iSUMO - integrative prediction of functionally relevant SUMOylation events

Xiaotong Yao[1,2], Shuvadeep Maity[1], Shashank Gandhi[1, 3], Marcin Imielenski[4], Christine Vogel[1]*

1 Center for Genomics and Systems Biology, New York University, New York, USA

2 Tri-Institutional Program in Computational Biology and Medicine, New York, USA

3 Division of Biology and Biological Engineering, California Institute of Technology, California, USA

4 New York Genome Center, New York, USA

*** Corresponding author:** cvogel@nyu.edu

Abstract

Post-translational modifications by the Small Ubiquitin-like Modifier (SUMO) are essential for diverse cellular functions. Large-scale experiment and sequence-based predictions have identified thousands of SUMOylated proteins. However, the overlap between the datasets is small, suggesting many false positives with low functional relevance. Therefore, we integrated ~800 sequence features and protein characteristics such as cellular function and protein-protein interactions in a machine learning approach to score likely functional SUMOylation events (iSUMO). iSUMO is trained on a total of 24 large-scale datasets, and it predicts 2,291 and 706 SUMO targets in human and yeast, respectively. These estimates are five times higher than what existing sequence-based tools predict at the same 5% false positive rate. Protein-protein and protein-nucleic acid interactions are highly predictive of protein SUMOylation, supporting a role of the modification in protein complex formation. We note the marked prevalence of SUMOylation amongst RNA-binding proteins. We validdate iSUMO predictions by experimental or other evidence. iSUMO therefore represents a comprehensive tool to identify high-confidence, functional SUMOylation events for human and yeast.

Keywords

SUMOylation; machine learning; computational biology; prediction; post-translational modification; bioinformatics

Introduction

The covalent attachment of Small Ubiquitin-like Modifier (SUMO) is, based on its common occurrence and wide array of functions in eukaryotic cells, one of the most important post-translational modifications. SUMOylation has been studied from numerous perspectives since its discovery in 1997 [1]. It is widely conserved across eukaryotes [2-4], and in many cases essential for the organismal viability [5]. SUMOylation resembles ubiquitination in terms of structure, enzymatic pathway, and it has a broad functional spectrum, ranging from chromatin organization [6], DNA damage repair [7], regulation of transcription [8], ribosome biogenesis [9-10], messenger RNA (mRNA) processing [11-12], nucleus-cytoplasm transport [13], to protein localization [14], proteolysis (where it cross-talks with ubiquitination)[15], stress response [16]and other functions [17].

Several computational approaches exist that predict SUMOylation based on the conserved amino acid sequence motif Ψ-K-X-D/E, where Ψ is a hydrophobic residue, K is the lysine being modified, X is any amino acid, and D/E is an acidic residue [18-21]. However, these sequence-based predictions have many false positives and false negatives: when comparing them to experimental data, the intersection is only small. For example, half of the human proteins contain the above SUMOylation motif in their sequence, but the modification is verified for only a small fraction. In addition, recent experimental data suggests that SUMOylation may also act on motifs other than the one described above [22], highlighting the need for methods that move beyond use of sequence alone.

Several experimental methods have been developed to identify SUMO-targets. For example, immunoprecipitation using antibodies against SUMOylated proteins reveals SUMO conjugation with high confidence [23], but the assay only works with a small number of proteins at a time. In comparison, mass spectrometry based methods sample a large fraction of the proteome and have by now identified thousands of SUMO targets in yeast and human (**Figure 1**). However, it is often unclear what the false-positive and false-negative identification rates of these approaches are. For example, a recent and in-depth screen of human SUMOylation targets using advanced technology identified only 1,606 proteins [22], and overlap between this and other studies is small (**Figure 1**). In comparison, when using prediction tools that are sole based on sequence features, many more targets have been identified, e.g. 9,173 of 16,849 human proteins (54%) are predicted to have a SUMOylation motif in their sequence.

Overall, these findings suggest that interpreting large scale identification results remains challenging, as it is likely enriched in many false-positive identifications. To achieve a unified interpretation, we developed iSUMO, a random forest-based predictive model, exploiting the putative SUMOylated proteins discovered in large-scale experiments in their relationship to various protein annotations. iSUMO's results not only provide a set of SUMOylation events that are likely functional, they also highlight properties common to SUMOylated proteins. In total, these results suggest that ~20% and ~10% of human and yeast proteins are SUMOylated.

Materials and methods

Training data sets of experimentally observed SUMOylated proteins

We assembled the results from 18 and 6 large-scale, experimental studies in human [22, 24-32] and yeast [33-38], respectively, which mapped SUMOylated proteins using mass spectrometry. **Figure 1** summarizes the composition of consensus or unique findings among the studies, and detailed description of the data can be found in **Supplementary Tables S1a and b**. We obtained a total of 2,216 and 555 distinct, well-annotated human and yeast proteins, respectively, to construct a highly quality training dataset.

To integrate the different datasets, we downloaded both reference proteomes from UniProt Reference Proteome database (Homo sapiens, UP000005640; Saccharomyces cerevisiae S288c, UP000002311), and only retained the subset that has been reviewed by Swiss-Prot database. We then filtered both reference proteomes to contain only one protein per gene, restricting unique identity across the combination of Ensembl Gene ID, National Center for Biotechnology Information (NCBI) Gene ID, and UniProt Knowledge Base (UniProtKB)[39] Accession ID. To avoid introducing biases with dubious function annotations for human genes, we created a high-confidence dataset for the human proteome (>11,000 proteins, **Suppl. Table S6a**). We also present iSUMO predictions with the extended dataset (>16,000 proteins, **Suppl. Table S6b**). For yeast, we used ~6,700 proteins. Within these reference proteomes, we labeled the proteins as "SUMOylated" if they are discovered as SUMOylated in any of the 24 studies described above.

Sequence-based prediction of protein SUMOylation

For genome-wide prediction of SUMOylation based on protein sequence, we used the Group-based Prediction System-SUMO (GPS-SUMO) _[19, 40] to predict consensus and non-consensus SUMOylation motifs and SUMO-interaction motifs (SIM) with the threshold set to 'high'.

Function enrichment analysis

GO enrichment analysis was carried out using gProfileR package [41-42]. The query lists were defined as the union of all lists of observed SUMOylated proteins and the background lists are the respective reference proteomes. We set additional parameters to: 1) excluding the GO associations based only on electronic annotation; 2) using adjusted p-values that correct for multiple hypotheses testing; 3) non-ordered query list, thus adopting hypergeometric test. Next we only retained the significant enriched GO terms (false discovery rate<0.05), whose total size are less than 2,000 for human and size within 5 to 1,000 for yeast. Finally, we strictly excluded any term containing "sumo" in its name, case-insensitively, from subsequent modelling steps.

Attribute selection for predictive modeling

The attributes used for the modeling were taken from multiple sources. The full list of attributes used for predictive modeling is in **Supplementary Table S2**. The attributes include, for example, numbers of occurrences of sequence-based predictions of SUMOylation or SUMO-interaction motifs. These attributes are purely based on sequence predictions and were also used as baseline to compare our own iSUMO predictions. iSUMO also includes attributes derived from associations with Gene Ontology (GO) terms (**Table 1**). Besides, we also incorporated protein-protein interaction data from the STRING database using only the "binding" type activity. We simplified its rich structure into the degree of proteins within this network [43-45]. Since the taxonomy id 559292 is absent in STRING, we used the equivalent data of 4932 for yeast. Moreover, using the annotations from CORUM database of mammalian protein complex and Costanzo et al., 2016 for yeast [46], we counted the number of protein complexes to which each protein belongs and the average size of these complexes (size defined as the number of distinct protein species per complex). Finally, we included other PTM sites of a protein based on the annotations in PhosphoSitePlus for human and dbPTM for yeast [47-48], both experimentally supported only, summarizing the number of amino acids that are phosphorylated, acetylated, methylated, or ubiquitylated.

Fitting, evaluating, and applying random forest model

iSUMO models SUMOylation state of proteins as a binary classification task based on training data, which comprises categorical and numerical attributes described above. We used the H2O open source machine learning platform with its R interface. The core classification method is random forest. We split the full dataset into 65% training, 20% validation, and 15% testing. We attempted different settings for the training,

such as hard setting number of trees versus dynamic termination rule, and restricting different max numbers of tree depth. The model selection is based on the performances with validation set and 10-fold cross validation, considering metrics including area under the receiver operating curve (AUROC), and maximum F1 and F2 values. As training labels were strongly biased towards non-SUMOylated proteins, we also tried balancing the training examples by over sampling of the minority label. Plus, H2O measures relative importance of variables in random forest models by calculating how often a variable is selected at a split of a tree and the amount of improvement in squared error over all trees (http://www.h2o.ai).

