SUMO maintains the chromatin environment of human induced pluripotent stem cells.

Barbara Mojsa[1], Michael H. Tatham[1], Lindsay Davidson[2], Magda Liczmanska[1], Jane E. Wright[1], Nicola Wiechens[1], Marek Gierlinski[3], Tom Owen-Hughes[1] and Ronald T. Hay[1*]

[1]Division of Gene Regulation and Expression, [2]Division of Cell and Developmental Biology, [3]Division of Computational Biology, School of Life Sciences, University of Dundee, Dundee, UK

[*]Corresponding author r.t.hay@dundee.ac.uk

Abstract

Pluripotent stem cells represent a powerful system to identify the mechanisms governing cell fate decisions during early mammalian development. Covalent attachment of the Small Ubiquitin Like Modifier (SUMO) to proteins has emerged as an important factor in stem cell maintenance. Here we show that SUMO is required to maintain stem cells in their pluripotent state and identify many chromatin-associated proteins as bona fide SUMO substrates in human induced pluripotent stem cells (hiPSCs). Loss of SUMO increases chromatin accessibility and expression of long non-coding RNAs and human endogenous retroviral elements, indicating a role for the SUMO modification of SETDB1 and a large TRIM28 centric network of zinc finger proteins in silencing of these elements. While most protein coding genes are unaffected, the Preferentially Expressed Antigen of Melanoma (PRAME) gene locus becomes more accessible and transcription is dramatically increased after inhibition of SUMO modification. When PRAME is silent, a peak of SUMO over the transcriptional start site overlaps with ChIP-seq peaks for cohesin, RNA pol II, CTCF and ZNF143, with the latter two heavily modified by SUMO. These associations suggest that silencing of the PRAME gene is maintained by the influence of SUMO on higher order chromatin structure. Our data indicate that SUMO modification plays an important role in hiPSCs by repressing genes that disrupt pluripotency networks or drive differentiation.

Introduction

Pluripotent cells display the property of self-renewal and have the capacity to generate all of the different cells required for the development of the adult organism. The pluripotent state is defined by the gene expression programme of the cells and is driven by expression of the core transcription factors OCT4, SOX2 and NANOG[1] that sustain their own expression by virtue of a positive linked autoregulatory loop while activating genes required to maintain the pluripotent state and repressing expression of the transcription factors for lineage specific differentiation[2]. Once terminally differentiated, somatic cell states are remarkably stable. However, forced expression of key pluripotency transcription factors that are highly expressed in embryonic stem cells (ESCs), including OCT4, SOX2 and NANOG, leads to reprogramming back to the pluripotent state[3-5]. As the efficiency of reprogramming is very low, it is clear that there are roadblocks to reprogramming designed to safeguard cell fates[6,

[7]. Small Ubiquitin like Modifier (SUMO) has emerged as one such roadblock and reduced SUMO expression decreases the time taken and increases the efficiency of reprogramming in mouse cells[8-10]. Three SUMO paralogues, known as SUMO1, SUMO2 and SUMO3 are expressed in vertebrates. Based on almost indistinguishable functional and structural features SUMO2 and SUMO3 are collectively termed SUMO2/3 and share only about 50% amino acid sequence identity with SUMO1. SUMOs are conjugated to lysine residues in a large number of target proteins and as a consequence influence a wide range of biological processes. SUMOs are initially translated as inactive precursors that require a precise proteolytic cleavage carried out by SUMO specific proteases (SENPs) to expose the terminal carboxyl group of a Gly-Gly sequence that ultimately forms an isopeptide bond with the ε-amino group of a lysine residue in the substrate protein. The heterodimeric E1 SUMO Activating Enzyme (SAE1/SAE2) uses ATP to adenylate the C-terminus of SUMO before forming a thioester with a cysteine residue in a second active site of the enzyme and releasing AMP. SUMO is then trans-esterified on to a cysteine residue in the single E2 SUMO conjugating enzyme Ubc9. Assisted by a small group of E3 SUMO E3 ligases, including the PIAS proteins, RanBP2 and ZNF451 the SUMO is transferred directly from Ubc9 onto target proteins[11]. Modification of target proteins may be short-lived, with the SUMO being removed by SENPs. Together this creates a highly dynamic SUMO cycle where the net SUMO modification status of individual proteins is determined by the rates of SUMO conjugation and deconjugation[12]. Preferred sites of SUMO modification conform to the consensus ψKxE, where ψ represents a large hydrophobic residue[13, 14]. A conjugation consensus is present in the N-terminal sequence of SUMO2 and SUMO3 and thus permits self-modification and the formation of SUMO2/3 chains[15]. As a strict consensus is absent from SUMO1, it does not form chains as readily as SUMO2/3[12]. Once linked to target proteins SUMO allows the formation of new protein-protein interactions as the modification can be recognised by proteins containing a short stretch of hydrophobic amino acids termed a SUMO interaction motif[16].

Stem cell lines are excellent models to study mechanisms controlling self-renewal and pluripotency. Mouse ESCs have been widely used as they can also be used for *in vivo* studies by making chimeras in mouse blastocysts, however, they do display different characteristics from human ESCs. Mouse ESCs require leukemia inhibitory factor (LIF) and bone morphogenetic protein (BMP) signalling to maintain their self-renewal and pluripotency[17, 18]. In contrast LIF does not support self-renewal and BMPs induce differentiation in hESCs[19-21].

The maintenance of the pluripotent state of hESCs requires basic fibroblast growth factor (bFGF, FGF2) and activin/nodal/TGF-β signalling along with inhibition of BMP signalling[22, 23]. These differences may reflect the developmental stages at which ESC lines are established *in vitro* from mouse and human blastocysts, or may be due to differences in early embryonic development[24]. As hESCs are derived from embryos their use is limited, in contrast human Induced Pluripotent Stem Cells (hiPSCs) are derived by reprogramming normal somatic cells and display most of the characteristics of hESCs[3]. As a result, hiPSCs are now widely used to study self-renewal and pluripotency in humans.

To determine the role of SUMO modification in hiPSCs we made use of ML792, a highly potent and selective inhibitor of the SUMO Activating Enzyme[25]. Treatment of hiPSCs with this inhibitor rapidly blocks *de novo* SUMO modification allowing endogenous SENPs to strip SUMO from targets. When used over the course of 48 hours hiPSCs treated with ML792 lose the majority of SUMO conjugation but show no large-scale changes to the cellular proteome nor loss of viability, although markers of pluripotency are reduced. By inhibiting SUMO conjugation ML792 reduces chromatin-associated SUMO and increases DNA accessibility. This results in increased transcription of a group of long non-coding RNAs (lncRNAs) and human endogenous retrovirus (HERV) elements, while protein coding genes are largely unaffected. One important exception is the Preferentially Expressed Antigen of Melanoma (PRAME) gene. SUMO modification inhibition increases the accessibility of the PRAME locus and leads to a large increase in transcription and accumulation of PRAME protein. SUMO site and paralogue specific proteomic analysis of hiPSCs reveals extensive SUMO modification of proteins involved in transcriptional repression, RNA splicing and ribosome biogenesis. Specifically, SUMO modification of the boundary and looping elements CTCF and ZNF143 and their colocalisation with cohesin components suggest an important role for SUMO in organising the higher order chromatin architecture in hiPSCs.

Results

Inhibition of SUMO modification leads to loss of select pluripotency markers in hiPSCs. To determine the role of SUMO modification in the maintenance of pluripotency in hiPSCs we used ML792[25]. This inhibitor has been reported to block proliferation of cancer cells, particularly those overexpressing Myc[25], but has not been evaluated in hiPSCs. To address the

role of SUMO modification in ChiPS4, we established that 400nM ML792 effectively reduced SUMO modification after 4 hrs with minimal effects on cell viability in longer treatments. We restricted our analyses to ML792 treatment times that did not exceed 48 hrs, such that the immediate effects of SUMO modification inhibition could be evaluated. Microscopic examination revealed that ML792 treatment caused morphological changes with the ChiPS4 cells becoming larger and flatter (Fig. 1a). The rate of proliferation was unchanged after 24 hrs but was slightly reduced after 48 hrs (Fig. 1b). DNA staining of the cells and analysis by flow cytometry indicated that the cell cycle distribution after 24hrs was unaltered by ML792 treatment but after 48 hrs displayed an increased proportion of cells in G2 phase and cells with increased DNA content suggesting endoreplication (Fig. 1c). Western blot analysis revealed a loss of high molecular weight SUMO1 and SUMO2 conjugates and concomitant increase in free SUMOs in ChiPS4 cells (Fig. 1d) caused by rapid removal of SUMO from modified proteins by SENPs. Analysis of the protein levels of key pluripotency markers indicated that while OCT4 and SOX2 were unchanged, inhibition of SUMOylation resulted in a decrease in NANOG protein (Fig. 1d). This appeared to be a consequence of reduced transcription as determination of mRNA levels by reverse transcriptase quantitative polymerase chain reaction (RT-qPCR) after ML792 treatment demonstrated a reduction in *NANOG* mRNA. This was also apparent for *KLF4*, but consistent with Western blotting, the levels of *OCT4* and *SOX2* mRNA were unchanged (Fig. 1e). To further investigate the nature and causes of the observed morphological changes ChiPS4 ML792 treated cells were analysed by phenotypic screening using cell painting[26] (Fig. 2a). Principle component analysis (PCA) indicated that there are clear differences between cells treated with ML792 for 48h and untreated/vehicle (DMSO) treated. The main differences were found in the nuclear compartment (Supplementary Fig. 1). Feature extraction identified changes in the global size (Area of nuclei) and shape (Nuclei form factor) of the nucleus and the structure of the nucleolus (Nuclei: FITC texture correlation) (Fig. 2a). These findings were validated using traditional immunofluorescence (IF) approaches. NANOG expression as well as the size and shape of the nucleus are both affected by ML792 treatment. NOP58 was used as a marker of the nucleolus, which undergoes a dramatic increase in size and shape. The classic punctate nuclear localisation pattern of SUMO1 and SUMO2 is altered by their deconjugation from substrates, becoming more diffuse and less tightly associated with the nucleus (Fig. 2b).

To evaluate the effect of inhibition of SUMO modification on the global proteome in hiPSCs, proteins from ChiPS4 cells either untreated or treated with ML792 for 24 or 48 hours were analysed by label-free quantitative proteomics. 4741 proteins were identified and quantified in all replicates of at least one experimental group (Supplementary Data File 1). PCA of the proteomic data showed a progressive trend of changing cellular proteome during SUMOylation inhibition (Supplementary Fig. 2a), although individual protein fold changes compared at both time-points showed little evidence for large-scale global shifts in protein abundance (Supplementary Fig. 2b, c). Known pluripotency markers were progressively reduced during ML792 exposure (Fig. 2c), and linker histones were reduced in abundance after 48 hours (Fig. 2d). Linker histones and the GOCC group 'Collagen-associated extracellular matrix' were the only two categories to show any significant co-regulation according to STRING analysis, (Supplementary Fig. 2d, e). Thus, global proteome changes do not seem to provide an explanation for an observed morphological change of this magnitude. It therefore seems likely that the observed changes in nuclear structure are due to direct consequences of removal of SUMO from critical factors that contribute to chromatin structure and function.

Removal of SUMO in hiPSCs increases chromatin accessibility. To determine the chromosomal landscape of SUMO1 modification Chromatin Immunoprecipitation coupled to high throughput sequencing (ChIP-seq) was conducted on ChiPS4 cells. SUMO1 bound chromatin was enriched from cross-linked cell extracts using an antibody with previously confirmed specificity for SUMO1[27] and utility in ChIP analysis[28]. Total genomic DNA was sequenced to obtain reference input profiles. SUMO peaks were usually less than 1kb and typically 300bp in length (Supplementary Fig. 3a) and were clearly enriched above the background (Supplementary Fig. 3b) and in promoter, introns and intergenic regions (Supplementary Fig. 3c). Similar to the situation reported in mouse ESCs[10, 29], SUMO peaks were over-represented on endogenous retroviruses (ERVs) and other non-viral long terminal repeats (LTRs) (Supplementary Fig. 3d). Of note, SUMO1 accounts for a high proportion of protein SUMOylation in ChiPS4 cells when assessed by mass spectrometry and Western blotting. SUMO peaks overlapped with over 10% of KAP1/TRIM28 and SETDB1 peaks, but over 50% of peaks for CTCF, ZNF143 and cohesion components RAD21 and SMC3 (Supplementary Fig. 3e, f).

Having used SUMO1 ChIP-seq to determine the precise location of SUMO1-modified proteins on chromatin in ChiPS4 cells we used ML792 to facilitate the removal of SUMO from these sites and used ATAC-seq to monitor changes in chromatin accessibility over time. For that, ChiPS4 cells were treated with DMSO vehicle or ML792 for 4, 8, 24, 48h leading to time dependent release of SUMO from high molecular weight material and a reduction in NANOG expression (Supplementary Fig. 4a). The loss of SUMO was accompanied by a general increase in chromatin accessibility across the genome as determined by an increase in number of ATAC-seq peaks over time (Fig. 3a, Supplementary Data File 5). ATAC-seq peaks that are lost do not increase in the same way and after 48h exposure to the inhibitor there are at least three times more ATAC-seq peaks gained then lost (Fig. 3a). Comparing the ATAC-seq and SUMO1 ChIP-seq data suggests that removing SUMO leads to an increase in chromatin accessibility at the sites previously occupied by SUMO as around 20% of gained ATAC-seq peaks at all time points overlap with a pre-existing SUMO1 ChIP-seq peak, while less that 5% of lost ATAC-seq peaks overlap with a pre-existing SUMO1 ChIP-seq peak (Fig. 3b).

To determine if specific genomic regions change their accessibility after removal of SUMO, the ATAC-seq data was analysed using HOMER[30], allowing various types of genomic regions to be annotated and classified into functionally related groups. Chromatin accessibility was gained mainly in repetitive DNA sequences, such as non-LTR and LTR retrotransposons which together account for around 80% of all gained ATAC-seq peaks (Fig. 3c). When compared to the types of genomic regions represented in non-changing ATAC-seq peaks, LTR-retrotransposons are strongly enriched in gained ATAC-seq peaks or gained ATAC-seq peaks overlapping with SUMO1 peaks throughout the treatment with ML792 (Fig. 3c, d). It thus suggests an important role for SUMO in maintaining these viral elements of the hiPSCs genome in a compact chromatin environment. There is little enrichment for any particular type of genomic region in the ATAC-seq peaks or ATAC-seq peaks overlapping with SUMO1 that are lost after removal of SUMO (Supplementary Fig. 4b, c).

The dynamic behaviour of each opening locus can be assessed using density plots (Fig. 3e). Regions that lost chromatin accessibility did not show any significant enrichments or patterns (Supplementary Fig. 4d). As indicated above, 20% of gained ATAC-seq peaks at all time points after treatment with ML792 overlap with a pre-existing SUMO ChIP-seq peak present in untreated cells (Fig. 3b). These are also the regions that respond quickly to ML792 mediated chromatin opening (higher intensity SUMO1 ChIP peaks overlap with ATAC-seq

peaks that appear sooner during the time course) and once open the chromatin state is maintained throughout the time course (Fig. 3e). It is also evident that the SUMO overlapping ATAC-seq peaks that are classified as 'nascent gained' at a later time point (e.g. 48h) already demonstrate a trend for chromatin relaxation at earlier time points, but do not reach the necessary threshold. These SUMO overlapping ATAC-seq peaks are strongly associated with repetitive DNA sequences, particularly retrotransposons (Fig. 3d). Based on the association of SUMO with repressed chromatin and the proteins associated with repression of viral elements in the genome, a similar peak overlap procedure was performed with transcription factor ChIP-seq data obtained from the ENCODE project[31]. An unbiased analysis of 161 factors was undertaken, but those showing the highest overlap with SUMO ChIP-seq and gained ATAC-seq peaks were TRIM28, SETDB1, CBX3 (Fig. 3f) that are known SUMO-modified silencers of viral DNA elements. The overlap between those factors and SUMO1 ChIP-seq or gained ATAC-seq peaks is significantly higher than that calculated for non-changing or lost ATAC-seq peaks (Fig. 3f). These data suggest an important role for SUMO in maintaining a compact chromatin environment around LTR elements in hiPSCs.

Inhibition of SUMO modification in hiPSCs selectively alters transcription. To determine if chromatin associated SUMO regulates transcription in hiPSCs, ChiPS4 cells were treated as for ATAC-seq and analysed by RNA-seq (Supplementary Fig. 5). After 48h of treatment with ML792 996 RNAs displayed an increase in transcription while 281 RNAs were decreased (Fig. 4a, Supplementary Data File 6). Although the observed effect on protein coding mRNAs was rather modest, the expression of lncRNAs was significantly affected (both increased and decreased) by inhibition of SUMO modification (Fig. 4b, Supplementary Fig. 6a). Interestingly, stage specific expression of certain lncRNAs e.g. LINC-ROR is important for the maintenance of the pluripotency network[32]. While protein coding genes as a group were not significantly enriched during the time-course, it is possible that the aggregation of incremental changes could exert a specific biological response. For example, consistent with the protein level analysis (Fig. 2c), mRNA levels of various pluripotency markers decrease globally during the treatment (Supplementary Fig. 6b). To determine if there were any functional patterns to the protein-coding genes regulated by ML792 treatment, a data-dependent clustering analysis was undertaken (Supplementary Data File 4). Hierarchical clustering of protein-coding RNAs based on response to ML792 allowed separation into 9 clusters (Fig. 4c). Cluster A contained

the least responsive RNAs and represented over 96% of the entries (Fig. 4c-f). A number of GSEA categories were depleted in cluster A with high significance, all of which were previously identified in ES cells as being regulated by histone methylation or by protein complexes themselves regulating histone methylation such as PRC2 (Fig. 4f). Indeed, these categories were among the most significantly enriched in the two largest clusters of ML792-sensitive RNAs (clusters B and C). Clusters B and C show opposing responses to ML792 treatment, therefore these data imply that although promoter methylation is a common feature of regulated genes the outcomes are not qualitatively the same for all of them (Supplementary Fig. 7). Thus, inhibition of SUMO modification increases transcription of a group of lncRNA genes, but with notable exceptions, has limited impact on transcription of protein coding genes.

SUMO silences the PRAME gene in human stem cells. One important exception is the PRAME gene, transcription of which significantly increases in response to ML792 treatment in a time-dependent manner. This increase is evident at the very earliest time point (4h) and continues to increase up to 48h (Fig. 5a). Inspection of the SUMO1 ChIP-seq data from untreated ChiPS4 cells where *PRAME* expression is silenced reveals a prominent SUMO peak located over the transcriptional start site (TSS) (Fig. 5b). Analysis of the ATAC-seq data indicates that while this region is minimally accessible in untreated cells, removal of SUMO leads to a time dependent increase in chromatin accessibility. At the earliest time point (4h) this increase is confined to the TSS, but over time the region of accessibility spreads in towards the coding body of the gene (Fig. 5b). This is highly specific for the *PRAME* gene as the same SUMO1 peak overlaps with a TSS of the divergently transcribed *LL22NCO3-63E9.3* lncRNA gene which displays neither an increase in chromatin accessibility, nor an increase in transcription in response to ML792 treatment (Fig. 5b). Analysis of ChIP-seq data from other hESC lines reveals that CTCF, ZNF143 and the cohesion subunit RAD21 almost precisely overlap with this SUMO ChIP-seq peak (Supplementary Fig. 8a). All of these factors are known to be associated with higher order chromatin interactions. Comparison of the protein level and transcriptional changes after 48 hours ML792 exposure shows a single outlier in PRAME (Fig. 5c, Supplementary Fig. 8d). PRAME is both the single most upregulated transcript and the most elevated protein upon SUMOylation inhibition. IF analysis further revealed that in untreated cells PRAME expression was not above background, but after 48 hours of ML792 treatment PRAME was

highly expressed and localised in the nuclei of ChiPS4 cells (Supplementary Fig. 8b). Likewise, Western blotting demonstrated that PRAME was accumulated after SUMO modification inhibition (Fig. 5d). To establish that the SUMO regulated expression of the *PRAME* gene was not unique to ChiPS4 cells a number of other hiPSC and hESC lines were exposed to ML792. PRAME expression was also robustly induced after 48h of treatment with ML792 (Supplementary Fig. 8c), indicating that SUMO-dependent silencing of PRAME is a common feature of all tested human pluripotent stem cells (hPSCs). While ML792 is a highly selective and potent inhibitor of SAE[25] it was important to assess SUMO regulated expression of *PRAME* by an orthogonal approach. Rather than inhibiting SUMO modification, conjugated SUMO can be directly removed by expression of an exogenous SUMO specific protease. We previously used such an approach to demonstrate SUMO dependent regulation of an integrated reporter gene[33]. Capped and polyadenylated mRNA encoding the catalytic domain of SENP1 was electroporated into ChiPS4 cells and PRAME expression was monitored by Western blotting and RT-qPCR. Expression of the protease effectively reduced global SUMO modification and increased both PRAME protein and mRNA (Fig 5e, f). Thus, silencing of the *PRAME* gene in hiPSCs appears to be directly mediated by SUMO modification. Enrichment analysis indicated that transcription of metallothionein genes was reduced in response to SUMOylation inhibition, while transcription of five further members of the *PRAME* gene family was elevated (Supplementary Fig. 8e, f), revealing a common link between SUMO and transcription of *PRAME* genes.

SUMO modification restricts HERV expression in hiPSCs. ATAC-seq and ChIP-seq analyses revealed that repetitive DNA elements including LTR retrotransposons are primary targets of SUMOylation-mediated repression. Indeed, expression of lncRNAs is often controlled by LTRs, which have been hijacked by cellular machinery to function as stage specific promoters. Typically for RNA-seq experiments such repetitive DNA sequences are mostly removed during data processing. To analyse HERV expression, an independent data alignment file was created based on the available Human Endogenous Retrovirus Database, which contains two major data sets: elements (contiguous sequences) and entities (loci in the human genome consisting of one or more elements)[34, 35]. Inhibition of SUMO modification leads to a general increase in HERV expression with the number of significantly increased elements being about three times higher than those reduced at each time point tested (Fig. 6a, Supplementary Fig. 9a, b). One

of the best examples of a complex LTR-containing HERV element that is highly and rapidly induced by deSUMOylation is ERV_4326325 (Fig. 6b). Inspection of the SUMO1 ChIP-seq indicates that this chromosomal location contains a pre-existing SUMO1 peak. Indeed, ATAC-seq data show that increase in chromatin accessibility is initiated from the site of the SUMO1 peak and spreads in towards the HERV locus (Fig. 6c). The increase in RNA expression follows the changes in chromatin structure and is already obvious at 24h (Fig, 6b, c). To investigate the global landscape of these changes, significantly affected HERVs were used for clustering analysis. Three independent clusters were obtained, in which all HERVs with a rapid increase in expression were found in cluster 1 (Fig. 6d, e, Supplementary Fig. 9c, d). Correlation of the SUMO ChIP-seq data with the HERV loci in each of the three clusters revealed that HERVs in cluster 1 have a significantly higher overlap with SUMO1 ChIP-seq peaks when compared to the proportion observed for all HERVs (Fig. 6f). These data suggest that SUMO modification maintains HERV loci in a compact chromatin state that facilitates transcriptional repression of these viral elements in hiPSCs.

