Supplementary Note

Statistical validation of reproducibility

Pearson correlations (R) were calculated with Perseus, using scatter plot analysis on the unfiltered site and protein data and individually comparing all biological replicates to each other on a per-site or per-protein basis. Internal correlation refers to correlation within same-condition replicates, and external correlation refers to correlation between different-condition samples. Principle Component Analysis was performed using Perseus, using the filtered and imputed data as input. For heatmap generation, the filtered and imputed data was Z-scored using Perseus, and then subjected to Euclidean hierarchical clustering.

IceLogo generation

For SUMOylation site analysis of all identified sites, amino acid sequence windows of 15 amino acids downstream as well as 15 amino acids upstream of the modified lysine were extracted from the corresponding proteins, generating 31 amino acid (31AA) sequence windows. IceLogo software version 1.2 [8] was used to overlay 31AA SUMOylation site sequence windows in order to generate a consensus sequence, and compensated against the expected random occurrence frequencies of amino acids across all human proteins (iceLogo). Alternatively, subsets of modification sites were compared directly to other subsets of modification sites, generating consensus sequences showing differential occurrence of amino acids between the subsets (SubLogo). Heat maps were generated in a similar fashion to IceLogos. For IceLogos, SubLogos and heat maps, all amino acids shown as enriched or depleted are significant with $P < 0.05$, as determined by IceLogo software version 1.2.

Secondary structure analysis

31AA sequence windows corresponding to SUMOylation sites were analyzed using NetSurfP version 1.1 [9] in order to predict localized protein surface accessibility and secondary structure. For each amino acid within the 31AA windows, probabilities for alpha-helix, beta-strand, and coil were calculated. Additionally, amino acids were predicted to be buried or solvent-exposed, with a threshold of 25% exposure. In addition to calculating the properties of the central lysine, the average properties of the central lysine +/- 5 AA were calculated, as well as the average properties of the entire 31AA sequence windows. As a reference set, random selection of lysines from SUMOylated proteins was performed by dividing all SUMOylated proteins into 60 amino acid sequences, and selecting the first lysine in each sequence. 31AA sequence windows were assigned to half of these random sites (5,726 out

of 11,451). Duplicate sequence windows were discarded. The reference set was processed identically as compared to the SUMOylation sites set.

SUMOylation and PTM site overlap analysis

For comparative analysis, all 103 SUMOylation sites identified by Matic et al. [5], all 202 SUMOylation sites identified by Schimmel et al. [6], and all 1,002 SUMOylation sites identified by Tammsalu et al. [4] were assigned to matching Uniprot IDs and 31AA sequence windows were parsed. Furthermore, all unique human PhosphoSitePlus (PSP; PhosphoSitePlus®, www.phosphosite.org, [10]) SUMOylation sites were used (639), including MS/MS-identified sites as well as identifications made with low-throughput methodology. Additionally, 26,345 MS/MS-identified ubiquitylation sites, 7,463 acetylation sites, and 812 lysine-methylation sites were extracted from PSP, and 31AA sequence windows were assigned. From within each data set, duplicate sequence windows were removed. Perseus software was used to generate a matrix where all sequence windows from all PTMs were cross-referenced to each other. Corresponding parental proteins were assigned, and multiple modifications targeting the same lysines within the same proteins were further investigated using STRING network analysis.

Term enrichment analysis

Statistical enrichment analysis for protein and gene properties was performed using Perseus software. The human proteome was annotated with Gene Ontology terms [11], including Biological Processes (GOBP), Molecular Functions (GOMF), and Cellular Compartments (GOCC). Additional annotation was performed with the Kyoto Encyclopedia of Genes and Genomes (KEGG)[12], Protein families (Pfam)[13], Gene Set Enrichment Analysis (GSEA)[14], Keywords, and Comprehensive Resource of Mammalian Protein Complexes (CORUM)[15] terms for comparative enrichment analysis. SUMOylated proteins or subgroups of SUMOylated proteins were compared by annotation terms to the entire human proteome, using Fisher Exact Testing. Lysines modified by SUMO or other PTMs were compared against a background of the total amount of lysines in the human proteome, calculated from the average lysine occurrence multiplied by the amount of proteins and the average protein size in the human proteome, using Fisher Exact Testing. Benjamini and Hochberg FDR was applied to *P* values to correct for multiple hypotheses testing, and final corrected *P* values were filtered to be less than 2%. In the figures, categories are either ranked by the negative $\log_{10}$ of the *P* value, or by a relative score calculated as follows: $\log_{10}$(*P* value) * ($\log_2$(enrichment ratio))^5.

