# iSUMO - integrative prediction of functionally relevant SUMOylated proteins

Xiaotong Yao[1,2], Rebecca Bish[1], 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

* Corresponding author: cvogel@nyu.edu

## Abstract

Post-translational modifications by the Small Ubiquitin-like Modifier (SUMO) are essential for many eukaryotic cellular functions. Several large-scale experimental datasets and sequence-based predictions exist that identify SUMOylated proteins. However, the overlap between these datasets is small, suggesting many false positives with low functional relevance. Therefore, we applied machine learning techniques to a diverse set of large-scale SUMOylation studies combined with protein characteristics such as cellular function and protein-protein interactions, to provide integrated SUMO predictions for human and yeast cells (iSUMO). Protein-protein and protein-nucleic acid interactions prove to be 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 predict 1,596 and 492 SUMO targets in human and yeast, respectively (5% false positive rate, FPR), which is five times more than what existing sequence-based tools predict at the same FPR. One third of the predictions are validated by an independent, high-quality dataset. iSUMO therefore represents a comprehensive SUMO prediction tool for human and yeast with a high probability for functional relevance of the predictions.

## 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, 6). SUMOylation resembles ubiquitination in terms of structure, enzymatic pathway, and it has a broad functional spectrum, ranging from chromatin organization (7), DNA damage repair (8), regulation of transcription (9), ribosome biogenesis (10), messenger RNA (mRNA) processing (11, 12), nucleus-cytoplasm transport (13), to protein localization (14), proteolysis (where it cross-talks with ubiquitination)(15), stress responses (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 sequence-based predictions, many more SUMO-targets have been identified, e.g. 8,272 of 17,741 human proteins (47%) are predicted to have a SUMOylation motif in their sequence.

Overall, these findings suggest that our current computational and experimental methods contain large numbers of both false positives and negatives, without high confidence in their functional relevance.

Therefore, we present an approach to integrate various dataset to provide a comprehensive picture of SUMOylation that distinguishes between functional and non-functional SUMOylation events. We developed iSUMO that integrates protein sequence and functional annotations into a comprehensive prediction strategy. We describe iSUMO's approach, its validation, and application to both yeast and human for a genome-wide assessment of SUMOylation.

## Materials and Methods

### Training data sets of experimentally observed SUMOylated proteins

We assembled the results from 14 and six large-scale, experimental studies in human (24-37) and yeast (38-43), respectively, which mapped SUMOylated proteins using mass spectrometry. **Figure 1** summarizes the relationships between the datasets, and detailed description of the data can be found in **Supplementary Tables S1**. We obtained a total of 1,860 and 555 distinct human and yeast proteins, respectively.

For the Gene Ontology (GO) enrichment analysis, the human reference proteome was downloaded from European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) release 2014_04, based on Universal Protein Resource (UniProt) release 2014_04, Ensembl release 75, and Ensembl Genome release 21 (http://www.ebi.ac.uk/reference_proteomes). The yeast reference proteome data was from Saccharomyces Genome Database (SGD), based on S288c reference genome release R64-1-1. We then filtered both reference proteomes to contain only one protein per gene, restricting the protein status in UniProt to 'reviewed'. The filtering resulted in 22,242 and 6,619 distinct genes/proteins for human and yeast, respectively.

To integrate the various datasets, we adopted different gene identifiers and filtered the reference lists to retain only the genes and proteins that had unique identity across the combination of Ensembl Gene ID, National Center for Biotechnology Information (NCBI) Gene ID, and UniProt Knowledge Base (UniProtKB) Accession ID. Also, we restricted the proteins' review scores in UniProtKB to five out of five, thus ensuring reliability of the gene annotations. As a result, the final datasets comprised 17,741 and 6,609 human and yeast proteins, respectively. Of these proteins, 1,742 and 530 were labeled as experimentally observed SUMOylated, respectively (**Figure 1**).

### 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, 44) to predict both canonical SUMOylation motifs and SUMO-

interaction motifs (SIM). The threshold was set to 'high', allowing for detection of both Type-I (covalent modification motif) and Type-II (SIM) motifs.

## Function enrichment analysis

GO enrichment analysis was carried out using gProfileR tool (45, 46). The background lists were defined as the union of all lists of observed SUMOylated proteins as described above. We set additional parameters to: 1) only report p-values smaller than E-10; 2) use adjusted p-values that correct for multiple hypotheses testing.

## 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, sequence-based predictions that identify SUMOylation-specific conserved sequence motifs type I and II which are grouped by type and are reported both as a binary event (predicted yes/no) or a quantitative event (number of predicted sites per protein). The resulting four 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 Gene Ontology (GO) terms, i.e. those terms with significant enrichment (**Table 1**). Enriched GO terms were filtered for redundancy, removing terms with a pairwise Jaccard indices larger than 0.75.

For the human data, we also incorporated information from the BioGrid interaction database (47) which records protein-protein interactions extracted from both high- and low-throughput experimental studies. We simplified its rich structure into one binary attribute called 'isBioGridInteractor' (yes/no) to represent the presence of a protein in any of the recorded physical interactions (48-50). Similarly, we constructed an attribute called 'isCORUMsubunit' (yes/no) to indicate a protein's membership in stable protein complex (51, 52). Moreover, we included phosphorylation or acetylation of a protein based on the annotations in UniProtKB, resulting in the two attributes 'phosphoProt' and 'acetylProt'. Finally, we also used probable evolutionary origin of the human proteins and constructed a corresponding feature 'ancestry' as obtained from reference (53).

## Alternating decision tree-based multi-step ensemble learning strategy

iSUMO models SUMOylated proteins as a binary classification task based on training data, which is a mixture of categorical and numerical attributes described above. For all learning, we used the Waikato Environment for Knowledge Analysis (WEKA), version 3.6.1 and the R-WEKA interface(54). The core classification method is an alternating decision tree induced by logistic boosting algorithm, implemented in WEKA as LADTree. Alternating decision trees, as suggested by the name, consist of two types of nodes,

decision nodes and prediction nodes, connected in alternating fashion. An instance traverses all the possible nodes that satisfy feature values and sums up the prediction scores as the output. In the LADTree model, the output score is between 0 and 1, and in our case, a prediction score of 1 represents 'SUMOylated', while 0 represents a 'non-SUMOylated' protein. As training labels were strongly biased towards non-SUMOylated proteins, we used a 'split and recombine' strategy to balance true positives and negatives. We randomly partitioned human non-SUMOylated proteins into nine non-overlapping subsets, each of the similar size of the set of SUMOylated proteins, and learned the model nine times independently, using the method described above. The prediction scores from these models were averages as described in the Results section. For yeast, every step is carried out in the same way, except we divided the negative examples into eleven random subsets.

Each model from the 'split and recombine' strategy was built and evaluated using ten-fold cross-validation using ten independent random re-samplings of the training data. In other words, after balancing the original data with respect to positive and negative instances, we applied bagging LADTree to each of the balanced partition of the data and used the average prediction scores as the final output. A representative example tree is shown in **Figure 4**.

We recorded the average measures of accuracy, precision, recall, and receiver operating characteristics (ROC) using the ROCR package(55). We compared the proteins' prediction scores given by each model to ensure agreement across sub-datasets. All R or Python scripts are available upon request; input and output data files are provided in the **Supplementary Material**.

## Estimation of the total number of SUMylated proteins

This calculation is independent of iSUMO predictions, as it only takes experimental datasets into account. In brief, the method uses an approximation of the hypergeometric distribution to estimate the total size of the population (SUMO-'ome') from the sizes and intersections of two 'samples' (experiments) drawn from the population. We adapted this approach from reference (56), in which the total number of protein-protein interactions was estimated from several large-scale data. When we estimated the total number of SUMOylated proteins in the entire human genome, the pairwise analysis of all 14 human datasets provided a median of 1,241 and a mean of 1,610 SUMOylated proteins.