Identification of SUMO1 and SUMO2 targets in hiPSCs. Our data suggest an important role for SUMO in maintaining the chromatin state of hiPSCs. To identify the proteins responsible and establish the sites of SUMO modification on these factors we used a SUMO site proteomic approach that allows sites modified by SUMO1 and SUMO2/3 to be identified[36].To enable this analysis in hiPSCs the ChiPS4 cell line was engineered to stably express 6His-SUMO-mCherry constructs for either SUMO1 or SUMO2 (Supplementary Fig. 10a, b) that incorporated the TGG to KGG mutations to facilitate GlyGly-K peptide immunoprecipitation and identification[36]. As mCherry is linked to the C-terminus of SUMO the expressed fusion protein will be processed by endogenous SUMO proteases, release free mCherry and expose the C-terminal GlyGly sequence for conjugation. Western blotting of single cell clones indicated that His-tagged SUMO-KGG paralogues were conjugated to substrates in response to heat shock (Supplementary Fig. 10c). Cells expressing SUMO1-KGG and SUMO2-KGG had normal cell cycle profiles (Supplementary Fig. 11a), expressed levels of pluripotency markers comparable to wild type ChiPS4 cells (Supplementary Fig. 11b, c) and retained the ability to differentiate into endoderm, ectoderm and mesoderm (Supplementary Fig. 11d). Analysis by proteomics (Supplementary Fig. 12a, b) identified the expected exogenous mCherry, SUMO1 and SUMO2 peptides (Supplementary Fig. 12c) while analysis of common peptides suggested that the

exogenous versions of SUMO were conjugated to substrates at roughly similar levels to their endogenous counterparts (Supplementary Fig. 12d). Whole cell proteomics (Supplementary Fig. 12e) confirmed that the engineered cell lines did not significantly change their expressed proteome (Supplementary Fig. 12f, Supplementary Data File 2). Thus, expression of SUMO mutants did not disrupt the normal pluripotent state or differentiation potential of ChiPS4 cells.

The workflow for the identification of SUMO targets incorporates proteomic analysis at three levels (Supplementary Fig. 13a). The experiment involves analysis of whole cell extracts (Supplementary Fig. 13b) to monitor total protein levels, analysis of Nickel NTA-affinity purified proteins (Supplementary Fig. 13c) to monitor SUMO modified proteins and analysis of GG-K immunoprecipitations (Supplementary Fig. 13d) to identify sites of SUMO modification. Across the two experimental runs a total of 976 SUMO sites were identified in 427 proteins. Approximately 84% of these had already been described in at least one of four large-scale SUMO2 site proteomics studies totalling 49768 unique sites of non-STEM cell origin (Fig. 7a, Supplementary Data file 3). DNA methyl transferase DNMT3B and the key embryonic stem cell transcription factor SALL4 were among a small group of proteins with at least three novel sites in this study (Fig. 7a). Based on GG-K peptide intensity SALL4 is the 6th most modified SUMO substrate in hiPSCs and contains 17 sites of modification (Fig. 7b). DNMT3B contains 12 sites and is the 7th most modified substrate while the methyl DNA binding protein MBD1 contains 8 sites and is also in the top 10 SUMO substrates (Fig. 7b). TRIM28 and TRIM24 are highly modified substrates and SUMO appears to play an important role in their ability to repress retroviral elements[29]. CTCF is heavily modified with SUMO and this is consistent with the overlap of SUMO and CTCF ChIP-seq peaks over the TSS of the PRAME gene that appears to be silenced by SUMO modification in ChiPS4 cells (Fig. 7b). There is also evidence for extensive SUMO chain formation as the branch points from SUMO2/3 chains are amongst the most abundant GG-K peptides (Fig. 7b). Indeed, SUMOylation of a number of these heavily SUMO modified proteins could be detected directly in total cell lysates from control ChiPS4 cells, but not ML792 treated hiPSCs (Fig. 7c).

An advantage to using the SUMO1 and SUMO2 KGG mutants for site-level proteomics is that both paralogues leave the same Gly-Gly remnant on substrates after LysC digestion. Thus, site-specific SUMO preference can be compared. To date this has not been undertaken on a large scale and remains an important question in the SUMO field. The proteomic

experimental design allowed these comparisons at multiple stages of the purification process (Supplementary Fig. 13a-d): SUMO1/SUMO2 ratios from crude hiPSC extracts shows there to be few differences at the whole proteome level (0.03% significant - Supplementary Fig. 13e). There are also surprisingly few differences between NiNTA purifications from the two cell types (7.8% significant - Supplementary Fig. 13f). Exceptions include the well-documented SUMO1 substrate RanGAP1, along with TRIM24 and TRIM33 which all show similar levels of SUMO1 preference (Supplementary Fig. 13f). In contrast, over half of the GGK-containing peptides quantified showed large and significant difference between SUMO1 and SUMO2 cells (Supplementary Fig. 13g). Extreme examples of SUMO1 preferential sites include RanGAP1 K524 and TRIM33 K776. Conversely, TRIM28 contains two of the most SUMO2-preferntial sites at K507 and K779, and lysine 48 and 63 from ubiquitin are among the extreme SUMO2 acceptors. This was confirmed by Western blot analysis from NiNTA purifications (Supplementary Fig. 13h). Thus, when considering net modification of a protein, the bulk of SUMO modified proteins do not appear to display SUMO paralogue specificity, while this difference is clear at the site level.

STRING enrichment analysis of the 427 modified proteins created a network consisting of 3 clusters of proteins that could be broadly categorised as having functions in ribosome biogenesis, RNA splicing, and regulation of gene expression (Fig. 7d). Despite forming extensive protein networks (Fig. 7e and f), proteins involved in ribosome biogenesis and RNA splicing represented only approximately 5% of the total GGK peptide intensity (Fig. 7b insert). The majority of the remainder have roles in transcription and chromatin structure or are closely linked to these functions (Fig. 7b insert and d). There is a prominent network of zinc-finger transcription factors, closely associated with TRIM28 (Fig. 7g) which contains many of the most heavily SUMOylated proteins identified, which play a key role in silencing retroviral elements. Histone proteins themselves, including H1, form a small cluster of SUMO substrates (Fig. 7h) in the centre of the gene regulation region of the whole network (Fig. 7d), and could potentially act as a direct link between SUMO and chromatin structure. The transcriptional regulators themselves form a bipolar network with the smaller sub-cluster consisting mainly of apparently weakly modified ribosomal proteins and the larger sub-cluster containing many heavily modified chromatin associated proteins (Fig. 7i). Strikingly, many members of chromatin remodelling complexes such as PRC2 (Fig. 7j), BAF (Fig. 7k) and NURD

(Fig. 7l) are among this group, potentially providing a link between SUMO and chromatin structure and remodelling.

Discussion

Our studies highlight an important role for SUMO in maintaining the pluripotent state of hiPSCs. Using a potent and highly specific inhibitor of the SUMO E1 ML792[25] we blocked *de novo* SUMO modification and allowed endogenous SENPs to remove SUMO from previously modified proteins. In response to short-term SUMOylation inhibition ChiPS4 cells showed no loss of viability, but underwent clear morphological changes, losing markers of pluripotency (NANOG, KLF4, LINC-ROR), without displaying large-scale changes to the cellular proteome (Figs. 1 and 2). Upon SUMOylation inhibition ATAC-seq analysis demonstrated that sites previously occupied by SUMO (SUMO1 ChIP-seq) became more accessible (Fig. 3, Supplementary Fig. 3), indicating a role for SUMO in maintaining a compact chromatin environment. About 80% of these sites were associated with non-LTR and LTR retrotransposons (Fig. 3), while RNA-seq analysis indicated that a subset of HERV increased their transcription in response to SUMO modification inhibition (Fig. 6). ChIP-seq analysis indicated that the peak of SUMO located close to these HERVs also overlapped with the ChIP-seq derived locations of TRIM28, SETDB1 and CBX3[31]. These proteins along with SUMO have previously been shown to function in HERV silencing in mESCs[29, 37] and adult human cells[38, 39] and this is consistent with our proteomic analysis that indicates that all three of these proteins are heavily SUMO modified (Fig. 7). Moreover, TRIM28 co-repressor functions by interacting with DNA bound Kruppel type zinc finger proteins, which are also heavily SUMO modified in our proteomic studies. In fact, they form a large TRIM28 centric network of SUMO modified proteins that also includes the histone methyl transferase SETDB1. It is suggested that a number of developmental genes are repressed by TRIM28/KRAB-ZNFs through deposition of H3K9me3 and *de novo* DNA methylation of their promoter regions[40], thus making TRIM28/ZNFs a crucial link in maintenance of pluripotency in human stem cells.

Several families of HERVs have been found to show stage specific expression in the preimplantation embryo and in hESCs *in vitro*[41]. These HERVs have been implicated in the maintenance of pluripotency in hESCs, are associated with the binding sites of pluripotency associated transcription factors (including OCT4, SOX2 and NANOG), and produce stage-specific lncRNAs that are required for the maintenance of the pluripotent state[42-44].

Furthermore, HERV-H expression is dynamically regulated during transcription factor-mediated reprogramming and the acquisition of appropriate stage-specific expression of HERV-H is required for the re-establishment of pluripotency in hiPSCs[45]. Recently, HERVs have also been implicated in the regulation of Topologically Associating Domains (TADs) in hPSCs as deletion of HERV-H elements eliminates their corresponding boundaries and reduces the expression of upstream genes, while *de novo* insertion of HERV-H sequences can create new TAD boundaries[46]. These observations suggest that proper control of HERV expression is required for the maintenance of pluripotency in hESCs and hiPSCs, and our data suggest that SUMO modification may play a role in defining the HERVs that are expressed in these cells.

Consistent with this observation, the RNA-seq analysis revealed major changes in the expression of lncRNAs, but with limited changes to expression of protein coding genes in response to ML792 treatment (Fig. 4). However, a notable exception to this was the *PRAME* gene, which showed a massive increase in transcription after treatment with ML792. Increased transcription could be detected at the earliest time point analysed (4h) suggesting this was a direct effect of inhibiting SUMO modification (Fig. 5). An increase in *PRAME* expression was also observed when SUMO was removed from substrates by expression of the catalytic domain of SENP1. After removal of SUMO the chromatin around the *PRAME* locus becomes accessible and transcription of the gene increases over 10,000 fold, which would be to date, the clearest example of a protein coding gene that is negatively regulated by SUMOylation. Although SUMO has long been implicated in the repression of protein coding genes, most of the previous work has used artificial promoters[33] and data showing SUMO mediated regulation of endogenous genes is rather sparse.

Inspection of the *PRAME* gene locus indicates that the TSS of *PRAME* is adjacent to the TSS of *LL22NCO3-63E9.3* lncRNA gene that is transcribed in the opposite direction. While SUMOylation inhibition increases transcription of the PRAME gene it does not lead to transcription of the neighbouring gene. ChIP-seq data indicates that in absence of ML792 treatment a peak of SUMO is present over the TSS of *PRAME* and overlaps with ChIP-seq peaks for cohesin, CTCF, ZNF143 and RNA pol II. Our SUMO site proteomic data indicate that CTCF is heavily modified by SUMO as is ZNF143. Cohesin, CTCF and ZNF143 are all associated with maintaining higher order chromatin structure[47-49] particularly loops or TADs and supports the hypothesis that silencing of the *PRAME* gene is maintained by the influence of SUMO on higher order chromatin structure. ZNF143 is thought to control transcription from

bidirectional promoters[50] and regulate the density of promoter-proximal paused RNA polymerase[51], which has been associated with silenced genes that can be rapidly activated. The co-location of RNA pol II with cohesion, CTCF, ZNF143 and SUMO is suggestive of such a scenario at the PRAME gene locus. This idea is supported by the observation that the most highly SUMO modified protein in our proteomic analysis is GTF2i/TFII-I, is a component of the RNA pol II transcriptional complex. Thus, while transcriptional repression of the PRAME gene also involves SUMO, it appears to be mediated by a very different mechanism from silencing of HERV genes as TRIM28, SETDB1 and CBX that are associated with HERV silencing, are not present at the PRAME locus. While our data at HERV entities are consistent with a recent SUMO proteomic analysis in mESCs[52], the tight SUMO mediated regulation of the PRAME locus appears to be specific to hPSCs. Although the *PRAME* gene is frequently over-expressed in tumours[53, 54] it appears to play a role in the differentiation of hPSCs into mesenchymal stem cells[55]. PRAME appears to function as the substrate adapter of a Cul2 E3 ubiquitin ligase that is targeted to chromatin and associates with active NFY promoters[56]. While the targets of PRAME mediated ubiquitination have yet to be identified it has been shown to associate with the highly conserved EKC/KEOPS complex on chromatin[57].

Analysis of the SUMO proteome of ChiPS4 cells shows that well-defined groups of protein are modified. Aside from the TRIM28/ZNF network mentioned above, proteins involved in “ribosome biogenesis” and “splicing” are SUMO modified and this likely impacts on the normal growth and self-renewal of the hiPSCs. However, the largest network of proteins falls into the category of ‘negative regulation of transcription’ with many chromatin remodellers, chromatin modification and DNA modification enzymes identified as SUMO substrates. The increases in transcription observed after SUMO modification inhibition indicate that SUMO modification plays an important role in maintaining pluripotency of hiPSCs by repressing genes that either disrupt pluripotency or drive differentiation.

Acknowledgements

We thank Linnan Shen for the kind gift of purified Tn5 transposase, Adel Ibrahim for GNb-SENP1 DNA and Alwyn Dady for the paPX1 vector. We would like to acknowledge the invaluable help from various facilities at the University of Dundee: Flow Cytometry and Cell Sorting, National Phenotypic Screening Centre, Dundee Imaging Facility, Human Pluripotent Stem Cell Facility. This work was supported by an Investigator Award from Wellcome (217196/Z/19/Z) and a Programme grant from Cancer Research UK (C434/A21747) to RTH. ML is part of the Ubi-Code European Training Network that received funding from the European Union Horizon 2020 research and innovation programme under the Marie Curie grant agreement No. 765445. BM was supported by the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 704989. JEW was supported by a Marie Skłodowska-Curie Actions Individual Fellowship from the European Commission grant agreement No. 625253. TOH and NW were supported by the MRC grant (MR/S021647/1).

Author Contributions

BM cloned expression vectors, generated cell lines, designed and performed most experiments using hiPSCs and interpreted data. JEW contributed to the initial design of the research and performed SUMO1 ChIP-seq. ML performed NiNTA purifications and Western blotting analyses. LD was in charge of cell culture, cell line derivation and quality control and was consulted over experimental design. RTH generated capped and polyadenylated RNAs. MHT consulted over proteomic experimental design, acquired and processed MS data and conducted bioinformatic and statistical analyses. NW performed the ATAC-seq experiments, processed and analysed the data. MG performed the bioinformatic analysis of RNA-seq and ChIP-seq data sets. BM, MG, NW, MHT, LD, TOH and RTH contributed to data analysis. BM, RTH, MHT and LD wrote the paper. RTH conceived the project.

Data availability

Data underlying all Figures and Supplementary Figures are available in the source data file. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE[58] partner repository with following datasets identifiers: PXD023241, PXD023257. Source data for ATAC-seq, RNA-seq and SUMO1 ChIP-seq are

available from EBI at https://www.ebi.ac.uk/ under accession numbers: E-MTAB-9961, E-MTAB-9962 and E-MTAB-9960 respectively. All other data are available from the corresponding authors on reasonable request.

Supplementary data files

Supplementary data file 1. Summary of the quantitative data from the proteomics experiment to study changes to the cellular proteome during ML792 treatment of ChiPS4 cells.

Supplementary data file 2. Summary of the quantitative data from the proteomics experiment to study differences in the cellular proteome among wild type ChiPS4 cells and cells expressing 6His-SUMO1-KGG-mCherry or 6His-SUMO2-KGG-mCherry.

Supplementary data file 3. Summary of the quantitative data from the proteomics experiment to identify SUMO1 and SUMO2 targets from ChiPS4 cells.

Supplementary data file 4. Merge of the RNA-seq and proteomic data for ChiPS4 cells treated with ML792 for 0h, 24h and 48h.

Supplementary data file 5. ATAC-seq data file for ChiPS4 cells treated with ML792 for 4h, 8h, 24h and 48h.

Supplementary data file 6. RNA-seq edger data file for ChiPS4 cells treated with ML792 for 4h, 8h, 24h and 48h.

Code availability

Code used for data analysis is included in the source data file and can be found at https://github.com/bartongroup/MG_SumoDiff2.

Author Information

Correspondence and requests for materials should be addressed to R.T.H (r.t.hay@dundee.ac.uk)

Methods

Antibodies and inhibitors. Rabbit antibodies against TRIM28 (4124S, 4123S), CTCF (3418S), OCT4A (2890S), SOX2 (23064S), NANOG (3580S), KLF4 (12173S) and mouse antibodies against TRA-1-60 (4746T), TRA-1-81(4745T), SSEA-4 (4755T) and SMA (D4K9N) were from Cell Signalling Technology. The anti-SALL4 (ab29112), anti-TRIM24 (ab70560), anti-NOP58 (ab155556), anti-PRAME (ab219650), anti-TRIM24 (ab70560), anti-NESTIN (ab196908) were from Abcam. Mouse antibody against α-Tubulin was from Bethyl Laboratories and mouse anti-LaminA/C antibody was from Sigma (SAB4200236), rabbit anti-mCherry (PA5-34974), rabbit anti-PRAME (PA5-83761), rabbit anti-TRIM33 (PA5-82152) and mouse anti-HIS (34650) were from Invitrogen and Qiagen respectively. Anti-Cytokeratin17 was a gift from R. Hickerson (University of Dundee). Sheep antibodies against SUMO1, SUMO2, and SENP1[27] and chicken antibodies against PML[59] were generated in-house. Secondary antibodies conjugated with HRP and Alexa fluorophores were from Sigma and Invitrogen, respectively. MG132 (474787) and N-ethylmaleimide (E3876) were from Sigma Aldrich. ML792 was from UbiQ. Protease Inhibitor cocktail (11836170001) was from Roche. Propidium iodide, Cy5 Cell Mask and DAPI were from Life Technologies.

Cloning. SUMO1-KGG-mCherry and SUMO2-KGG-mCherry PiggyBac expression vectors were generated by GATEWAY cloning. Briefly, SUMO1, SUMO2 and mCherry fragments were PCR amplified using the following resources: 6His SUMO1 T95K (300nt) from pSCAI88 and 6His SUMO2 T90K (300nt) from pSCAI89 with a common forwards primer (5'-CACCatgcatcatcatcatcatgct-3') and set of specific mCherry fusing primers (5'-TCACCATACCCCCCTTTTGTTCCTG-3' and 5'-TCACCATACCTCCCTTCTGCTGCT-3'); mCherry from pRHAI4 CMV-OsTIR1-mCherry2-PURO (700 nt) with a set of common overlapping oligos (5'-GGTATGGTGAGCAAGGGCG-3' and 5'-TTATTACTTGTACAGCTCGTCCATG-3'). Subsequently, PCR fragments were fused together using overlap extension PCR and TOPO cloned into pENTR™/D-TOPO™ (Invitrogen) and verified by DNA sequencing. The assembled SUMO1-KGG-mCherry and SUMO2-KGG-mCherry sequences were then sub-cloned from the pENTR vector into the destination PiggyBac GATEWAY expression vector paPX1 using LR clonase II (ThermoFisher Scientific).

mRNA synthesis and purification. The catalytic domain of the SUMO specific protease SENP1[60] was fused at its N-terminus to a nano body directed against GFP[61] to form GNb-SENP1. The DNA encoding GNb-SENP1 was amplified by PCR using a 5' primer containing the sequence of the T7 RNA polymerase promoter. Amplified DNA was purified on a MinElute Gel Extraction kit (Qiangen). 4 µg of the eluted DNA was used as template for the production of capped and poly adenylated mRNA by in vitro transcription using an mMessage mMachine T7 Ultra kit (ThermoFisher) as described by the manufacturer. RNA was purified using a MegaClear kit (ThermoFisher) as described. Purified RNA was quantified by NanoDrop and analysed on a TapeStation (Agilent).

Human Induced pluripotent stem cells (hiPSCs) culture and transfection protocols. Human ESC lines (SA121 and SA181) were obtained from Cellartis / Takara Bio Europe. All work with hESCs was approved by the UK Stem cell bank steering committee (Approval reference: SCSC17-14). Human iPSC lines were obtained from Cellartis / Takara Bio Europe (ChiPS4) or the HipSci consortium (bubh3, oaqd3, ueah1 and wibj2). Cell lines were maintained in TESR medium[62] containing FGF2 (Peprotech, 30 ng/ml) and noggin (Peprotech, 10 ng/ml) on growth factor reduced geltrex basement membrane extract (Life Technologies, 10 µg/cm2) coated dishes at 37°C in a humidified atmosphere of 5% CO_2 in air. Cells were routinely passaged twice a week as single cells using TrypLE select (Life Technologies) and replated in TESR medium that was further supplemented with the Rho kinase inhibitor Y27632 (Tocris, 10 µM). Twenty four hours after replating Y27632 was removed from the culture medium. To make SUMO1-KGG-mCherry and SUMO2-KGG-mCherry expressing stable cell lines ChiPS4 cells were transfected using a Neon electroporation system (Thermo Fisher Scientific) using 10 µl tips. Briefly, ChiPS4 cells were dispersed to single cells as described above then $1x10^6$ cells were collected by centrifugation at 300xg for 2 minutes and resuspended in 11 µl of electroporation buffer R containing 1 µg of either paPX1-SUMO1-KGG-mCherry or paPX1-SUMO2-KGG-mCherry PiggyBac expression vectors along with 0.2 µg of Super PiggyBac transposase (System Biosciences). Electroporation was performed at 1150 V, 1 pulse, 30 mSec and cells plated in mTESR containing Y27632. 5 days after electroporation, mCherry positive cells were selected by fluorescence activated cell sorting (FACS) using an SH800 cell sorter (Sony). Monoclonal cell lines were prepared from the bulk sorted population by plating at low density on geltrex coated dishes and individual clones picked using 3.2 mm cloning discs

(Sigma Aldrich) soaked in TrypLE select. Cell lines were then expanded and analysed to check for expression of mCherry and His-SUMO1/2. Transfection of SENP1 mRNAs was performed using the same protocol.

