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

Rune Matthiesen, Manuel S Rodriguez, Ana Sofia Carvalho

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

Rune Matthiesen[1*], Manuel S. Rodriguez[2] and Ana Sofia Carvalho[1]

[1]iNOVA4Health, NOVA Medical School, Faculdade de Ciências Médicas, NMS, FCM, Universidade Nova de Lisboa; Lisboa, Portugal.

[2]Laboratoire de Chimie de Coordination (LCC)-CNRS, UPS, Toulouse, France.

*) Corresponding author: rune.matthiesen@nms.unl.pt; https://orcid.org/0000-0002-6353-2616

Running title: Ubiquitin proteasome analysis in R

Key words: Ubiquitin proteasome system, ubiquitin enrichment, mass spectrometry, R programming language, ubiquitin branch statistics, proximity drugs, proteolysis targeting chimera

Abstract

Mass spectrometry data on ubiquitin and ubiquitin-like modifiers are becoming increasingly more accessible and the coverage progressively deepen as methodologies mature. This type of mass spectrometry data is linked to specific data analysis pipelines for ubiquitin. This chapter describes a computational tool to facilitate analysis of mass spectrometry data obtained on ubiquitin enriched samples. For example, the analysis of ubiquitin branch site statistics and functional enrichment analysis against ubiquitin proteasome system protein sets are completed with a few functional calls. We foresee that the proposed computational methodology can aid in proximity drug design by, for example, elucidating the expression of E3 ligases and other factors related to the ubiquitin proteasome system.

1. Introduction

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

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

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

Ubiquitin and ubiquitin-like modifiers are covalently bound to amino groups of lysine residues or to the N-terminal of the proteins that they modify. Ubiquitin and ubiquitin-like modifiers also contain lysine residues that, in turn, can be modified. Specific branch linkages of ubiquitin have been linked to specific molecular functions. Although, the system is far more complex, having heterogeneous linkage structures that mix ubiquitin with ubiquitin-like modifiers. Furthermore, ubiquitin can also undergo other modifications such as acetylation and phosphorylation ***(3)***. The proposed R package contains functions that automatically estimate the overall average level of linkages in each sample for cross comparisons. This tool was developed based on our past research on tandem ubiquitin binding entities (TUBEs) combined with MS ***(5-7)***. However, it can be applied to basically all the current ubiquitin enrichment strategies that currently are being combined with MS identification (*see* **Figure 1**). These technologies TUBEs ***(8)***, UbiSite ***(9)***, His_6-tag enrichment ***(10)*** and *in vivo* biotinylation of ubiquitin ***(11)***. Each of these technologies can be combined with additional di-glycine enrichment. The R package "ubiquitin proteasome system in R based on mass spectrometry" (UPSRM) is demonstrated here with a data set obtained using a recent technology called UBISITE for large scale enrichment of ubiquitin modified peptides combined with MS identification ***(12)***. However, UPSRM works for all enrichment strategies that end with identification of di-glycine by MS.

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

2. Materials

1. R Cran (https://cran.r-project.org/)
2. R studio (optional - https://www.rstudio.com/)
3. R packages: ggpubr, plyr, gplots, ComplexHeatmap.
4. MaxQuant ***(13)*** data: “GlyGly (K)Sites.txt” and “proteinGroups.txt” (*see* **Note 1**). For this tutorial the data set PXD027328 in ProteomeXchange was used. Only the two text files “GlyGly (K)Sites.txt” and “proteinGroups.txt” are necessary to download.

3. Methods

3.1 Software setup and installation of R packages

1. Follow the instructions in the provided links to install R Cran.
2. Run the below commands in an R session to install UPSRM:

```
install.packages("ggpubr")
install.packages("plyr")
install.packages("gplots")
install.packages("ComplexHeatmap")
install.packages("remotes")
library(remotes)
remotes::install_github("ruma1974/UPSRM@main")
```

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

```
library(UPSRM)
library(ComplexHeatmap)
```

3.2 Perform UPS protein expression analysis

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

1. Provide the directory with the MaxQuant text output files

```
Dir="[Directory]"
```

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

```
res=loadMaxQuant(Dir)
```

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

```
showUPSgroups()
```

Output (*see* **Note 3**):

"Proteasome", "SUMO", "DUBs", "immunoproteasome", "ProteasomeRegulators" "ProteasomeActivators", "Ub", "E1", "E1sumo", "E2", "E3", "Ubl"

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

```
showGenes("E3")
```

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

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

3.3 Visualize quantitative UPS data in R

The R package "ComplexHeatmap" provides useful functions for visualizing numeric matrix data.