## Results

### Integrating large-scale studies of SUMOylated proteome

To obtain a comprehensive training dataset of true-positive SUMOylated proteins, we integrated 14 and six large-scale, experimental studies for human and yeast, respectively (**Figure 1**). In human, these studies identified between eleven to 841 proteins, and a total of 1,862 SUMO targets. About half (45%) of these proteins were identified by two or more studies. In yeast, only one third (30%) of the 555 total SUMOylated proteins were identified by more than one study. The lack of overlap between individual studies indicates that the individual studies suggests that many have false-positive identifications which are not biologically functional.

### SUMOylated human and yeast proteins primarily have nucleic acid related functions

**Table 1** shows representative, highly significant Gene Ontology (GO) enrichments for both the human and yeast SUMOylated proteins at an FDR cutoff of E-40. The complete results are in the **Supplementary Tables**. Both human and yeast were enriched in functions related to DNA use and metabolism, 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 (57, 58). Interestingly, we found more than half of the 163 proteins that work in viral transcription annotated as SUMOylated, which is consistent with the notion that viruses take advantage of host cell SUMOylation to optimize viral gene expression (59).

In addition, GO terms relating to RNA use and metabolism were highly enriched, often even more significantly than those concerning DNA (**Table 1**). These 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 link to RNA functions at various levels. The function enrichment towards RNA metabolism was even stronger in human than in yeast, perhaps due to the expanded role of post-transcriptional regulation, e.g. via splicing, in human.

Further, SUMOylated proteins were preferably part of protein complexes, consistent with the hypothesis that SUMOylation helps with complex assembly and stabilization of protein-protein interactions (60). SUMOylation of protein complexes is discussed more in detail below (**Table 2**). We also observed enriched functions which have so far received less attention in connection with SUMOylation. 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.

## Large protein-RNA complexes in human are heavily SUMOylated

To test if SUMOylation is a property for all large complexes, we mapped stable human complexes reported in the CORUM (the Comprehensive Resource of Mammalian Protein Complexes) database to the SUMOylation data. **Table 2** (upper part) shows the complexes in descending order of significance of SUMO enrichment. The complexes function in, for example, the pre-rRNA complex, involved in ribosome biogenesis, splicing, translation, protein degradation (proteasome), and centromere chromatin complex (CEN). Interestingly, complexes which are depleted of SUMO (**Table 2**, lower part) often involve RNA polymerization which involves DNA-binding, but not necessarily RNA-binding; other non-SUMOylated complexes are localized to the mitochondria. When we plotted the number of RNA-binding proteins versus the number of SUMOylated subunits (**Figure 2**), we observed, with the exception of the mitochondrial ribosome, that the number of RNA-binding subunits and the number of SUMOylated subunits were strongly correlated. The probability to be SUMOylated was therefore linked to the number of protein interactions and the complex size.

## Predicting SUMOylation is improved by integration of diverse annotations

The wide range of characteristics of SUMOylated proteins highlight the need for tools that include more than sequence information to predict SUMOylation. We developed such a tool, called iSUMO, employing machine learning algorithms, i.e. boosting tree-based predictive models. This group of algorithms performs well with binary attributes which comprise much of our training set. iSUMO integrates a total of 105 and 77 attributes for human and yeast, respectively, which are listed in the **Supplementary Tables.** These attributes include the function biases discussed above, information on protein interactions, and sequence-based predictions derived from the GPS-SUMO tool(19). Since the training set (**Figure 1**) contains ten-fold more negatives than positives, i.e. non-SUMOylated and SUMOylated proteins, respectively, we employed a 'split and recombine' strategy that randomly split the set of negatives into multiple subsets of the same size as the positives, and then averaged over the resulting separate training results. Training and learning was carried out using bagging multiple LADTrees (61) as this algorithm outperformed other algorithms that we tested (*not shown*). For each dataset with equal instances of positives and negatives, we fitted a Bagging LADTree using the WEKA environment(54). To evaluate the success of the learning, we then performed ten-fold cross-validation which, in ten iterations, used 90% and 10% of the data for training and testing, respectively. The results of this testing were then presented in Receiver-Operator-Curves (ROC).

Overall, iSUMO showed a substantial improvement over predictions based on sequence alone (**Figure 2**). For example, iSUMO's average area underneath the ROC is 0.86 and 0.76 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 0.77 and 0.74 for human and yeast, respectively, which in turn predict 1,596 and 492 SUMOylated human and yeast proteins. The complete predictions, including average scores and standard deviations are available in the **Supplementary Tables**.

## Protein-protein interactions are predictive of SUMOylation

Next, we analyzed the iSUMO models for attributes that are highly predictive across separate round of learning (**Figure 3**). A representative example tree is shown in **Figure 4**. As the underlying model was a decision tree, the level of 'depth' at which an attribute occurred was indicative of its importance: the smaller the depth, the more important the attribute.

The decision tree in **Figure 4** shows the 'isBioGridInteractor' attribute which describes the protein having at least one protein interaction partner listed in the BioGrid database(47-50). This attribute nearly always occurred at depth 1 which confirmed the importance of protein-protein interactions when predicting SUMOylation events. A SUMOylated human protein is highly likely to interact with other proteins – and *vice versa*.

The next most common attribute was 'localization to exosomes' (**Figure 3**). This annotation is non-trivial to interpret, since two types of exosomes exist with very different functions. In one definition, an exosome is a multiprotein complex that exists in the cytosol and nucleus, and degrades RNA using different endo- and exonucleases. This complex therefore has many RNA-binding proteins which would explain the high degree of SUMOylation. In the second use, exosomes are membrane-enclosed microvesicles that are secreted and contain miRNA, RNA, and proteins. Recent work has shown that the RNA-binding protein hnRNPA2B1, which is responsible for sorting miRNAs into the exosome vesicles, is SUMOylated (62).

Other common attributes with high predictive power included 'RNA binding', nucleolus', 'nucleoplasm', 'chromosome', 'macromolecule complex subunit organization', 'isCORUMsubunit', and 'ancestry' (**Figure 3**). Notably, being part of a protein complex ('isCORUMsubunit') is different from a simple protein-protein interaction ('isBioGridInteractor'), as not all protein-protein interactions necessarily lead to stable complexes.

In comparison, the most common attribute in the yeast models was the sequence-based SUMO prediction 'countTypeIpred', which is the number of covalent SUMO modification motif predicted within the protein sequence using a software tool (19, 44). This observation suggested that sequence-based predictions in yeast

proteins are quite successful, but SUMOylation may have acquired more complex roles in humans that are not encoded in sequence.

### Validation and application of iSUMO predictions

To validate the iSUMO predictions independently, we compared the results to a publication by Hendriks et al. which reports SUMOylation sites for >1,600 human proteins (22). This study was not part of iSUMO's training dataset. **Figure 5** shows the kernel densities of the iSUMO prediction scores for all human proteins, separated into proteins observed by Hendriks et al. and those not observed in the independent study. The higher the iSUMO prediction score, the higher the fraction of proteins confirmed by Hendriks et al. For example, at an iSUMO prediction score cutoff of 0.77 (5% FPR), of the 1,596 proteins predicted to be SUMOylated, one third (458) were validated by the independent dataset.

**Table 3** lists ten human proteins with the highest iSUMO prediction scores, but which were not reported in the original 14 training datasets. Four of the ten proteins were validated by independent large-scale studies (**Table 3**), three are listed in the recent study by Hendriks et al. (PSMA1, DGCR8, NUFIP1)(22). Seven proteins are also ubiquitinated – and given that the ubiquitin and SUMO often co-occur, this observation strengthens the prediction. One of highly predicted proteins is microprocessor complex subunit DGCR8, an RNA-binding component of the microprocessor complex which is responsible for cleaving pri-miRNA to precursor miRNA. DGCR8's SUMOylation has recently been confirmed in a targeted study: SUMO at lysine 707 stabilizes the protein by preventing ubiquitination and subsequent degradation through ubiquitin proteosomal system (63). The same study also showed that SUMO affects DGCR8's affinity to pri-miRNAs, ensuring their repression of the mRNA targets.