***In vitro* differentiation assay**

For assessment of pluripotency *in vitro*, $1x10^4$ hiPSCs were seeded into the wells of v-bottomed 96 well plates in TESR medium supplemented with Y27632 and centrifuged at 300xg for 5 minutes. After 48 hours the resultant embryoid bodies were picked from the v-bottom plates using a pipette and seeded on gelatin coated dishes in knockout DMEM medium supplemented with 20% knockout serum replacement, 1x non-essential amino acids, 1x glutamax, 100 µM 2-mercaptoethanol. The medium was changed every 3 - 4 days and the cells were fixed with 4% formaldehyde and the expression of germ layer markers analysed by IF on day 20 of differentiation.

Flow cytometry for cell cycle assessment and pluripotency markers. For cell cycle analysis and staining for pluripotency markers ChiPS4 cells were harvested using standard procedures, washed and fixed with ice cold 70% ethanol or 4% formaldehyde for the analysis of cell cycle or NANOG staining respectively. Next cells were stained with propidium iodide or anti-NANOG primary antibody, followed by Alexa 488 conjugated secondary antibody and analysed by flow cytometry using a Canto analyser (Becton Dickson). Data was then analysed using FlowJo 10.

Immunofluorescence, cell painting assay and high content microscopy. For IF assays ChiPS4 cells were seeded on µ-Slide 8 Well (ibidi) or 96 well plates suitable for high content microscopy (Nunc). Standard IF procedure was used where appropriate. Briefly, following treatments cells were washed with PBS, fixed with 4% formaldehyde, blocked in 5% BSA in PBS-T and incubated with primary and Alexa conjugated secondary antibodies and co-stained with DAPI and/or Cy5 Cell Mask (Life Technologies). Cell painting was performed as described[26]. Imaging and subsequent analysis was performed using INCell Analyzer systems (GE Healthcare) and Spotfire (Tibco).

Protein sample preparation and Western blotting (WB). ChiPS4 were maintained in a stable culture as described before and treated with inhibitors for a stated time and dose, usually 400nM ML792 was used for 24h or 48h. For WB cells were washed with PBS +/+ and directly lysed in an appropriate volume of 2x Laemmli buffer (approximately 200 µl of buffer was used per 0.5x10^6 cells) (LD; [4% SDS; 20% Glycerol; 120mM 1 M Tris-Cl (pH 6.8); 0.02% w/v bromophenol blue]) and subsequently sonicated using Bioruptor Twin (Diagenode). Protein content was assessed using BCA protein assay (ThermoFisher Scientific) and for most purposes 15 µg of total protein was loaded per lane on SDS-PAGE gel (NuPage 4-12% polyacrylamide, Bis-Tris with MOPS buffer). Proteins were transferred to PVDF membrane using iBlot™ 2 Gel Transfer Device (Invitrogen). Membranes were blocked for 1h in 5% milk in TBS-T and incubated overnight with primary antibodies and 1h with secondary HRP conjugated antibodies before being developed using enhanced chemiluminescence (ThermoFisher Scientific).

NiNTA purification. Cells were washed with PBS and scraped in PBS containing 1mM N-ethylmaleimide. The cells were then collected by centrifugation at 300 xg for 5 minutes and the pellets weighed. An aliquot of the cells was lysed in 1.2x NuPage sample buffer (ThermoFisher Scientific) for analysis by Western blotting. The remaining cell pellets (approximately 1 g) were lysed with 5x the pellet weight of lysis buffer (6 M guanidine-HCl, 100 mM sodium phosphate buffer (pH 8.0), 10 mM Tris-HCl (pH 8.0), 10 mM imidazole and 5 mM 2-mercaptoethanol). DNA was sheared by sonication using a probe sonicator (3min, 35% amplitude, 20sec pulses, 20sec intervals on ice and the samples centrifuged at 4000 rpm for 15min at 4°C to remove insoluble material). The protein concentration of the lysate was determined using BCA assay and 6.5mg of total protein from each sample was then incubated overnight at 4˚C with 50 µl of packed pre-equilibrated Ni-NTA agarose beads. After the overnight incubation the supernatant was removed and the beads were washed once with 10 resin volumes of lysis buffer, followed by 1 wash with 10 resin volumes of 8 M urea, 100 mM sodium phosphate buffer (pH 8.0), 10 mM Tris-HCl (pH 8.0), 10 mM imidazole and 5 mM 2-mercaptoethanol, and then 6 washes with 10 resin volumes of 8 M urea, 100 mM sodium phosphate buffer (pH 6.3), 10 mM Tris-HCl (pH 8.0), 10 mM imidazole and 5 mM 2-

667 mercaptoethanol. Proteins were eluted from Ni-NTA agarose beads with 125 µl 1.2x NuPAGE
668 sample buffer for SDS-PAGE.
669
670 **Mass Spectrometry based proteomics and quantitative data analysis**
671 Three proteomic experiments are described in this study;
672 (1) Changes in total proteome of ChiPS4 cells during ML792 treatment
673 ChiPS4 cells were either DMSO treated (0 hours condition), or treated with 400nM ML792 for
674 24 hours or 48 hours. Four replicates of each condition were prepared. Crude cell extracts
675 were made to a protein concentration of between 1 and 2 mg/ml by addition of 1.2x NuPAGE
676 sample buffer to PBS washed cells followed by sonication. For each replicate 25µg protein
677 was fractionated by SDS-PAGE (NuPage 10% polyacrylamide, Bis-Tris with MOPS buffer—
678 Invitrogen) and stained with Coomassie blue. Each lane was excised into four roughly equally
679 sized slices and peptides were extracted by tryptic digestion[63] including alkylation with
680 chloroacetamide. Peptides were resuspended in 35 µL 0.1% TFA 0.5% acetic acid and 10uL of
681 each sample was analysed by LC-MS/MS. This was performed using a Q Exactive mass
682 spectrometer (Thermo Scientific) coupled to an EASY-nLC 1000 liquid chromatography system
683 (Thermo Scientific), using an EASY-Spray ion source (Thermo Scientific) running a 75 µm x 500
684 mm EASY-Spray column at 45ºC. A 240 minute elution gradient with a top 10 data-dependent
685 method was applied. Full scan spectra (m/z 300–1800) were acquired with resolution R =
686 70,000 at m/z 200 (after accumulation to a target value of 1,000,000 ions with maximum
687 injection time of 20 ms). The 10 most intense ions were fragmented by HCD and measured
688 with a resolution of R = 17,500 at m/z 200 (target value of 500,000 ions and maximum
689 injection time of 60 ms) and intensity threshold of $2.1x10^4$. Peptide match was set to
690 'preferred', a 40 second dynamic exclusion list was applied and ions were ignored if they had
691 unassigned charge state 1, 8 or >8. Data analysis used MaxQuant version 1.6.1.0[64]. Default
692 setting were used except the match between runs option was enabled, which matched
693 identified peaks among slices from the same position in the gel as well as one slice higher or
694 lower. The uniport human proteome database (downloaded 24/02/2015 - 73920 entries)
695 digested with Trypsin/P was used as search space. LFQ intensities were required for each slice
696 but LFQ normalization was switched off. Manual LFQ normalization was done by calculating
697 the LFQ ratio for each protein in each slice compared to the average LFQ for the same protein
698 across all equivalent slices in the other lanes. This was done only for proteins with LFQ

intensities reported in all 12 equivalent slices. The median protein Slice LFQ/Average LFQ ratio was used to normalize all protein LFQ values for each slice. The final protein LFQ intensity per lane (and therefore sample) was calculated by the sum of LFQ values for that protein intensity in all four slices. Downstream data processing used Perseus v1.6.1.1[65]. Proteins were only carried forward if an LFQ intensity was reported in all four replicates of at least one condition. Zero intensity values were replaced from log2 transformed data (default settings) and outliers were defined by 5% FDR from Student's t-test using an S0 value of 0.1. A summary of these data can be found in Supplementary Data File 1

(2) Characterisation of ChiPS4 cells stably expressing 6His-SUMO1-KGG-mCherry and 6His-SUMO2-KGG-mCherry.

Crude cell extracts were prepared in triplicate from ChiPS4 cells, ChiPS4-6His-SUMO1-KGG-mCherry and ChiPS4-SUMO2-KGG-mCherry cells and fractionated by SDS-PAGE as described above. In an almost identical manner gels were sectioned into four slices per lane, tryptic peptides prepared, peptides analysed by LC-MS/MS, and the resultant raw data processed by MaxQuant. The only exceptions being the inclusion of a second sequence database containing the two 6His-SUMO-KGG-mCherry constructs, and the use of MaxQuant LFQ normalization. Two MaxQuant runs were performed; the first aggregating all slices per lane into a single output ("by lane"), and the second considering each slice separately ("by slice"). The former was used to determine cell-specific changes in protein abundance from the proteinGroups.txt file, and the latter used the peptides.txt file to monitor differences in abundance of SUMO-specific peptides between samples, to infer overexpression levels. For the whole cell proteome change analysis only proteins with data in all three replicates of at least one condition were carried forward. In Perseus zero intensity values were replaced from log2 transformed data (default settings) and outliers were defined by 5% FDR from Student's t-test using an S0 value of 0.1. A summary of these data can be found in Supplementary Data File 2.

(3) Identification of SUMO1 and SUMO2 modified proteins from ChiPS4 cells.

Two repeats of this experiment were performed using approximately $0.5x10^8$ cells of ChiPS4-6HisSUMO1-KGG-mCherry and ChiPS4-6HisSUMO2-KGG-mCherry per replicate. Samples were taken at different steps of the protocol to assess different fractions. These were; crude cell extracts, NiNTA column elutions and GlyGly-K immunoprecipitations. The last being the source of SUMO-substrate branched peptides. The whole procedure was carried out as

described previously[66]. In brief, crude cell lysates were prepared of which approximately 100 μg was retained for whole proteome analysis as described for experiments 1 and 2 above. The remaining lysate (~20 mg protein) was used for NiNTA chromatographic enrichment of 6His-SUMO conjugates. Elutions from the NiNTA columns were digested consecutively with LysC then GluC, of which 7% of each was retained for proteomic analysis and the remainder for GlyGly-K immunoprecipitation. The final enriched fractions of LysC and LysC/GluC GG-K peptides were resuspended in a volume of 20 μl for proteomic analysis. Peptides from whole cell extracts were analysed once by LC-MS/MS using the same system and settings as described for experiments 1 and 2 above except a 180 minute gradient was used with a top 12 data dependent method. NiNTA elution peptides were analysed identically except a top 10 data dependent method was employed and maximum MS/MS fill time was increased to 120ms. GG-K immunoprecipitated peptides were analysed twice. Firstly, 4 μl was fractionated over a 90 minute gradient and analysed using a top 5 data dependent method with a maximum MS/MS fill time of 200ms. Secondly, 11 μl of sample was fractionated over a 150 minute gradient and analysed using a top 3 method with a maximum MS/MS injection time of 500ms. Data from WCE and NiNTA elutions were processed together in MaxQuant using Trypsin/P enzyme specificity (2 missed cleavages) for WCE samples and LysC (3 missed cleavages), or LysC+GluC_D/E (considering cleavage after D or E and 8 missed cleavages) for NiNTA elutions. GlyGly (K) and phospho (STY) modifications were selected. The human database and sequences of the two exogenous 6His-SUMO-KGG-mCherry constructs described above were used as search space. In all cases every raw file was treated as a separate 'experiment' in the design template such that protein or peptide intensities in each peptide sample were reported, allowing for manual normalization. Matching between runs was allowed but only for peptide samples from the same cellular fraction (WCE, NiNTA elution or GG-K IP), the same or adjacent gel slice, the same protease and the same LC elution gradient. For example, spectra from adjacent gel slices in the WCE fraction across all lanes were matched, and spectra from all GG-K IPs that were digested by the same enzymes were matched. Normalization followed a similar method as described above where 'equivalent' peptide samples (i.e. those from the same gel slice or 'equivalent' peptide samples) from different replicates were compared with one another. Manual normalization used a similar method described above. For each protein or peptide common to all equivalent peptide

762 samples the ratio of intensity in that sample to the average across all equivalent samples was
763 calculated. The median of that ratio of was used to normalize all protein or peptide intensities
764 for each sample. The final protein or peptide intensity per replicate was calculated by the sum
765 of all normalized intensities in samples derived from that replicate. Importantly, peptide
766 samples derived from SUMO1 and SUMO2 cells were considered equivalent for normalization
767 purposes, which assumes largely similar abundances of proteins or peptides across cell types.
768 Zero intensity values were replaced from log2 transformed data (default settings) and outliers
769 were defined by 5% FDR from Student's t-test using an S0 value of 0.1. A summary of these
770 data can be found in Supplementary Data File 3.
771
772 **Bioinformatic analysis of the SUMO site proteomics**. 429 proteins identified with at least one
773 SUMO1 or SUMO2 modification site were uploaded to STRING[67] for network analysis. Only
774 proteins associated by a minimum STRING interaction score of 0.7 (high confidence) were
775 included in the final network. Disconnected nodes were removed. Selected groups of
776 functionally related proteins were resubmitted to STRING to create smaller sub-networks.
777 These were visualised in Cytoscape v 3.7.2[68] allowing the graphical display of numbers of sites
778 identified and total GG-K peptide intensity into the protein networks.
779
780 **SUMO1 ChIP-seq.** Cells were dispersed with TrypLE select as previously described, cross-
781 linked with 1% formaldehyde for 10 min, then quenched for 5 min with 125 mM glycine at
782 room temperature. Fixed cells were washed twice with ice-cold PBS, frozen in liquid nitrogen
783 and stored at -80°C. Frozen cell pellets were thawed on ice and resuspended in lysis buffer (5
784 mM PIPES pH 8; 85 mM KCl; 0.5% Igepal CA-630) supplemented with complete Protease
785 Inhibitor Cocktail without EDTA (Roche) and 20mM iodoacetamide. Nuclear extraction by
786 sonication[69] was performed using a Bioruptor Twin sonicator on low power with 2-3 cycles
787 (15 secs on and 15secs off). Nuclei were collected by centrifugation at 2500 xg for 5 min at
788 4°C, washed once in 1 ml of lysis buffer, and visualised by microscopy to confirm the release
789 of nuclei from cells. Nuclei were resuspended to a final concentration of 2 x 10^7 nuclei/ml in
790 nuclei lysis buffer (50 mM Tris–HCl, pH 8.0, 10 mM EDTA, 1% SDS) plus protease inhibitors.
791 To shear chromatin to fragments of approximately 500bp (range 100–800 bp) in length,
792 samples were sonicated in a volume of 300 µl for 20 cycles (10 min total sonication time)
793 using the Bioruptor Twin on high power. Sonicated lysates were then clarified by

794 centrifugation for 15000 xg for 10 minutes at 4°C. Input DNA was purified using an IPURE kit
795 (Diagenode) according to the manufacturer's instructions. For each IP, 15-µL aliquots of
796 Protein G Dynabeads (10 mg/mL, Dynal) were washed in 500 µl PBS then 5 µg of sheep anti-
797 SUMO-1 antibody was bound in 1ml of PBS containing 0.1% IgG free BSA and protease
798 inhibitor cocktail for 1 - 4hr at 4°C with agitation. 25 µg of chromatin was premixed with 8
799 volumes of IP dilution buffer (20 mM Tris-HCl at pH 7.6, 0.625% Triton X-100, 182.5 mM NaCl)
800 then added to the antibody on Dynabeads and left overnight at 4°C with agitation. Samples
801 were then centrifuged for 1min at 500 xg, placed in a magnetic separation rack, and the beads
802 were washed once with 1ml of IP wash buffer 1 (20 mM Tris-HCl at pH 8.0, 2 mM EDTA, 0.2%
803 SDS, 0.5% Triton X-100, and 150 mM NaCl), once with 1ml of IP wash buffer 2 (20 mM Tris-
804 HCl at pH 8.0, 2 mM EDTA, 0.2% SDS, 0.5% Triton X-100, and 500 mM NaCl) once with 1ml of
805 IP wash buffer 3 (0.25 M LiCl, 1 % NP-40, 1 % deoxycholate, 1 mM EDTA, 10 mM Tris pH 8.1),
806 and once with 1ml of TE buffer (1 mM EDTA, 10 mM Tris pH 8.1). Immunoprecipitated
807 material was eluted from the beads and DNA purified using Diagenode IPURE kit according to
808 manufacturer's protocol. The resulting DNA was used for Illumina library preparations.
809 Libraries from SUMO1 ChIP DNA and Input DNA were prepared using the NEBNext
810 Ultra II DNA library prep kit with sample purification beads according to the manufacturer's
811 instruction. Barcoding of the samples was performed using NEBNext Multiplex oligos Index
812 set 1 and 2. Chromatin size distribution was measured on Agilent Tapestation and sample
813 concentration quantified with Qubit® dsDNA HS (High Sensitivity) Assay Kit and Qubit®
814 Fluorometer (ThermoFisher Scientific). The sequencing was performed on a NextSeq 500
815 (Illumina): paired end, high throughput 2x75bp run. Chromatin input seq and SUMO1 ChIP-
816 seq data analysis were performed as described here:
817 http://www.compbio.dundee.ac.uk/user/mgierlinski/sumodiff/doc/analysis.html. Briefly,
818 ChIP-seq reads were mapped to human genome reference GRCh38 (repeat-masker filtered)
819 using bwa version 0.7.15. Then, peak calling was done with MACS2 version 2.1.0. Data was
820 also mapped and processed for the hg17 genome reference to be used for overlaps and
821 alignments with ENCODE TF database.
822
823 **ATAC-seq**. ATAC-seq libraries were generated following the Omni-ATAC protocol[70] without
824 enrichment for viable cells. Amplified barcoded DNA fragments were purified using NEB
825 Monarch PCR purification kit. The DNA concentration was measured with the Qubit High

826 Sensitivity assay (ThermoFisher) and the fragment sizes were determined on the TapeStation
827 Bioanalyser (Agilent). Samples were pooled in equimolar ratios. The pooled library was
828 subjected to a dual size selection using Promega Pronex beads 1.2/0.4 x beads:sample ratio
829 to enrich for fragments between 180 bp and 800 bp. Multiplexed libraries were sequenced
830 with 2x150 bp paired-end reads by Novogene with ~20 Mio reads per sample using NovaSeq
831 6000 S4 (Illumina). Fastq files were trimmed using trimmomatic-0.36 (CROP: 66) and aligned
832 to the human GRCh38 genome using bowtie2 with the parameter –X 1000. Peaks were called
833 using MACS2 *callpeak* function with the following parameters: -t "$1".bam -f BAMPE -n
834 "$1"_MACS -g 2.7e9 -q 0.05 --broad -B. Differential peaks were obtained using *DiffBind*,
835 doing pair-wise comparison of two time points to control DMSO treated samples. When
836 performing *dba.count*, a minOverlap was set to 3, requiring a peak to be observed in at least
837 3 datasets in order to be retained. Differential peaks were called using the *edgeR* method
838 during *dba.analyze*. Of the differentially called peaks, a second filtering step was performed
839 to retain only peaks that met an FDR < 0.00001 and a scores.fold > 0.58 (equal to a fold change
840 >1.5). A summary of these data can be found in Supplementary Data File 5. Non-changing
841 peaks were obtained from the *DiffBind* consensus peak set, with all differential peaks
842 removed (all timepoints, no extra thresholding). The non-differential peaks were randomly
843 subsampled to the same sample-size as differential peaks. The *bedtools intersect* function was
844 used to call overlap of ATAC-seq peaks with ChIP-seq data and genomic regions were
845 annotated using the *annotatePeaks.pl* command from the *HOMER* software.
846
847 **RNA preparation and real-time quantitative PCR (RT-qPCR)**. Total RNA was extracted using
848 RNeasy Mini Kit (Qiagen) and treated with the on-column RNase-Free DNase Set (Qiagen)
849 according to the manufacturer's instructions. RNA concentration was then measured using
850 NanoDrop and 1μg of total RNA per sample was subsequently used to perform a two-step
851 reverse transcription polymerase chain reaction (RT-PCR) using random hexamers and First
852 Strand cDNA Synthesis Kit (ThermoFisher Scientific). Each qPCR reaction contained PerfeCTa
853 SYBR Green FastMix ROX (Quantabio), forward and revers primer mix (200 nM final
854 concentration) and 6 ng of analysed cDNA and was set up in triplicates in MicroAmp™ Fast
855 Optical 96-Well or 384-Well Reaction Plates with Barcodes (Applied Biosystems™). The
856 sequences of primers used were as follows: NANOG (hNANOG_FOR624
857 ACAGGTGAAGACCTGGTTCC; hNANOG_REV722 GAGGCCTTCTGCGTCACA), SOX2 (hSOX2

_FOR907 TGGACAGTTACGCGCACAT; hSOX2_REV1121 CGAGTAGGACATGCTGTAGGT), OCT4A (hOCT4A_FOR825 CCCACACTGCAGCAGATCA and hOCT4A_REV1064 ACCACACTCGGACCACATCC), KLF4 (hKLF4_FOR1630 GGGCCCAATTACCCATCCTT and hKLF4_REV1706 GGCATGAGCTCTTGGTAATGG), TBP (hTBP_FOR896 TGTGCTCACCCACCAACAAT; hTBP_REV1013 TGCTCTGACTTTAGCACCTGTT), PRAME (hPRAME_F1661 TACCTGGAAGCTACCCACCT and hPRAME_R1892 GTGCCTGAGCAACTGATCCA). Data were collected using QuantStudio™ 6 Flex Real-Time PCR Instrument and analysed using a corresponding software (Applied Biosystems™). Relative amounts of specifically amplified cDNA were calculated using TBP amplicons as normalizers.