1. Rename columns to remove redundant information.

```
colnames(M)=c("DMSO1", "DMSO2", "DMSO3", "MG1", "MG2", "MG3", "PR1", "PR2", "PR3",
"TAK1", "TAK2", "TAK3")
```

2. Define sample conditions

Fac=c("DMSO", "DMSO", "DMSO", "MG", "MG", "MG", "PR", "PR", "PR", "TAK", "TAK", "TAK")

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

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

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

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

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

We observe that E3 ligase levels are heavily affected by treatment conditions in the same cell line. This illustrates the need to profile E3 ligases in specific cell conditions for proximity drug design.

3.4 Comparative average branch topology for ubiquitin

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

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

UPSRM::showUBls()

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

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

Fac=c("DMSO", "DMSO", "DMSO", "MG", "MG", "MG", "PR", "PR", "PR", "TAK", "TAK", "TAK")

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

M=getUbQdata(res,"Ub",Fac,comp,"Q_Ubsites.pdf")

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

M=getUbQdata(res,"Ub",Fac,comp,"Q_SUMO1sites.pdf")

4. Notes

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

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

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

Acknowledgment

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

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

References

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

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

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

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

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

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

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

8. Hjerpe R, Aillet F, Lopitz-Otsoa F, et al (2009) Efficient protection and isolation of ubiquitylated proteins using tandem ubiquitin-binding entities. EMBO reports 10 (11):1250-1258. doi:10.1038/embor.2009.192

9. Akimov V, Barrio-Hernandez I, Hansen SVF, et al (2018) UbiSite approach for comprehensive mapping of lysine and N-terminal ubiquitination sites. Nature structural & molecular biology 25 (7):631-640. doi:10.1038/s41594-018-0084-y

10. Lee KA, Hammerle LP, Andrews PS, et al (2011) Ubiquitin ligase substrate identification through quantitative proteomics at both the protein and peptide levels. The Journal of biological chemistry 286 (48):41530-41538. doi:10.1074/jbc.M111.248856

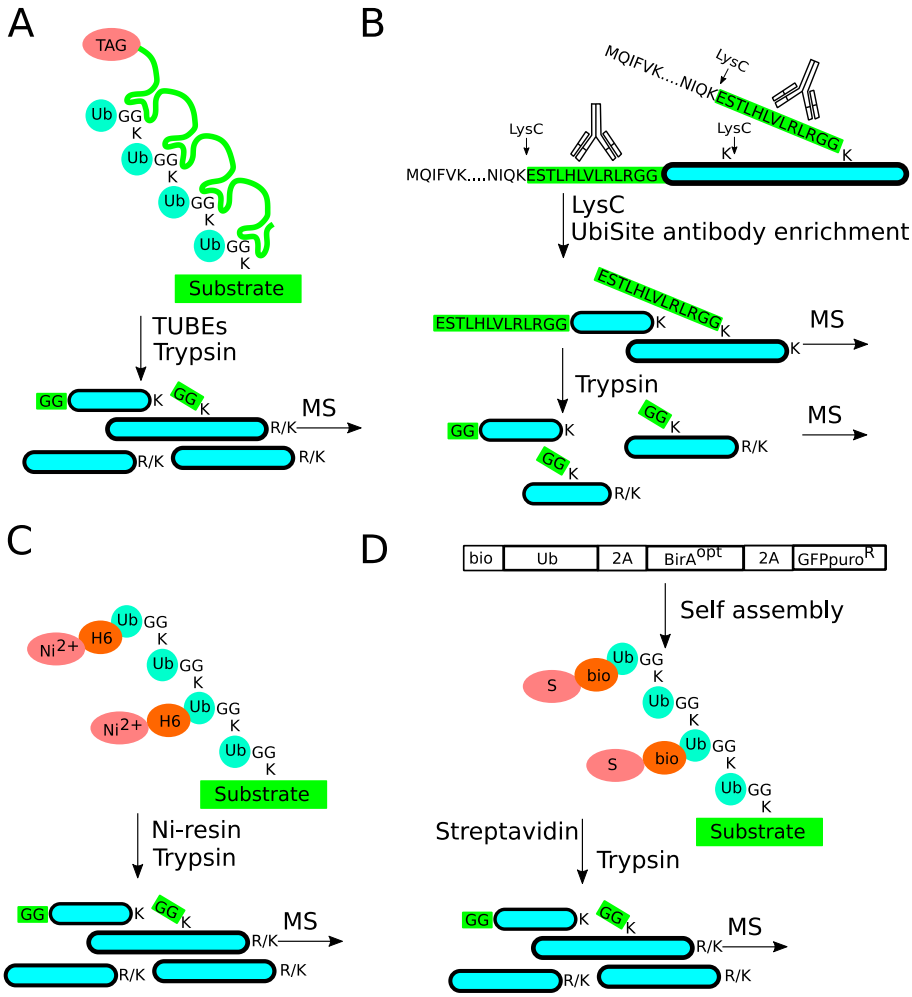
11. Pirone L, Xolalpa W, Sigurethsson JO, et al (2017) A comprehensive platform for the analysis of ubiquitin-like protein modifications using in vivo biotinylation. Scientific reports 7:40756. doi:10.1038/srep40756

