Supplementary information

Plasmid constructs

The following plasmids were used in mammalian cell transfections. pEGR-1-Luc (kindly provided by Ian Stratford), pCH110, pG5-E1B-Luc and pLexA-G5-E1a-Luc, were described previously (Yang et al., 2004). pAS883 [GAL-MEF2A(266-413)], pAS2076 [pCDNA3-Flag-MEF2A(1-498)] (Yang et al., 1999), pAS383 [pCMV5-Flag-Elk-1], pAS2054 [pCMV5-Flag-Elk-1(K230,249R)], pAS1553 [pCMV-GAL-Elk(223-428)], pAS2066 [pCMV-GAL-Elk(223-428)(K230,249R)] (Yang et al., 2003 and 2004), pcDNA3-HA-SUMO-1 and pCDNA3-ubc9(C35S) [DN-ubc9] (kindly provided by R. Hay), pSHP-luc (containing the *SHP* promoter linked to a luciferase gene), pCDNA3-myc-LRH-1(WT) and pCDNA3-myc-LRH-1(K224R) (kindly provided by I. Talianidis; Chalkiadaki and Talianidis, 2005) were described previously. pFlag-ARIP3(1-572) (PIASxα), (kindly provided by J. Palvimo), was described previously (Kotaja et al., 2002).

pAS2420 [pCDNA3-Flag-MEF2A(1-498)(S400A/D404A)], pAS2279 [pCDNA3-Flag-MEF2A(1-498)(S400A)], pAS2426 [pCDNA3-Flag-MEF2A(1-498)(S400A/R403E)], pAS2418 [pCMV5-Flag-Elk-1(K254R)], pAS2419 [pCMV5-Flag-Elk-1(E3A)], pAS2421 [pCDNA3-myc-LRH-1(D229A)] and pAS2422 [pCDNA3-myc-LRH-1(D229A/E236A)] were constructed by Quikchange mutagenesis (Stratagene) using the following primer pairs-template combinations ADS1394/ADS1395 & pAS2279, ADS1536/ADS1537 & pAS2076, ADS1385/ADS1386 & pAS2076, ADS1542/ADS1543 & pAS383, ADS1544/ADS1545 & pAS383, ADS1538/ADS1539 & pCDNA3-myc-LRH-1 and ADS1540/ADS1541 & pAS2421 respectively. pAS2437 [pCDNA3(myc6)-SUMO-1]

was constructed by ligating a XhoI/XbaI cleaved PCR fragment (created with primer pairs ADS1578/ADS1579 and template pGEX2T-SUMO-1{kindly provided by Ron Hay}) into the same sites of pCDNA3.

The following series of plasmids based on pCMV-GAL-Elk(223-428), with the indicated mutations were created by Quikchange mutagenesis using the indicated primer-template combinations: pAS2403(K230R), ADS1546/ADS1547 & pAS1553; pAS2404(K249R), ADS1548/ADS1549 & pAS1553; pAS2405(K254R), ADS1550/ADS1551 & pAS1553; pAS2406(K249,254R), ADS1550/ADS1551 & pAS2404; pAS2407(K230,254R), ADS1550/ADS1551 & pAS2403; pAS2408(E3A), ADS1544/ADS1545 & pAS1553; pAS2409(K230R/E3A), ADS1546/ADS1547 & pAS2408; pAS2410(K249R/E3A), ADS1548/ADS1549 & pAS2408; pAS2411(K230,249R/E3A), ADS1546/ADS1547 & pAS2410.

The following series of plasmids based on pCMV-GAL-MEF2A(266-413), with the indicated mutations were created by Quikchange mutagenesis using the indicated primer pairs on the template pAS883: pAS2269(S400A), ADS1375/ADS1376; pAS2270(S400E), ADS1377/ADS1378; pAS2277(S400A/R403E), ADS1385/ADS1386; pAS2273(R403E), ADS1383/ADS1384.