## Discussion

Despite the availability of several large-scale experimental datasets that identify SUMOylated proteins in human and yeast cells, the studies only partially intersect with each other, suggesting many false positive identifications (**Figure 1**). To predict the true-positive SUMOylation events that are likely biologically functional, we used 14 and 6 large-scale studies from human and yeast, respectively, and integrated the data with sequence-based SUMOylation predictions and other protein characteristics. Using these overrepresented characteristics and the experimental training data, we constructed iSUMO, an integrated search engine which predicts about five times more proteins in the training data than sequence-based predictions alone (at 5% false positive rate) – and one third of these predictions are validated by an independent, high-quality study that has been published recently (**Figure 5**).

iSUMO predicted a total of 1,596 SUMO targets in human - a number which is very close to the total size of the human SUMO-'ome' which we estimated based on a published method that analyzes the overlap between different experimental datasets (56). This similarity in numbers suggests that we are perhaps close to having identified the entirety of SUMOylated proteins, and that iSUMO balances prediction depth (coverage) with accuracy.

The validity of iSUMO predictions is further illustrated when examining human proteins that were not part of our training set (of 14 experimental studies), but scored highly in our framework (**Table 3**). Seven of the ten proteins have reported ubiquitination events – a modification that often coincides with SUMOylation. Four of the proteins in **Table 3** were reported in other SUMOylation studies, three of these in a recent high-quality dataset by Hendriks et al. (22). Most excitingly, one protein has been validated by a targeted experimental study – namely DGCR8 whose SUMOylation at lysine 707 stabilizes the protein by preventing ubiquitination at the same residue (63).

In addition to providing high-quality predictions, our study also highlighted several characteristics that appear to be 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 (64) and RNA-binding proteins are major components of RNA-protein granules whose aggregation forms the molecular bases of many neurodegenerative disorders (65). Therefore, SUMOylation may act to prevent such aggregation in these densely packed cellular structures – a hypothesis supported by some experimental work (66).

These considerations also link to the biomedical relevance of SUMOylation, in particular with respect to neuronal diseases (67-69). Indeed, we find that abnormalities of the nervous system amongst SUMOylated proteins (HPO (70), p-value < 1E-7). Further, one of the highest-scoring iSUMO predictions in human, NUFIP1, interacts with a major neuronal regulator protein, the Fragile X Mental Retardation protein 1 (FXR1), and might therefore be linked to this neuronal disease (**Table 3**).

## 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).

## References


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

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

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

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

5. Li, S. J., and Hochstrasser, M. (2003) The Ulp1 SUMO isopeptidase: distinct domains required for viability, nuclear envelope localization, and substrate specificity. *J Cell Biol* 160, 1069-1081

6. Saracco, S. A., Miller, M. J., Kurepa, J., and Vierstra, R. D. (2007) Genetic analysis of SUMOylation in Arabidopsis: conjugation of SUMO1 and SUMO2 to nuclear proteins is essential. *Plant Physiol* 145, 119-134

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

8. Zhao, X., and Blobel, G. (2005) A SUMO ligase is part of a nuclear multiprotein complex that affects DNA repair and chromosomal organization. *Proc Natl Acad Sci U S A* 102, 4777-4782

9. Girdwood, D. W., Tatham, M. H., and Hay, R. T. (2004) SUMO and transcriptional regulation. *Seminars in cell & developmental biology* 15, 201-210

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

11. Vethantham, V., Rao, N., and Manley, J. L. (2007) Sumoylation modulates the assembly and activity of the pre-mRNA 3' processing complex. *Mol Cell Biol* 27, 8848-8858

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., and Wu, Y. (2014) Important role of SUMOylation of Spliceosome factors in prostate cancer cells. *J Proteome Res* 13, 3571-3582

13. Meier, I. (2012) mRNA export and sumoylation-Lessons from plants. *Biochim Biophys Acta* 1819, 531-537

14. Truong, K., Lee, T. D., Li, B., and Chen, Y. (2012) Sumoylation of SAE2 C terminus regulates SAE nuclear localization. *J Biol Chem* 287, 42611-42619

15. Schimmel, J., Larsen, K. M., Matic, I., van Hagen, M., Cox, J., Mann, M., Andersen, J. S., and Vertegaal, A. C. (2008) The ubiquitin-proteasome system is a key component of the SUMO-2/3 cycle. *Mol Cell Proteomics* 7, 2107-2122

16. Guo, C., and Henley, J. M. (2014) Wrestling with stress: roles of protein SUMOylation and deSUMOylation in cell stress response. *IUBMB Life* 66, 71-77

17. Geiss-Friedlander, R., and Melchior, F. (2007) Concepts in sumoylation: a decade on. *Nat Rev Mol Cell Biol* 8, 947-956

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

19. Zhao, Q., Xie, Y., Zheng, Y., Jiang, S., Liu, W., Mu, W., Liu, Z., Zhao, Y., Xue, Y., and Ren, J. (2014) GPS-SUMO: a tool for the prediction of sumoylation sites and SUMO-interaction motifs. *Nucleic Acids Res* 42, W325-330

20. Teng, S., Luo, H., and Wang, L. (2012) Predicting protein sumoylation sites from sequence features. *Amino Acids* 43, 447-455

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

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

23. Sarge, K. D., and Park-Sarge, O. K. (2009) Detection of proteins sumoylated in vivo and in vitro. *Methods Mol Biol* 590, 265-277

24. Lamoliatte, F., Bonneil, E., Durette, C., Caron-Lizotte, O., Wildemann, D., Zerweck, J., Wenshuk, H., and Thibault, P. (2013) Targeted identification of SUMOylation sites in human proteins using affinity enrichment and paralog-specific reporter ions. *Molecular & cellular proteomics : MCP* 12, 2536-2550

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

26. Galisson, F., Mahrouche, L., Courcelles, M., Bonneil, E., Meloche, S., Chelbi-Alix, M. K., and Thibault, P. (2011) A novel proteomics approach to identify SUMOylated proteins and their modification sites in human cells. *Molecular & cellular proteomics : MCP* 10, M110 004796

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

28. Matic, I., Schimmel, J., Hendriks, I. A., van Santen, M. A., van de Rijke, F., van Dam, H., Gnad, F., Mann, M., and Vertegaal, A. C. (2010) Site-specific identification of SUMO-2 targets in cells reveals an inverted SUMOylation motif and a hydrophobic cluster SUMOylation motif. *Molecular cell* 39, 641-652

29. Grant, M. M. (2010) Identification of SUMOylated proteins in neuroblastoma cells after treatment with hydrogen peroxide or ascorbate. *BMB reports* 43, 720-725

30. Blomster, H. A., Imanishi, S. Y., Siimes, J., Kastu, J., Morrice, N. A., Eriksson, J. E., and Sistonen, L. (2010) In vivo identification of sumoylation sites by a signature tag and cysteine-targeted affinity purification. *The Journal of biological chemistry* 285, 19324-19329

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

32. Blomster, H. A., Hietakangas, V., Wu, J., Kouvonen, P., Hautaniemi, S., and Sistonen, L. (2009) Novel proteomics strategy brings insight into the prevalence of SUMO-2 target sites. *Molecular & cellular proteomics : MCP* 8, 1382-1390

33. Schimmel, J., Larsen, K. M., Matic, I., van Hagen, M., Cox, J., Mann, M., Andersen, J. S., and Vertegaal, A. C. (2008) The ubiquitin-proteasome system is a key component of the SUMO-2/3 cycle. *Molecular & cellular proteomics : MCP* 7, 2107-2122