RNA-seq. RNA samples were collected and prepared as for standard RNA extraction procedure. Samples were quality controlled using Qubit (Thermo Fisher Scientific) and TapeStation (Agilent) and sent for further analysis to Novogen, who prepared the Illumina Library using NEB Next® Ultra™ RNA Library Prep Kit. These libraries were sequenced using the Illumina NovaSeq 6000 S4 (Illumina PE150, Q30 ≥ 80% delivering 6G raw data per sample). Following data QC, RNA-seq reads were mapped to human genome reference GRCh38 using *STAR* version 2.7.3a. Ensembl gene annotations release 99 were used. For HERV expression annotations of 519,060 loci from Human Endogenous Retrovirus Database (https://herv.img.cas.cz) were used (database accessed on 26 June 2020). Read counts per gene/HERV were found in the same *STAR* run. Features with at least 10 counts in at least one sample were selected. Downstream analysis was performed in RStudio using *R* version 4.0.2. The code is available at GitHiub (https://github.com/bartongroup/MG_SumoDiff2). Differential expression was performed using *edgeR* version 3.30.3. Gene/HERV profiles were calculated as a $\log_2$ ratio of normalised counts at time point 4, 8, 24 and 48 hours versus DMSO. These profiles were used for clustering. A summary of these data can be found in Supplementary Data File 6. A notebook with details of RNA-seq, ChIP-seq and ATAC-seq data analyses is available at http://www.compbio.dundee.ac.uk/user/mgierlinski/sumodiff2/doc/analysis.html.

RNA seq protein-coding genes clustering analysis. Protein-coding gene data were cross-referenced to protein names using UniProt mapping (uniport.org) leaving 15333 entries. These along with fold change values at each of the time-points of ML792 treatment were

890 uploaded to Perseus (v1.6.1.1) and each protein coding gene annotated with GOBP, GOMF,
891 GOCC, KEGG, Pfam, GSEA, Keywords, Corum, PRINTS, Prosite, SMART and Reactome terms
892 using the UniProt reference. Based on the entire time-course quantitative data hierarchical
893 clustering was performed using an unconstrained Euclidian distance method pre-processed
894 with k-means. 300 clusters were considered with 20 iterations and 5 restarts. Based on the
895 hierarchical clustering, multiple rounds of cluster definition were made using a range of
896 fbetween 5 and 10 clusters. 9 clusters gave the highest number of significantly enriched or
897 depleted categorical terms (ontologies) according to Fisher's exact test employing a 2% FDR
898 truncation. They were labelled A-I and carried forward for functional enrichment analysis. To
899 reduce the quantitative data to a single metric (for STRING analysis) a slope value (log_2 fold
900 change per hour) for each entry was calculated based on all time-point fold change values in
901 addition to a zero value at 0h. These RNA-seq data and the ML792 proteomics data
902 (experiment 1) were combined using gene names as cross-reference. This gave 4526 entries
903 with data in both the proteomics and RNA-seq experiments. A summary of these data can be
904 found in Supplementary Data File 4.
905

References

1. Chambers, I. *et al.* Nanog safeguards pluripotency and mediates germline development. *Nature* **450**, 1230-1234 (2007).
2. Young, R.A. Control of the embryonic stem cell state. *Cell* **144**, 940-954 (2011).
3. Takahashi, K. *et al.* Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* **131**, 861-872 (2007).
4. Takahashi, K. & Yamanaka, S. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* **126**, 663-676 (2006).
5. Yu, J. *et al.* Induced pluripotent stem cell lines derived from human somatic cells. *Science* **318**, 1917-1920 (2007).
6. Vierbuchen, T. *et al.* Direct conversion of fibroblasts to functional neurons by defined factors. *Nature* **463**, 1035-1041 (2010).
7. Apostolou, E. & Hochedlinger, K. Chromatin dynamics during cellular reprogramming. *Nature* **502**, 462-471 (2013).
8. Cheloufi, S. *et al.* The histone chaperone CAF-1 safeguards somatic cell identity. *Nature* **528**, 218-224 (2015).
9. Borkent, M. *et al.* A Serial shRNA Screen for Roadblocks to Reprogramming Identifies the Protein Modifier SUMO2. *Stem Cell Reports* **6**, 704-716 (2016).
10. Cossec, J.C. *et al.* SUMO Safeguards Somatic and Pluripotent Cell Identities by Enforcing Distinct Chromatin States. *Cell Stem Cell* **23**, 742-757 e748 (2018).
11. Flotho, A. & Melchior, F. Sumoylation: a regulatory protein modification in health and disease. *Annu Rev Biochem* **82**, 357-385 (2013).
12. Hay, R.T. SUMO: a history of modification. *Mol Cell* **18**, 1-12 (2005).
13. Rodriguez, M.S., Dargemont, C. & Hay, R.T. SUMO-1 conjugation in vivo requires both a consensus modification motif and nuclear targeting. *J Biol Chem* **276**, 12654-12659 (2001).
14. Sampson, D.A., Wang, M. & Matunis, M.J. The small ubiquitin-like modifier-1 (SUMO-1) consensus sequence mediates Ubc9 binding and is essential for SUMO-1 modification. *J Biol Chem* **276**, 21664-21669 (2001).
15. Tatham, M.H. *et al.* Polymeric chains of SUMO-2 and SUMO-3 are conjugated to protein substrates by SAE1/SAE2 and Ubc9. *J Biol Chem* **276**, 35368-35374 (2001).
16. Song, J., Durrin, L.K., Wilkinson, T.A., Krontiris, T.G. & Chen, Y. Identification of a SUMO-binding motif that recognizes SUMO-modified proteins. *Proc Natl Acad Sci U S A* **101**, 14373-14378 (2004).
17. Williams, R.L. *et al.* Myeloid leukaemia inhibitory factor maintains the developmental potential of embryonic stem cells. *Nature* **336**, 684-687 (1988).
18. Ying, Q.L., Nichols, J., Chambers, I. & Smith, A. BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3. *Cell* **115**, 281-292 (2003).
19. Xu, R.H. *et al.* BMP4 initiates human embryonic stem cell differentiation to trophoblast. *Nat Biotechnol* **20**, 1261-1264 (2002).
20. Xu, R.H. *et al.* Basic FGF and suppression of BMP signaling sustain undifferentiated proliferation of human ES cells. *Nat Methods* **2**, 185-190 (2005).
21. Humphrey, R.K. *et al.* Maintenance of pluripotency in human embryonic stem cells is STAT3 independent. *Stem Cells* **22**, 522-530 (2004).
22. Xu, C. *et al.* Basic fibroblast growth factor supports undifferentiated human embryonic stem cell growth without conditioned medium. *Stem Cells* **23**, 315-323 (2005).

23. James, D., Levine, A.J., Besser, D. & Hemmati-Brivanlou, A. TGFbeta/activin/nodal signaling is necessary for the maintenance of pluripotency in human embryonic stem cells. *Development* **132**, 1273-1282 (2005).
24. Rossant, J. Mouse and human blastocyst-derived stem cells: vive les differences. *Development* **142**, 9-12 (2015).
25. He, X. *et al.* Probing the roles of SUMOylation in cancer cell biology by using a selective SAE inhibitor. *Nature Chemical Biology* **13**, 1164-1171 (2017).
26. Bray, M.A. *et al.* Cell Painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nat Protoc* **11**, 1757-1774 (2016).
27. Tatham, M.H. *et al.* RNF4 is a poly-SUMO-specific E3 ubiquitin ligase required for arsenic-induced PML degradation. *Nat Cell Biol* **10**, 538-546 (2008).
28. Seifert, A., Schofield, P., Barton, G.J. & Hay, R.T. Proteotoxic stress reprograms the chromatin landscape of SUMO modification. *Sci Signal* **8**, rs7 (2015).
29. Yang, B.X. *et al.* Systematic identification of factors for provirus silencing in embryonic stem cells. *Cell* **163**, 230-245 (2015).
30. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38**, 576-589 (2010).
31. Gerstein, M.B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91-100 (2012).
32. Wang, Y. *et al.* Endogenous miRNA sponge lincRNA-RoR regulates Oct4, Nanog, and Sox2 in human embryonic stem cell self-renewal. *Dev Cell* **25**, 69-80 (2013).
33. Girdwood, D. *et al.* P300 transcriptional repression is mediated by SUMO modification. *Mol Cell* **11**, 1043-1054 (2003).
34. Paces, J., Pavlicek, A. & Paces, V. HERVd: database of human endogenous retroviruses. *Nucleic Acids Res* **30**, 205-206 (2002).
35. Paces, J. *et al.* HERVd: the Human Endogenous RetroViruses Database: update. *Nucleic Acids Res* **32**, D50 (2004).
36. Tammsalu, T. *et al.* Proteome-wide identification of SUMO2 modification sites. *Sci Signal* **7**, rs2 (2014).
37. Wolf, D. & Goff, S.P. TRIM28 Mediates Primer Binding Site-Targeted Silencing of Murine Leukemia Virus in Embryonic Cells. *Cell* **131**, 46-57 (2007).
38. Tie, C.H. *et al.* KAP1 regulates endogenous retroviruses in adult human cells and contributes to innate immune control. *EMBO Rep* **19** (2018).
39. Schmidt, N. *et al.* An influenza virus-triggered SUMO switch orchestrates co-opted endogenous retroviruses to stimulate host antiviral immunity. *Proceedings of the National Academy of Sciences* **116**, 17399-17408 (2019).
40. Oleksiewicz, U. *et al.* TRIM28 and Interacting KRAB-ZNFs Control Self-Renewal of Human Pluripotent Stem Cells through Epigenetic Repression of Pro-differentiation Genes. *Stem Cell Reports* **9**, 2065-2080 (2017).
41. Goke, J. *et al.* Dynamic transcription of distinct classes of endogenous retroviral elements marks specific populations of early human embryonic cells. *Cell Stem Cell* **16**, 135-141 (2015).
42. Glinsky, G.V. Transposable Elements and DNA Methylation Create in Embryonic Stem Cells Human-Specific Regulatory Sequences Associated with Distal Enhancers and Noncoding RNAs. *Genome Biol Evol* **7**, 1432-1454 (2015).

43. Santoni, F.A., Guerra, J. & Luban, J. HERV-H RNA is abundant in human embryonic stem cells and a precise marker for pluripotency. *Retrovirology* **9**, 111 (2012).
44. Lu, X. *et al.* The retrovirus HERVH is a long noncoding RNA required for human embryonic stem cell identity. *Nat Struct Mol Biol* **21**, 423-425 (2014).
45. Ohnuki, M. *et al.* Dynamic regulation of human endogenous retroviruses mediates factor-induced reprogramming and differentiation potential. *Proc Natl Acad Sci U S A* **111**, 12426-12431 (2014).
46. Zhang, Y. *et al.* Transcriptionally active HERV-H retrotransposons demarcate topologically associating domains in human pluripotent stem cells. *Nat Genet* **51**, 1380-1388 (2019).
47. Merkenschlager, M. & Nora, E.P. CTCF and Cohesin in Genome Folding and Transcriptional Gene Regulation. *Annu Rev Genomics Hum Genet* **17**, 17-43 (2016).
48. Rowley, M.J. & Corces, V.G. Organizational principles of 3D genome architecture. *Nat Rev Genet* **19**, 789-800 (2018).
49. Bailey, S.D. *et al.* ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nat Commun* **2**, 6186 (2015).
50. Anno, Y.N. *et al.* Genome-wide evidence for an essential role of the human Staf/ZNF143 transcription factor in bidirectional transcription. *Nucleic Acids Res* **39**, 3116-3127 (2011).
51. Sathyan, K.M. *et al.* An improved auxin-inducible degron system preserves native protein levels and enables rapid and specific protein depletion. *Genes Dev* **33**, 1441-1455 (2019).
52. Theurillat, I. *et al.* Extensive SUMO Modification of Repressive Chromatin Factors Distinguishes Pluripotent from Somatic Cells. *Cell Rep* **32**, 108146 (2020).
53. Ikeda, H. *et al.* Characterization of an antigen that is recognized on a melanoma showing partial HLA loss by CTL expressing an NK inhibitory receptor. *Immunity* **6**, 199-208 (1997).
54. Kilpinen, S. *et al.* Systematic bioinformatic analysis of expression levels of 17,330 human genes across 9,783 samples from 175 types of healthy and pathological tissues. *Genome Biol* **9**, R139 (2008).
55. Zhang, L. *et al.* MSX2 Initiates and Accelerates Mesenchymal Stem/Stromal Cell Specification of hPSCs by Regulating TWIST1 and PRAME. *Stem Cell Reports* **11**, 497-513 (2018).
56. Costessi, A. *et al.* The tumour antigen PRAME is a subunit of a Cul2 ubiquitin ligase and associates with active NFY promoters. *EMBO J* **30**, 3786-3798 (2011).
57. Costessi, A. *et al.* The human EKC/KEOPS complex is recruited to Cullin2 ubiquitin ligases by the human tumour antigen PRAME. *PLoS One* **7**, e42822 (2012).
58. Perez-Riverol, Y. *et al.* The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res* **47**, D442-D450 (2019).
59. Hands, K.J., Cuchet-Lourenco, D., Everett, R.D. & Hay, R.T. PML isoforms in response to arsenic: high-resolution analysis of PML body structure and degradation. *J Cell Sci* **127**, 365-375 (2014).
60. Shen, L. *et al.* SUMO protease SENP1 induces isomerization of the scissile peptide bond. *Nature Structural & Molecular Biology* **13**, 1069-1077 (2006).
61. Ibrahim, A.F.M. *et al.* Antibody RING-Mediated Destruction of Endogenous Proteins. *Mol Cell* **79**, 155-166 e159 (2020).

62. Ludwig, T.E. *et al.* Derivation of human embryonic stem cells in defined conditions. *Nat Biotechnol* **24**, 185-187 (2006).
63. Shevchenko, A., Tomas, H., Havlis, J., Olsen, J.V. & Mann, M. In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nat Protoc* **1**, 2856-2860 (2006).
64. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol* **26**, 1367-1372 (2008).
65. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat Methods* **13**, 731-740 (2016).
66. Tammsalu, T. *et al.* Proteome-wide identification of SUMO modification sites by mass spectrometry. *Nat Protoc* **10**, 1374-1388 (2015).
67. Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* **47**, D607-D613 (2019).
68. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504 (2003).
69. Arrigoni, L. *et al.* Standardizing chromatin research: a simple and universal method for ChIP-seq. *Nucleic Acids Res* **44**, e67 (2016).
70. Corces, M.R. *et al.* Omni-ATAC-seq: Improved ATAC-seq protocol. *Protocol Exchange* (2017).

Figure 1.

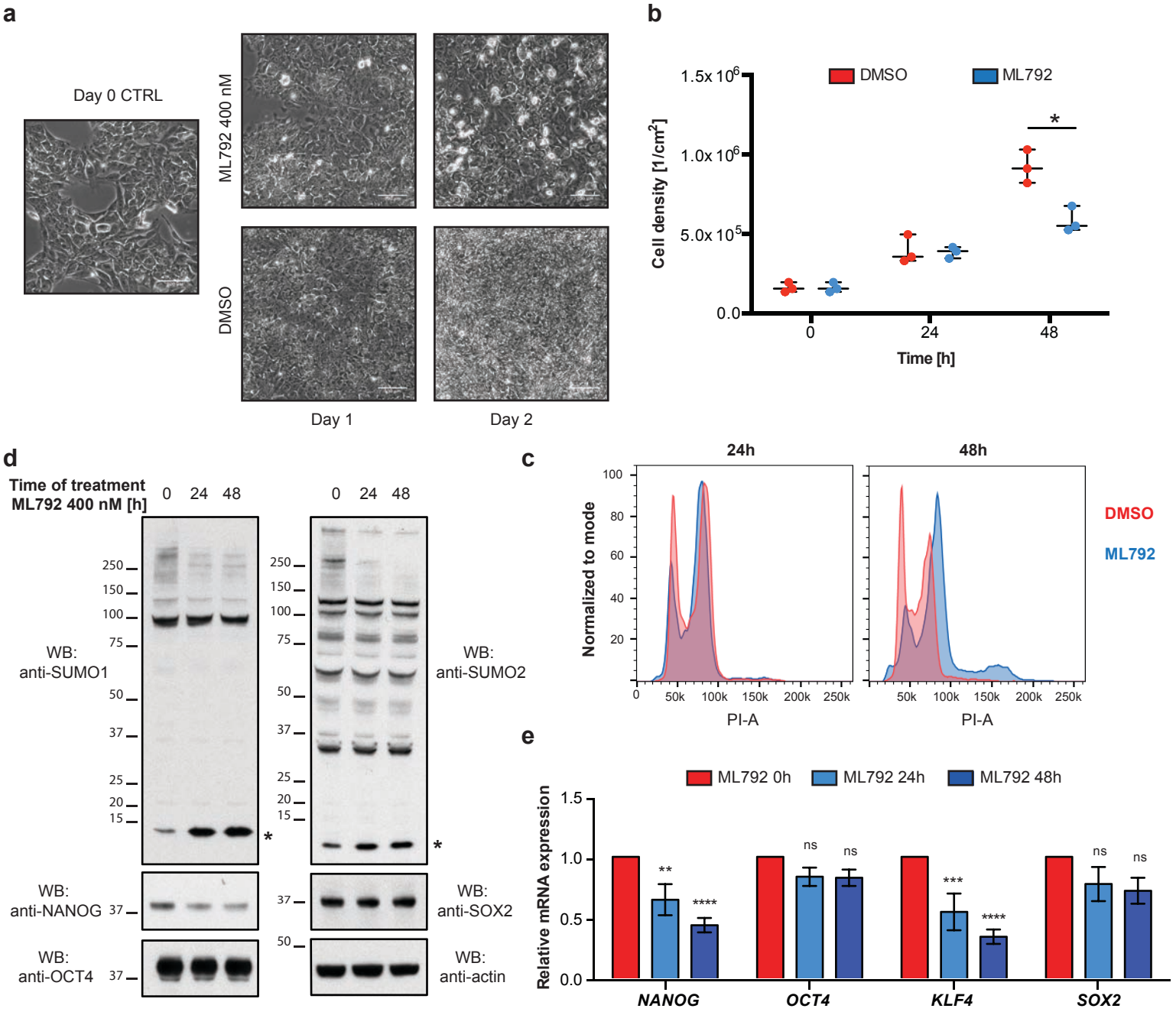


Figure 1. Inhibition of SUMO modification leads to loss of select pluripotency markers. ChiPS4 cells were treated with ML792 (400 nM) or DMSO vehicle for the indicated time and analysed by various approaches. **a.** Cell morphology was assessed using phase contrast microscopy (all images contain a 100 μm scale bar). **b.** For proliferation assessment, ChiPS4 cells were seeded at a standard density of $3x10^5$ cells/cm^2 in triplicate for each time point. The following day cells were treated with DMSO vehicle or ML792 and every 24h they were harvested using TrypLE select and counted. Data are plotted as mean cell density (line) with individual replicates (dots) shown N=3. Statistical significance was calculated with t-tests corrected for multiple comparisons using Holm-Sidak's method (* P<0.05 significantly different from the corresponding DMSO control) **c.** To analyse cell cycle distribution, cells were collected as in **b**, fixed, stained with propidium iodide (PI) and analysed by flow cytometry. Plots are a representative of three independent experiments d. Protein samples were analysed by Western blotting to determine conjugation levels of SUMO1, SUMO2/3 and abundance of key pluripotency markers NANOG, SOX2 and OCT4 using appropriate antibodies. Anti-Actin western blot was used as a loading control. * represents a band corresponding to free SUMO1 or SUMO2/3. **e.** Cells were treated with ML792 and after the indicated time they were lysed and total RNA was extracted. mRNA levels of *NANOG*, *OCT4*, *SOX2* and *KLF4* were determined by qPCR. Relative mRNA expression levels normalized to *TBP* were plotted as means ± SEM of four independent experiments. **P <0.01; ***P<0.001; ****P<0.0001 significantly different from the corresponding value for untreated control (two-way ANOVA followed by Sidak's multiple comparison test).

Figure 2.