Results

Many large-scale studies of the SUMOylated proteome exist, but their overlap is small

To obtain a comprehensive training dataset of true-positive SUMOylated proteins, we integrated 18 and six large-scale, experimental studies for human and yeast, respectively. These studies identified individually between ten to >1,600 SUMOylated human proteins, and 13 to 271 SUMOylated yeast proteins. When mapped to the selected reference proteome, we retained a total of 2,270 and 527 SUMO targets for the two species, respectively. About half (45%) of these proteins were identified by two or more studies. In yeast, only one third (30%) of the 527 total SUMOylated proteins were identified by more than one study (**Figure 1**). The lack of overlap between individual studies indicates that the individual studies many have false-positive identifications which are not biologically functional.

SUMOylated human and yeast proteins bind and process nucleic acids

Table 1 shows representative, highly significant Gene Ontology (GO) enrichments for both the human and yeast SUMOylated proteins at a (false discovery rate <e-40). The complete results are in the **Suppl. Table S2**. Functions related to DNA processing and metabolism were enriched in both human and yeast SUMOylated proteins, including chromatin organization, DNA damage response, mitosis cell cycle, and transcription; cellular compartments including nucleus, nucleoplasm, nuclear body, and nucleolus. These functions are consistent with our current understanding of SUMOylation's important role in gene expression regulation [49-50]. More than two thirds of the 174 human involved in viral gene expression are SUMOylated, which is consistent with the process in which viruses take advantage of host cell SUMOylation to optimize viral gene expression [51].

SUMOylated proteins were enriched in functions related to RNA processing, often more significantly than those concerning DNA processing (**Table 1**). The effect was stronger in human than in yeast. The enriched functions included RNA processing, translation elongation and termination, cellular compartments like nucleolus, ribonucleoprotein complex, spliceosomal complex, cytoplasmic ribosomes, and molecular functions like RNA binding – all related to various steps of RNA synthesis and processing.

Large protein-RNA complexes are extensively SUMOylated

One distinct pattern of SUMOylation was the simultaneous modification of several subunits, probably assisting with organized recruitment of complex components, for instance in proteins in DNA double strand break repair by homologous recombination [52]. Thus, we first tested if SUMOylated proteins are more often part of any stable protein complex than non-SUMOylated complexes. Indeed, this is the case in both human and yeast (unadjusted hypergeometric test p-value at 1E-132 and 9E-58, respectively).

Next, we tested as if the complexes were enriched in specific functions (**Table 2**). Since most complexes have few subunits, we limited this analysis to the complexes with size larger than 20. The top SUMO-enriched complexes function in, for example, the pre-rRNA complex, involved in ribosome biogenesis, splicing, translation, protein degradation (proteasome), and centromere chromatin complex. We also observed cases whose SUMOylation has so far received little attention. For example, three-quarters (86/110) of the human proteins annotated as part of the signal recognition particle and its co-translational protein targeting to membranes are SUMOylated – which has, to the best of our knowledge, not yet been reported in literature.

For both organisms, we found the sizes of complexes tend to be significantly larger when the complex contains both RNA binding and SUMOylated proteins (**Figure 2A**). We then plot the number of SUMOylated subunits against the number of RNA binding proteins per complex, with the size of points corresponding to the size of the complexes (**Figure 2B**). While there are some outliers, there is a clear positive correlation between number of RNA binding and SUMOylated proteins among the largest complexes for both organisms.

One of the outlying groups is high in SUMOylation but has few RNA binding proteins, represented by human PA700-20S-PA28 complex, proteasome, and yeast RSC (chromatin structure remodeling) complex (**Figure 2B**). They both consist of multiple different subunits which may legitimize the extensive SUMOylation observed across studies, while neither has RNA binding function. Another outlier contains

many RNA-binding proteins but little SUMOylation. An example is the human 55S mitochondrial ribosome, contrasted by the 60S cytoplasmic ribosome, which is almost entirely SUMOylated. This may be the result of non-random subcellular localization of SUMOylation enzymes, or systematic bias in the high-throughput capture assays.

Integrating diverse protein annotations substantially boosts prediction of SUMOylation (iSUMO)

The wide range of characteristics of SUMOylated proteins (see above) along with just a handful of SUMO E3 ligases highlight the need for tools that include more than sequence information to predict the SUMOylation state of proteins. To this end, we developed iSUMO, which employs a random forest model to make use of the intricate information in various protein annotations, ranging from basic molecular weight, SUMO-specific sequence motif, to biological functions, and connectivity in the protein-protein interaction network. This group of algorithms performs well with binary attributes which comprise much of the training set. In total, iSUMO integrates 638 and 293 attributes for human and yeast proteins, respectively (**Suppl. Table 1**). To show the benefit of incorporating protein annotations, we compared iSUMO to the same training method but with only sequence motif predictions from the well-established GPS-SUMO tool [53]. We split the data into 65% training, 20 % validation, and 15% testing subsets, and recorded training, 10-fold cross validation, validation performance metrics for model selection. Training methods varied with respect to termination condition, fixed number of trees, the maximum depth of trees, and whether we applied balancing of the training labels. We then chose the model with the highest area under the receiver-operator-characteristic (ROC) curve, and high F1, F2 values in validation and cross validation. In case of a tie, we chose the model with the shortest training time. To avoid overfitting, we also examined the model's performance on test set. We accepted the model choice if it was close to training and validation metrics.

Overall, iSUMO showed a substantial improvement over predictions based on sequence alone (**Figure 3A**). For example, iSUMO's average area underneath the ROC is 0.88 and 0.84 for human and yeast, respectively, compared to the sequence-based areas of 0.58 and 0.58. Further, at a 5% false positive rate (FPR), iSUMO's true positive rate is about five-fold higher than that of the sequence-based predictions in human with 53% vs. 9%, respectively. This 5% FPR corresponds to an iSUMO score cutoff of 0.19 and 0.18 for human and yeast, respectively, which in turn predict 2,291 and 706 SUMOylated human and yeast proteins. The complete predictions, including for the extended human dataset, are available in the **Suppl. Table S6**.

Predictive features of SUMOylation in yeast and human

Next, we analyzed the iSUMO models for attributes that are highly predictive (**Figure 3B**). From the random forest model training, we can estimate the relative importance of individual features by the frequency of being chosen at the split and the reduction of squared error it introduces. First, a protein's connectivity in the PPI network ("ppiDegree") ranks top in human model and fourth in yeast. Similarly, the complex size ("avgCompSz") and number of complex memberships of a protein ("nComp") rank fourth and fifth in the yeast model, respectively, and "avgCorumSz" and "nCorum" rank seventh and eleventh in the human model. These observations are consistent with the putative role of SUMOylation in supporting non-covalent protein-protein interactions and complex assembly and is likely common across eukaryotic organisms.

Second, the most important GO terms in both models are centered around DNA or RNA processing pathways, molecular function, or cellular localization (**Figure 3B**). As before, RNA binding and related functions are much more important in the human than in the yeast model. This finding suggests an increasing role of SUMOylation from transcriptional to post-transcriptional regulation in higher eukaryotes.

Third, unique to the human model, the numbers of ubiquitylation, acetylation, and phosphorylation sites in a protein are three highly predictive features, while they are much less important in the yeast model. This result is in line with numerous pieces of evidence on the extensive cross-talk between SUMOylation and other PTMs [22, 54]. However, the pattern is less prevalent in yeast, perhaps because more data is available for human than yeast and the extent of post-translational modifications increases in higher organisms.

Fourth and last, basic molecular characteristics like length and weight as well as sequence motifs (consensus, non-consensus, and interaction) are far more dominant features in the yeast than in the human model (**Figure 3B**). This result gives a hint on how SUMOylation achieves it substrate specificity with a limited number of E3 ligases compared to, for example, kinases and ubiquitination enzymes. It is tempting to speculate that in the simpler yeast, SUMOylation only requires sequence specificity, but in the more complex mammalian system, SUMOylation requires cross-talk with other proteins and other protein modifications. Again, the trend might also be due to human data being complemented by a large number of additional information.