34. Vertegaal, A. C., Andersen, J. S., Ogg, S. C., Hay, R. T., Mann, M., and Lamond, A. I. (2006) Distinct and overlapping sets of SUMO-1 and SUMO-2 target proteins revealed by quantitative proteomics. *Molecular & cellular proteomics : MCP* 5, 2298-2310

35. Rosas-Acosta, G., Russell, W. K., Deyrieux, A., Russell, D. H., and Wilson, V. G. (2005) A universal strategy for proteomic studies of SUMO and other ubiquitin-like modifiers. *Molecular & cellular proteomics : MCP* 4, 56-72

36. Vertegaal, A. C., Ogg, S. C., Jaffray, E., Rodriguez, M. S., Hay, R. T., Andersen, J. S., Mann, M., and Lamond, A. I. (2004) A proteomic study of SUMO-2 target proteins. *The Journal of biological chemistry* 279, 33791-33798

37. Manza, L. L., Codreanu, S. G., Stamer, S. L., Smith, D. L., Wells, K. S., Roberts, R. L., and Liebler, D. C. (2004) Global shifts in protein sumoylation in response to electrophile and oxidative stress. *Chemical research in toxicology* 17, 1706-1715

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

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

40. Denison, C., Rudner, A. D., Gerber, S. A., Bakalarski, C. E., Moazed, D., and Gygi, S. P. (2005) A proteomic strategy for gaining insights into protein sumoylation in yeast. *Mol Cell Proteomics* 4, 246-254

41. Zhou, W., Ryan, J. J., and Zhou, H. (2004) Global analyses of sumoylated proteins in Saccharomyces cerevisiae induction of protein sumoylation by cellular stresses. *Journal of Biological Chemistry* 279, 32262-32268

42. Wohlschlegel, J. A., Johnson, E. S., Reed, S. I., and Yates, J. R. (2004) Global analysis of protein sumoylation in Saccharomyces cerevisiae. *Journal of Biological Chemistry* 279, 45662-45668

43. Panse, V. G., Hardeland, U., Werner, T., Kuster, B., and Hurt, E. (2004) A proteome-wide approach identifies sumoylated substrate proteins in yeast. *J Biol Chem* 279, 41346-41351

44. Xue, Y., Zhou, F., Fu, C., Xu, Y., and Yao, X. (2006) SUMOsp: a web server for sumoylation site prediction. *Nucleic Acids Res* 34, W254-257

45. Reimand, J., Arak, T., and Vilo, J. (2011) g:Profiler--a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Res* 39, W307-315

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

47. 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., and Tyers, M. (2013) The BioGRID interaction database: 2013 update. *Nucleic Acids Res* 41, D816-823

48. Stark, C., Breitkreutz, B. J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M. S., Nixon, J., Van Auken, K., Wang, X., Shi, X., Reguly, T., Rust, J. M., Winter, A., Dolinski, K., and Tyers, M. (2011) The BioGRID Interaction Database: 2011 update. *Nucleic Acids Res* 39, D698-704

49. Breitkreutz, B. J., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., Oughtred, R., Lackner, D. H., Bahler, J., Wood, V., Dolinski, K., and Tyers, M. (2008) The BioGRID Interaction Database: 2008 update. *Nucleic Acids Res* 36, D637-640

50. Stark, C., Breitkreutz, B. J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006) BioGRID: a general repository for interaction datasets. *Nucleic Acids Res* 34, D535-539

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

52. Ruepp, A., Brauner, B., Dunger-Kaltenbach, I., Frishman, G., Montrone, C., Stransky, M., Waegele, B., Schmidt, T., Doudieu, O. N., Stumpflen, V., and Mewes, H. W. (2008) CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res* 36, D646-650

53. Alvarez-Ponce, D., and McInerney, J. O. (2011) The human genome retains relics of its prokaryotic ancestry: human genes of archaebacterial and eubacterial origin exhibit remarkable differences. *Genome Biol Evol* 3, 782-790

54. Mark Hall, E. F., Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009) The WEKA Data Mining Software: An Update. *SIGKDD Explorations* 11

55. Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005) ROCR: visualizing classifier performance in R. *Bioinformatics* 21, 3940-3941

56. Hart, G. T., Ramani, A. K., and Marcotte, E. M. (2006) How complete are current yeast and human protein-interaction networks? *Genome Biol* 7, 120

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

58. Ouyang, J., Valin, A., and Gill, G. (2009) Regulation of transcription factor activity by SUMO modification. *Methods in molecular biology* 497, 141-152

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

60. Widagdo, J., Taylor, K. M., Gunning, P. W., Hardeman, E. C., and Palmer, S. J. (2012) SUMOylation of GTF2IRD1 regulates protein partner interactions and ubiquitin-mediated degradation. *PLoS One* 7, e49283

61. Holmes, G., Bernhard Pfahringer, Richard Kirkby, Eibe Frank, Mark Hall (2002) Multiclass alternating decision trees. *Machine learning: ECML 2002*, 161-172

62. Kunadt, M., Eckermann, K., Stuendl, A., Gong, J., Russo, B., Strauss, K., Rai, S., Kugler, S., Falomir Lockhart, L., Schwalbe, M., Krumova, P., Oliveira, L. M., Bahr, M., Mobius, W., Levin, J., Giese, A., Kruse, N., Mollenhauer, B., Geiss-Friedlander, R., Ludolph, A. C., Freischmidt, A., Feiler, M. S., Danzer, K. M., Zweckstetter, M., Jovin, T. M., Simons, M., Weishaupt, J. H., and Schneider, A. (2015) Extracellular vesicle sorting of alpha-Synuclein is regulated by sumoylation. *Acta neuropathologica* 129, 695-713

63. Zhu, C., Chen, C., Huang, J., Zhang, H., Zhao, X., Deng, R., Dou, J., Jin, H., Chen, R., Xu, M., Chen, Q., Wang, Y., and Yu, J. (2015) SUMOylation at K707 of DGCR8 controls direct function of primary microRNA. *Nucleic Acids Res* 43, 7945-7960

64. Marblestone, J. G., Edavettal, S. C., Lim, Y., Lim, P., Zuo, X., and Butt, T. R. (2006) Comparison of SUMO fusion technology with traditional gene fusion systems: enhanced expression and solubility with SUMO. *Protein Sci* 15, 182-189

65. Ramaswami, M., Taylor, J. P., and Parker, R. (2013) Altered ribostasis: RNA-protein granules in degenerative disorders. *Cell* 154, 727-736

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

67. Henley, J. M., Craig, T. J., and Wilkinson, K. A. (2014) Neuronal SUMOylation: mechanisms, physiology, and roles in neuronal dysfunction. *Physiol Rev* 94, 1249-1285

68. Martin, S., Wilkinson, K. A., Nishimune, A., and Henley, J. M. (2007) Emerging extranuclear roles of protein SUMOylation in neuronal function and dysfunction. *Nat Rev Neurosci* 8, 948-959

69. Scheschonka, A., Tang, Z., and Betz, H. (2007) Sumoylation in neurons: nuclear and synaptic roles? *Trends Neurosci* 30, 85-91

70. Robinson, P. N., and Mundlos, S. (2010) The human phenotype ontology. *Clin Genet* 77, 525-534

71. Watts, F. Z., Baldock, R., Jongjitwimol, J., and Morley, S. J. (2014) Weighing up the possibilities: Controlling translation by ubiquitylation and sumoylation. *Translation (Austin)* 2, e959366

72. Ustrell, V., Hoffman, L., Pratt, G., and Rechsteiner, M. (2002) PA200, a nuclear proteasome activator involved in DNA repair. *EMBO J* 21, 3516-3525

