub>-tag enrichment (10) and *in vivo*  
66 biotinylation of ubiquitin (11). Each of these technologies can be combined with additional di-glycine  
67 enrichment. The R package “ubiquitin proteasome system in R based on mass spectrometry”  
68 (UPSRM) is demonstrated here with a data set obtained using a recent technology called UBISITE  
69 for large scale enrichment of ubiquitin modified peptides combined with MS identification (12).  
70 However, UPSRM works for all enrichment strategies that end with identification of di-glycine by  
71 MS.

72 The output results of R package UPSRM compare ubiquitin occupancy branch sites on ubiquitin and  
73 ubiquitin like modifiers in different conditions. For example, cell treatment with protease inhibitors.  
74 Overall, these comparisons can potentially highlight the main ubiquitin branches affected by a  
75 specific drug targeting ubiquitin ligases, deubiquitinases or more recently the proteolysis targeting  
76 chimera (PROTAC) drugs targeting protein degradation using the ubiquitin code.

77

78

79 **2. Materials**

80

81 1. R Cran (<https://cran.r-project.org/>)

82 2. R studio (optional - <https://www.rstudio.com/>)

83 3. R packages: ggpubr, plyr, gplots, ComplexHeatmap.

84 4. MaxQuant (**13**) data: “GlyGly (K)Sites.txt” and “proteinGroups.txt” (*see Note 1*). For this tutorial

85 the data set PXD027328 in ProteomeXchange was used. Only the two text files “GlyGly (K)Sites.txt”

86 and “proteinGroups.txt” are necessary to download.

87

88 **3. Methods**

89

90 **3.1 Software setup and installation of R packages**

91 1. Follow the instructions in the provided links to install R Cran.

92 2. Run the below commands in an R session to install UPSRM:

93 `install.packages("ggpubr")`

94 `install.packages("plyr")`

95 `install.packages("gplots")`

96 `install.packages("ComplexHeatmap")`

97 `install.packages("remotes")`

98 `library(remotes)`

99 `remotes::install_github("ruma1974/UPSRM@main")`

100 3. Load all packages needed (*see Note 2*):

101 `library(UPSRM)`

102 `library(ComplexHeatmap)`

103

104 **3.2 Perform UPS protein expression analysis**

105 The example analysis is performed on a recent data set in PXD027328. The data set consists of U2OS  
106 cell lines treated with DMSO, MG132 (proteasome inhibitor), PR619 (DUB inhibitor) or TAK243  
107 (UAE inhibitor). The below steps demonstrate the R commands to run and the expected output.

108

109 1. Provide the directory with the MaxQuant text output files

110 `Dir="[Directory]"`

111 2. The command below automatically loads the two data files into R.

112 `res=loadMaxQuant(Dir)`

113 3. Run the command below to see the protein groups available.

114 `showUPSgroups()`

115 Output (*see Note 3*):

116 "Proteasome", "SUMO", "DUBs", "immunoproteasome", "ProteasomeRegulators"

117 "ProteasomeActivators", "Ub", "E1", "E1sumo", "E2", "E3", "Ubl"

118 4. Annotated genes in a specific UPS group can be retrieved with for example:

119 `showGenes("E3")`

120 5. Quantitative LFQ values for the E3 ligases can be extracted with:

121 `M=getQdata(res,"E3")`

122

### 123 ***3.3 Visualize quantitative UPS data in R***

124 The R package “ComplexHeatmap” provides useful functions for visualizing numeric matrix data.

125

126 1. Rename columns to remove redundant information.

127 `colnames(M)=c("DMSO1", "DMSO2", "DMSO3", "MG1", "MG2", "MG3", "PR1", "PR2", "PR3",`

128 `"TAK1", "TAK2", "TAK3")`

129 2. Define sample conditions

```
130 Fac=c("DMSO", "DMSO", "DMSO", "MG", "MG", "MG", "PR", "PR", "PR", "TAK", "TAK",  
131 "TAK")
```

132 3. Use sample condition information to generate a heatmap annotation.

```
133 ha = ComplexHeatmap::HeatmapAnnotation(Sample = Fac,col = list(Sample = c(DMSO = "red",  
134 MG = "green",PR="brown", TAK = "blue")),na_col = "black")
```

135 4. Plot the expression heatmap of E3 ligases to generate **Figure 2**.

```
136 x11(w=20,h=10) # only need on linux systems
```

```
137 ComplexHeatmap::Heatmap(log2(M+1), col=gplots::redgreen(255), name = "LFQ", top_annotation  
138 =ha,show_row_names =FALSE)
```

139

140 We observe that E3 ligase levels are heavily affected by treatment conditions in the same cell line.

141 This illustrates the need to profile E3 ligases in specific cell conditions for proximity drug design.

142

### 143 ***3.4 Comparative average branch topology for ubiquitin***

144 The below code demonstrates how UPSRM convince functions can be used to explore average  
145 ubiquitin branch topology across sample conditions. It assumes that the MaxQuant data is already  
146 loaded (see section 3.2).