Validation and application of iSUMO predictions

We used iSUMO with the high-quality reference proteomes that we created (**Method**) to provide a quantitative prediction of SUMOylation events in human and yeast that are highly enriched in functional SUMOylation events (**Suppl. Table 6a-c**). For high-confidence predictions, we restricted the analysis to

well-annotated human genes and extracted predictions with the false positive rate (FPR) smaller than 0.05. The FPR was determined based on the model's performance with a e part of data unseen during training. At the 5% FPR threshold, iSUMO predicts 2,291 and 706 in human and yeast, respectively (**Suppl. Tables S6a, S6c**). Interestingly, the number is very similar to the combined number of experimentally determined SUMOylation events in the two organisms (2,270 and 527, respectively, **Suppl. Table S7**). Given a total size of the proteomes used for the study, ~19% and ~11% of the human and yeast proteome appears to be SUMOylated, respectively. This number is robust to the use of an extended dataset in human (**Suppl. Table 6b**).

At 5% FPR, iSUMO predicts SUMOylation of 467 human and 303 yeast proteins that were never observed in any of the 24 experimental datasets used for this study (**Suppl. Table S7**). To further validate these predictions, we searched other databases and literature for evidence of these SUMOylation of the top best predictions in human (**Table 3**). All ten proteins show some evidence for the prediction to be true. The yeast homolog of the ribosomal protein RRP1 is known to be SUMOylated [55], and for four proteins (ISY1, HIST2H2AB, TUBA3C, NSD1), other large-scale databases report SUMOylation events (Table 3). For other proteins, we found indirect evidence. For example, three of the proteins (SF3A3, ISY1, SNRPA) are part of RNA splicing, and the spliceosomal complex is known to depend on SUMOylation [56]. Other proteins are part of chromatin reorganization and segregation during cell division (CHAF1B, ANAPC5, TUBA3C), which also involves many SUMOylation events [57]. SUMOylation also plays a role in rRNA maturation and assembly of the 60S subunit [9], supporting the predicted SUMOylation of UTP15. Finally, plenty of literature supports SUMOylation of histones and histone methyltransferases [58-59], supporting the predicted modification of HIST2H2AB and NSD1. In sum, we argue that even these new predictions which had a high iSUMO score but were absent from the experimental training data, are highly enriched in true positives, i.e. functional modification events.

Discussion

A major question arising from these comparative studies concerns the total number of SUMOylated proteins. Here, we present iSUMO, an integrated SUMOylation prediction framework that outperforms methods based on sequence features alone (**Figure 3**). We apply iSUMO to both human and yeast proteins, predicting 2,291(19%) and 706 (11%) of SUMOylated proteins, respectively. Encouragingly, this number is close to a

recent estimate by Hendriks et al. which, combining several experimental studies, predicted a total set of ~3,600 (15%) of SUMOylated human proteins [60].

However, a remaining challenge in these analyses is the small overlap between the individual large-scale studies, suggesting that each experiment produces many false positive identifications (**Figure 1**). Alternatively, our current methods might be unable to capture the complete SUMO space and only detect small subsets. Therefore, the next goal should be to identify the true positive SUMOylation events amongst the many detected ones. iSUMO offers a solution to this problem by integrating sequence and protein functional features to learn computationally the properties of SUMOylated proteins. It outperforms prediction tools that are solely based on sequence features and produces a list of high-confidence predictions. In total, 1,824 (15%) and 403 (6%) proteins in human and yeast, respectively, are positive iSUMO predictions *and* occur in the experimental datasets (**Suppl. Table S7**) – these proteins are most likely to be truly functional.

In addition, our study also highlighted protein characteristics strongly connected to SUMOylation. For example, SUMOylated proteins often bind nucleic acids (e.g. RNA or DNA) and are part of large complexes (**Table 2**)[22]. Specifically, there is strong correlation between SUMOylation, the size of a complex, and the number of subunits that are RNA-binding (**Figure 2**). Overall, two fifths (654 of 1,536) of the human RNA-binding proteins are SUMOylated, while this is the case for less than 10% of the total human proteome -- suggesting that SUMOylation might play a role specifically in mediating protein-RNA binding, beyond its known function as a facilitator of protein-protein interactions. Whether SUMOylation modifies the structure of the RNA-binding protein, or affects its surface charge to enable the interaction with the nucleic acid remains subject to future studies.

It is tempting to speculate on the reasons for the prevalence for SUMOylation amongst RNA-binding proteins. Perhaps, with the expansion of RNA-based regulatory pathways in mammals compared to yeast, the well-established, extensive role of SUMOylation of DNA-binding proteins was simply transferred. Alternatively, SUMOylation might be essential for the correct assembly of large complexes, which are very often involved in RNA-related processes, and the prevalence of SUMOylation for RNA-binding proteins might be a side-effect of its role in complexes. A third intriguing hypothesis arises from two observations: SUMO is one of the most soluble of all known proteins [61] and RNA-binding proteins are major components of RNA-protein granules whose aggregation forms the molecular bases of many neurodegenerative disorders

[62]. Therefore, SUMOylation may act to prevent such aggregation in these densely packed cellular structures – a hypothesis supported by some experimental work [63].

Acknowledgements

C.V. acknowledges funding by the NIH (Ro1 GM113237), the NSF EAGER grant, the DOD (Hypothesis Testing Award PC121532), and the NYU Whitehead Foundation. X.Y. acknowledges funding by the Biology Department at New York University (Master's Research Grant).

Supporting Materials

Supplementary Tables in Excel file.

S1a. Human proteome-wide SUMOylation data

S1b. Yeast proteome-wide SUMOylation data

S2a. Enriched GO categories, human

S2b. Enriched GO categories, yeast

S3. Protein complexes, human

S4. Features used in modeling

S5. Model selection, human and yeast

S6a. Final prediction, human - high-confidence dataset

S6b. Final prediction, human - extended dataset

S6c. Final prediction, yeast

S7. Numbers

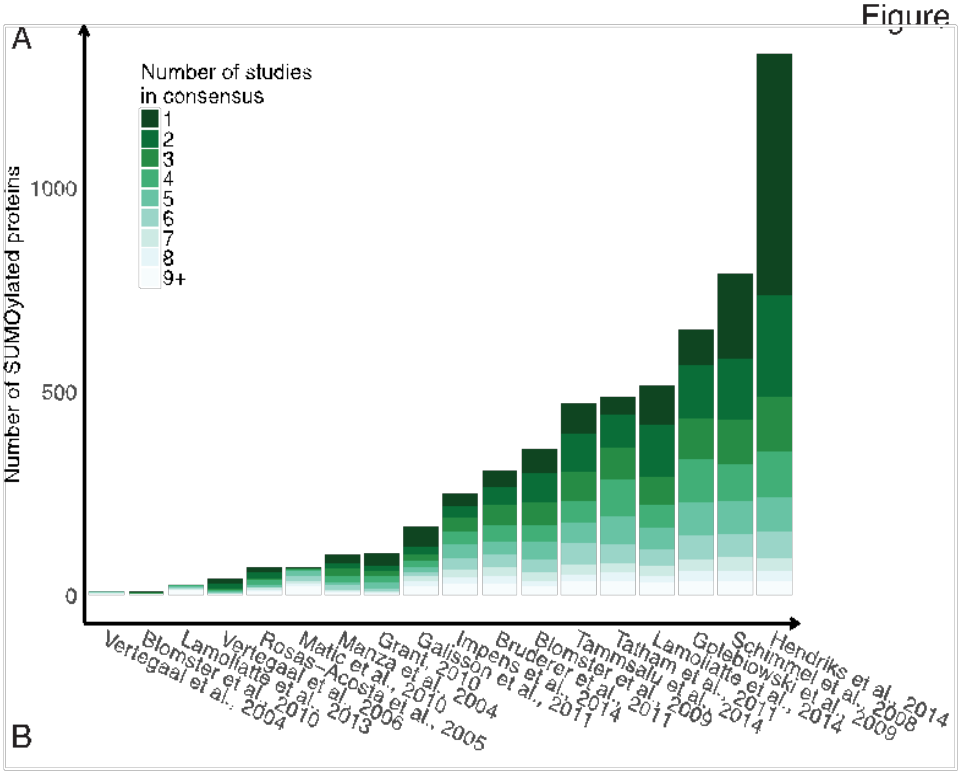
Figures and Tables

(high resolution figures are submitted separately)

Figure 1. Large-scale experimental datasets for SUMOylated proteins.