For bacterial expression, pAS2069 [encoding GST-Elk-1(201-260)] (Yang and Sharrocks, 2004), pAS2097 (encoding GST fused to full-length PIASxα) (Yang and Sharrocks, 2006), pGEX2T-RanGAP-1(418-587) (Tatham et al., 2003 {kindly provided by Ron Hay}) were described previously. pAS2423 [encoding GST-Elk-1(201-260)(E3A)] was constructed by ligating a BamHI/XbaI cleaved PCR fragment (created with primer pairs ADS1534/ADS1535 and template pAS728) into the same sites of pGEX-KG. pAS2438 [encoding GST-LRH-1(134-257)] and pAS2439

[encoding GST-LRH-1(134-257)(D229A/E236A)] were created by inserting BamHI/XbaI-cleaved PCR fragments [generated using primers ADS1623/ADS1624 and templates pCDNA3-myc-LRH-1(WT) and pAS2422 respectively] into the same sites of pGEX-KG.

pAS2425 [encoding GST-myc_6-Ubc9] was created by cloning the BamHI/XbaI fragment from pAS2424 into the same sites in pGEXKG. pAS2424 [encoding CMV driven myc_6-Ubc9] was constructed by ligating an EcoRI/XbaI-cleaved PCR fragment (created with primer pairs ADS1560/ADS1561 and template pGEX2T-Ubc9 {kindly provided by Ron Hay}) into the same sites in pCDNA3$(myc)_6$ (kindly provided by Alan Whitmarsh). pGEX2T-Ubc9 has been described previously (Desterro et al., 1997). The following untagged Ubc9 mutants were created by Quikchange mutagenesis using the indicated primer pairs on the template pGEX2T-Ubc9: pAS2427(R13E/K14E), ADS1568/ADS1569; pAS2428(R17E/K14E), ADS1570/ADS1571; pAS2429(K48E/K49E), ADS1572/ADS1573; pAS2430(R59E/K61E), ADS1574/ADS1575. pAS2435 [encoding GST-Ubc9(K146A/R147A)] and pAS2436 [encoding GST-Ubc9(R153A/K154A)] were created by ligating BamHI/StuI-cleaved products from pGEX2T-Ubc9 into the same sites of pAS2433 and pAS2434 respectively. pAS2433 [encoding GST-myc_6-Ubc9(R146A/K147A)] and pAS2434 [encoding GST-myc_6-Ubc9(R153A/K154A)] were created by ligating BamHI/XbaI-cleaved products from pAS2431 and pAS2432 respectively into the same sites of pGEX-KG. pAS2431 [encoding myc_6-Ubc9(R146A/K147A)] and pAS2432 [encoding myc_6-Ubc9(R153A/K154A)] were created by ligating EcoRI/XbaI-cleaved PCR products (ADS1560/ADS1576 and ADS1560/ADS1577 respectively on template pAS2424) into the same sites of pCDNA3$(myc)_6$.

Details of PCR primers are available on request.

Bioinformatics approaches

The total number of potential SUMO sites within human proteins conforming to ψKxE (where ψ is F, I, L, M or V) was determined using ScanProsite (Gattiker et al 2002) to search the Swiss-Prot database (Version 50.1). A control data set was derived by searching with 10 shuffled variations of the core ψKxE motif ψExK, xψKE, xψEK, KψEx, KxψE, KxEψ, KEψx, EψKx, ExKψ, EKψX) and averaging the occurrence of particular amino acids downstream from these motifs.

Swiss-Prot was also searched using the NDSM extended SUMO module (the ψKxE motif with at least two acidic amino acids in the 10 amino acid downstream region, at least one of which is between positions 3-6).

For searching the Gene Ontology (GO) database, GOTermFinder (http://go.princeton.edu/cgi-bin/GOTermFinder/GOTermFinder; Boyle et al 2004) was used to find GO terms that were significantly enriched. We compared the dataset derived using the NDSM, with that derived using an alternative dataset generated using 10 shuffled variations of the core ψKxE motif (ψExK, xψKE, xψEK, KψEx, KxψE, KxEψ, KEψx, EψKx, ExKψ, EKψX) with at least two amino acids in the 10 amino acid downstream region, at least one of which is between positions 3-6. A chi-squared test was used to determine gene ontology categories that were significantly upregulated in the NDSM data set relative to this “shuffled” dataset.

Supplementary references

Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, Sherlock G (2004) GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics*. **20:** 3710-3715.

Desterro JM, Thomson J, Hay RT (1997) Ubch9 conjugates SUMO but not ubiquitin. *FEBS Lett*. **417**: 297-300.

Gattiker A, Gasteiger E, Bairoch A (2002) ScanProsite: a reference implementation of a PROSITE scanning tool. *Appl Bioinformatics*. **1:** 107-108.

Yang SH, Galanis A, Sharrocks AD (1999) Targeting of p38 mitogen-activated protein kinases to MEF2 transcription factors. *Mol Cell Biol*. **19:** 4028-4038.