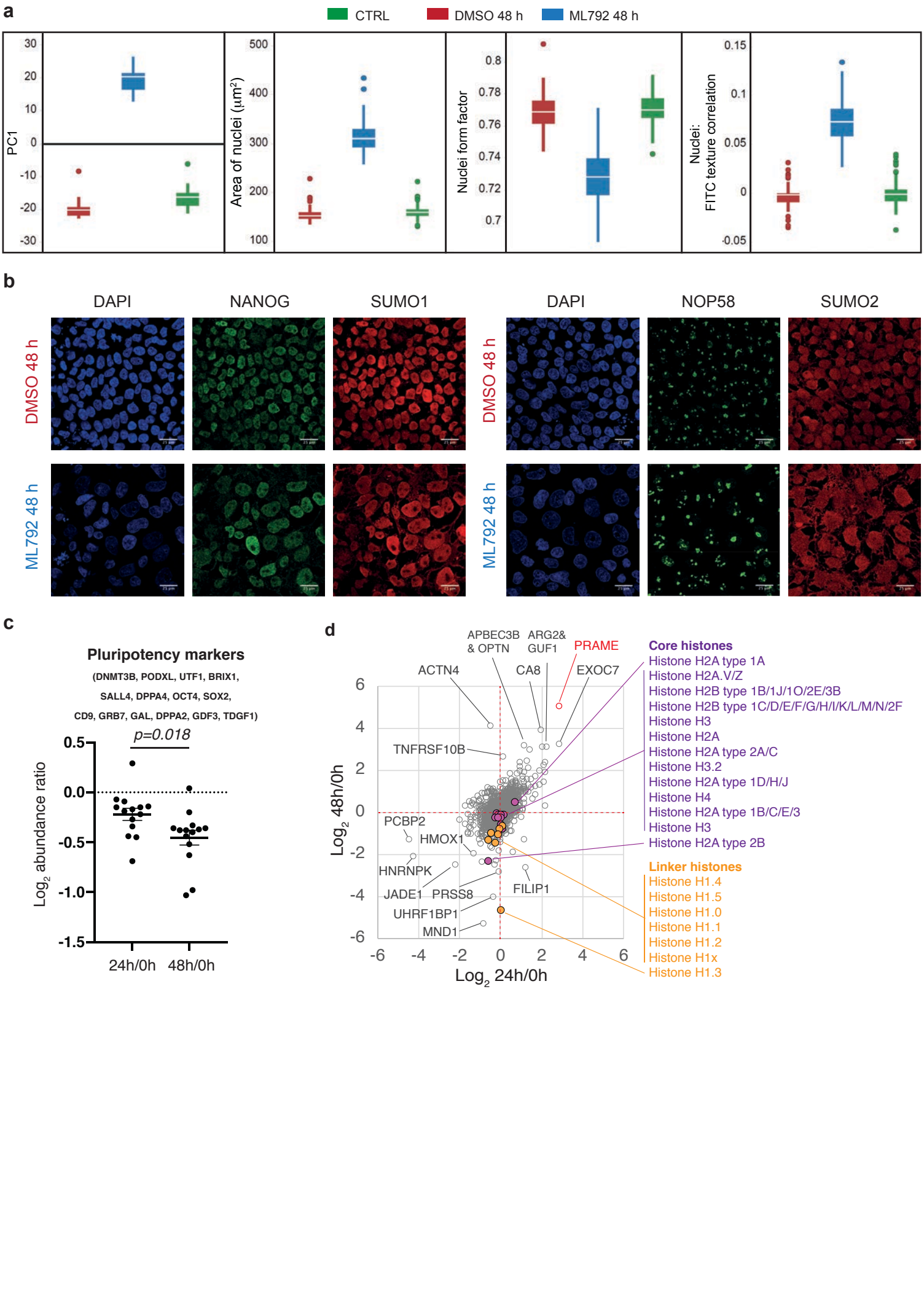


Figure 2. Change in morphology, but unchanged proteome in hiPSCs in presence of ML792. **a.** Cell painting analysis. ChiPS4 cells were treated with PBS, DMSO vehicle or 400 nM ML792 for 48 h. Cells were then stained, fixed and analysed using high content microscopy. The experiment was performed three times with 8 replicates per condition. Information extracted from cell painting analysis was focused on subcellular compartments most affected by ML792 treatment. Most variation between treatments and controls in PCA was captured by PC1. Selected graphs represent quantitation of individual measures contributing to the difference observed in PCA: area of nuclei; nuclei form factor (size and shape of nucleus); nuclei FITC texture correlation (size of nucleolar structures). **b.** ChiPS4 cells were treated for 48 h with DMSO vehicle or 400 nM ML792, fixed and stained with DAPI (blue), anti-SUMO1 or anti-SUMO2 (red) and anti-NANOG or anti-NOP58 (green) antibodies. IF images were obtained using a Leica SP8 confocal microscope and a 60x water immersion lens. All images contain 25 μm scale bar. **c.** Log_2 abundance ratio data extracted from whole cell proteomic analysis for the 14 indicated markers of pluripotency comparing 24 h and 48 h ML792 exposure to untreated cells. The plot shows individual data points and mean with standard error of the mean for the entire set. The result of a paired two-tailed student's t-test is shown. **d.** Scatter plot of Log_2 24h/0h and log_2 48h/0h abundance change for the entire 4741 protein whole cell proteomic dataset. Extreme outliers are indicated. All identified core and linker histones are represented by coloured markers, others are in grey. Linker histones were identified by STRING analysis as a functionally related group of proteins that are significantly reduced in abundance at 48h compared with 0h. Core histone proteins are indicated for reference.

Figure 3.

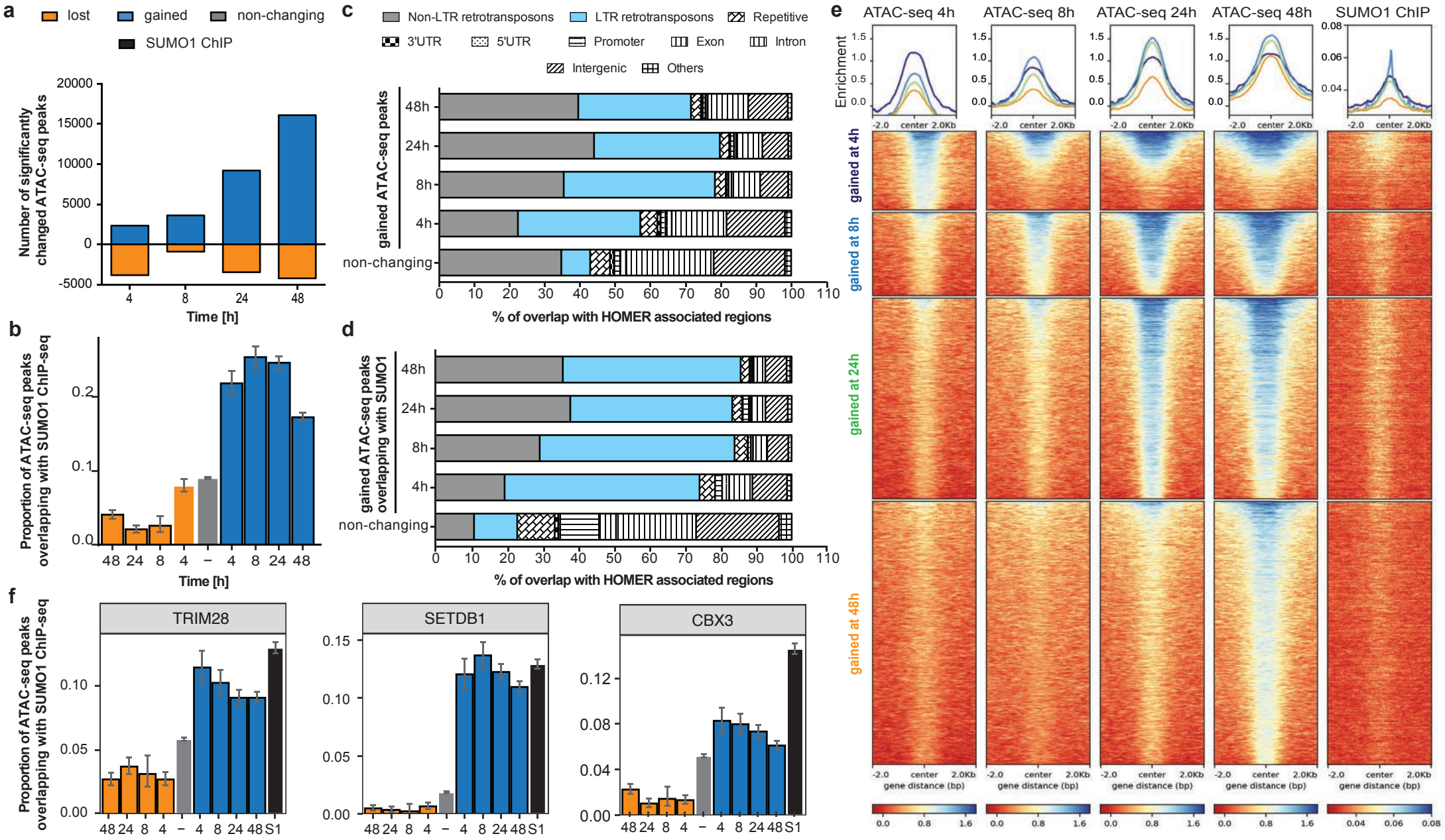

a
lost
gained
non-changing
SUMO1 ChIP
Number of significantly changed ATAC-seq peaks
Time [h]
b
Proportion of ATAC-seq peaks overlapping with SUMO1 ChIP-seq
Time [h]
c
Non-LTR retrotransposons
LTR retrotransposons
Repetitive
3'UTR
5'UTR
Promoter
Exon
Intron
Intergenic
Others
gained ATAC-seq peaks
non-changing
% of overlap with HOMER associated regions
d
gained ATAC-seq peaks overlapping with SUMO1
% of overlap with HOMER associated regions
e
ATAC-seq 4h
ATAC-seq 8h
ATAC-seq 24h
ATAC-seq 48h
SUMO1 ChIP
Enrichment
gained at 4h
gained at 8h
gained at 24h
gained at 48h
gene distance (bp)
f
TRIM28
SETDB1
CBX3
Proportion of ATAC-seq peaks overlapping with SUMO1 ChIP-seq

Figure 3. Removal of SUMO in hiPSCs increases chromatin accessibility. a. Numbers of significantly changed ATAC-seq peaks (applied criteria: $|\log_2 FC|>1.5$, minimal overlap of 3) at each time point in ChiPS4 cells treated with ML792 when compared to DMSO vehicle. Gained peaks are shown in blue, lost in orange. **b.** Proportion of ATAC-seq peaks gained or lost at different times of ML792 treatment overlapping with SUMO-1 ChIP-seq peaks (found in untreated cells). Overlap between non-changing ATAC-seq peaks and SUMO1 ChIP-seq is shown as a reference (grey). Thick black borders mark statistically significant changes (Fisher's exact test, corrected for multiple tests using Benjamini-Hochberg method), error bars are 95% confidence intervals (CI) of the proportion. **c.** Percentage of overlap between ATAC-seq peaks gained at each time point of ML792 treatment or non-changing ATAc-seq peaks found in DMSO vehicle control with HOMER-based annotations of chromatin regions (indicated). **d.** Percentage of overlap between peaks common for SUMO1 ChIP-seq/gained ATAC-seq peaks at each time point of ML792 treatment and HOMER-based annotations of chromatin regions. Different chromatin regions are represented as in the legend in **c**. **e.** Density plots for gained ATAC-seq changes at each time point following ML792 treatment. The sites are ordered by the time at which a change of >1.5 fold is first detected. The same order of genomic locations has been plotted for SUMO ChIP-seq peaks. Scale used for each density plot is based on the $\log_2$ ratio of ATAC-seq signal at each time point to the signal detected in DMSO control. Graphs at the top represent summary plots for the ATAC-seq signal at changing sites for the indicated time points. **f.** Proportion of ATAC-seq peaks gained or lost at different time points of ML792 treatment overlapping with various transcription factor ChIP-seq peaks (TRIM28, SETDB1, CBX3; data obtained from ENCODE database). Overlap between non-changing ATAC-seq peaks or SUMO1 ChIP-seq peaks are shown as references (grey and black bars respectively). Thick black borders mark statistically significant changes (Fisher's exact test, corrected for multiple tests using Benjamini-Hochberg method), error bars are 95% CI of the proportion.

Figure 4.

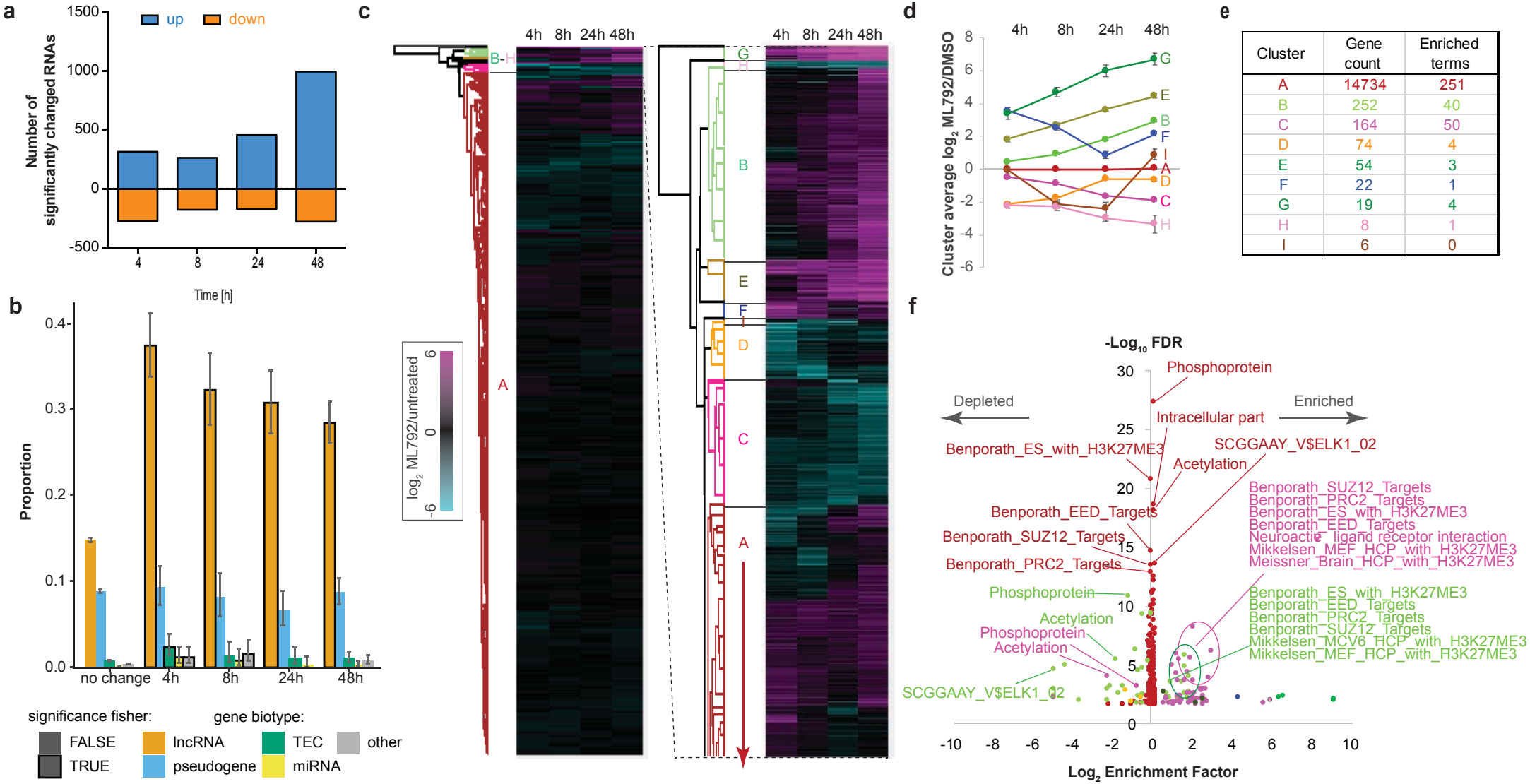

a
up
down
Number of significantly changed RNAs
1500
1000
500
0
-500
4
8
24
48
Time [h]
b
Proportion
0.4
0.3
0.2
0.1
0.0
no change
4h
8h
24h
48h
significance fisher:
FALSE
TRUE
gene biotype:
lncRNA
pseudogene
TEC
miRNA
other
c
4h 8h 24h 48h
B-H
A
G
B
E
F
I
D
C
log2 ML792/untreated
6
0
-6
d
Cluster average log2 ML792/DMSO
e
Cluster | Gene count | Enriched terms
A | 14734 | 251
B | 252 | 40
C | 164 | 50
D | 74 | 4
E | 54 | 3
F | 22 | 1
G | 19 | 4
H | 8 | 1
I | 6 | 0
f
-Log10 FDR
Depleted
Enriched
Phosphoprotein
Intracellular part
Benporath_ES_with_H3K27ME3
Acetylation
SCGGAAY_V$ELK1_02
Benporath_SUZ12_Targets
Benporath_PRC2_Targets
Benporath_ES_with_H3K27ME3
Benporath_EED_Targets
Neuroactive_ligand_receptor_interaction
Mikkelsen_MEF_HCP_with_H3K27ME3
Meissner_Brain_HCP_with_H3K27ME3
Benporath_EED_Targets
Benporath_SUZ12_Targets
Benporath_PRC2_Targets
Mikkelsen_MCV6_HCP_with_H3K27ME3
Mikkelsen_MEF_HCP_with_H3K27ME3
Log2 Enrichment Factor

Figure 4. Inhibition of SUMO modification in hiPSCs selectively alters transcription. a. Numbers of significantly changed RNAs in ChiPS4 cells (applied criteria: $|\log_2 FC| > 1.5$, FDR < 0.05, P < 0.01) at each time point of ML792 treatment (up regulated – blue, down regulated – orange) when compared to DMSO vehicle treated cells. **b.** Distribution of biotypes among non-protein coding genes that are significantly changed between DMSO and ML792 at a given time point and genes that do not change at any time point, with respect to DMSO. Error bars are 95% CI of a proportion. Black outlines indicate proportions significantly different (Fisher's exact test, $p < 0.05$) from the "no change" group. **c.** Hierarchical clustering analysis of protein-coding gene mRNAs during ML792 treatment. For each time-point data were represented as $\log_2$ fold change compared to DMSO treatment and clustered using a Euclidean distance function with linkage based on averages and k-means pre-processing. Data were binned into 9 row clusters (labelled A-H). The entire data set of 15333 entries is shown (left) and a zoom view of clusters B-H (right). **d.** Cluster-specific data shown as average and SEM at each time-point for all members of each group. **e.** Overview of functional group enrichment for each cluster relative to the entire dataset. Gene IDs were converted to protein IDs and the enrichment of different functional annotations was calculated by Fisher Exact Test with Benjamini-Hochberg truncation at 2% FDR. **f.** Scatter plot representation of all enriched categories for all clusters using $\log_2$ enrichment factor and $-\log_{10}$ FDR as co-ordinates.

Figure 5.

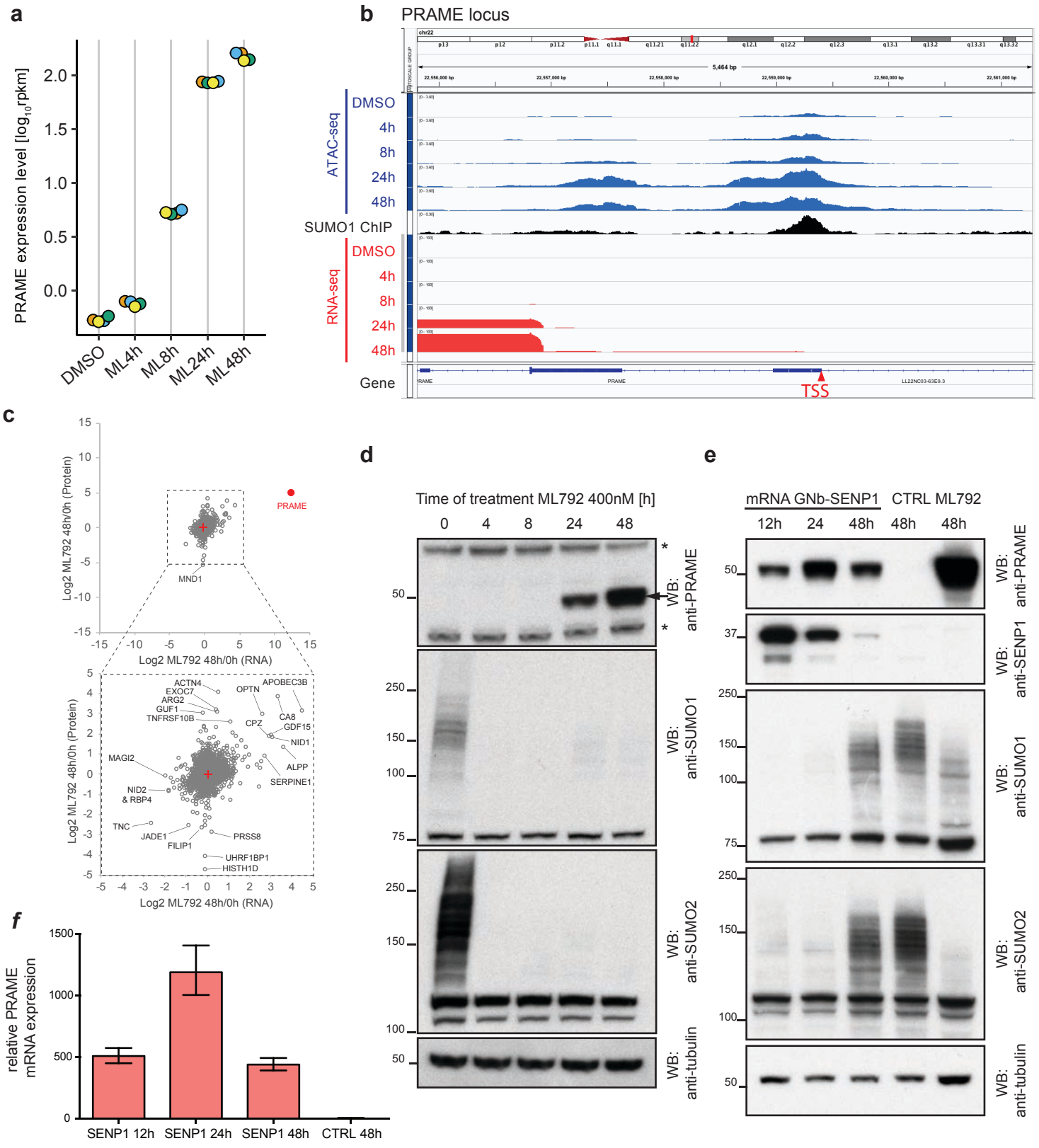


Figure 5. Inhibition of SUMOylation leads to expression of PRAME in hiPSCs. a. PRAME expression levels in RNA-seq samples at different time points were plotted as $\log_{10}$rpkm (reads per kilobase of transcript, per million mapped reads). Individual replicates are represented in different colours. **b.** Integrative Genomic Viewer was used to visualize changes in ATAC-seq (blue) and RNA-seq (red) occurring at the PRAME locus in response to ML792 treatment. SUMO1 ChIP-seq signal was also aligned and represented in black. All traces of the same type (ATAC-seq or RNA-seq) were normalized and scaled in the same way. **c.** Scatter-plot of RNA change versus protein-level change during 48 hours ML792 treatment (4498 common entries). Full scale is shown in the upper panel, and a smaller scale is shown below. Selected outliers are indicated. **d.** Western blot analysis of PRAME protein expression and SUMO1, SUMO2/3 conjugation levels after treatment of ChiPS4 cells with ML792 for the indicated times. Anti-tubulin was used as a loading control. * represents nonspecific bands, while the arrow indicates PRAME. **e**. ChiPS4 cells were transfected with capped and polyadenylated RNA encoding the catalytic domain of SUMO specific protease SENP1 (GNb-SENP1). Samples were collected at the times indicated after transfection and analysed by Western blotting for PRAME expression, SENP1 expression and conjugation levels of SUMO1 and SUMO2/3. Tubulin was used as loading control. **f**. ChiPS4 cells transfected with GNb-SENP1 were analysed by RT-qPCR to assess the relative expression of *PRAME* mRNA using *TBP* as a normalizing gene.

Figure 6.