147

148 1. The possible ubiquitin or ubiquitin like modifiers that can be analyzed can be retrieved by:

```
149 UPSRM::showUBIs()
```

150 2. Next we define the comparisons to perform (DMSO is used as reference in below example):

```
151 comp <- list( c("MG", "DMSO"), c("PR", "DMSO"), c("TAK", "DMSO") )
```

```
152 Fac=c("DMSO", "DMSO", "DMSO", "MG", "MG", "MG", "PR", "PR", "PR", "TAK", "TAK",  
153 "TAK")
```

154 3. Boxplot summarizing average ubiquitin branch sites across conditions is written to a pdf file by  
155 the below command (see **Figure 3**):

156 M=getUbQdata(res,"Ub",Fac,comp,"Q\_Ubsites.pdf")

157 4. Boxplot summarizing average SUMO1-ubiquitin branch sites across conditions is written to a pdf  
158 file by the command below (*see Figure 4*):

159 M=getUbQdata(res,"Ub",Fac,comp,"Q\_SUMO1sites.pdf")

160

#### 161 **4. Notes**

162

163 1. The file “GlyGly (K)Sites.txt” is optional. If only “proteinGroups.txt” is provided then expression  
164 analysis of UPS factors is still possible. The columns in “GlyGly (K)Sites.txt” must include  
165 “Intensity.[sample]”, “Gene.names”, “GlyGly..K..Probabilities”. The columns in “proteinGroups.txt”  
166 must contain “LFQ.[sample]” and “Gene.names”. If interest exists the authors will update the R  
167 package with support for other database dependent search engines.

168 2. The other installed R packages are used by UPSRM and will be loaded indirectly when using  
169 UPSRM functions.

170 3. The genes annotated to each UPS group will likely change overtime as more knowledge is  
171 generated on UPS. We consider the current annotations as core annotations but we will be pleased to  
172 receive user feedback to update UPSRM with additional UPS groups.

173

#### 174 **Acknowledgment**

175 R.M. is supported by Fundação para a Ciência e a Tecnologia (CEEC position, 2019–2025  
176 investigator). This article is a result of the projects (iNOVA4Health—UIDB/04462/2020), supported  
177 by Lisboa Portugal Regional Operational Programme (Lisboa2020), under the PORTUGAL 2020  
178 Partnership Agreement, through the European Regional Development Fund (ERDF). This work is  
179 also funded by FEDER funds through the COMPETE 2020 Programme and National Funds through  
180 FCT—Portuguese Foundation for Science and Technology under the projects number PTDC/BTM-  
181 TEC/30087/2017 and PTDC/BTM-TEC/30088/2017. This publication is based upon work from



182 COST Action, CA20113 'PROTEOCURE' supported by COST (European Cooperation in Science  
183 and Technology).

184

185

186 **References**

187 1. Matthiesen R, Azevedo L, Amorim A, et al (2011) Discussion on common data analysis strategies  
188 used in MS-based proteomics. *Proteomics* 11 (4):604-619. doi:10.1002/pmic.201000404

189 2. Schjoldager KT, Narimatsu Y, Joshi HJ, et al (2020) Global view of human protein glycosylation  
190 pathways and functions. *Nature reviews Molecular cell biology* 21 (12):729-749.  
191 doi:10.1038/s41580-020-00294-x

192 3. Swatek KN, Komander D (2016) Ubiquitin modifications. *Cell research* 26 (4):399-422.  
193 doi:10.1038/cr.2016.39

194 4. Bedford L, Lowe J, Dick LR, et al (2011) Ubiquitin-like protein conjugation and the ubiquitin-  
195 proteasome system as drug targets. *Nature reviews Drug discovery* 10 (1):29-46.  
196 doi:10.1038/nrd3321

197 5. Quinet G, Xolalpa W, Reyes-Garau D, et al (2022) Constitutive Activation of p62/Sequestosome-  
198 1-Mediated Proteaphagy Regulates Proteolysis and Impairs Cell Death in Bortezomib-Resistant  
199 Mantle Cell Lymphoma. *Cancers* 14 (4). doi:10.3390/cancers14040923