Figure and table legends

Figure S1. Over-representation of acidic residues in the tail region of the SUMO consensus module. (A) The number of acidic and basic residues found at each position in the tail region of 94 experimentally verified SUMO sites (left hand graph) compared to the tail regions of all sites conforming to the consensus ψKxE (right hand graph). The dotted line represents the expected number of residues based on the average distribution of basic (13.8%) and acidic (14.6%) residues in a control dataset derived from the tail regions of shuffled ψKxE motifs. (B) Net charge at each position in the tail region of 94 experimentally verified SUMO sites (left hand graph) compared to the tail regions of all sites conforming to the consensus ψKxE (right hand graph). The net charge was calculated by assigning a value of "-1" to acidic residues and "+1" to basic residues, and summing the total charge found at each position in the dataset. (C) The distribution of acidic regions in the tail region. The percentage of sites containing the indicated numbers of acidic residues in the tail region is shown for the dataset containing the 94 experimentally verified SUMO sites (green) and the dataset containing all predicted ψKxE sites (grey). (D) Amino acid composition of the tail region. The percentage contribution of each amino acid to the tail region of 94 experimentally verified SUMO sites (green) was compared to the datasets derived from all the human proteins found in the SWISSPROT database containing either all predicted ψKxE sites (dark grey), or the average of ten datasets containing shuffled versions of the ψKxE motif (white).

Figure S2. Elk-1 does not bind to SUMO. GST pulldown assays were carried out with GST-PIASxα, or the indicated GST-Elk-1(201-260) derivatives and SUMO-1.

All proteins were expressed and purified from bacteria. Inputs of the GST fusion proteins (bottom panel; anti-GST IB) and SUMO-1 (lane 1- 5%) are shown. Bound SUMO was detected by anti-SUMO-1 IB. PIASxα binds to SUMO but no detectable binding of to Elk-1 was observed.

Figure S3. The basic patch on Ubc9 is not involved in SUMO loading. SUMO loading onto wild-type and K59E/R61E mutant Ubc9 (100 ng) was monitored in the presence of E1 (150 ng) and SUMO (100 ng). Ubc9 bound to SUMO through a thioester linkage was revealed by the presence of a lower mobility band (SUMO-S-Ubc9). Modified and unmodified Ubc9 was detected by anti-Ubc9 IB. DTT was added where indicated to remove thio-ester linkages. Equivalent amounts of SUMO loading is apparent of wild-type Ubc9 and Ubc9(K59E/R61E)(compare lanes 3 and 5).

Table S1. List of verified SUMO modified sites in human transcriptional regulators that contain the core SUMO motif ψKxE. For inclusion in this list, experimental verification of modification of the SUMO site has to be obtained and a role of the protein in transcriptional regulation demonstrated. The location of the SUMO site modules is shown in columns C and D and their sequence is shown in column E, and in more detail in columns G-AE, where only the acidic (coloured red) and basic residues (coloured blue) in the tail region are depicted. In addition, serine or threonine residues that form part of an SP motif and hence are potentially phosphorylated by proline-directed kinases are also shown in orange.

Table S2. List of human proteins containing the core SUMO motif ψKxE. The location of the SUMO site modules is shown in columns C and D and their sequence is shown in column E, and in more detail in columns G-T, where only the acidic (coloured red) and basic residues (coloured blue) in the tail region are depicted. In addition, serine or threonine residues that form part of an SP motif and hence are potentially phosphorylated by proline-directed kinases are also shown in orange.

Table S3. List of human proteins containing the extended NDSM SUMO module. Human substrates containing the extended module consisting of the core SUMO site ψKxE and at least two downstream residues in the downstream 10 amino acid tail region (at least one of which is in the acidic patch between residues 3-6) are listed. The columns contain the same information as described in Table S2.

Table S4. Enrichment of Gene ontology categories in substrates containing the extended NDSM SUMO module. The percentage of proteins in a control dataset (total 6288 proteins) and in the NDSM-predicted dataset (total 2605 proteins) which fall within the indicated gene ontology categories are shown. The control dataset was generated with a σηυφφλεδ ψKxE motif with a comparable distribution of downstream acidic residues. The final column represents the enrichment of the NDSM dataset as fold increase over the control dataset. This ensures that there is no bias for amino acid composition of the compared datasets which is evident if compared to the total protein content encoded by the genome. P values are generated using a chi squared test. The data shown in red show those categories with statistically significant enrichment compared to the control dataset ($p<0.05$). The top four scores under each overarching GO term are shown in Fig. 7B.