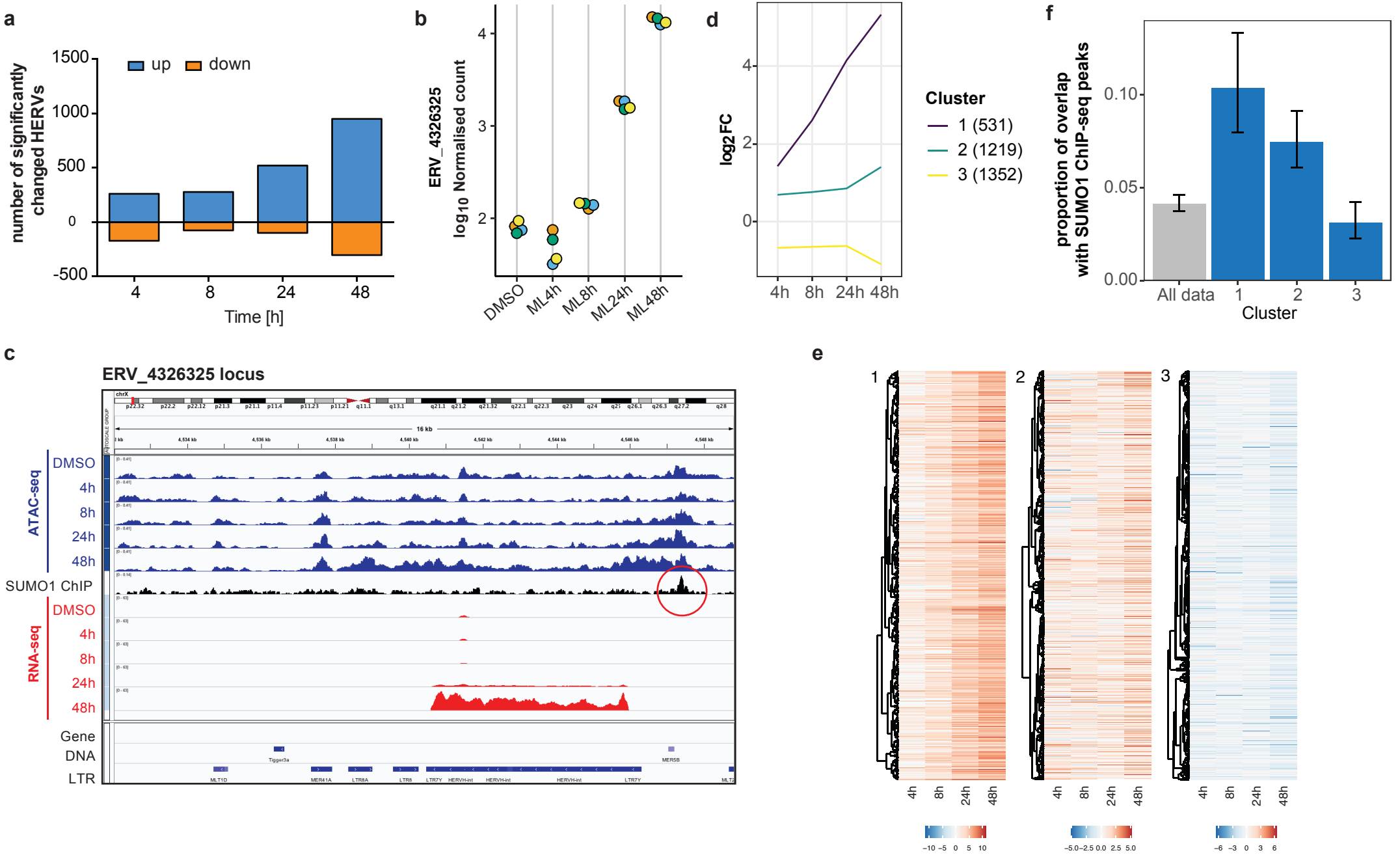

a
number of significantly changed HERVs
up
down
1500
1000
500
0
-500
4
8
24
48
Time [h]
b
ERV_4326325
log10 Normalised count
4
3
2
DMSO
ML4h
ML8h
ML24h
ML48h
d
log2FC
4
2
0
4h
8h
24h
48h
Cluster
1 (531)
2 (1219)
3 (1352)
f
proportion of overlap with SUMO1 ChIP-seq peaks
0.10
0.05
0.00
All data
1
2
3
Cluster
c
ERV_4326325 locus
chrX
16 kb
ATAC-seq
DMSO
4h
8h
24h
48h
SUMO1 ChIP
RNA-seq
DMSO
4h
8h
24h
48h
Gene
DNA
LTR
MLT1D
MER41A
LTR8A
LTR8
LTR7Y
HERVH-int
HERVH-int
HERVH-int
LTR7Y
MLT2
Tigger3a
MER5B
e
1
2
3
4h
8h
24h
48h
−10 −5 0 5 10
−5.0 −2.5 0.0 2.5 5.0
−6 −3 0 3 6

Figure 6. SUMO modification regulates HERV expression in hiPSCs. a. Numbers of significantly changed HERVs in ChiPS4 cells (applied criteria: $|\log_2 FC| > 1.5$, FDR < 0.05, P < 0.01) at each time point of ML792 treatment (up regulated – blue, down regulated – orange) when compared to DMSO vehicle treated cells. **b.** Expression levels of ERV_4326325 in RNA-seq samples at different time points following ML792 treatment of ChiPS4 cells plotted as $\log_{10}$normalized counts. Individual replicates are represented in different colours. **c.** Integrative Genomic Viewer display of changes in ATAC-seq (blue) and RNA-seq (red) occurring at the ERV_4326325 locus in response to ML792 treatment. SUMO1 ChIP-seq signal was also aligned and represented in black. All traces of the same type (ATAC-seq or RNA-seq) were normalized and scaled in the same way. **d.** Cluster centroids from clustering HERV profiles into 3 clusters, using k-means clustering. Profiles were created as $\log_2$ ratio between a given time point and DMSO normalised counts. Only HERVs with at least one statistically significant change, FDR < 0.05, between any time point and DMSO were selected for clustering. Numbers in brackets show the number of HERVs in each cluster. **e.** Content of each of the HERV clusters indicated in **d.** represented as a heatmap of $\log_2$ fold change between a given time point and DMSO. **f.** Overlap between HERVs with at least 10 counts detected in at least one sample and SUMO1 ChIP-seq peaks, for all HERV data and for each of the HERV clusters.

Figure 7.

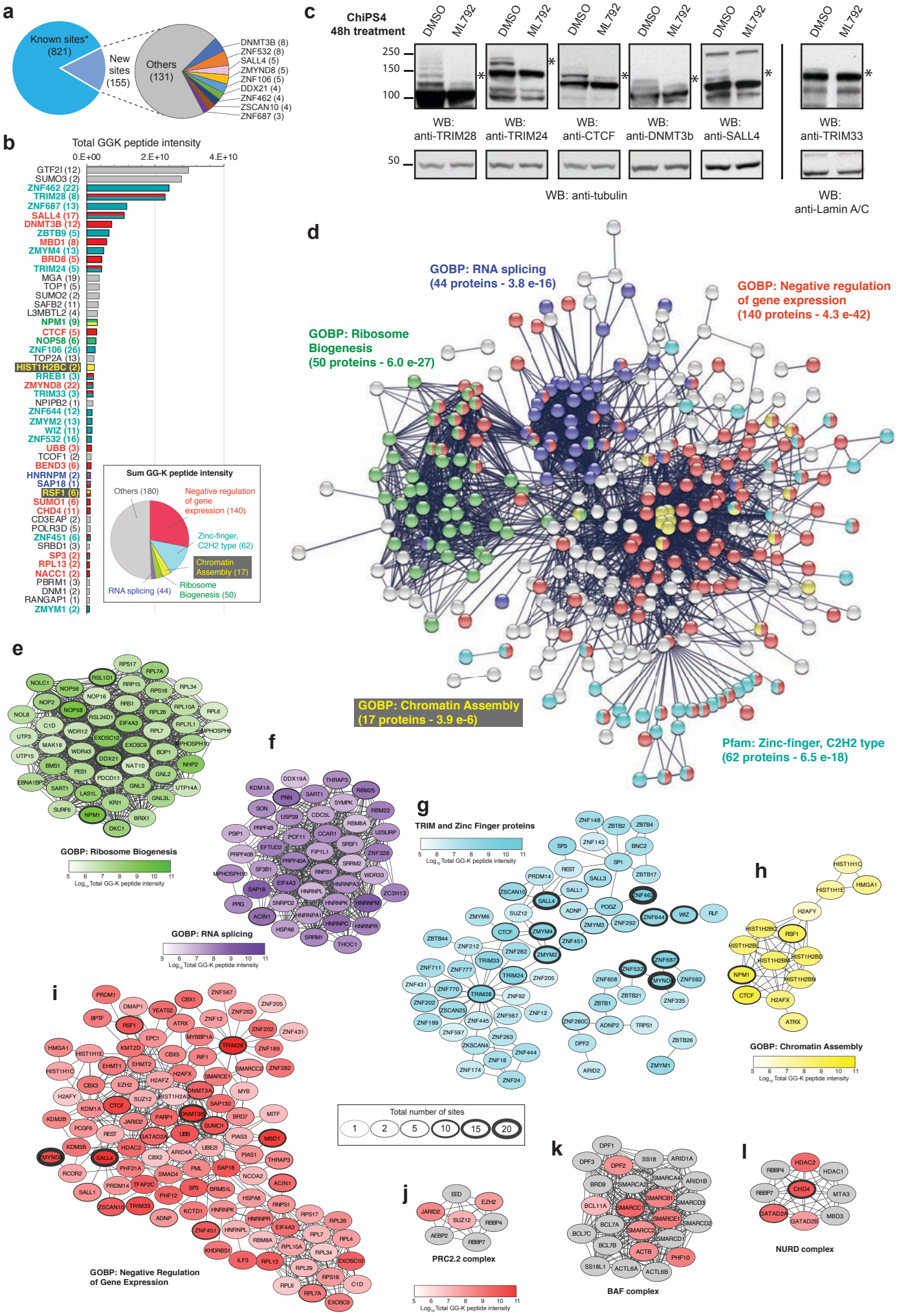


Figure 7. Identification of SUMO1 and SUMO2 targets in hiPSCs. a. 976 SUMO sites identified from 6His-SUMO1-KGG and 6His-SUMO2-KGG ChiPS4 cells, of which 155 were novel compared with previous high-throughput SUMO site proteomics studies (Supplementary Data File 2). Proteins with three or more novel sites are highlighted. **b.** Summary of the top 50 SUMO substrates by total GGK-peptide intensity for all identified sites. Gene names are shown with numbers of sites in brackets. Bars are colour coded by category shown in panel **d.** The insert shows contribution to total GGK peptide intensity of proteins from the categories shown in **d** (note categories are not mutually exclusive). **c.** Western blot analysis of ChiPS4 cells treated with ML792 or DMSO for 48h. Total protein extracts were probed with anti-TRIM28, anti-TRIM24, anti-CTCF, anti-DNMT3b, anti-SALL4, anti-TRIM33 and anti-tubulin or anti-Lamin A/C antibodies (loading controls). SUMO-modified proteins present above the band for unmodified proteins and disappearing in samples treated with ML792 are labelled with *. **d.** STRING interaction network of the 427 hiPSCs SUMO substrates. Only high confidence interactions were considered from 'Text mining', 'Experiments' and 'Databases' sources. Network PPI enrichment p-value $<1.0 \times 10^{-16}$. Nodes are coloured by functional or structural group as indicated. **e.-l.** Protein interaction networks derived from **d**. for the indicated functional groups. Node shade is proportional to $\log_{10}$ total GGK peptide intensity and border thickness indicates numbers of sites found. **j.-l.** shows individual network clusters for selected chromatin remodelling complexes. Grey nodes were not identified in the present study. TRIM proteins were included in **g.** to allow more complete network interactions.

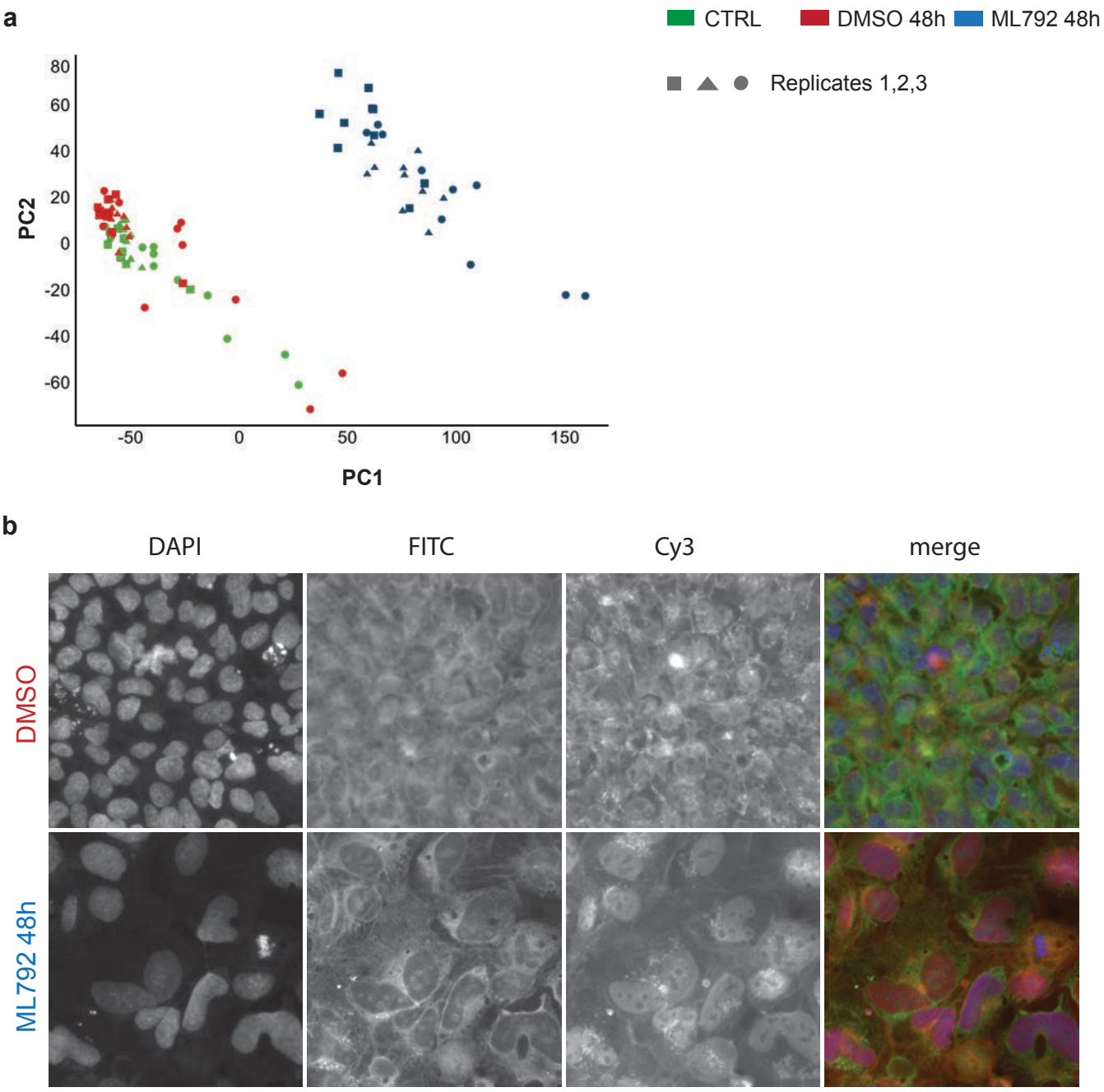
Supplementary information

SUMO maintains the chromatin environment of human induced pluripotent stem cells.

Barbara Mojsa[1], Michael H. Tatham[1], Lindsay Davidson[2], Magda Liczmanska[1], Jane E. Wright[1], Nicola Wiechens[1], Marek Gierlinski[3], Tom Owen-Hughes[1] and Ronald T. Hay[1]

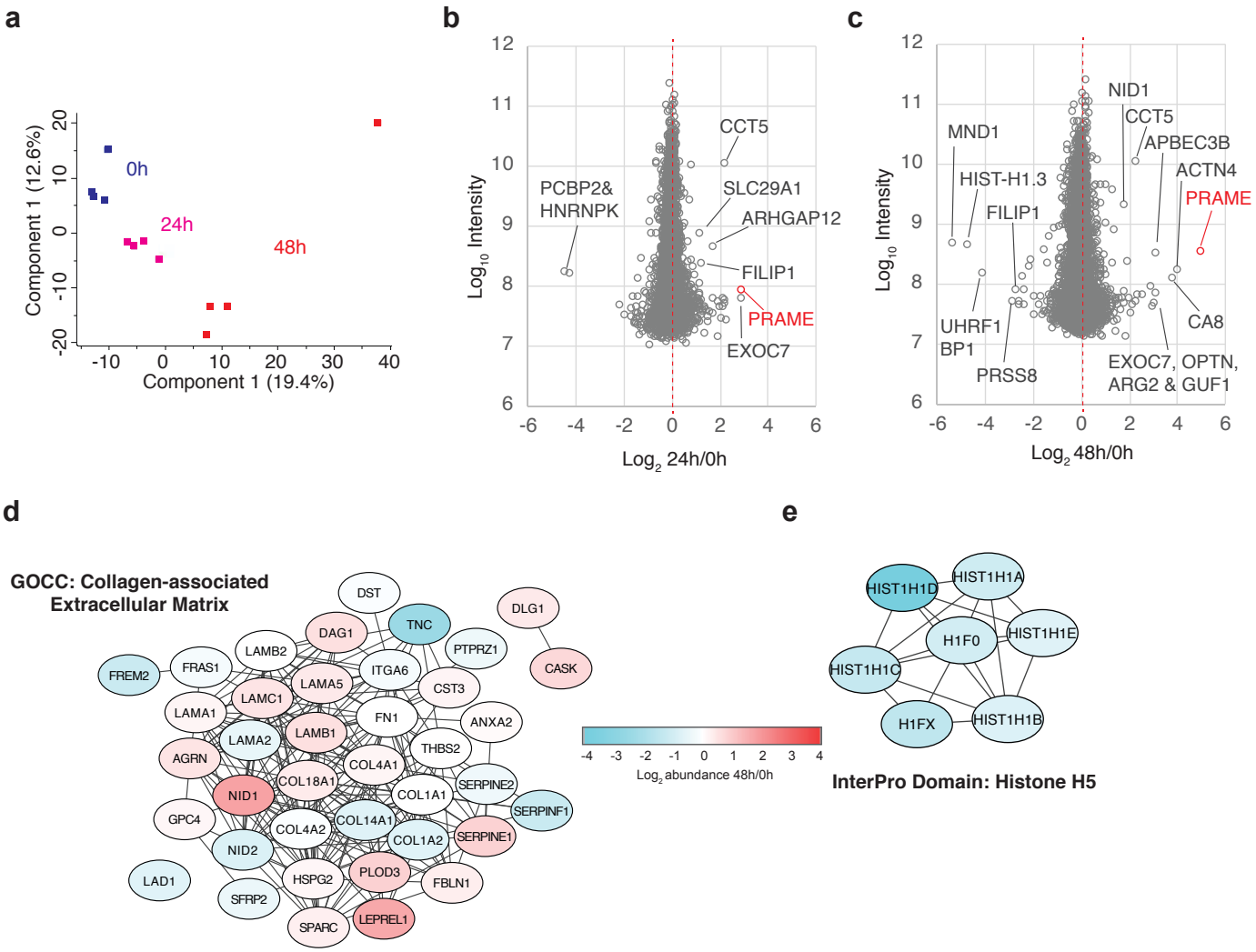
[1]Division of Gene Regulation and Expression, [2]Division of Cell and Developmental Biology, [3]Division of Computational Biology, School of Life Sciences, University of Dundee, Dundee, UK

Supplementary Figure 1. (S1)



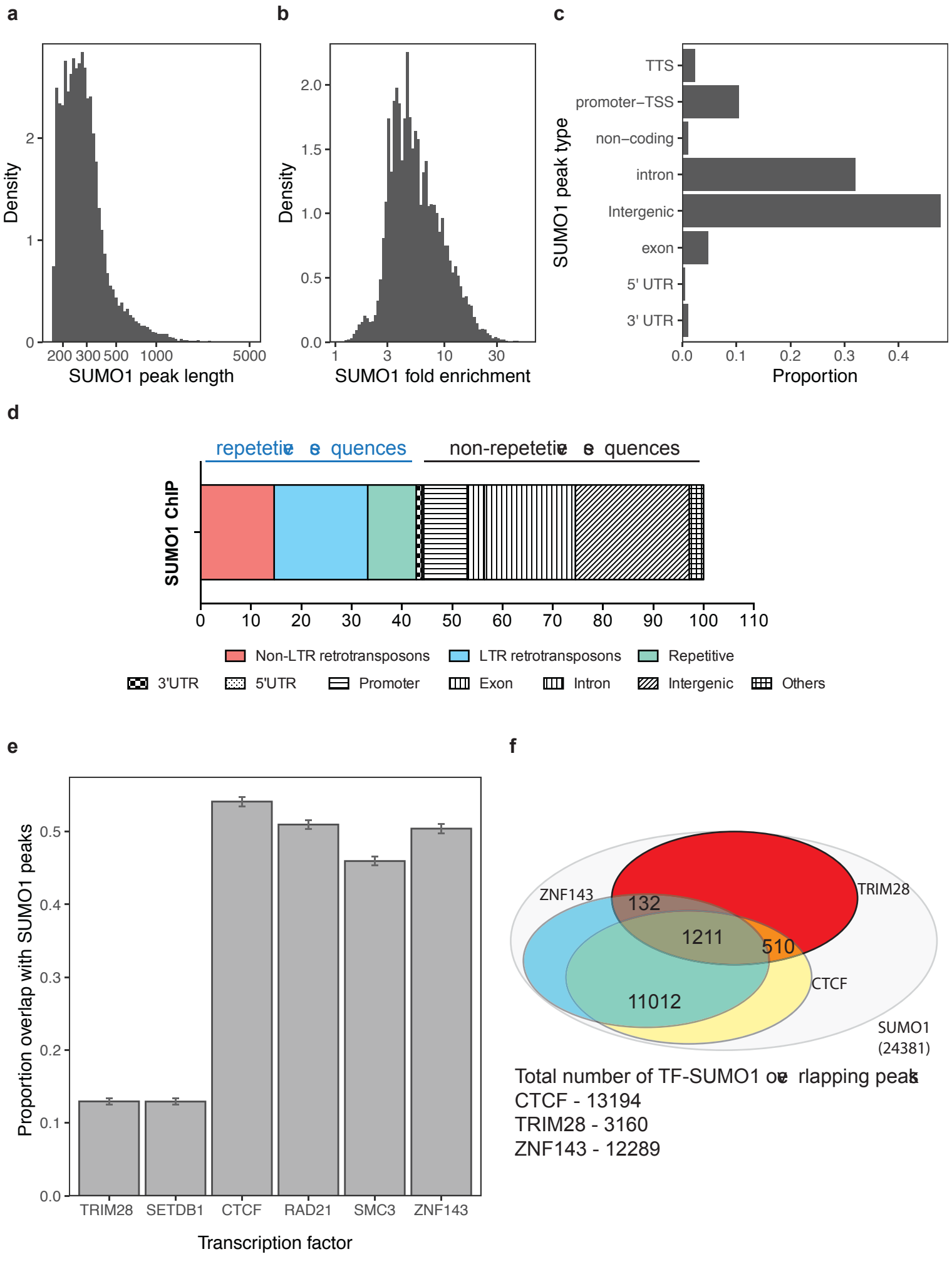
Supplementary Figure 1. Nuclear and nucleolar phenotypes related to ML792 treatment. **a.** Principal component analysis of three independent cell painting experiments (replicates 1,2,3) of ChiPS4 cells treated with 400nM ML792 (blue), DMSO (red) or untreated (green) for 48h. **b.** Representative sample images from the cell painting experiments showing the DAPI, FITC and Cy3 channels. Cells were stained with Hoechst 33342, to reveal nuclei, Concanavalin A, to reveal endoplasmic reticulum, SYTO14 to reveal nucleoli and cytoplasmic RNAs, Phalloidin to reveal F-actin, wheat-germ agglutinin to reveal Golgi and plasma membrane and MitoTracker to reveal mitochondria.