We assembled 18 and six datasets with mass spectrometry-based identifications of SUMOylated proteins in (A) human and (B) yeast, respectively, which are used as training data in the iSUMO prediction tool. Each column represents a published dataset. Colored entries represent proteins observed as SUMOylated by the respective dataset. Datasets are lists in full in **Suppl. Tables S1**.

Figure 1


A
Number of studies
in consensus
1
2
3
4
5
6
7
8
9+
Number of SUMOylated proteins
1000
500
0
Vertegaal et al., 2004
Blomster et al., 2010
Lamoliatte et al., 2013
Vertegaal et al., 2006
Rosas-Acosta et al., 2005
Matic et al., 2010
Manza et al., 2004
Grant, 2010
Galisson et al., 2011
Impens et al., 2014
Bruderer et al., 2011
Blomster et al., 2009
Tammsalu et al., 2014
Tatham et al., 2011
Lamoliatte et al., 2014
Goleblowski et al., 2009
Schimmel et al., 2008
Hendriks et al., 2014
B

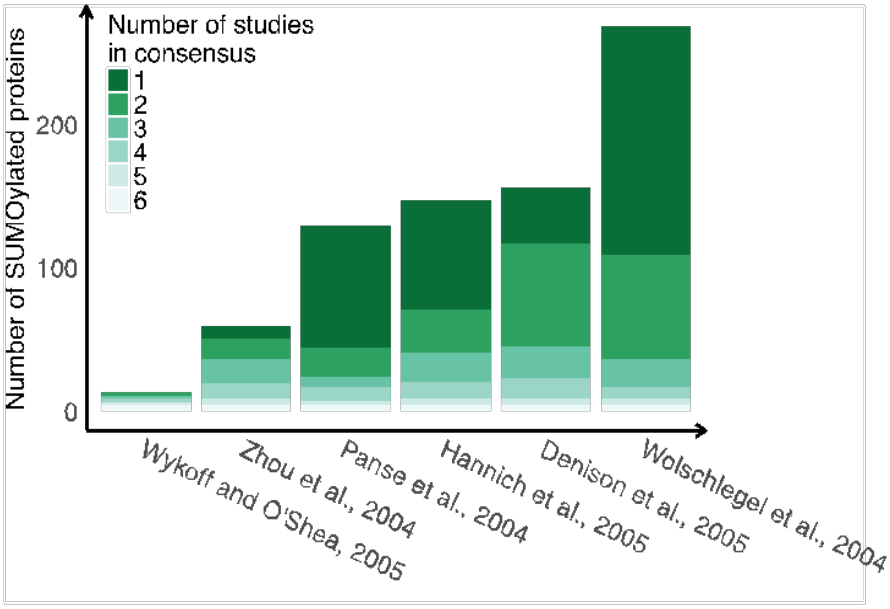

Number of studies
in consensus
1
2
3
4
5
6
Number of SUMOylated proteins
200
100
0
Wykoff and O'Shea, 2005
Zhou et al., 2004
Panse et al., 2004
Hannich et al., 2005
Denison et al., 2005
Wolschlegel et al., 2004

Figure 2. Correlation between protein SUMOylation, the total number of distinct subunits per complex, and the number subunits that bind RNA.

Complex information was taken from the CORUM database [64]. Dot size is proportional to the total number of distinct subunits of the complexes. Complexes that are both SUMOylated and contain many RNA-binding proteins are larger than other complexes. **A.** The columns are annotated as Is-RNA-binding-protein(TRUE/FALSE). Is-SUMOylated-protein(TRUE/FALSE). Protein complexes with RNA-binding members are more often SUMOylated than others. **B.** Relationship between SUMylation and RNA-binding proteins.

Figure 2

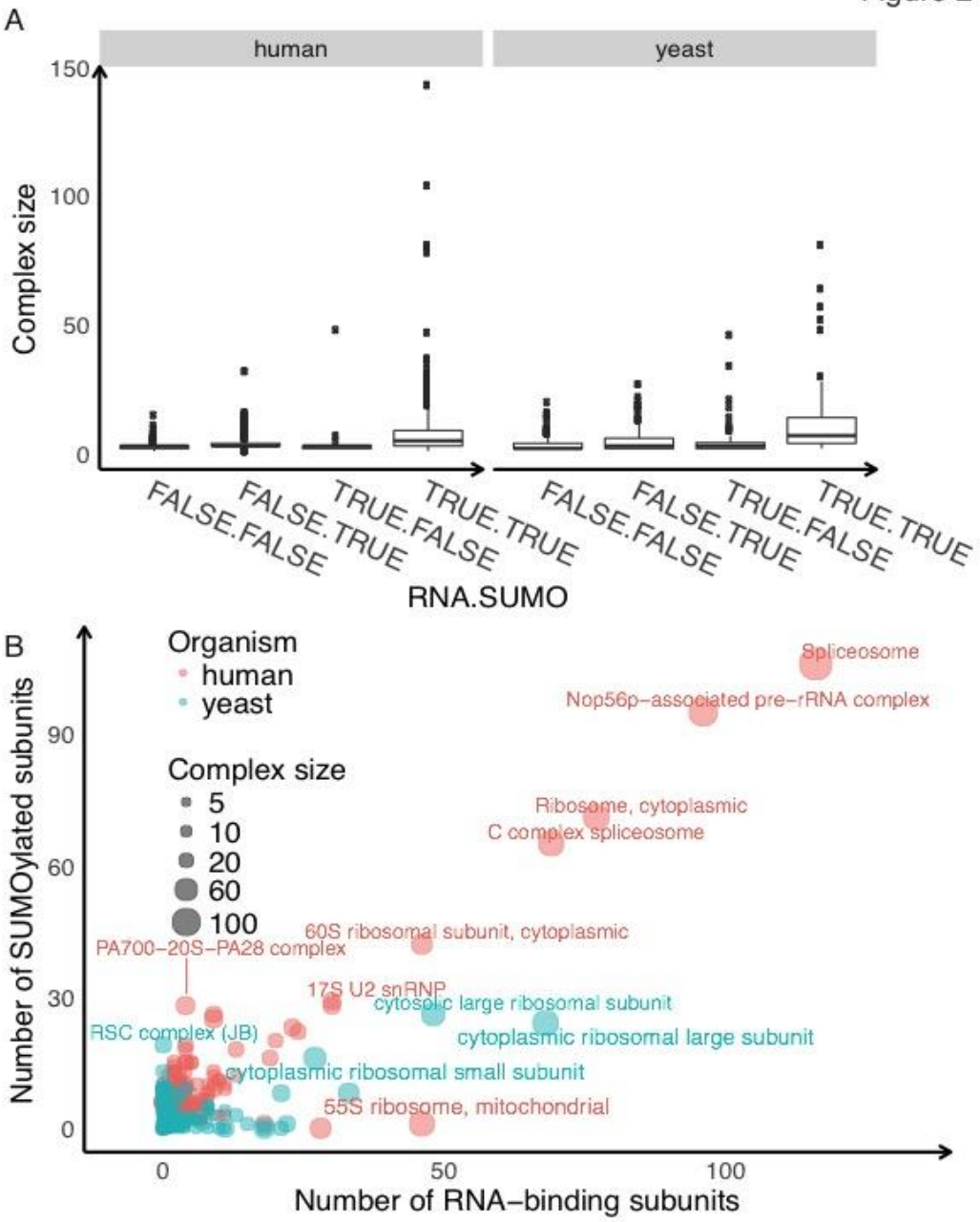

A
human
yeast
Complex size
150
100
50
0
FALSE.FALSE
FALSE.TRUE
TRUE.FALSE
TRUE.TRUE
FALSE.FALSE
FALSE.TRUE
TRUE.FALSE
TRUE.TRUE
RNA.SUMO
B
Organism
human
yeast
Complex size
5
10
20
60
100
Spliceosome
Nop56p−associated pre−rRNA complex
Ribosome, cytoplasmic
C complex spliceosome
60S ribosomal subunit, cytoplasmic
PA700−20S−PA28 complex
17S U2 snRNP
cytosolic large ribosomal subunit
RSC complex (JB)
cytoplasmic ribosomal large subunit
cytoplasmic ribosomal small subunit
55S ribosome, mitochondrial
Number of SUMOylated subunits
90
60
30
0
0
50
100
Number of RNA−binding subunits

Figure 3. iSUMO predictions outperform sequence-based predictions

A., B. Receiver operator characteristics of iSUMO predictions trained on integrated sequence and annotation-based features (red) versus sequence-based features only (blue). Gray lines are the original 10-fold cross validation runs for different sets of non-overlapping, randomly chosen true negative entries. Balancing the number of positive and negative labels ensures learning quality and fair ROC evaluation.