STRING network analysis

STRING network analysis was performed using the online STRING database [16], using all SUMOylated proteins or subgroups of SUMOylated proteins as input. Enrichment analysis was performed allowing network interactions at high or greater confidence (p>0.7). *P* values corresponding to the individual analyses were directly taken from the STRING database output. Protein interaction enrichment was performed based on the amount of interactions in the networks, as compared to the randomly expected amount of interactions, with both variables directly derived from the STRING database output. Network participation was measured as a percentage of the proteins which was connected to the core cluster. The total network interconnectivity score was calculated by multiplication of the interaction enrichment, the network participation, and the average STRING confidence of all individual interactions. Visualization of interaction networks was performed using Cytoscape version 3.0.2 [17], and highly interconnected sub-clusters were localized using the Cytoscape plugin Molecular Complex Detection (MCODE) version 1.4.0-beta2 [18]. For sub-cluster localization the following settings were used; a degree cut-off of 3, a node score cut-off of 0.1, a maximum depth of 2, a K-Core of 5, and haircut.

Phylogenetic conservation analysis

Perseus software was used to annotate the human proteome using phylogenetic conservation scores from the Ensembl Database [19], which contains phylogenetic orthologous information ranging over many eukaryotic species. Only transcripts and genes marked as 'known' were extracted from the Ensembl Database, and only unique results were retrieved. Phylogenetic conservation target scores from 62 eukaryotic species were mapped by their human Ensembl Protein ID to human Uniprot IDs. Subsequently, proteins identified as SUMOylated in this work under standard growth conditions were aligned to the annotated human proteome. All MS/MS-identified phosphorylation, acetylation, ubiquitylation and methylation sites were extracted from PSP, mapped to their respective parental proteins, and aligned to the annotated human proteome. Phylogenetic conservation scores were calculated for the entire human proteome as compared to all individual 62 eukaryotic species as a reference, and scores were separately calculated for SUMOylation in addition to all other PTMs. For phylogenetic conservation within orthologues, all human proteins lacking an orthologue were excluded from the analysis on a per-species basis. For phylogenetic conservation outside of orthologues, all human proteins lacking an orthologue were set to 0% conservation. Statistical significance between different phylogenetic conservation rates of the PTM and total protein groups was calculated using two-tailed paired Student's *t* test, using the average conservation scores of all 62 eukaryotic organisms. As such, the resulting *P* values are indicative of significance of difference across the entire eukaryotic dataset.

SUMO target protein overlap analysis

For SUMO target protein analysis, all proteins identified in this work with at least one SUMO site were selected. Proteins successfully identified by various peptides, but lacking a SUMO-peptide, were ignored. For comparative analysis, identified SUMO target proteins were compared to other studies. SUMO-2 target proteins from Becker et al. [3] were selected by a SUMO-2 / Control ratio of greater than 2, and additionally filtered for a SUMO-2 intensity of greater than 10% as compared to SUMO-1 and control. All SUMO target proteins (SUMO-1 and SUMO-2) from Becker et al. were selected by a SUMO-1 / Control ratio or a SUMO-2 / Control ratio of greater than 2. SUMO-2 target proteins from Golebiowski et al. [1] were selected for a SUMO-2 / Control SILAC ratio of greater than 1.5. Heat-shock inducible SUMO-2 target proteins from Golebiowski et al. were filtered by a SUMO-2-HEAT / SUMO-2 SILAC ratio of greater than 1.5, in addition to a (SUMO-2-HEAT / SUMO-2 SILAC ratio) * (SUMO-2 / Control SILAC ratio) of greater than 1.5. Putative poly-SUMO-modified proteins from Bruderer et al. [2] were selected as all identified proteins. Putative poly-SUMO-modified proteins from Bruderer et al. were additionally filtered for an observed molecular weight shift of 2 times the molecular weight of SUMO or greater, as compared to the expected protein molecular weight. SUMO-2 target proteins from Matic et al. [5] were considered to be all proteins in which at least one site of SUMOylation was identified. SUMO-2 target proteins from Schimmel et al. [6] were selected as all proteins with a SUMO-2 / Control SILAC ratio of over 2. SUMO-2 target proteins from Tammsalu et al. [4] were considered to be all proteins in which at least one SUMOylation site was identified. Where required, gene IDs were mapped to the corresponding Uniprot IDs. Additionally, where multiple Uniprot IDs were listed for a singular protein identification, a major Uniprot ID was selected by selecting the first Uniprot ID in the list starting with a P, or otherwise the first Uniprot ID starting with a Q, or otherwise the first Uniprot ID in the list. Perseus [20] software was used to generate a complete gene list for all known human proteins, and all identified SUMO target proteins from our study as well as the above-mentioned studies were aligned based on matching Uniprot IDs.