Supplementary Figure 2. (S2)



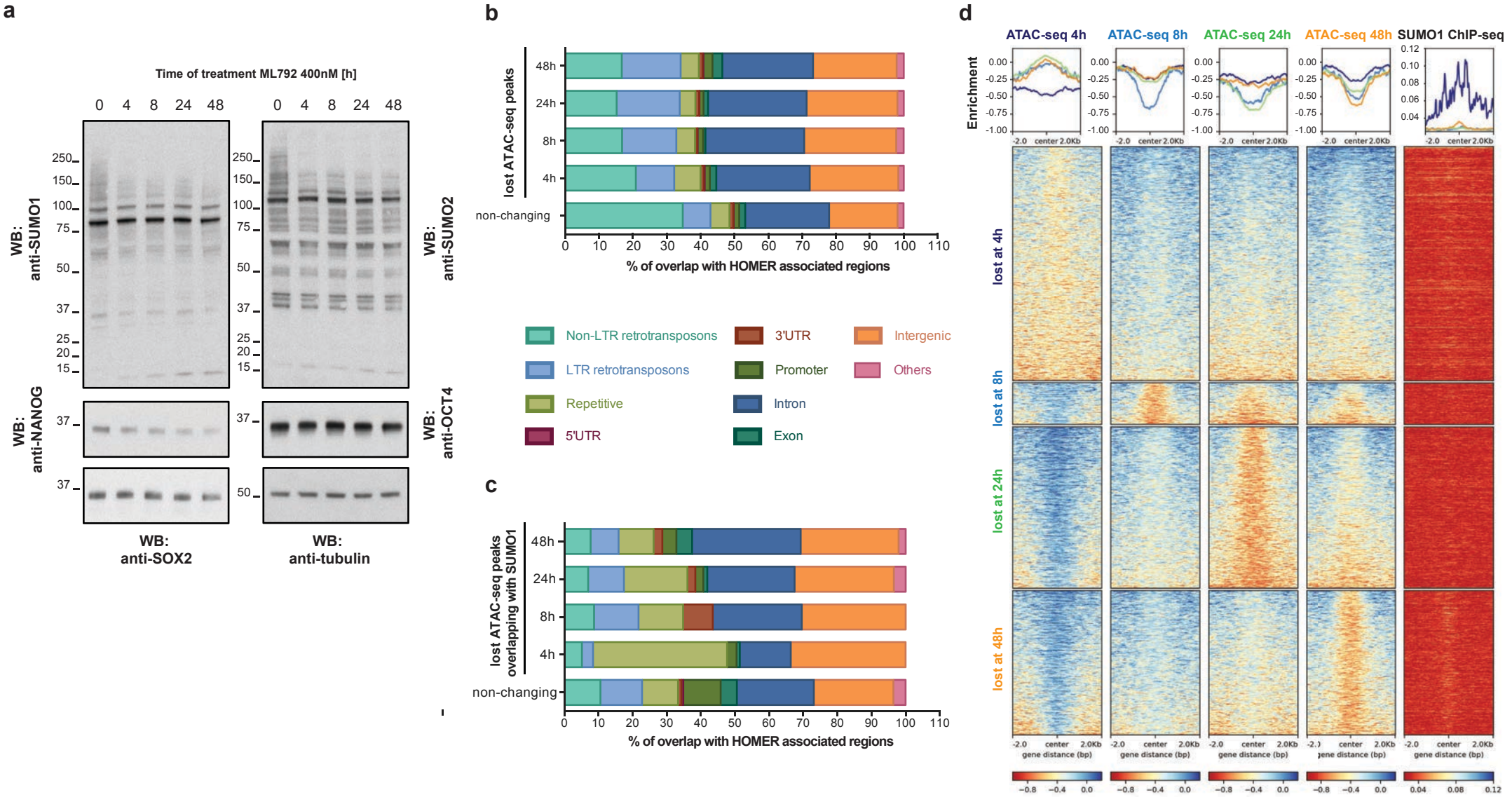
Supplementary Figure 2. Global deSUMOylation does not induce large changes in protein abundance in ChiPS4 cells over 48h. a. Principal Component Analysis of proteomic analysis using log_2 intensity values for 4741 proteins identified from crude cell extracts from ChiPS4 cells treated in quadruplicate with ML792 for the indicated times. **b.** Log_2 ratio and Log_{10} protein intensity data for 4741 proteins identified and quantified in crude extracts. Comparisons between untreated cells and 24h ML792. **c.** As in **b** but comparison between untreated cells and 48 h ML792. Selected outliers are indicated. **d.** Protein interaction network derived from the GOCC term 'Collagen-associated extracellular matrix', which was the only GOCC term significantly affected by 48 h ML792 treatment. All members are shown in the network which is colour-coded by log2 48h/0h ratio. No functional group clustering was identified by STRING for 24h/0h ratio data. **e.** Protein interaction network for the InterPro group 'Histone H5'. Log_2 48h/0h abundance ratio shown by colour as shown in the key.

Supplementary Figure 3. (S3)



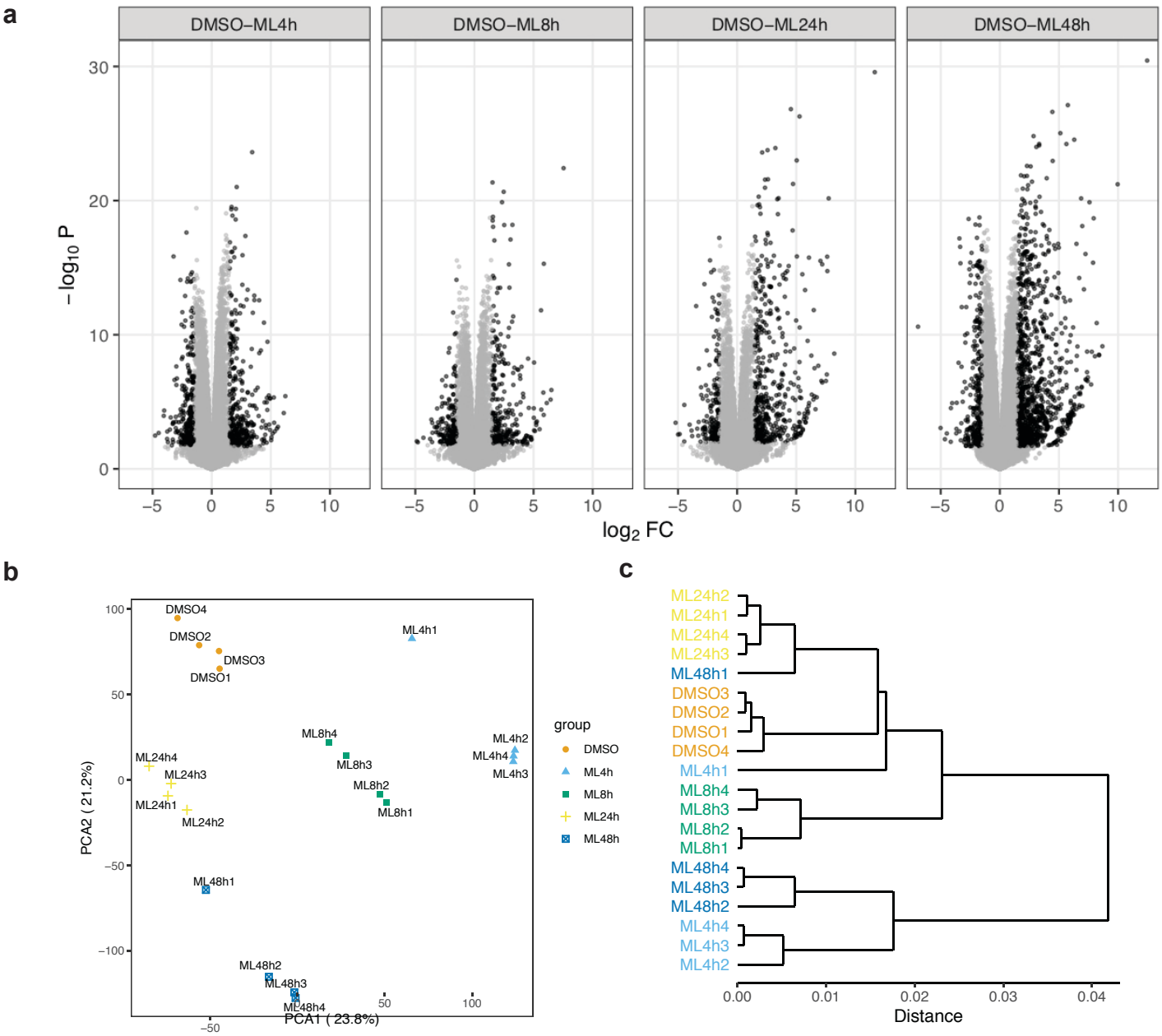
Supplementary Figure 3. SUMO1 ChIP-seq peaks in untreated ChiPS4 cells. a. Density plotted against SUMO1 peak lengths. **b.** Density plotted against fold enrichments calculated against the input samples. **c.** Proportion of SUMO1 ChIP peaks associated with various types of genomic locations were plotted based on HOMER annotations. **d.** Detailed analysis of repetitive and non-repetitive sequences using HOMER was used to calculate a percentage of overlap between those and SUMO1 ChIP-seq peaks. Different chromatin regions are represented by colours (repetitive) or grey patterns (non-repetitive) as indicated. **e.** Proportion of overlap of SUMO1 ChIP-seq peaks with various chromatin factor ChIP-seq peaks (TRIM28, SETDB1, CTCF, RAD21, SMC3, ZNF143 data for H1 hESCs were obtained from ENCODE http://genome.ucsc.edu/cgi-bin/hgTrackUi?db=hg19&g=wgEncodeRegTfbsClusteredV3). Error bars are 95% of confidence intervals of the proportion. **f.** Overlaps between ZNF143, CTCF and TRIM28 peaks overlapping with SUMO1 peaks were calculated using an intersect function in *bedtools* (at least 50% overlap) and plotted using Venn diagram. The exact number of peaks for each category are shown.

Supplementary Figure 4. (S4)



Supplementary Figure 4. ATAC-seq analysis of ChIPS4 cells treated with 400nM ML792. a. Western blot analysis of total protein samples from ChiPS4 cells treated with 400nM ML792 for the indicated times using antibodies against SUMO1, SUMO2, NANOG, SOX2, OCT4 and tubulin (loading control). **b.** Percentage of overlap between HOMER-based annotations of chromatin regions with lost ATAC-seq peaks at various time points and **c.** Percentage of overlap between HOMER-based annotations of chromatin regions with peaks common for SUMO1 CHIP-seq and lost ATAC-seq peaks at each time point. Different chromatin regions are represented by colours as indicated **d.** Density plots for lost ATAC-seq changes at each time point following ML792 treatment. The sites are ordered by the time at which a change of >1.5 fold is first detected. The same order of genomic locations has been plotted of SUMO ChIP-seq signal. Graphs at the top represent summary plots for the ATAC-seq peaks at changing sites for the indicated time points.

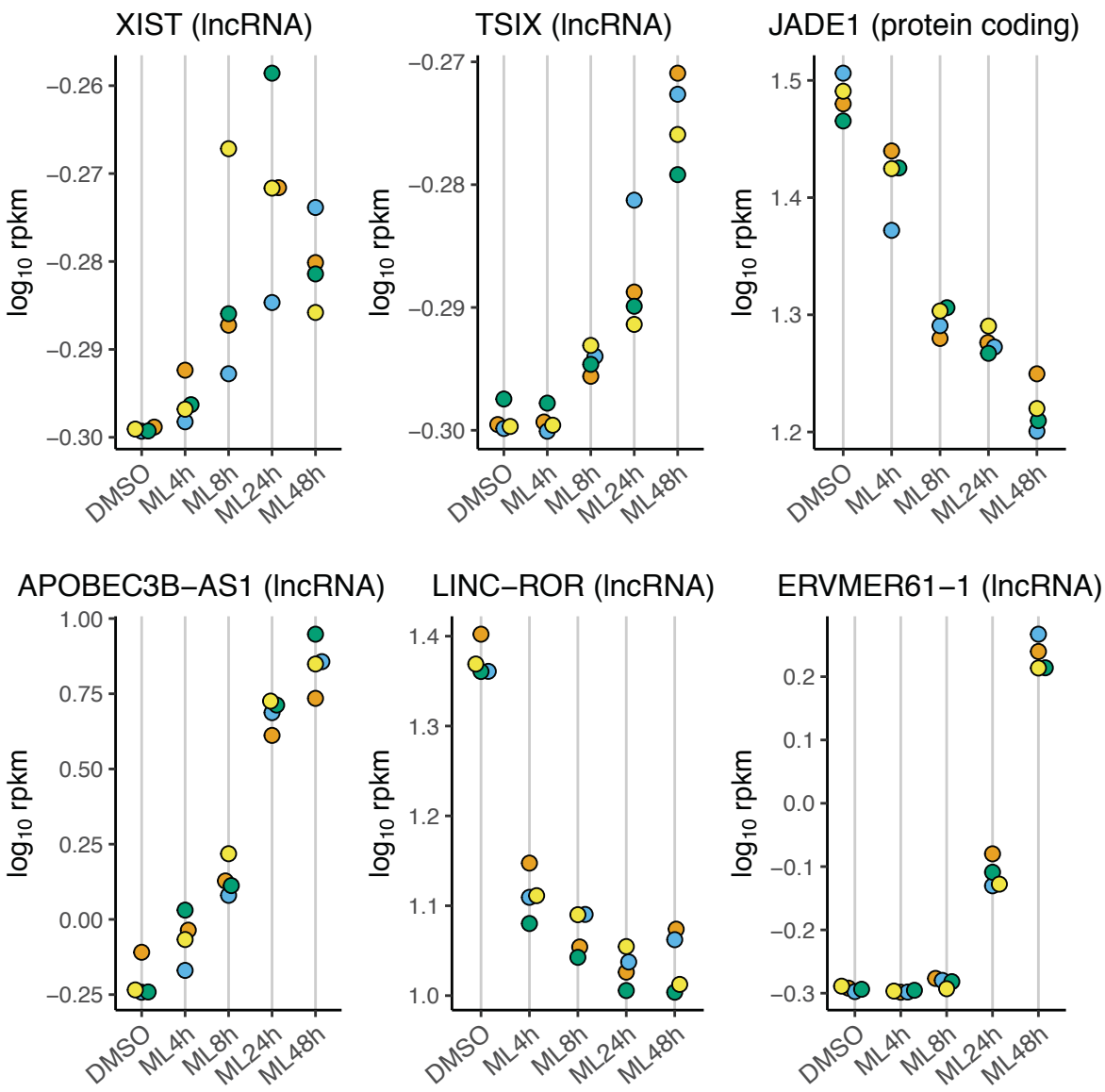
Supplementary Figure 5. (S5)



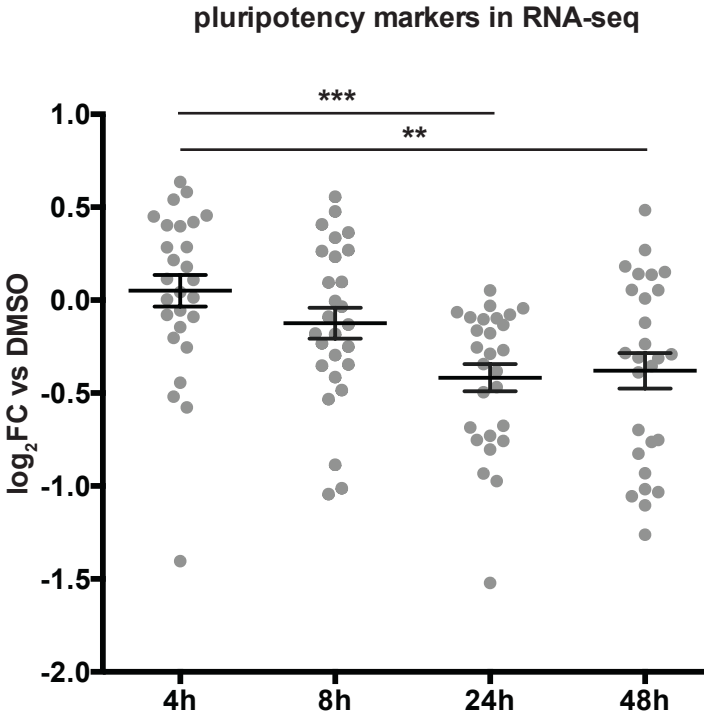
77 **Supplementary Figure 5. RNA-seq global data analysis. a.** Volcano plots showing differential
78 expression of RNA-seq data for each time point versus DMSO. Black dots indicate
79 "differentially expressed" genes, defined by FDR < 0.05 and $|\log_2 FC| > 1.5$. **b.** PCA
80 decomposition of RNA-seq RPKM data. **c.** Hierarchical clustering of RNA-seq RPKM data using
81 correlation distance.
82
83

Supplementary Figure 6. (S6)

a

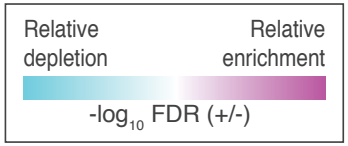


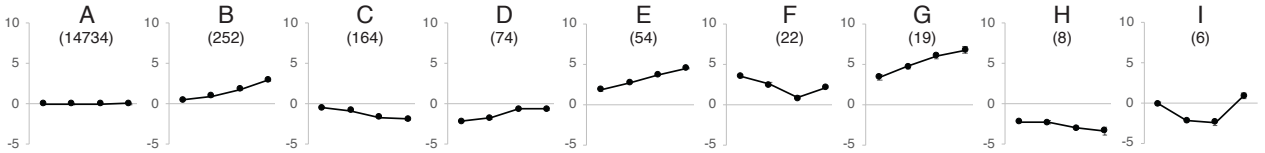
b



84 **Supplementary Figure 6. Expression of RNAs associated with pluripotency and ML792**
85 **response. a.** Expression profiles of selected genes, biotype indicated in brackets. Graphs
86 represent log_{10}rpkm at a given time point with all replicates shown in different colours. **b.**
87 Log_2 abundance ratio data for 27 markers of pluripotency comparing ML792 exposure to
88 DMSO treated cells. The plot shows individual data points and mean with standard error of
89 the mean for the entire set. The result of a one-way ANOVA adjusted for multiple comparisons
90 using Holm-Sidak's method is shown. ** P<0.01; *** P<0.001.
91
92

Supplementary Figure 7. (S7)