C., D. Frequencies of the most predictive attributes in human and yeast, measured as the number of occurrences in the different models. **Suppl. Table S4** lists all attributes used. 'Depth' marks the level in the decision tree and is displayed in different colors. The more frequent a feature is selected at low tree depth, the more predictive of SUMOylation it is.

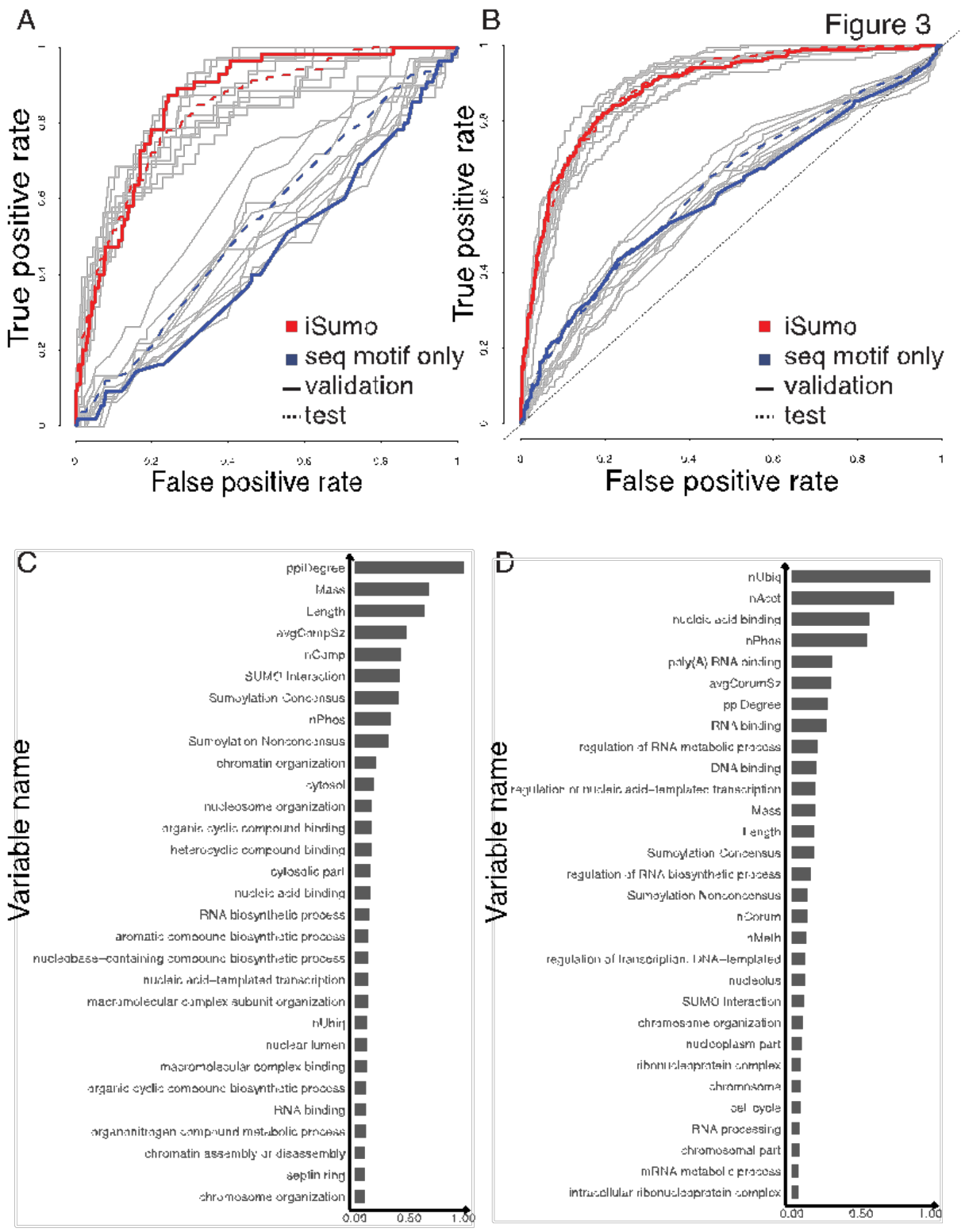


Figure 3
A
B
True positive rate
False positive rate
iSumo
seq motif only
validation
test
C
D
Variable name
ppiDegree
Mass
Length
avgCompSz
nComp
SUMO Interaction
Sumoylation Consensus
nPhos
Sumoylation Nonconsensus
chromatin organization
cytosol
nucleosome organization
organic cyclic compound binding
heterocyclic compound binding
cytosolic part
nucleic acid binding
RNA biosynthetic process
aromatic compound biosynthetic process
nucleobase-containing compound biosynthetic process
nucleic acid-templated transcription
macromolecular complex subunit organization
nUbiq
nuclear lumen
macromolecular complex binding
organic cyclic compound biosynthetic process
RNA binding
organonitrogen compound metabolic process
chromatin assembly or disassembly
septin ring
chromosome organization
nAcet
nucleic acid binding
nPhos
poly(A) RNA binding
avgComplSz
pp Degree
RNA binding
regulation of RNA metabolic process
DNA binding
regulation of nucleic acid-templated transcription
Mass
Length
Sumoylation Consensus
regulation of RNA biosynthetic process
Sumoylation Nonconsensus
nComp
nMeth
regulation of transcription, DNA-templated
nucleolus
SUMO Interaction
chromosome organization
nucleoplasm part
ribonucleoprotein complex
chromosome
cell cycle
RNA processing
chromosomal part
mRNA metabolic process
intracellular ribonucleoprotein complex
0.00
0.50
1.00

Table 1. SUMOylated proteins are biased in their functions.

We tested for function enrichment using hypergeometric tests of Gene Ontology (GO) term in human and yeast, respectively. 'Domain' codes indicate the three main branches of Gene Ontology: biological processes (BP), cellular compartment (CC), and molecular function (MF). 'Term size' refers to the total number of genes associated with the term in the GO database, and 'Intersection size' is the intersection with SUMOylated proteins. The 'Adjusted p-value' has been corrected for multiple hypotheses testing. The entries are sorted according to the adjusted p-value. Extended data is in **Suppl. Table S2.**

A. *Homo sapiens*

Domain	Term name	Term size	Intersection size	Adjusted P-value
BP	RNA processing	620	380	9.17E-119
	mRNA metabolic process	479	293	2.87E-89
	ribonucleoprotein complex biogenesis	311	213	5.90E-77
	ribosome biogenesis	202	149	1.18E-59
	rRNA processing	170	134	1.57E-59
	RNA splicing	294	187	4.30E-59
	mRNA splicing, via spliceosome	238	162	5.51E-57
	ncRNA metabolic process	382	216	8.65E-56
	chromatin organization	457	230	1.74E-47
	regulation of transcription, DNA-templated	1849	610	5.25E-47
	SRP-dependent cotranslational protein targeting to membrane	83	77	4.40E-44
CC	chromosome	604	337	4.22E-88
	ribonucleoprotein complex	470	288	6.77E-88
	nuclear chromosome	388	227	1.14E-62
	nucleolus	558	282	7.94E-60
	nucleoplasm part	491	258	1.02E-58
	chromatin	310	180	3.46E-48
MF	nucleic acid binding	1949	1014	4.45E-281
	poly(A) RNA binding	906	596	3.42E-220
	RNA binding	1093	665	1.07E-219
	DNA binding	963	428	3.95E-72

B. *Saccharomyces cerevisiae*

Domain	Term name	Term size	Intersection size	Adjusted P-value
BP	transcription, DNA-templated	678	163	3.98E-34
	organic cyclic compound biosynthetic process	918	194	1.64E-33
	nucleobase-containing compound biosynthetic process	823	181	4.25E-33
	chromatin organization	325	106	9.56E-33
	heterocycle biosynthetic process	883	188	1.24E-32
	macromolecular complex subunit organization	867	173	1.6E-25
CC	nuclear lumen	764	168	4.58E-30
	nucleoplasm	272	78	4.41E-19
	chromatin	141	52	2.74E-17
	SWI/SNF superfamily-type complex	59	33	4.82E-17
MF	nucleic acid binding	879	160	1.81E-18
	DNA binding	376	75	7.31E-09
	macromolecular complex binding	203	47	5.42E-07
	RNA binding	524	88	1.43E-06

poly(A) RNA binding	255	53	2.65E-06

Table 2. Large protein complexes are heavily SUMOylated.