12. Trulsson F, Akimov V, Robu M, et al (2022) Deubiquitinating enzymes and the proteasome regulate preferential sets of ubiquitin substrates. Nature communications 13 (1):2736. doi:10.1038/s41467-022-30376-7

13. Tyanova S, Temu T, Cox J (2016) The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nature protocols 11 (12):2301-2319. doi:10.1038/nprot.2016.136

Figures

227



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

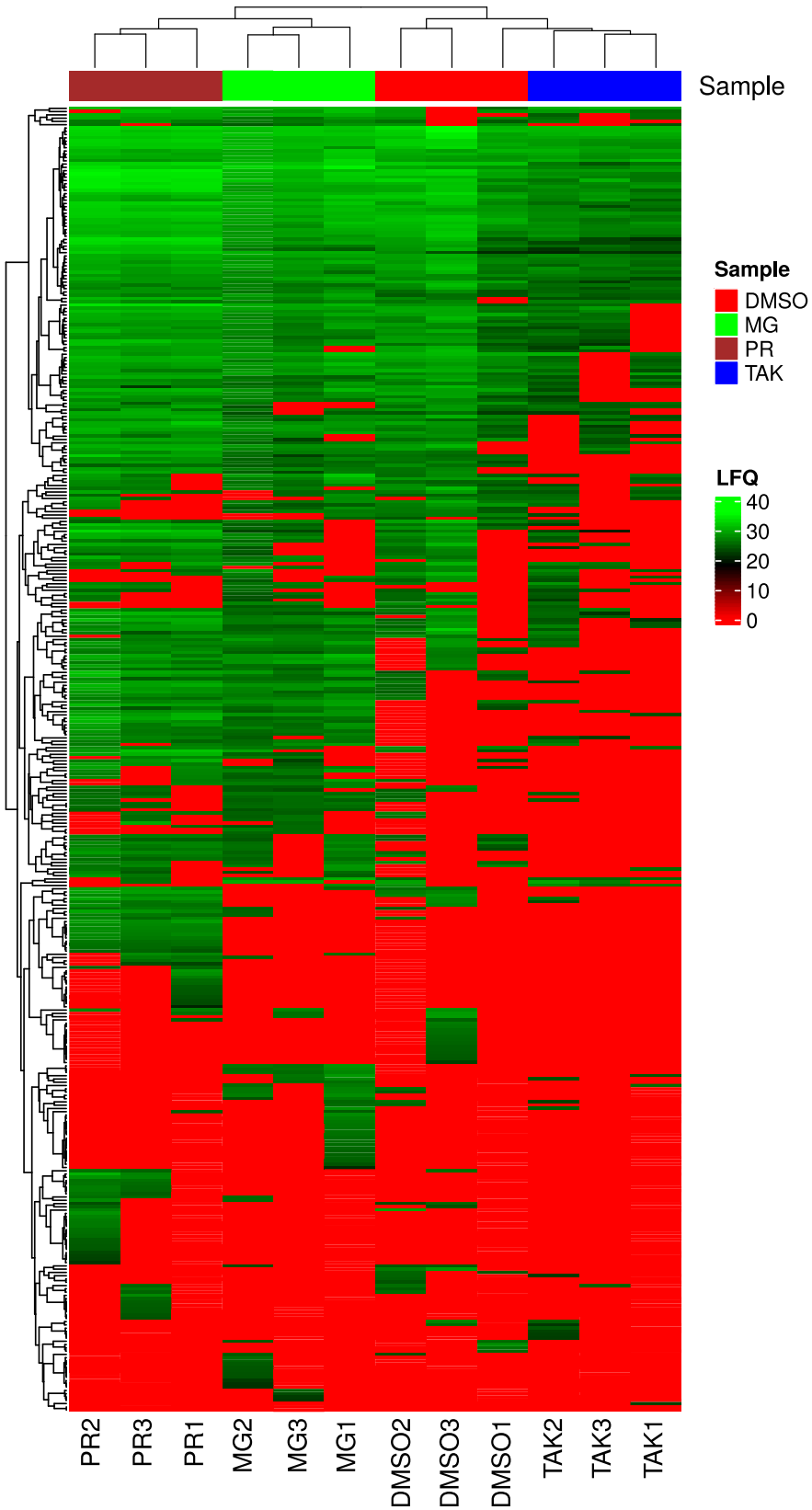


Figure 2. Expression of E3 ligases across sample treatment groups.

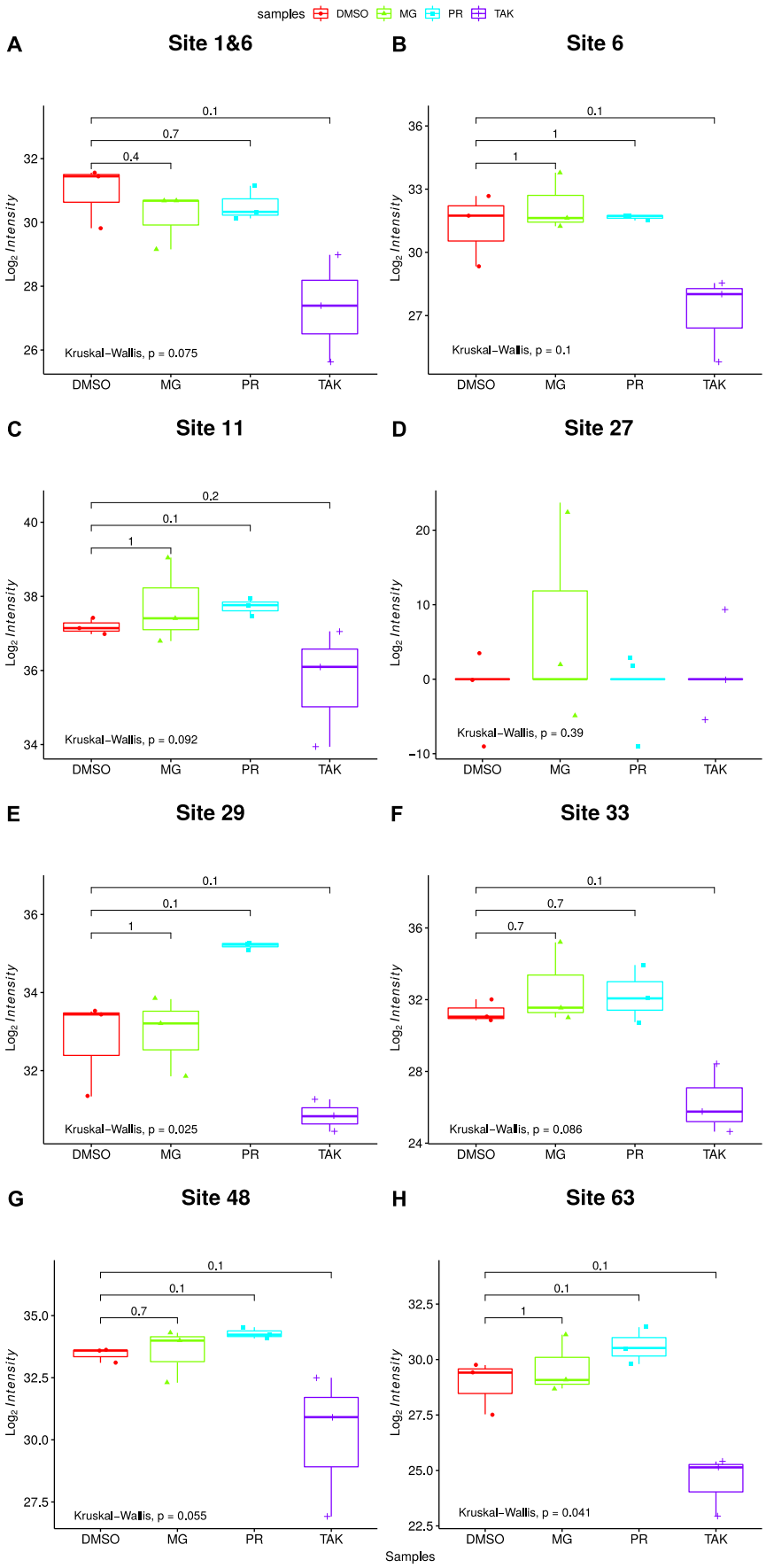


Figure 3. Average ubiquitin branch sites across sample conditions.