200 6. Mata-Cantero L, Azkargorta M, Aillet F, et al (2016) New insights into host-parasite ubiquitin  
201 proteome dynamics in *P. falciparum* infected red blood cells using a TUBEs-MS approach. *Journal*  
202 *of proteomics* 139:45-59. doi:10.1016/j.jprot.2016.03.004

203 7. Lopitz-Otsoa F, Rodriguez-Suarez E, Aillet F, et al (2012) Integrative analysis of the ubiquitin  
204 proteome isolated using Tandem Ubiquitin Binding Entities (TUBEs). *Journal of proteomics* 75  
205 (10):2998-3014. doi:10.1016/j.jprot.2011.12.001

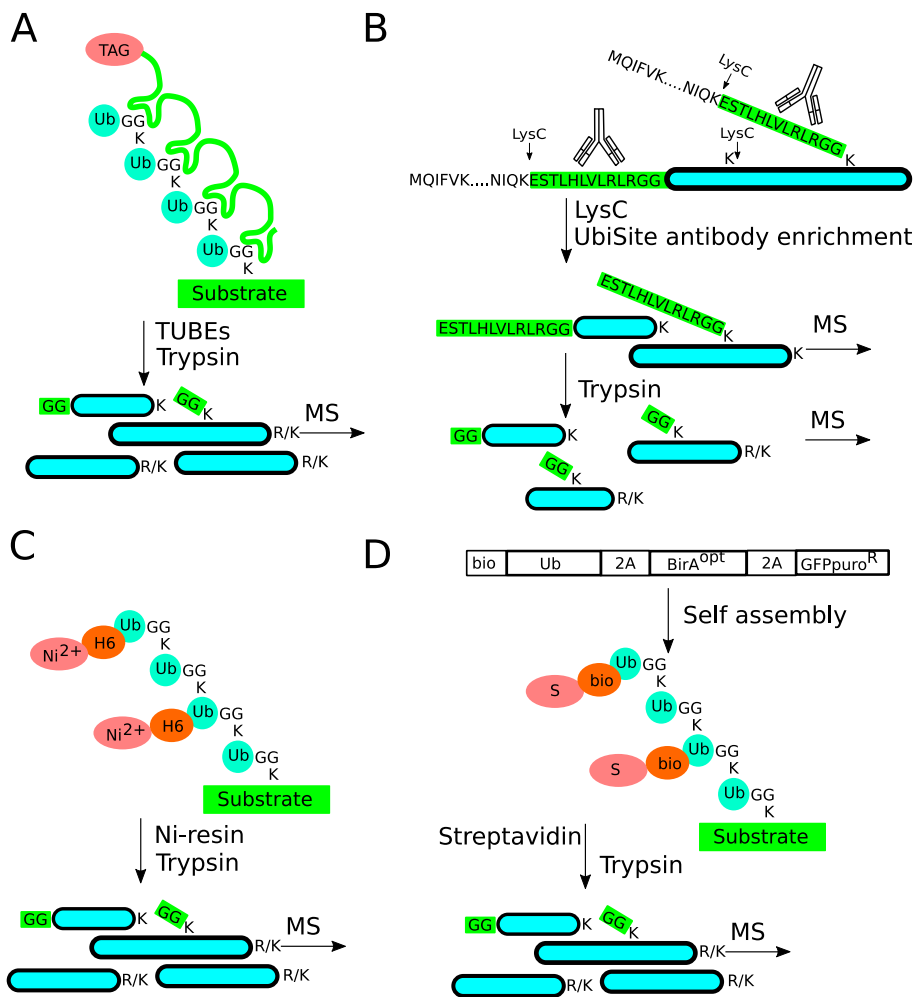
- 206 8. Hjerpe R, Aillet F, Lopitz-Otsoa F, et al (2009) Efficient protection and isolation of ubiquitylated  
207 proteins using tandem ubiquitin-binding entities. *EMBO reports* 10 (11):1250-1258.  
208 doi:10.1038/embor.2009.192
- 209 9. Akimov V, Barrio-Hernandez I, Hansen SVF, et al (2018) UbiSite approach for comprehensive  
210 mapping of lysine and N-terminal ubiquitination sites. *Nature structural & molecular biology* 25  
211 (7):631-640. doi:10.1038/s41594-018-0084-y
- 212 10. Lee KA, Hammerle LP, Andrews PS, et al (2011) Ubiquitin ligase substrate identification through  
213 quantitative proteomics at both the protein and peptide levels. *The Journal of biological chemistry*  
214 286 (48):41530-41538. doi:10.1074/jbc.M111.248856
- 215 11. Pirone L, Xolalpa W, Sigurethsson JO, et al (2017) A comprehensive platform for the analysis of  
216 ubiquitin-like protein modifications using in vivo biotinylation. *Scientific reports* 7:40756.  
217 doi:10.1038/srep40756
- 218 12. Trulsson F, Akimov V, Robu M, et al (2022) Deubiquitinating enzymes and the proteasome  
219 regulate preferential sets of ubiquitin substrates. *Nature communications* 13 (1):2736.  
220 doi:10.1038/s41467-022-30376-7
- 221 13. Tyanova S, Temu T, Cox J (2016) The MaxQuant computational platform for mass spectrometry-  
222 based shotgun proteomics. *Nature protocols* 11 (12):2301-2319. doi:10.1038/nprot.2016.136