The table shows the subunit compositions of stable protein complexes in human. The upper and lower tables show the largest complexes that are SUMOylated and not SUMOylated, respectively, as determined by the adjusted p-value. The complexes are sorted by descending total number of subunits. The adjusted p-value reports the bias with respect to SUMOylation of the complex subunits. Even complexes with only few RNA-binding subunits, like PA700-20S-PA28 can be SUMOylated.

Complex name	Complex size	Number of RNA-binding subunits	Number of SUMOylated subunits	Adjusted P-value, SUMO enrichment
SUMOylated complexes				
Spliceosome	143	116	106	4.37E-29
Nop56p-associated pre-rRNA complex	104	96	95	1.46E-41
Ribosome, cytoplasmic	81	77	71	4.83E-28
C complex spliceosome	80	69	65	1.52E-21
60S ribosomal subunit, cytoplasmic	47	46	42	4.44E-17
CEN complex	37	9	25	1.09E-04
PA700-20S-PA28 complex	36	4	28	7.12E-08
17S U2 snRNP	33	30	29	4.27E-11
40S ribosomal subunit, cytoplasmic	31	30	28	1.82E-11
CDC5L complex	30	23	23	4.68E-06
Non-SUMOylated complexes				
55S ribosome, mitochondrial	78	46	1	1.00E+00
39S ribosomal subunit, mitochondrial	48	28	0	1.00E+00
Mediator complex	32	0	6	1.00E+00
28S ribosomal subunit, mitochondrial	30	18	1	1.00E+00
MLL1-WDR5 complex	27	4	15	7.86E-01
BRCA1-RNA polymerase II complex	26	9	9	1.00E+00
TNF-alpha/NF-kappa B signaling complex 5	25	4	10	1.00E+00
RNA polymerase II holoenzyme complex	24	8	8	1.00E+00
Emerin complex 32	22	2	14	1.25E-01
RNA polymerase II complex	19	2	6	1.00E+00

Table 3. Highest-scoring, newly predicted SUMOylation targets are likely true events.

The table lists human proteins that have high iSUMO scores, but were not part of the training dataset. The annotation - through independent databases and literature search - shows that their SUMOylation is highly likely. The protein names are the primary common names as used by UniProtKB (http://www.ebi.ac.uk/reference_proteomes). The average iSUMO prediction scores range from 0 to 1, and is the higher, the higher the probability for SUMOylation. Function annotation is from GeneCards.org. * - SUMO site reported in PhosphoSitePlus [65]; # - SUMO site reported in BioGrid [66]. **Suppl. Table S6** shows the complete version of this table.

Gene name	iSUMO score	Protein name	Protein function	Evidence for SUMO from literature
SF3A3 (SAP61)	0.82	Splicing factor 3A subunit 3	Subunit of the splicing factor SF3A (spliceosomal complex) required for 'A' complex assembly.	SUMOylation of spliceosomal protein complex is required for function [56].
RRP1 (NNP1, RRP1A)	0.82	Ribosomal RNA processing protein 1 homolog A	Plays a critical role in the generation of 28S rRNA.	Yeast homolog is SUMOylated [55].
UTP15	0.71	U3 small nucleolar RNA-associated protein 15 homolog	Ribosome biogenesis factor, involved in nucleolar processing of pre-18S ribosomal RNA.	SUMOylation, in particular via SENP3, plays a role in rRNA maturation and assembly of the 60S subunit [9].
ISY1 (KIAA1160)	0.71#	May play a role in pre-mRNA splicing	May play a role in pre-mRNA splicing.	SUMOylation of spliceosomal protein complex is required for function [56]. Sumoylation detected under heat stress [67].
CHAF1B (CAF1A, MPP7)	0.57	Chromatin Assembly Factor 1 Subunit B	Complex that is thought to mediate chromatin assembly in DNA replication and DNA repair.	SUMO-2 target identified in screen. SUMOylation essential spindle organization, chromosome congression, and chromosome segregation [57 50].
SNRPA	0.55	Small Nuclear Ribonucleoprotein Polypeptide A	Component of the spliceosomal U1 snRNP, which is essential for recognition of the pre-mRNA 5' splice-site and the subsequent assembly of the spliceosome.	SUMOylation of spliceosomal protein complex is required for function [56]. SUMOylation known for SNRPA1.

HIST2H2AB	0.54#*	Histone H2A type 2-B	Core component of nucleosome.	Histones are heavily SUMOylated, in particular histone H3 and H4 [58].
ANAPC5 (APC5)	0.53	Anaphase-promoting complex subunit 5	Tetratricopeptide repeat-containing component of the anaphase promoting complex/cyclosome (APC/C), a large E3 ubiquitin ligase that controls cell cycle progression.	SUMOylation essential for spindle organization, chromosome congression, and chromosome segregation [57] . SUMO regulates mitosis and APC function in yeast [68].
TUBA3C (TUBA2)	0.53*	Tubulin alpha-3C/D chain	Major constituent of microtubules.	SUMOylation (especially also of tubulin) essential for spindle organization, chromosome congression, and chromosome segregation during cell division [57 69 57].
NSD1	0.51*	Histone-lysine N-methyltransferase	Methylates 'Lys-36' of histone H3 and 'Lys-20' of histone H4 (in vitro).	SUMO regulates some methyl-transferases [59].

References

1. Mahajan, R.; Delphin, C.; Guan, T.; Gerace, L.; Melchior, F., A small ubiquitin-related polypeptide involved in targeting RanGAP1 to nuclear pore complex protein RanBP2. *Cell* **1997,** *88* (1), 97-107.

2. Su, H.-L.; Li, S. S. L., Molecular features of human ubiquitin-like SUMO genes and their encoded proteins. *Gene* **2002,** *296* (1-2), 65-73.

3. Jones, D.; Crowe, E.; Stevens, T. A.; Candido, E. P. M., Functional and phylogenetic analysis of the ubiquitylation system in Caenorhabditis elegans: ubiquitin-conjugating enzymes, ubiquitin-activating enzymes, and ubiquitin-like proteins. *Genome Biol.* **2002,** *3* (1), RESEARCH0002.

4. Francis, O.; Han, F.; Adams, J. C., Molecular phylogeny of a RING E3 ubiquitin ligase, conserved in eukaryotic cells and dominated by homologous components, the muskelin/RanBPM/CTLH complex. *PLoS One* **2013,** *8* (10), e75217.

5. Saracco, S. A.; Miller, M. J.; Kurepa, J.; Vierstra, R. D., Genetic analysis of SUMOylation in Arabidopsis: conjugation of SUMO1 and SUMO2 to nuclear proteins is essential. *Plant Physiol* **2007,** *145* (1), 119-34.

6. Eskiw, C. H.; Dellaire, G.; Bazett-Jones, D. P., Chromatin contributes to structural integrity of promyelocytic leukemia bodies through a SUMO-1-independent mechanism. *J. Biol. Chem.* **2004,** *279* (10), 9577-9585.

7. Zhao, X.; Blobel, G., From The Cover: A SUMO ligase is part of a nuclear multiprotein complex that affects DNA repair and chromosomal organization. *Proceedings of the National Academy of Sciences* **2005,** *102* (13), 4777-4782.

8. Girdwood, D. W. H.; Tatham, M. H.; Hay, R. T., SUMO and transcriptional regulation. *Semin. Cell Dev. Biol.* **2004,** *15* (2), 201-210.

9. Finkbeiner, E.; Haindl, M.; Raman, N.; Muller, S., SUMO routes ribosome maturation. *Nucleus* **2011,** *2* (6), 527-532.

10. Vethantham, V.; Rao, N.; Manley, J. L., Sumoylation regulates multiple aspects of mammalian poly(A) polymerase function. *Genes Dev.* **2008,** *22* (4), 499-511.

11. Vethantham, V.; Rao, N.; Manley, J. L., Sumoylation modulates the assembly and activity of the pre-mRNA 3' processing complex. *Mol Cell Biol* **2007,** *27* (24), 8848-58.

12. Wen, D.; Xu, Z.; Xia, L.; Liu, X.; Tu, Y.; Lei, H.; Wang, W.; Wang, T.; Song, L.; Ma, C.; Xu, H.; Zhu, W.; Chen, G.; Wu, Y., Important role of SUMOylation of Spliceosome factors in prostate cancer cells. *J Proteome Res* **2014,** *13* (8), 3571-82.