73. Qian, M. X., Pang, Y., Liu, C. H., Haratake, K., Du, B. Y., Ji, D. Y., Wang, G. F., Zhu, Q. Q., Song, W., Yu, Y., Zhang, X. X., Huang, H. T., Miao, S., Chen, L. B., Zhang, Z. H., Liang, Y. N., Liu, S., Cha, H., Yang, D., Zhai, Y., Komatsu, T., Tsuruta, F., Li, H., Cao, C., Li, W., Li, G. H., Cheng, Y., Chiba, T., Wang, L., Goldberg, A. L., Shen, Y., and Qiu, X. B. (2013) Acetylation-mediated proteasomal degradation of core histones during DNA repair and spermatogenesis. *Cell* 153, 1012-1024

74. Niskanen, E. A., Malinen, M., Sutinen, P., Toropainen, S., Paakinaho, V., Vihervaara, A., Joutsen, J., Kaikkonen, M. U., Sistonen, L., and Palvimo, J. J. (2015) Global SUMOylation on active chromatin is an acute heat stress response restricting transcription. *Genome Biol* 16, 153

75. Satpathy, S., Guerillon, C., Kim, T. S., Bigot, N., Thakur, S., Bonni, S., Riabowol, K., and Pedeux, R. (2014) SUMOylation of the ING1b tumor suppressor regulates gene transcription. *Carcinogenesis* 35, 2214-2223

76. Bonacci, T., Audebert, S., Camoin, L., Baudelet, E., Bidaut, G., Garcia, M., Witzel, II, Perkins, N. D., Borg, J. P., Iovanna, J. L., and Soubeyran, P. (2014) Identification of new mechanisms of cellular response to chemotherapy by tracking changes in post-translational modifications by ubiquitin and ubiquitin-like proteins. *J Proteome Res* 13, 2478-2494

## Figures and Tables

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

We assembled 14 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; each row one human or yeast protein. Grey entries represent proteins observed as SUMOylated by the respective dataset. Numbers at the top of the column are the total number of SUMOylated proteins reported by the study. Datasets are lists in full in **Supplementary Tables S1**.

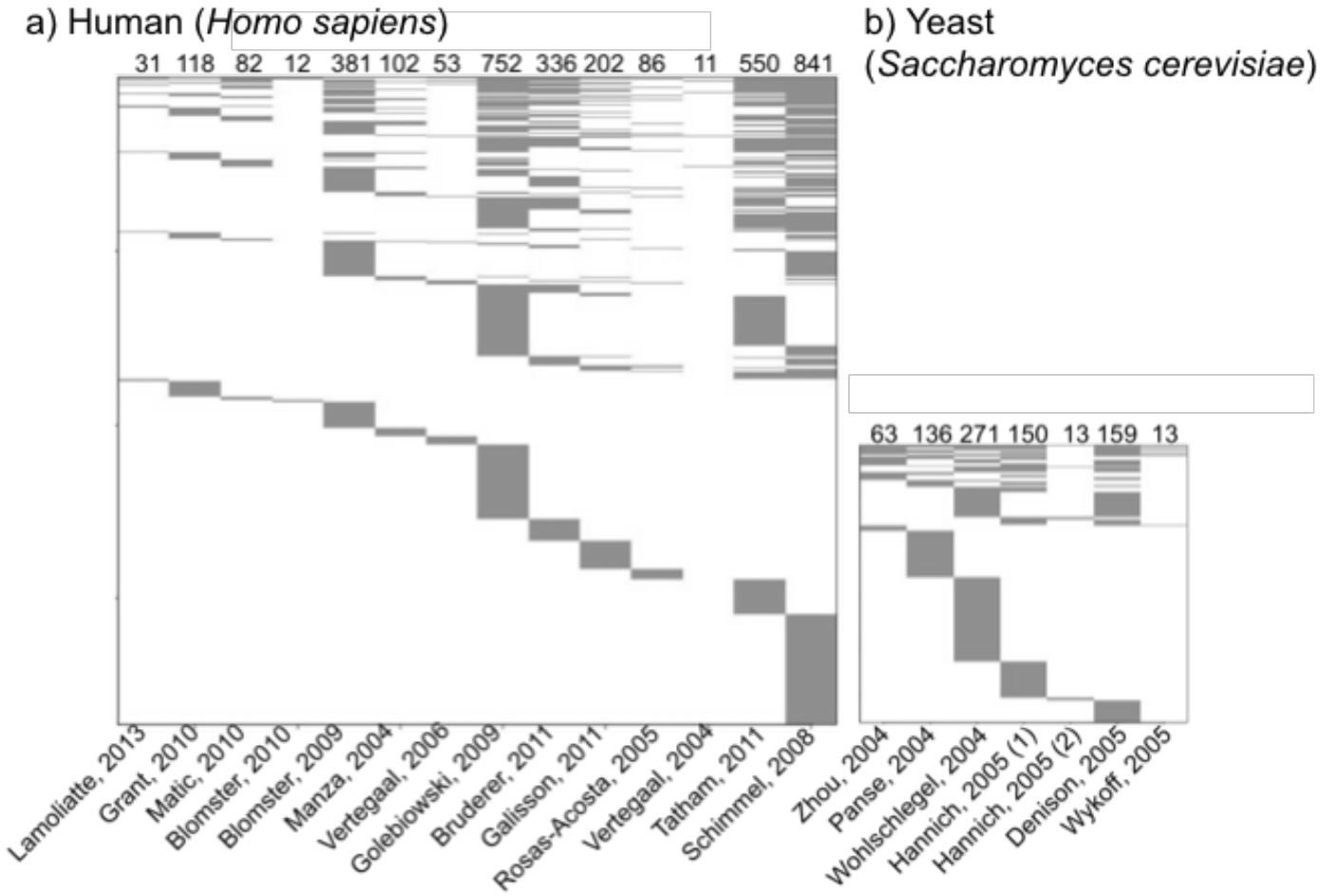

**Figure 2. Correlation between protein SUMOylation, the total number of distinct subunits per complex, and the number subunits that bind RNA.**

Complex information was taken from the CORUM database (52). Dot size is proportional to the total number of distinct subunits of the complexes.