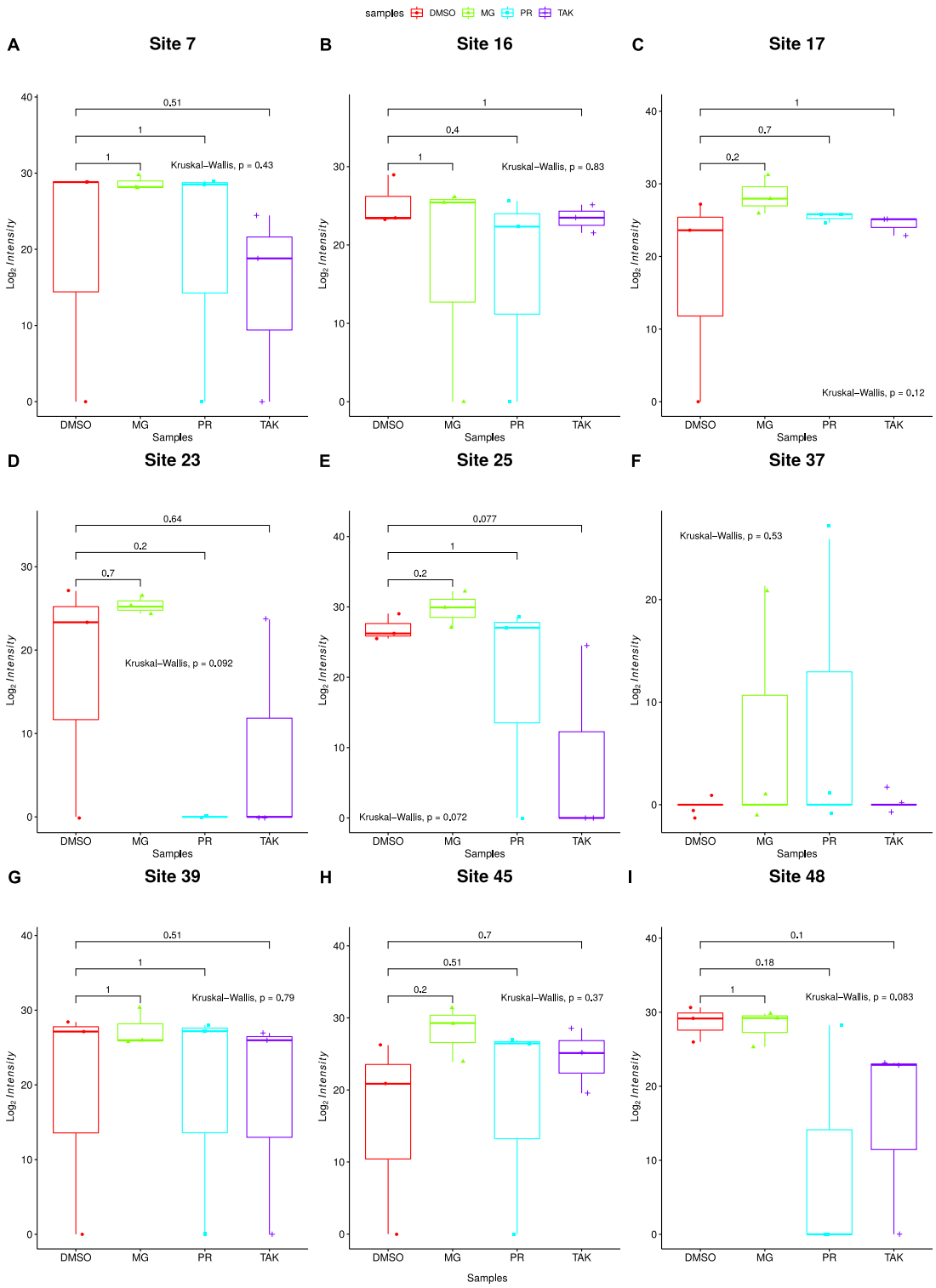


Figure 4. Average ubiquitin-SUMO1 branch sites across sample conditions.