223

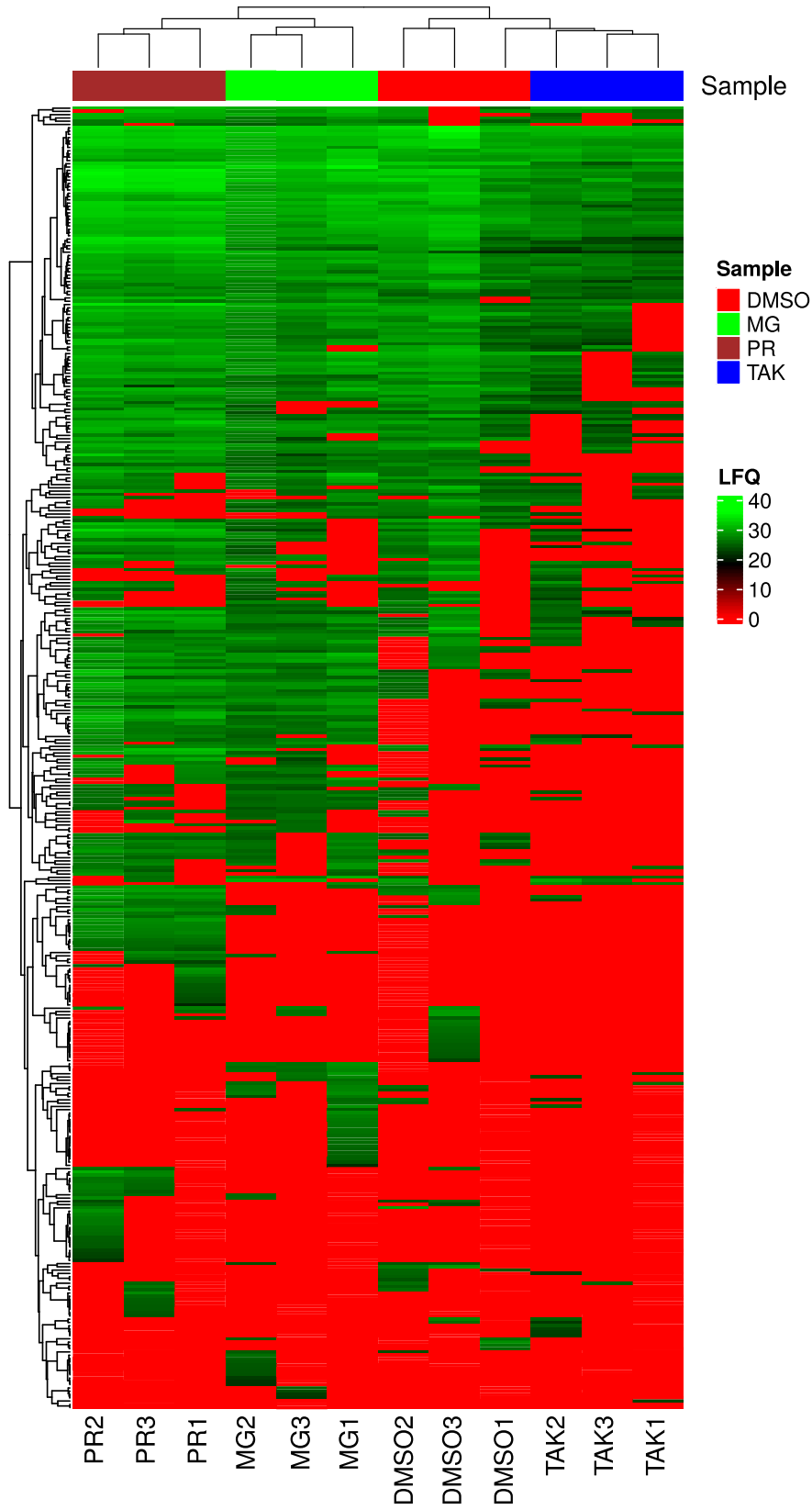
224

225

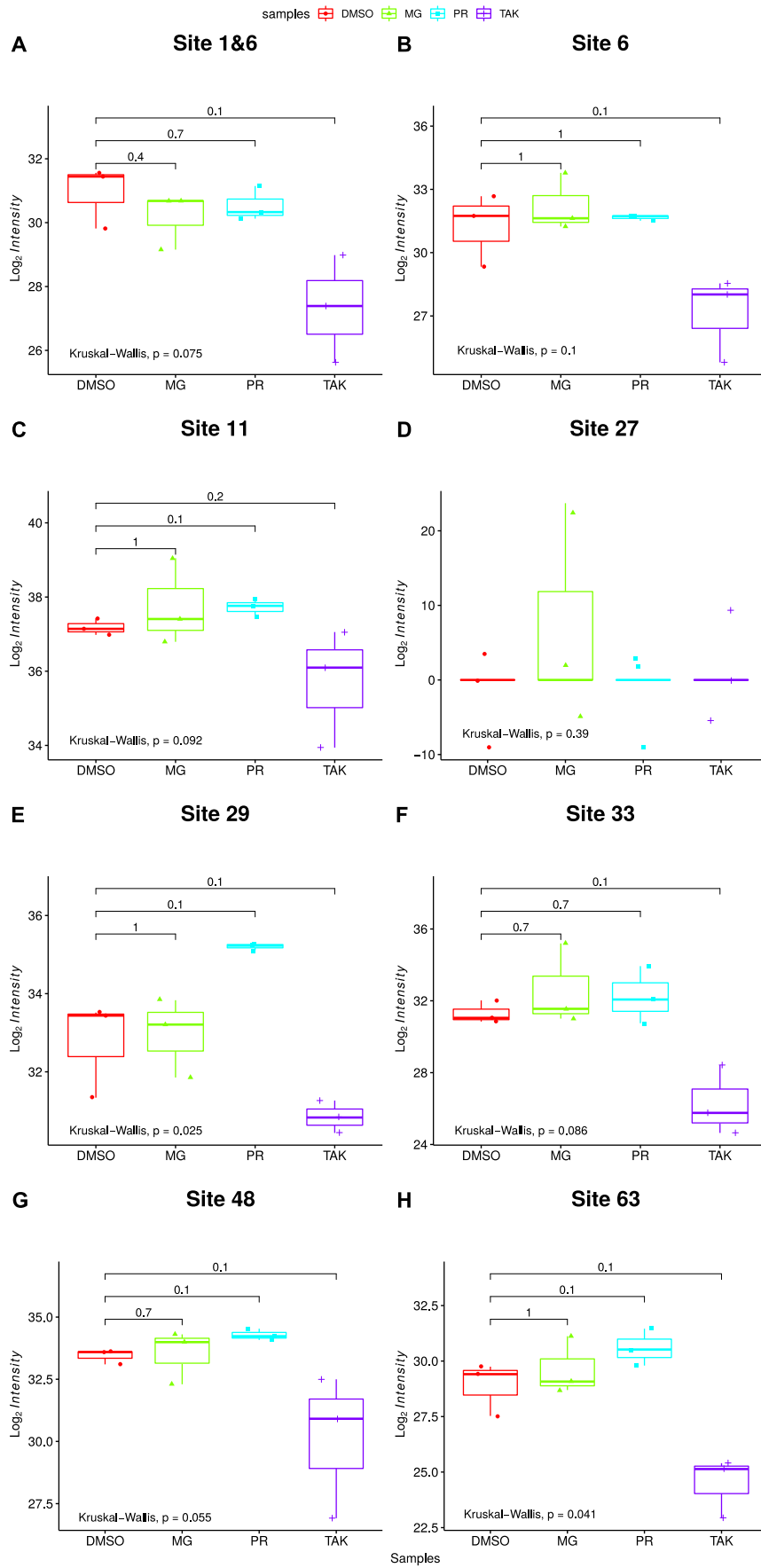
226 **Figures**



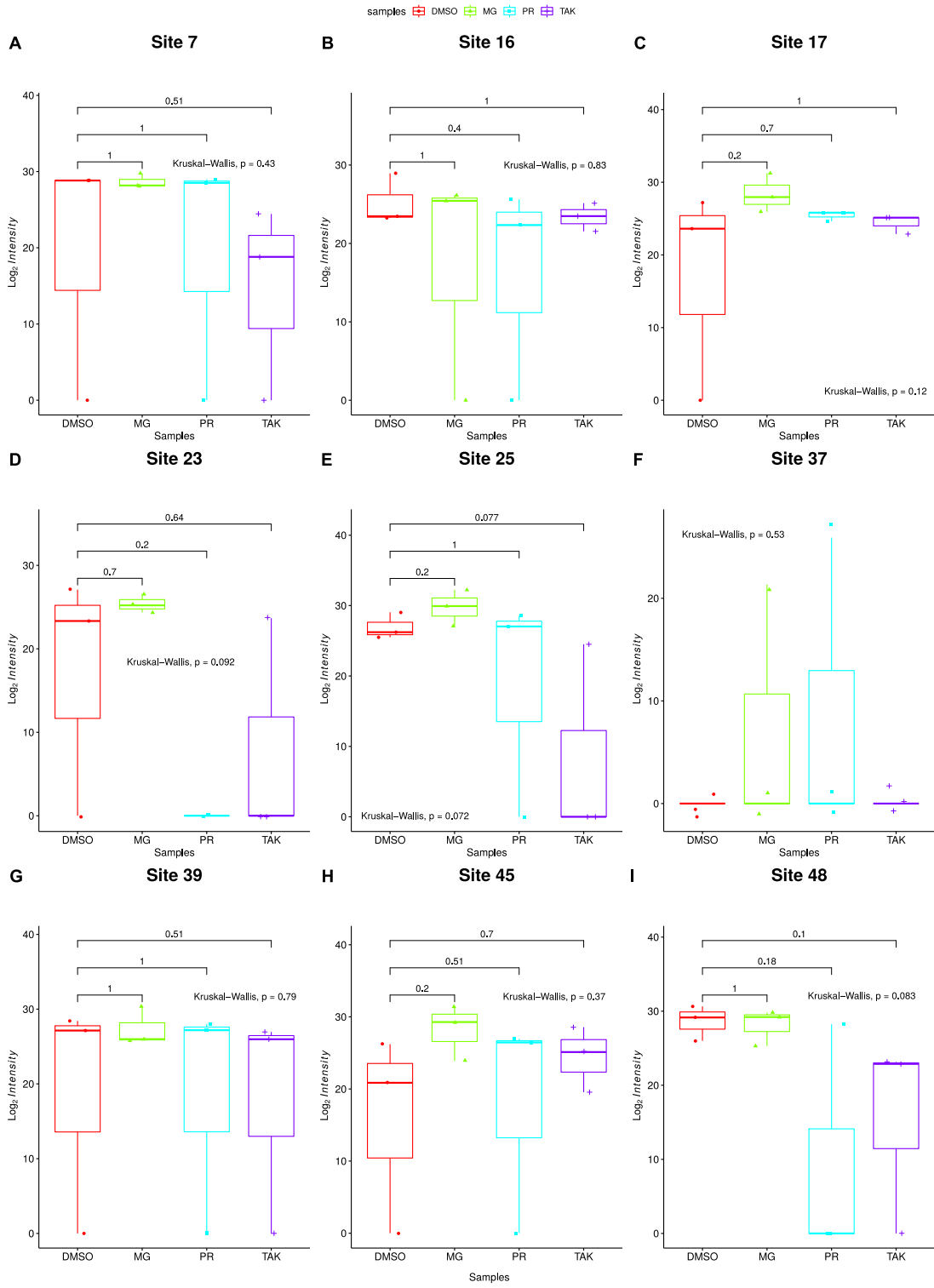
228 **Figure 1. Different enrichment strategies of ubiquitinated proteins and peptides.** A) Strategy  
 229 based on tandem ubiquitin binding entities, B) UbiSite, C) Ni<sup>2+</sup> His6-tag affinity enrichment and D)  
 230 *in vivo* biotinylation-Ub followed by streptavidin enrichment.



232 **Figure 2. Expression of E3 ligases across sample treatment groups.**



**Figure 3. Average ubiquitin branch sites across sample conditions.**



**Figure 4. Average ubiquitin-SUMO1 branch sites across sample conditions.**