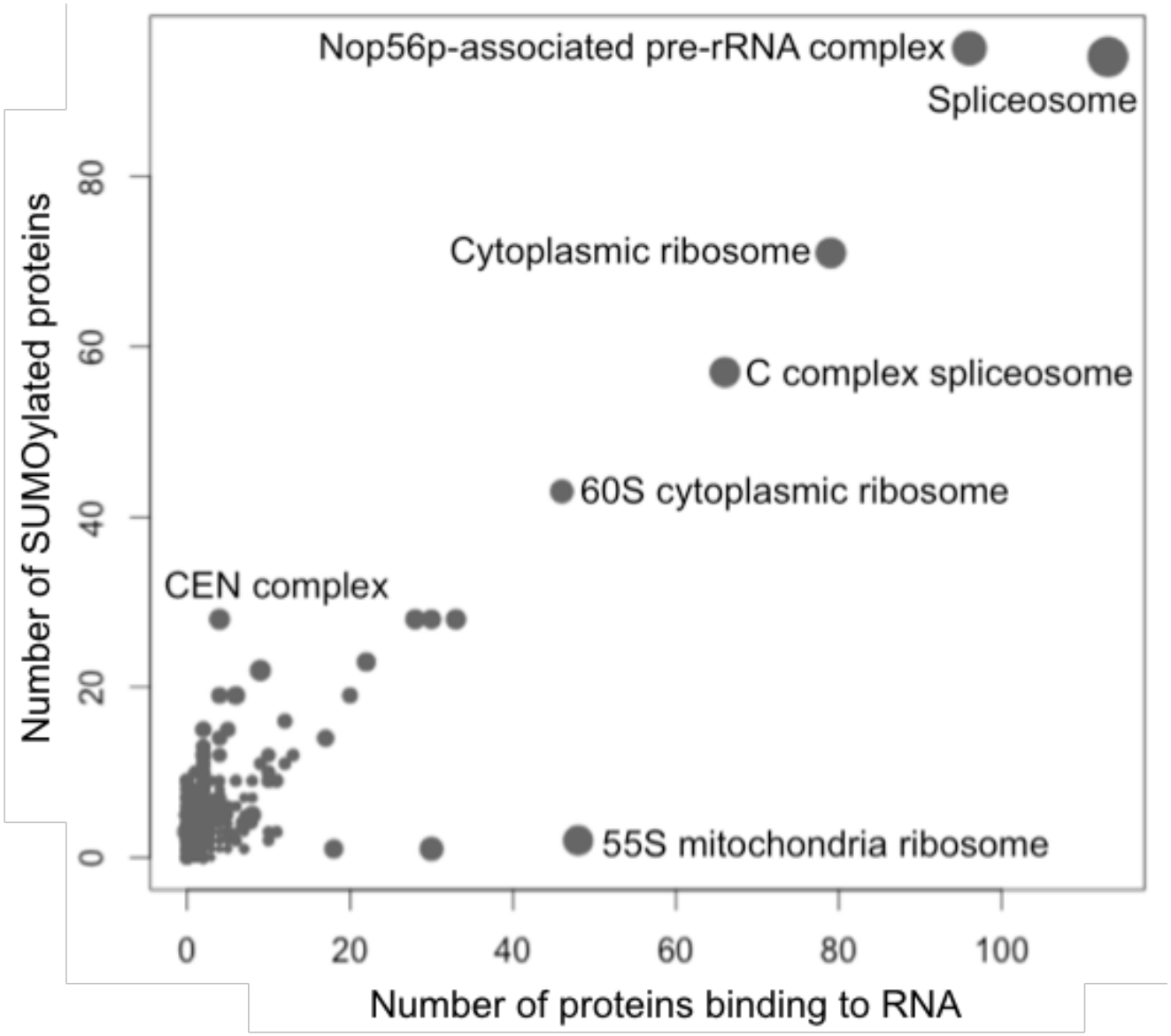

**Figure 3. iSUMO predictions outperform sequence-based predictions**

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 (c) human and (d) yeast, measured as the number of occurrences in the different models. '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.

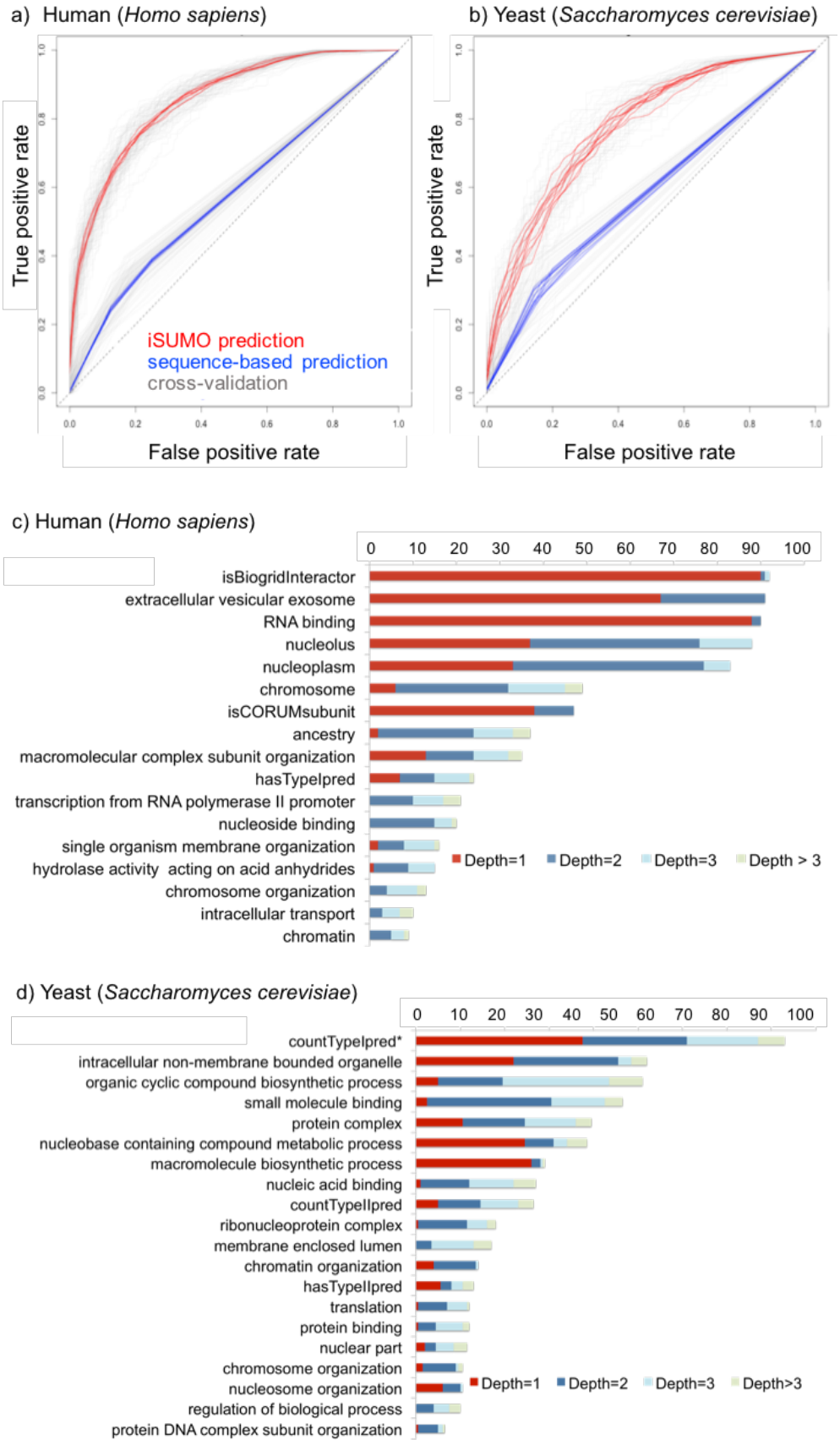
a) Human (Homo sapiens)
b) Yeast (Saccharomyces cerevisiae)
True positive rate
False positive rate
iSUMO prediction
sequence-based prediction
cross-validation
c) Human (Homo sapiens)
isBiogridInteractor
extracellular vesicular exosome
RNA binding
nucleolus
nucleoplasm
chromosome
isCORUMsubunit
ancestry
macromolecular complex subunit organization
hasTypeIpred
transcription from RNA polymerase II promoter
nucleoside binding
single organism membrane organization
hydrolase activity acting on acid anhydrides
chromosome organization
intracellular transport
chromatin
Depth=1
Depth=2
Depth=3
Depth > 3
d) Yeast (Saccharomyces cerevisiae)
countTypeIpred*
intracellular non-membrane bounded organelle
organic cyclic compound biosynthetic process
small molecule binding
protein complex
nucleobase containing compound metabolic process
macromolecule biosynthetic process
nucleic acid binding
countTypeIIpred
ribonucleoprotein complex
membrane enclosed lumen
chromatin organization
hasTypeIIpred
translation
protein binding
nuclear part
chromosome organization
nucleosome organization
regulation of biological process
protein DNA complex subunit organization
Depth=1
Depth=2
Depth=3
Depth>3

**Figure 4. Representative iSUMO decision tree to predict SUMOylation**

Example of an alternating decision tree. The red circles are decision nodes, each containing an attribute name. The blue rectangles are the prediction nodes, which generate scores whenever an instance satisfies an attribute value. The total scores are reported as the iSUMO prediction scores which range between 0 (non-SUMOylated) and 1 (SUMOylated).