Supplementary References

1. Golebiowski,F. *et al.* System-wide changes to SUMO modifications in response to heat shock. *Sci. Signal.* **2**, ra24 (2009).

2. Bruderer,R. *et al.* Purification and identification of endogenous polySUMO conjugates. *EMBO Rep.* **12**, 142-148 (2011).

3. Becker,J. *et al.* Detecting endogenous SUMO targets in mammalian cells and tissues. *Nat. Struct. Mol. Biol.* **20**, 525-531 (2013).

4. Tammsalu,T. *et al.* Proteome-Wide Identification of SUMO2 Modification Sites. *Sci. Signal.* **7**, rs2 (2014).

5. Matic,I. *et al.* Site-specific identification of SUMO-2 targets in cells reveals an inverted SUMOylation motif and a hydrophobic cluster SUMOylation motif. *Mol. Cell* **39**, 641-652 (2010).

6. Schimmel,J. *et al.* Uncovering SUMOylation Dynamics during Cell-Cycle Progression Reveals FoxM1 as a Key Mitotic SUMO Target Protein. *Mol. Cell*(2014).

7. Roukens,M.G. *et al.* Identification of a new site of sumoylation on Tel (ETV6) uncovers a PIAS-dependent mode of regulating Tel function. *Mol. Cell Biol.* **28**, 2342-2357 (2008).

8. Colaert,N., Helsens,K., Martens,L., Vandekerckhove,J., & Gevaert,K. Improved visualization of protein consensus sequences by iceLogo. *Nat. Methods* **6**, 786-787 (2009).

9. Petersen,B., Petersen,T.N., Andersen,P., Nielsen,M., & Lundegaard,C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC. Struct. Biol.* **9**, 51 (2009).

10. Hornbeck,P.V. *et al.* PhosphoSitePlus: a comprehensive resource for investigating the structure and function of experimentally determined post-translational modifications in man and mouse. *Nucleic Acids Res.* **40**, D261-D270 (2012).

11. Ashburner,M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25-29 (2000).

12. Kanehisa,M., Goto,S., Sato,Y., Furumichi,M., & Tanabe,M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* **40**, D109-D114 (2012).

13. Punta,M. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **40**, D290-D301 (2012).

14. Subramanian,A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A* **102**, 15545-15550 (2005).

15. Ruepp,A. *et al.* CORUM: the comprehensive resource of mammalian protein complexes. *Nucleic Acids Res.* **36**, D646-D650 (2008).

16. Franceschini,A. *et al.* STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res.* **41**, D808-D815 (2013).

17. Shannon,P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498-2504 (2003).

18. Bader,G.D. & Hogue,C.W. An automated method for finding molecular complexes in large protein interaction networks. *BMC. Bioinformatics.* **4**, 2 (2003).

19. Flicek,P. *et al.* Ensembl 2013. *Nucleic Acids Res.* **41**, D48-D55 (2013).

20. Cox,J. & Mann,M. 1D and 2D annotation enrichment: a statistical method integrating quantitative proteomics with complementary high-throughput data. *BMC. Bioinformatics.* **13 Suppl 16**, S12 (2012).