13. Meier, I., mRNA export and sumoylation-Lessons from plants. *Biochim Biophys Acta* **2012,** *1819* (6), 531-7.

14. Truong, K.; Lee, T. D.; Li, B.; Chen, Y., Sumoylation of SAE2 C terminus regulates SAE nuclear localization. *J Biol Chem* **2012,** *287* (51), 42611-9.

15. Schimmel, J.; Larsen, K. M.; Matic, I.; van Hagen, M.; Cox, J.; Mann, M.; Andersen, J. S.; Vertegaal, A. C. O., The Ubiquitin-Proteasome System Is a Key Component of the SUMO-2/3 Cycle. *Mol. Cell. Proteomics* **2008,** *7* (11), 2107-2122.

16. Guo, C.; Henley, J. M., Wrestling with stress: roles of protein SUMOylation and deSUMOylation in cell stress response. *IUBMB Life* **2014,** *66* (2), 71-7.

17. Geiss-Friedlander, R.; Melchior, F., Concepts in sumoylation: a decade on. *Nat Rev Mol Cell Biol* **2007,** *8* (12), 947-56.

18. Yavuz, A. S.; Sezerman, O. U., Predicting sumoylation sites using support vector machines based on various sequence features, conformational flexibility and disorder. *BMC genomics* **2014,** *15 Suppl 9*, S18.

19. Zhao, B.; Villhauer, E. B.; Bhuripanyo, K.; Kiyokawa, H.; Schindelin, H.; Yin, J., SUMO-mimicking peptides inhibiting protein SUMOylation. *Chembiochem : a European journal of chemical biology* **2014,** *15* (18), 2662-6.

20. Teng, S.; Luo, H.; Wang, L., Predicting protein sumoylation sites from sequence features. *Amino acids* **2011,** *43* (1), 447-455.

21. Ren, J.; Gao, X.; Jin, C.; Zhu, M.; Wang, X.; Shaw, A.; Wen, L.; Yao, X.; Xue, Y., Systematic study of protein sumoylation: Development of a site-specific predictor of SUMOsp 2.0. *Proteomics* **2009,** *9* (12), 3409-3412.

22. Hendriks, I. A.; D'Souza, R. C. J.; Yang, B.; Verlaan-de Vries, M.; Mann, M.; Vertegaal, A. C. O., Uncovering global SUMOylation signaling networks in a site-specific manner. *Nat. Struct. Mol. Biol.* **2014,** *21* (10), 927-936.

23. Sarge, K. D.; Park-Sarge, O.-K., Detection of Proteins Sumoylated In Vivo and In Vitro. In *Methods in Molecular Biology*, 2009; pp 265-277.

24. Tammsalu, T.; Matic, I.; Jaffray, E. G.; Ibrahim, A. F.; Tatham, M. H.; Hay, R. T., Proteome-wide identification of SUMO2 modification sites. *Sci Signal* **2014,** *7* (323), rs2.

25. Lamoliatte, F.; Caron, D.; Durette, C.; Mahrouche, L.; Maroui, M. A.; Caron-Lizotte, O.; Bonneil, E.; Chelbi-Alix, M. K.; Thibault, P., Large-scale analysis of lysine SUMOylation by SUMO remnant immunoaffinity profiling. *Nat Commun* **2014,** *5*, 5409.

26. Tatham, M. H.; Matic, I.; Mann, M.; Hay, R. T., Comparative proteomic analysis identifies a role for SUMO in protein quality control. *Science signaling* **2011,** *4* (178), rs4.

27. Galisson, F.; Mahrouche, L.; Courcelles, M.; Bonneil, E.; Meloche, S.; Chelbi-Alix, M. K.; Thibault, P., A novel proteomics approach to identify SUMOylated proteins and their modification sites in human cells. *Mol. Cell. Proteomics* **2011,** *10* (2), M110.004796.

28. Bruderer, R.; Tatham, M. H.; Plechanovova, A.; Matic, I.; Garg, A. K.; Hay, R. T., Purification and identification of endogenous polySUMO conjugates. *EMBO Rep.* **2011,** *12* (2), 142-148.

29. Grant, M. M., Identification of SUMOylated proteins in neuroblastoma cells after treatment with hydrogen peroxide or ascorbate. *BMB Rep* **2010,** *43* (11), 720-5.

30. Blomster, H. A.; Hietakangas, V.; Wu, J.; Kouvonen, P.; Hautaniemi, S.; Sistonen, L., Novel proteomics strategy brings insight into the prevalence of SUMO-2 target sites. *Mol. Cell. Proteomics* **2009,** *8* (6), 1382-1390.

31. Vertegaal, A. C.; Andersen, J. S.; Ogg, S. C.; Hay, R. T.; Mann, M.; Lamond, A. I., Distinct and overlapping sets of SUMO-1 and SUMO-2 target proteins revealed by quantitative proteomics. *Molecular & cellular proteomics : MCP* **2006,** *5* (12), 2298-310.

32. Vertegaal, A. C.; Ogg, S. C.; Jaffray, E.; Rodriguez, M. S.; Hay, R. T.; Andersen, J. S.; Mann, M.; Lamond, A. I., A proteomic study of SUMO-2 target proteins. *J Biol Chem* **2004,** *279* (32), 33791-8.

33. Hannich, J. T.; Lewis, A.; Kroetz, M. B.; Li, S.-J.; Heide, H.; Emili, A.; Hochstrasser, M., Defining the SUMO-modified proteome by multiple approaches in Saccharomyces cerevisiae. *Journal of Biological Chemistry* **2005,** *280* (6), 4102-4110.

34. Denison, C.; Rudner, A. D.; Gerber, S. A.; Bakalarski, C. E.; Moazed, D.; Gygi, S. P., A Proteomic Strategy for Gaining Insights into Protein Sumoylation in Yeast. *Mol. Cell. Proteomics* **2004,** *4* (3), 246-254.

35. Wykoff, D. D.; O'Shea, E. K., Identification of sumoylated proteins by systematic immunoprecipitation of the budding yeast proteome. *Mol Cell Proteomics* **2005,** *4* (1), 73-83.

36. Zhou, W.; Ryan, J. J.; Zhou, H., Global analyses of sumoylated proteins in Saccharomyces cerevisiae. Induction of protein sumoylation by cellular stresses. *J Biol Chem* **2004,** *279* (31), 32262-8.

37. Wohlschlegel, J. A.; Johnson, E. S.; Reed, S. I.; Yates, J. R., Improved Identification of SUMO Attachment Sites Using C-Terminal SUMO Mutants and Tailored Protease Digestion Strategies. *J. Proteome Res.* **2006,** *5* (4), 761-770.

38. Panse, V. G.; Hardeland, U.; Werner, T.; Kuster, B.; Hurt, E., A proteome-wide approach identifies sumoylated substrate proteins in yeast. *J Biol Chem* **2004,** *279* (40), 41346-51.

39. Bursteinas, B.; Britto, R.; Bely, B.; Auchincloss, A.; Rivoire, C.; Redaschi, N.; O'Donovan, C.; Martin, M. J., Minimizing proteome redundancy in the UniProt Knowledgebase. *Database* **2016,** *2016*, baw139.

40. Xue, Y.; Zhou, F.; Fu, C.; Xu, Y.; Yao, X., SUMOsp: a web server for sumoylation site prediction. *Nucleic Acids Res* **2006,** *34* (Web Server issue), W254-7.

41. Reimand, J.; Arak, T.; Adler, P.; Kolberg, L.; Reisberg, S.; Peterson, H.; Vilo, J., g:Profiler-a web server for functional interpretation of gene lists (2016 update). *Nucleic Acids Res.* **2016,** *44* (W1), W83-9.

42. Reimand, J.; Kull, M.; Peterson, H.; Hansen, J.; Vilo, J., g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Res* **2007,** *35* (Web Server issue), W193-200.

43. Chatr-Aryamontri, A.; Breitkreutz, B.-J.; Heinicke, S.; Boucher, L.; Winter, A.; Stark, C.; Nixon, J.; Ramage, L.; Kolas, N.; O'Donnell, L.; Reguly, T.; Breitkreutz, A.; Sellam, A.; Chen, D.; Chang, C.; Rust, J.; Livstone, M.; Oughtred, R.; Dolinski, K.; Tyers, M., The BioGRID interaction database: 2013 update. *Nucleic Acids Res.* **2013,** *41* (Database issue), D816-23.