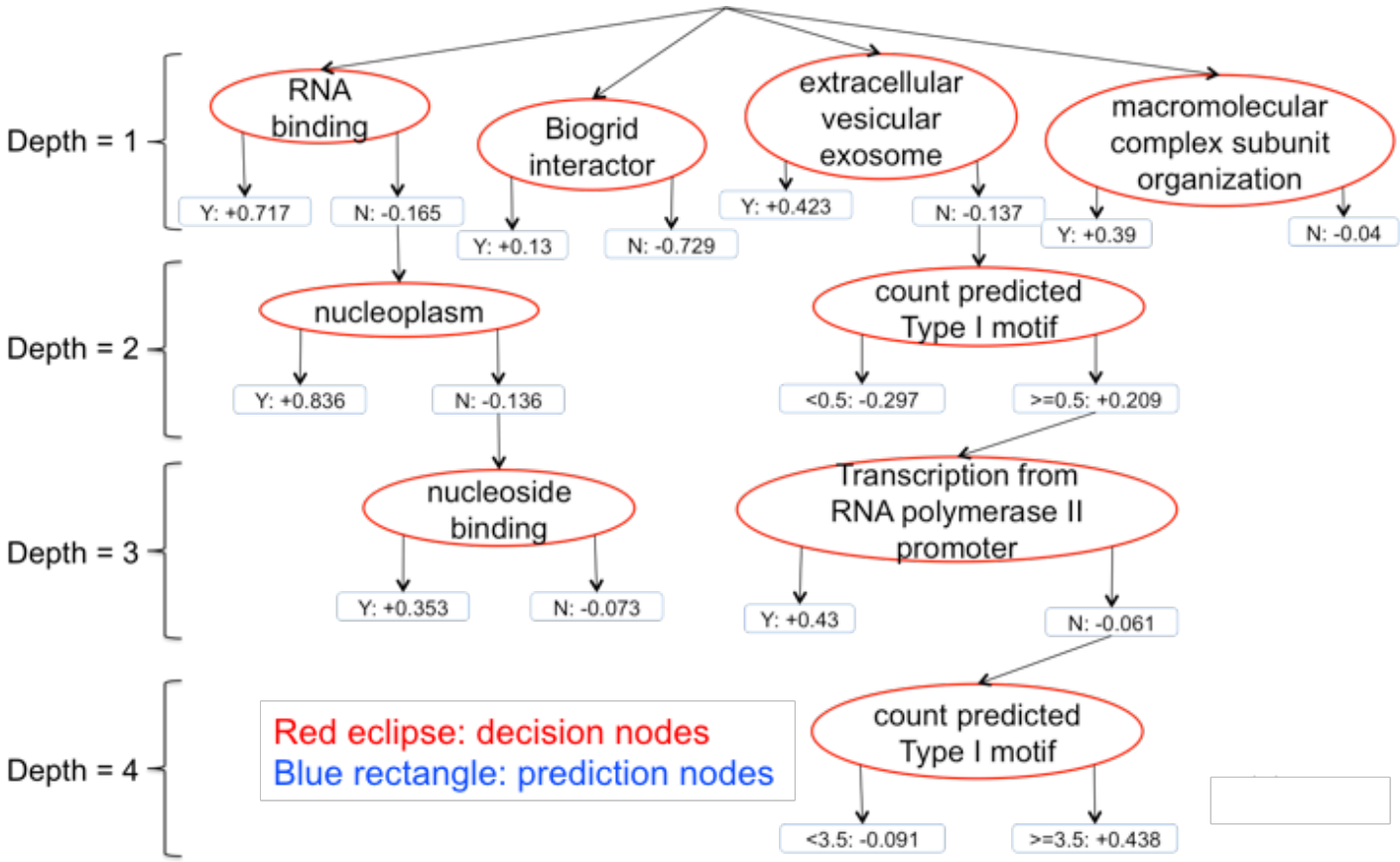

**Figure 5. iSUMO predictions are independently validated.**

The plot validates iSUMO predictions with independent, high-quality data by Hendriks et al. which was not part of the training data used here (22)(solid red line). Importantly, at high iSUMO scores, the fraction of proteins that iSUMO predicts to be SUMOylated but are not part of the Hendriks dataset (blue dashed line) is very small. A 5% false positive rate (FPR) corresponds to an iSUMO prediction score of 0.77, as indicated by the black dotted line. Of the 1,596 proteins predicted to be SUMOylated at this threshold, one third (458) were validated by the Hendriks study (22).

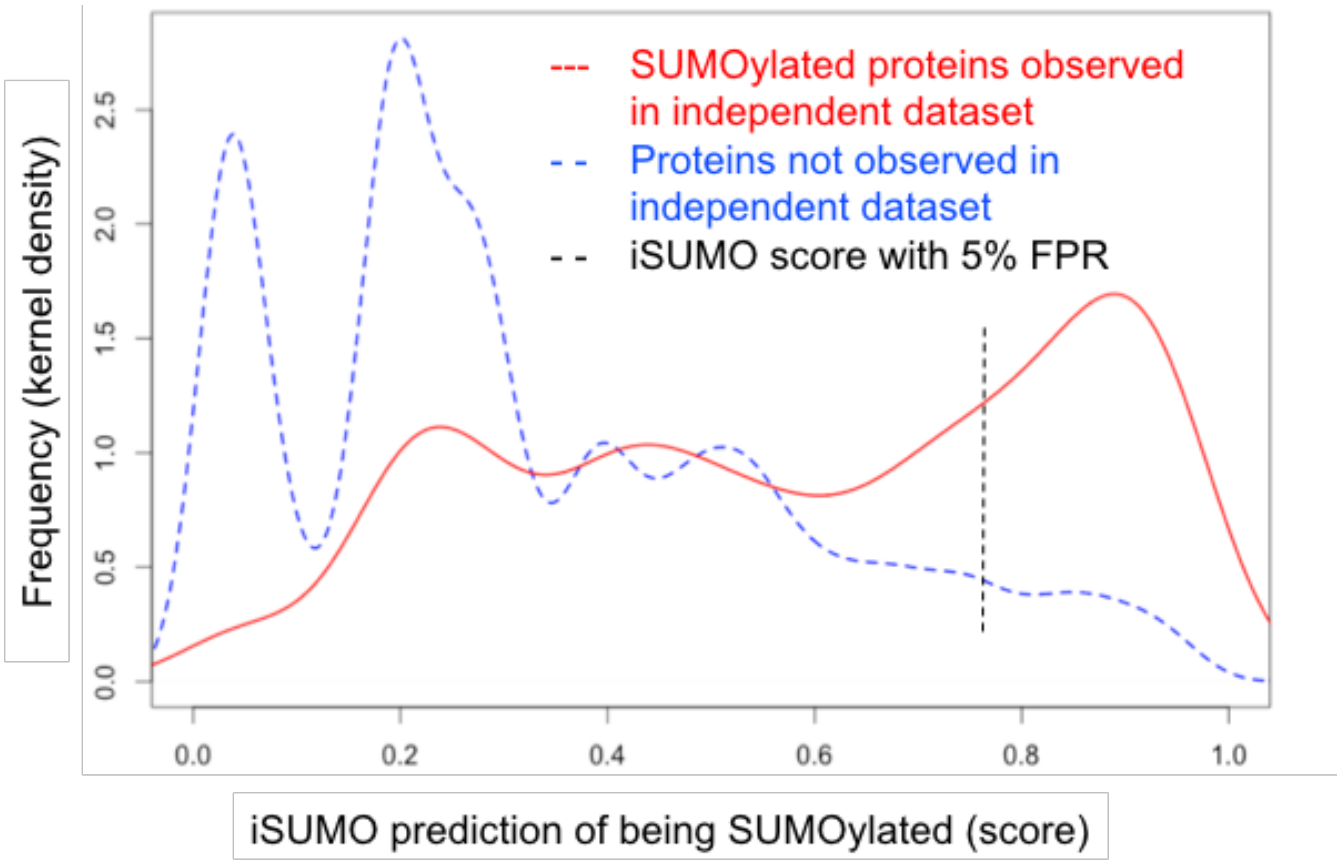

**Table 1. SUMOylated proteins have significant function biases.**

We tested for function enrichment using hypergeometric tests of Gene Ontology (GO) term in (a) human and (b) 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 information is in the **Supplementary Table S2.**