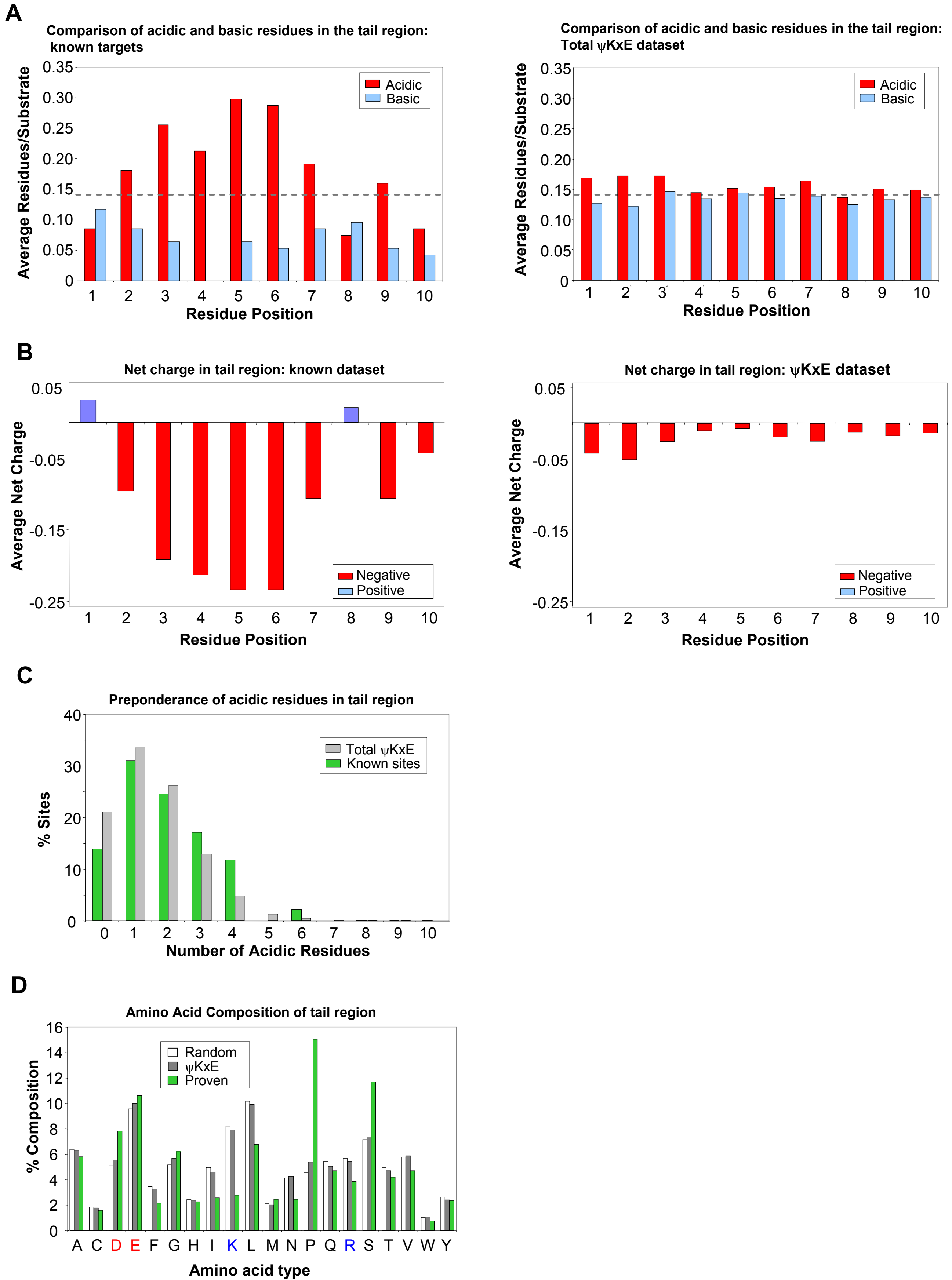

Fig. S1

Fig. S2

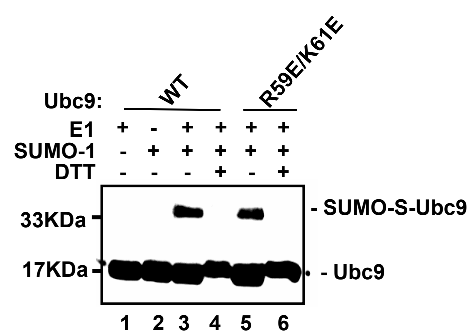

Fig. S3