44. Winter, A. G.; Wildenhain, J.; Tyers, M., BioGRID REST Service, BiogridPlugin2 and BioGRID WebGraph: new tools for access to interaction data at BioGRID. *Bioinformatics* **2011,** *27* (7), 1043-1044.

45. Stark, C.; Breitkreutz, B.-J.; Reguly, T.; Boucher, L.; Breitkreutz, A.; Tyers, M., BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **2006,** *34* (Database issue), D535-9.

46. Ruepp, A.; Waegele, B.; Lechner, M.; Brauner, B.; Dunger-Kaltenbach, I.; Fobo, G.; Frishman, G.; Montrone, C.; Mewes, H. W., CORUM: the comprehensive resource of mammalian protein complexes--2009. *Nucleic Acids Res* **2010,** *38* (Database issue), D497-501.

47. Hornbeck, P. V.; Kornhauser, J. M.; Tkachev, S.; Zhang, B.; Skrzypek, E.; Murray, B.; Latham, V.; Sullivan, M., PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.* **2011,** *40* (D1), D261-D270.

48. Lu, C.-T.; Huang, K.-Y.; Su, M.-G.; Lee, T.-Y.; Bretaña, N. A.; Chang, W.-C.; Chen, Y.-J.; Chen, Y.-J.; Huang, H.-D., DbPTM 3.0: an informative resource for investigating substrate site specificity and functional association of protein post-translational modifications. *Nucleic Acids Res.* **2013,** *41* (Database issue), D295-305.

49. Texari, L.; Stutz, F., Sumoylation and transcription regulation at nuclear pores. *Chromosoma* **2015,** *124* (1), 45-56.

50. Ouyang, J.; Gill, G., SUMO engages multiple corepressors to regulate chromatin structure and transcription. *Epigenetics* **2009,** *4* (7), 440-4.

51. Everett, R. D.; Boutell, C.; Hale, B. G., Interplay between viruses and host sumoylation pathways. *Nat. Rev. Microbiol.* **2013,** *11* (6), 400-411.

52. Psakhye, I.; Jentsch, S., Protein group modification and synergy in the SUMO pathway as exemplified in DNA repair. *Cell* **2012,** *151* (4), 807-20.

53. Zhao, Q.; Xie, Y.; Zheng, Y.; Jiang, S.; Liu, W.; Mu, W.; Liu, Z.; Zhao, Y.; Xue, Y.; Ren, J., GPS-SUMO: a tool for the prediction of sumoylation sites and SUMO-interaction motifs. *Nucleic Acids Res* **2014,** *42* (Web Server issue), W325-30.

54. Lawrence, R. T.; Searle, B. C.; Llovet, A.; Villén, J., Plug-and-play analysis of the human phosphoproteome by targeted high-resolution mass spectrometry. *Nat. Methods* **2016,** *13* (5), 431-434.

55. Costanzo, M.; VanderSluis, B.; Koch, E. N.; Baryshnikova, A.; Pons, C.; Tan, G.; Wang, W.; Usaj, M.; Hanchard, J.; Lee, S. D.; Pelechano, V.; Styles, E. B.; Billmann, M.; van Leeuwen, J.; van Dyk, N.; Lin, Z.-Y.; Kuzmin, E.; Nelson, J.; Piotrowski, J. S.; Srikumar, T.; Bahr, S.; Chen, Y.; Deshpande, R.; Kurat, C. F.; Li, S. C.; Li, Z.; Usaj, M. M.; Okada, H.; Pascoe, N.; San Luis, B.-J.; Sharifpoor, S.; Shuteriqi, E.; Simpkins, S. W.; Snider, J.; Suresh, H. G.; Tan, Y.; Zhu, H.; Malod-Dognin, N.; Janjic, V.; Przulj, N.; Troyanskaya, O. G.; Stagljar, I.; Xia, T.; Ohya, Y.; Gingras, A.-C.; Raught, B.; Boutros, M.; Steinmetz, L. M.; Moore, C. L.; Rosebrock, A. P.; Caudy, A. A.; Myers, C. L.; Andrews, B.; Boone, C., A global genetic interaction network maps a wiring diagram of cellular function. *Science* **2016,** *353* (6306).

56. Pozzi, B.; Bragado, L.; Will, C. L.; Mammi, P.; Risso, G.; Urlaub, H.; Lührmann, R.; Srebrow, A., SUMO conjugation to spliceosomal proteins is required for efficient pre-mRNA splicing. *Nucleic Acids Res.* **2017**.

57. Cubeñas-Potts, C.; Matunis, M. J., SUMO: a multifaceted modifier of chromatin structure and function. *Dev. Cell* **2013,** *24* (1), 1-12.

58. Hendriks, I. A.; Vertegaal, A. C., A comprehensive compilation of SUMO proteomics. *Nat Rev Mol Cell Biol* **2016,** *17* (9), 581-95.

59. Nayak, A.; Viale-Bouroncle, S.; Morsczeck, C.; Muller, S., The SUMO-specific isopeptidase SENP3 regulates MLL1/MLL2 methyltransferase complexes and controls osteogenic differentiation. *Mol. Cell* **2014,** *55* (1), 47-58.

60. Hendriks, I. A.; Vertegaal, A. C. O., A comprehensive compilation of SUMO proteomics. *Nat. Rev. Mol. Cell Biol.* **2016,** *17* (9), 581-595.

61. Marblestone, J. G.; Edavettal, S. C.; Lim, Y.; Lim, P.; Zuo, X.; Butt, T. R., Comparison of SUMO fusion technology with traditional gene fusion systems: enhanced expression and solubility with SUMO. *Protein Sci* **2006,** *15* (1), 182-9.

62. Ramaswami, M.; Paul Taylor, J.; Parker, R., Altered Ribostasis: RNA-Protein Granules in Degenerative Disorders. *Cell* **2013,** *154* (4), 727-736.

63. Krumova, P.; Meulmeester, E.; Garrido, M.; Tirard, M.; Hsiao, H.-H.; Bossis, G.; Urlaub, H.; Zweckstetter, M.; Kügler, S.; Melchior, F.; Bähr, M.; Weishaupt, J. H., Sumoylation inhibits alpha-synuclein aggregation and toxicity. *J. Cell Biol.* **2011,** *194* (1), 49-60.

64. Ruepp, A.; Brauner, B.; Dunger-Kaltenbach, I.; Frishman, G.; Montrone, C.; Stransky, M.; Waegele, B.; Schmidt, T.; Doudieu, O. N.; Stümpflen, V.; Mewes, H. W., CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res.* **2008,** *36* (Database issue), D646-50.

65. Hornbeck, P. V.; Zhang, B.; Murray, B.; Kornhauser, J. M.; Latham, V.; Skrzypek, E., PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res* **2015,** *43* (Database issue), D512-20.

66. Chatr-Aryamontri, A.; Breitkreutz, B. J.; Heinicke, S.; Boucher, L.; Winter, A.; Stark, C.; Nixon, J.; Ramage, L.; Kolas, N.; O'Donnell, L.; Reguly, T.; Breitkreutz, A.; Sellam, A.; Chen, D.; Chang, C.; Rust, J.; Livstone, M.; Oughtred, R.; Dolinski, K.; Tyers, M., The BioGRID interaction database: 2013 update. *Nucleic Acids Res* **2013,** *41* (Database issue), D816-23.

67. Golebiowski, F.; Matic, I.; Tatham, M. H.; Cole, C.; Yin, Y.; Nakamura, A.; Cox, J.; Barton, G. J.; Mann, M.; Hay, R. T., System-wide changes to SUMO modifications in response to heat shock. *Science signaling* **2009,** *2* (72), ra24.

68. Dieckhoff, P.; Bolte, M.; Sancak, Y.; Braus, G. H.; Irniger, S., Smt3/SUMO and Ubc9 are required for efficient APC/C-mediated proteolysis in budding yeast. *Mol. Microbiol.* **2004,** *51* (5), 1375-1387.

69. Yuan, Y.-F.; Zhai, R.; Liu, X.-M.; Khan, H. A.; Zhen, Y.-H.; Huo, L.-J., SUMO-1 plays crucial roles for spindle organization, chromosome congression, and chromosome segregation during mouse oocyte meiotic maturation. *Mol. Reprod. Dev.* **2014,** *81* (8), 712-724.