(a) *Homo sapiens*

| Domain | Term name | Term size | Intersection size | Adjusted p-value |
|---|---|---|---|---|
| BP | RNA processing | 680 | 267 | 3E-102 |
| | SRP-dependent co-translational protein targeting to Membrane | 110 | 86 | 1E-64 |
| | DNA metabolic process | 971 | 267 | 1E-62 |
| | Translation | 111 | 83 | 2E-59 |
| | Cytoplasmic transport | 824 | 224 | 1E-50 |
| | Viral transcription | 163 | 93 | 1E-50 |
| | Protein complex disassembly | 194 | 98 | 6E-47 |
| | Chromatin organization | 620 | 184 | 7E-47 |
| | Mitotic cell cycle | 890 | 223 | 7E-44 |
| CC | Ribonucleoprotein complex | 625 | 275 | 2E-120 |
| | Nucleolus | 685 | 252 | 1E-88 |
| | Chromosome | 695 | 243 | 8E-80 |
| | Nucleoplasm part | 643 | 214 | 4E-65 |
| | Cytosolic ribosome | 95 | 71 | 2E-50 |
| | Spliceosomal complex | 145 | 84 | 3E-46 |
| | Nuclear body | 303 | 121 | 9E-45 |
| MF | Poly(A) RNA binding | 1155 | 580 | 0 |
| | RNA binding | 1536 | 654 | 4E-304 |
| | DNA binding | 2356 | 474 | 1E-67 |

(b) *Saccharomyces cerevisiae*

| Domain | Term name | Term size | Overlap size | Adjusted p-value |
|---|---|---|---|---|
| BP | Transcription | 474 | 108 | 2E-23 |
| | Chromatin modification | 93 | 36 | 8E-15 |
| | Protein-DNA complex subunit organization | 61 | 24 | 7E-10 |
| | Nucleosome organization | 31 | 17 | 1E-9 |
| | Translation | 345 | 62 | 4E-08 |
| CC | Nucleoplasm | 122 | 39 | 8E-13 |
| | Protein complex | 901 | 129 | 3E-10 |
| | Chromatin | 70 | 26 | 4E-10 |
| | Ribonucleoprotein complex | 458 | 78 | 3E-09 |
| | Nucleolus | 155 | 33 | 8E-06 |
| MF | Nucleic acid binding | 910 | 158 | 1E-21 |
| | DNA binding | 460 | 96 | 1E-17 |
| | Protein binding | 879 | 120 | 6E-08 |

**Table 2. 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.

| CORUM protein complex | Total number of subunits | Number of SUMOylated subunits | Number of RNA-binding subunits | Adjusted p-value |
|---|---|---|---|---|
| **High SUMOylation** | | | | |
| Spliceosome | 143 | 94 | 113 | 1E-64 |
| Nop56p-associated pre-rRNA complex | 104 | 95 | 96 | 2E-91 |
| Ribosome, cytoplasmic | 81 | 71 | 79 | 5E-65 |
| CEN complex | 37 | 22 | 9 | 3E-13 |
| PA700-20S-PA28 complex | 36 | 28 | 4 | 2E-22 |
| 17S U2 snRNP | 33 | 28 | 28 | 1E-24 |
| CDC5L complex | 30 | 23 | 22 | 5E-18 |
| 26S proteasome | 22 | 19 | 4 | 5E-17 |
| Large Drosha complex | 20 | 19 | 20 | 3E-19 |
| SNW1 complex | 18 | 16 | 12 | 8E-15 |
| **Low SUMOylation** | | | | |
| 55S ribosome, mitochondrial | 78 | 2 | 48 | 1 |
| 39S ribosomal subunit, mitochondrial | 48 | 1 | 30 | 1 |
| Respiratory chain complex I (holoenzyme), mitochondrial | 44 | 2 | 1 | 1 |
| Mediator complex | 32 | 3 | 0 | 1 |
| 28S ribosomal subunit, mitochondrial | 30 | 1 | 18 | 1 |
| BRCA1-RNA polymerase II complex | 26 | 5 | 8 | 1 |
| RNA polymerase II holoenzyme complex | 24 | 4 | 7 | 1 |

**Table 3. Highest-scoring iSUMO-predicted SUMOylation targets that are not reported in training datasets.**

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. None of the proteins listed here have been reported as SUMOylated in the 14 training datasets. Function annotation is from GeneCards.org. * - Independent validation of SUMOylation by observation in Hendriks study (22). ** - SUMOylation observed in other large-scale studies as reported by BIOGRID (47). # - Ubiquitinated as reported by BIOGRID (47). The extended version of this table is presented in **Supplementary Table S6**. Ref. - references

| Protein common names | iSUMO score | Function | Evidence for SUMOylation | Ref. |
|---|---|---|---|---|
| **40S ribosomal protein S19 RPS19** | 0.97 | Part of ribosome complex | No direct evidence for SUMOylation. However, during ribosome biogenesis, pre-ribosomal particles are SUMOylated - mainly during 60S, but also 40S biogenesis. Many translation initiation and elongation factors are also SUMOylated. | (71) |
| **Proteasome subunit alpha type-1*,**,#, PSMA1** | 0.96 | Part of proteasome complex | No direct evidence for SUMOylation. The proteasome is not known to interact with nucleic acids directly, but one particle (PA200 in human, Blm10 in yeast) regulates DNA/expression via interaction with histones. | (12, 47, 72, 73) |
| **E3 ubiquitin-protein ligase#, HUWE1** | 0.96 | Protein ubiquitination for proteasomal degradation | No direct evidence for SUMOylation. HUWE1 interacts with the proteasome and the DNA-binding CTCF complex. CTCF is SUMOylated. | (74) |
| **Cleavage stimulation factor subunit 3#, CSTF3** | 0.96 | Part of cleavage factor stimulation complex promoting polyadenylation and cleavage of pre-mRNAs | No direct evidence for SUMOylation, neither for CSTF3, nor for other complex members. | n/a |
| **Renal Carcinoma Antigen#, NY-REN-24, CACTIN** | 0.96 | Part of spliceosome complex, RNA binding | Not direct evidence for SUMOylation. However, spliceosome and splicing factors are known to be SUMOylated. | (12, 47) |
| **Microprocessor complex subunit*, DGCR8** | 0.95 | Part of microprocessor complex which mediates biogenesis of miRNAs | Direct validation of SUMOylation of DGCR8 in a recent publication: DGCR8 is modified at K707 by SUMO1 which is thought to stabilize the protein via blocking ubiquitination at the same site and preventing degradation. SUMOylation also enhances the affinity of DGCR8 to pri-miRNAs. DGCR8 SUMOylation is linked to tumorignesis and tumor cell migration. | (63, 75) |
| **Cyclin-T1**,#, CCNT1/CDK9** | 0.95 | Part of transcription elongation factor p-TEFb (complex) | No direct evidence for SUMOylation. However, SUMOylation affects many transcription factors and parts of the transcription initiation and elongation complex. | (47, 76) |
| **Exportin-5#, XPO5** | 0.95 | Nuclear export of small RNAs and RNA-binding proteins | No direct evidence for SUMOylation, but many nuclear export factors are SUMOylated. | (12, 47) |

| | | | | |
|---|---|---|---|---|
| **Nuclear fragile X mental retardation-interacting protein 1*,#, NUFIP1** | 0.95 | Nuclear, RNA-binding protein that interacts with the fragile X mental retardation protein | Not direct evidence for SUMOylation. | n/a |
| **40S ribosomal protein S28, RPS28** | 0.95 | Part of ribosome complex | See RPS19 above. | n/a |