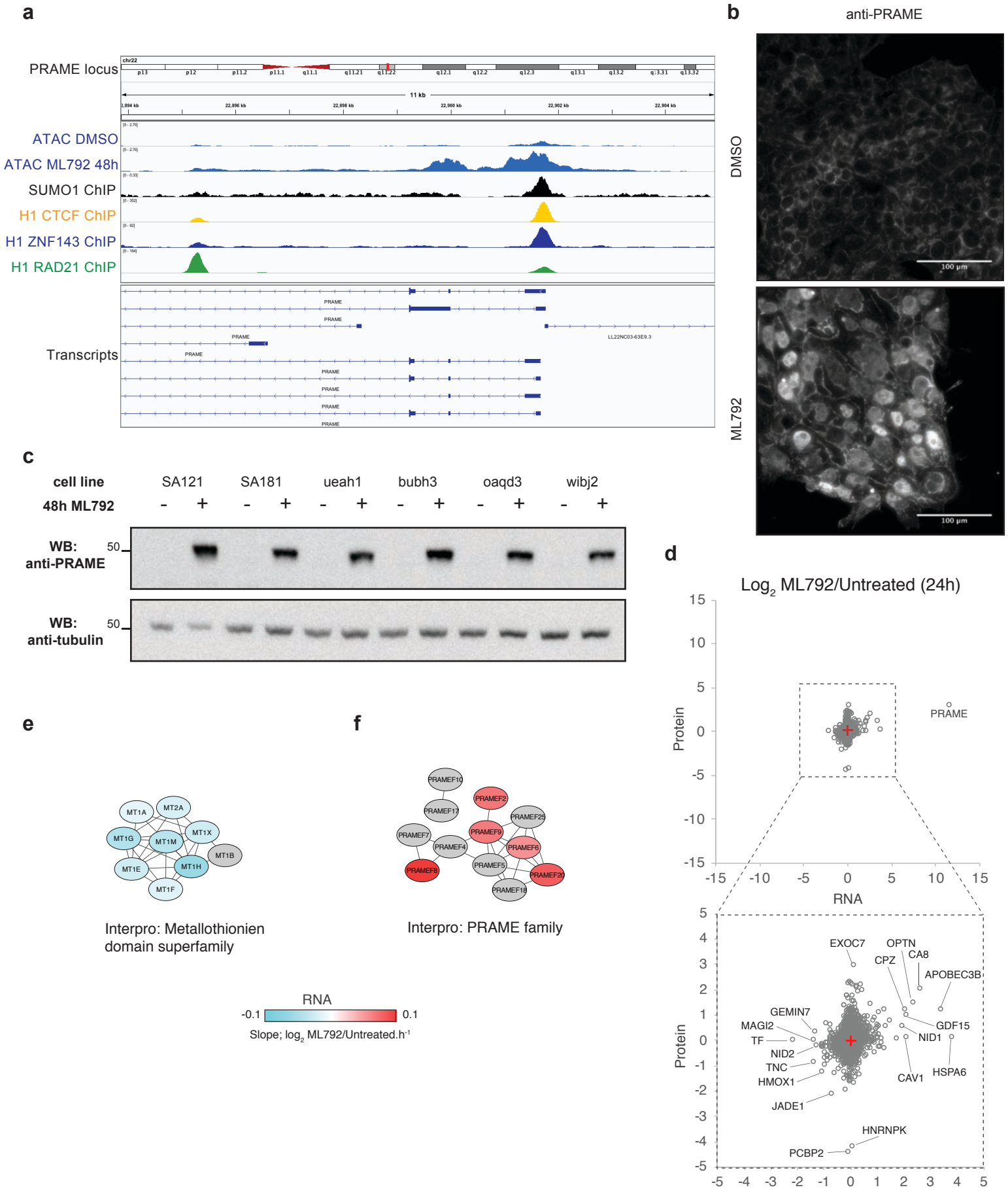




Functional group	-log FDR_A	-log FDR_B	-log FDR_C	-log FDR_D	-log FDR_E	-log FDR_F	-log FDR_G	-log FDR_H	-log FDR_I
Keywords:Phosphoprotein	27.45	-10.93	-3.23	-2.95	-1.54	-1.55	-0.62	-0.32	-0.12
GSEA:BENPORATH_ES_WITH_H3K27ME3	-20.78	5.91	8.34	0.88	0.00	0.00	0.00	0.88	0.00
GOCC name:intracellular part	18.71	-9.36	-0.93	-2.33	-0.76	-0.13	-0.18	0.00	-0.12
Keywords:Acetylation	18.24	-5.55	-4.12	-1.45	-1.09	-0.27	0.00	0.00	0.00
GSEA:BENPORATH_EED_TARGETS	-14.80	4.99	5.67	0.49	0.04	0.00	0.00	0.00	0.00
GSEA:BENPORATH_SUZ12_TARGETS	-13.57	3.59	4.48	0.47	0.39	0.00	0.00	0.95	0.00
Keywords:Signal	-11.17	3.70	6.05	0.00	1.03	0.12	0.00	0.00	0.00
GSEA:BENPORATH_PRC2_TARGETS	-13.00	4.15	2.72	1.04	0.14	0.00	0.00	0.12	0.00
GSEA:SCGGAAGY_V$ELK1_02	13.65	-4.74	-2.37	-0.10	-0.13	0.00	0.00	0.00	0.00
Keywords:Disulfidebond	-11.33	2.38	5.56	0.09	0.99	0.28	0.16	0.00	0.00
GOCC name:cytoplasmic part	12.22	-6.44	-0.56	-1.05	-0.47	0.00	0.00	0.00	0.00
Keywords:Glycoprotein	-9.75	2.72	5.09	0.00	0.69	0.00	0.05	0.00	0.44
GSEA:PILON_KLF1_TARGETS_DN	12.62	-2.64	-1.66	-1.42	-0.28	0.00	0.00	0.00	0.00
GSEA:DIAZ_CHRONIC_MEYLOGENOUS_LEUKEMIA_UP	11.37	-5.12	-0.81	-0.28	-0.31	0.00	0.00	0.00	0.00
GOCC name:organelle part	10.60	-4.58	-1.03	-0.74	0.00	0.00	-0.21	0.00	0.00
Keywords:Cellmembrane	-9.36	1.28	4.33	0.62	0.96	0.00	0.42	0.00	0.00
GSEA:MIKKELSEN_MEF_HCP_WITH_H3K27ME3	-10.12	2.71	3.77	0.00	0.00	0.10	0.00	0.11	0.00
GOCC name:intracellular organelle part	10.21	-4.58	-0.90	-0.65	0.00	0.00	-0.18	0.00	0.00
Keywords:Receptor	-9.38	2.34	2.04	1.57	1.00	0.00	0.00	0.00	0.00
KEGG name:Neuroactive ligand-receptor interaction	-9.08	0.06	6.29	0.00	0.05	0.00	0.00	0.00	0.00
Keywords:Referenceproteome	5.53	-9.46	0.00	-0.10	0.00	0.00	0.00	0.00	0.00
Keywords:Completeproteome	5.51	-9.40	0.00	-0.10	0.00	0.00	0.00	0.00	0.00
GSEA:MIKKELSEN_MCV6_HCP_WITH_H3K27ME3	-8.39	2.82	2.39	0.11	0.00	0.21	0.00	0.00	0.00
GSEA:GRAESSMANN_APOPTOSIS_BY_DOXORUBICIN_DN	8.88	-3.24	-0.63	-0.27	0.00	0.00	0.00	0.00	0.00
Keywords:Secreted	-6.17	1.98	2.47	-0.05	1.36	0.49	0.00	0.00	0.00
Keywords:Transducer	-7.71	2.34	0.99	0.57	0.45	0.00	0.18	0.00	0.00
GSEA:JOHNSTONE_PARVB_TARGETS_3_DN	8.22	-3.04	-0.77	-0.16	0.00	0.00	0.00	0.00	0.00
GSEA:PUJANA_BRCA1_PCC_NETWORK	8.00	-2.99	-0.37	-0.39	-0.09	0.00	0.00	0.00	0.00
Keywords:G-proteincoupledreceptor	-7.53	2.08	1.36	0.34	0.20	0.00	0.25	0.01	0.00
GSEA:REACTOME_GPCR_LIGAND_BINDING	-6.97	1.42	2.87	0.00	0.02	0.00	0.06	0.00	0.00
Keywords:Alternativesplicing	5.26	-4.83	-0.10	-0.57	0.01	-0.26	-0.19	-0.03	-0.06
GSEA:ZWANG_TRANSIENTLY_UP_BY_2ND_EGF_PULSE_ONLY	-7.35	1.34	2.40	0.07	0.00	0.00	0.09	0.00	0.00
GOCC name:organelle	6.57	-2.51	-0.06	-2.02	0.00	0.00	0.00	0.00	0.00
GOCC name:intracellular organelle	6.55	-2.58	-0.03	-1.97	0.00	0.00	0.00	0.00	0.00
GSEA:MARTENS_TRETINOIN_RESPONSE_UP	-6.66	0.51	1.46	0.28	2.20	0.00	0.00	0.00	0.00
GSEA:KEGG_NEUROACTIVE_LIGAND_RECEPTOR_INTERACTION	-6.05	0.33	3.95	0.00	0.46	0.00	0.00	0.00	0.00
GSEA:PUJANA_CHEK2_PCC_NETWORK	7.71	-2.43	-0.51	-0.08	0.00	0.00	0.00	0.00	0.00
GOBP name:cellular macromolecule metabolic process	6.95	-1.66	-1.49	-0.51	0.00	0.00	0.00	0.00	0.00
GOCC name:cytosol	6.59	-3.25	-0.47	0.00	-0.13	0.00	0.00	0.00	0.00
GOBP name:cellular metabolic process	5.77	-1.91	-0.48	-0.86	0.00	0.00	0.00	0.00	0.00
GOCC name:cell part	3.49	-5.48	0.00	0.00	0.00	0.00	0.00	0.00	0.00
GSEA:DODD_NASOPHARYNGEAL_CARCINOMA_DN	6.06	-1.68	-0.83	-0.16	-0.20	0.00	0.00	0.00	0.00
Keywords:Palmitate	-4.95	1.64	1.40	0.81	0.00	0.00	0.00	0.00	0.00
Keywords:Transmembranehelix	-4.74	0.41	2.28	0.33	0.08	0.08	0.78	0.00	0.00
GSEA:GRAESSMANN_RESPONSE_TO_MC_AND_DOXORUBICIN_DN	6.20	-1.46	-0.93	-0.06	0.00	0.00	0.00	0.00	0.00
Keywords:Transmembrane	-4.63	0.39	2.24	0.33	0.08	0.08	0.78	0.00	0.00
GOMF name:receptor activity	-6.06	1.02	1.18	0.00	0.21	0.00	0.00	0.00	0.00
GSEA:LASTOWSKA_NEUROBLASTOMA_COPY_NUMBER_DN	6.37	-1.51	-0.47	-0.07	0.00	0.00	0.00	0.00	0.00
Keywords:Cytoplasm	5.02	-2.40	-0.30	-0.32	-0.08	-0.02	-0.06	-0.02	0.00
GSEA:REACTOME_GPCR_DOWNSTREAM_SIGNALING	-5.21	0.20	2.63	0.15	0.00	0.00	0.00	0.00	0.00

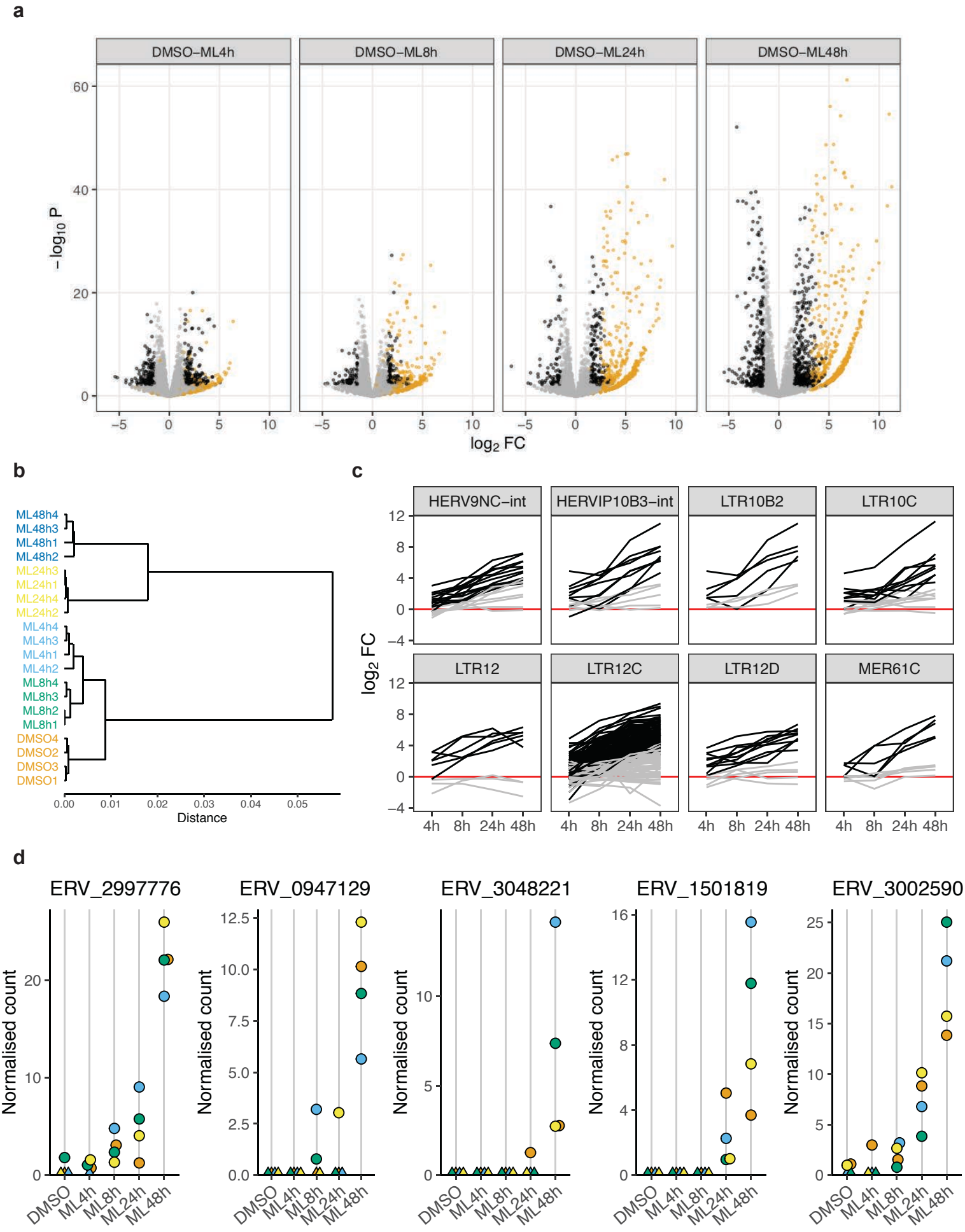
Supplementary Figure 7. Summary of functional group enrichment in different RNA-seq clusters. Summary of functional group enrichment in the RNAseq clusters sorted by the groups showing the most enrichment or depletion across all categories. This is the top 50 most enriched/depleted groups. Negative values are depleted and positive enriched. Colour coded by degree of depletion or enrichment within each group. This analysis only uses 'protein coding' genes selected from the entire RNA-seq experiment. Source data can be found in Supp. Data File 4.

Supplementary Figure 8: (S8)



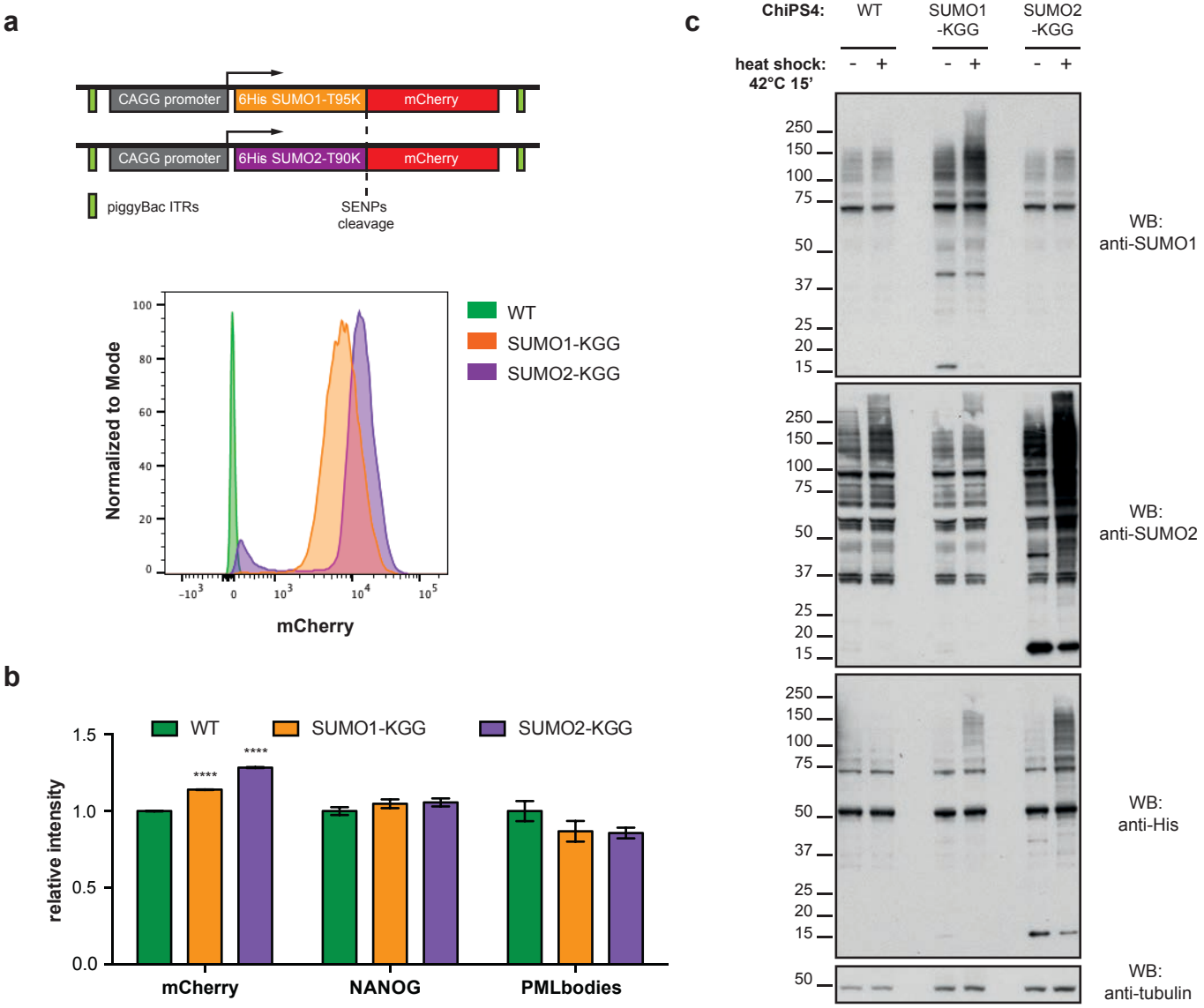
Supplementary Figure 8 (S8). PRAME is strongly induced by ML792 treatment in hiPSCs. a. Integrative Genomic Viewer was used to visualize changes in ATAC-seq (blue) at the PRAME locus in response to ML792 treatment. SUMO1 ChIP-seq signal in untreated ChiPS4 was also aligned and represented in black. ChIP-seq signals for CTCF (yellow), ZNF143 (dark blue) and RAD21 (green) from H1 hESC line were imported from ENCODE dataset and aligned with the same detailed genomic annotations. All traces of the same type were normalized and scaled in an identical way. **b.** ChiPS4 cells were treated for 48h with DMSO or 400 nM ML792, fixed and stained using DAPI and anti-PRAME antibody. IF images were obtained using a Leica DM-IRB microscope equipped with a Hamamatsu CCD camera and 20x 0.3C-Plan lens. All images contain 100 μm scale bar. **c.** Various human stem cell lines (hESC lines: SA121, SA181; hiPSC lines: ueah1, bubh3, oaqd3, wibj2) were treated for 48h with DMSO or 400 nM ML792 and analysed by Western blot using anti-PRAME and anti-tubulin (loading control) antibodies. **d.** Scatter-plot of RNA change versus protein-level change during 24 hours ML792 treatment (4498 common entries). Full scale is shown in the upper panel, and a smaller scale is shown below. Selected outliers are indicated. **e.-f.** STRING networks for the indicated functional groups with colouring based on slopes of the RNA-seq data only. Grey nodes were not measured. Slopes were calculated using all 4 time-point ratios plus 0h = 0 fold change.

Supplementary Figure 9. (S9)



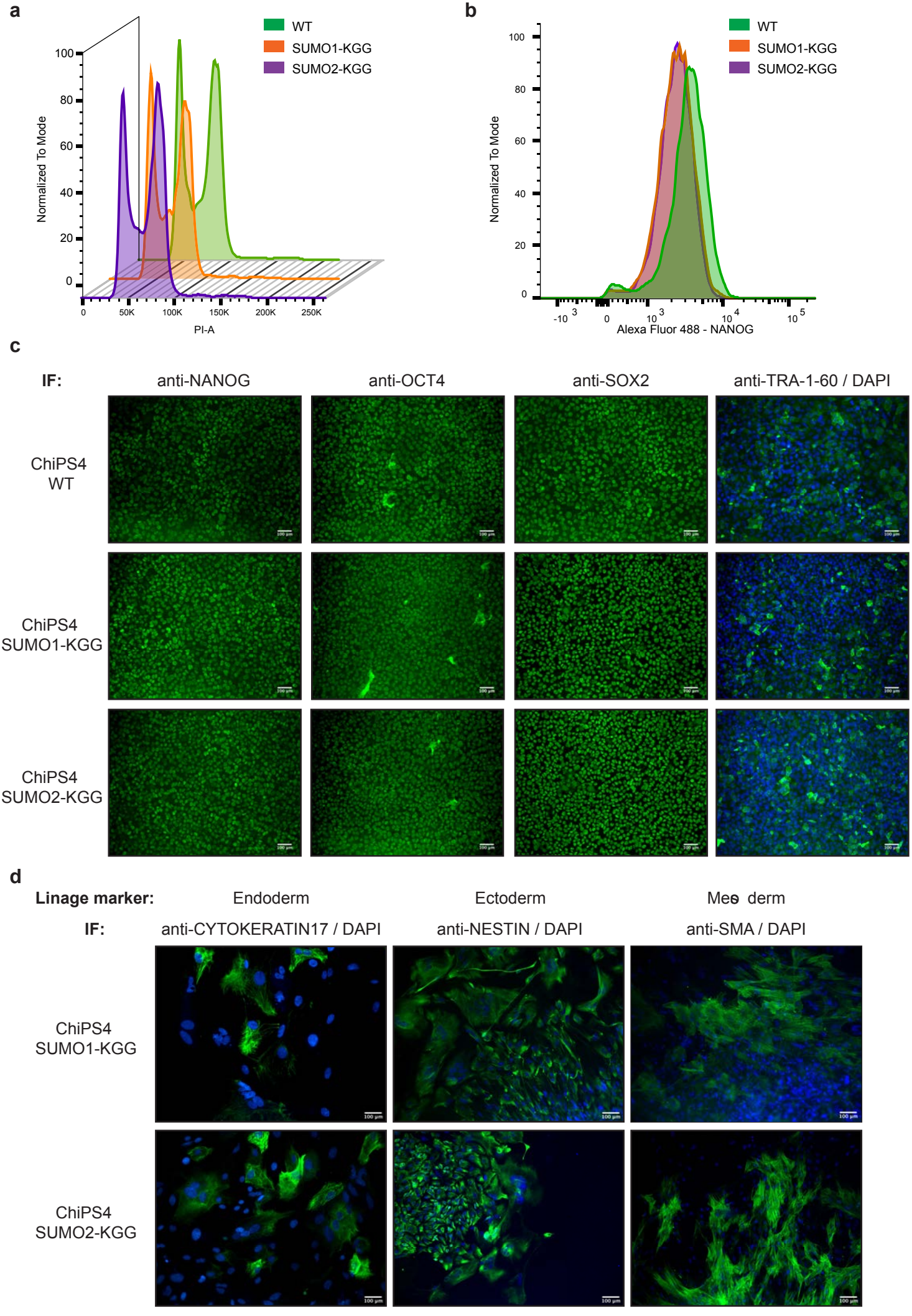
Supplementary Figure 9. RNA-seq analysis of HERVs expression. The GTF file was created with HERV loci and *STAR* was used to map RNA-seq reads to the genome without repeat masker filtering generating count reads per HERV. Of all mapped loci, 108,607 have non-zero count in at least one sample but in our analysis only HERVs (8,422) with at least 10 counts in at least one sample were considered. Differential expression of *STAR* count data was performed with *edgeR*. **a.** Volcano plots showing differential expression of HERV data for each time point versus DMSO. Black dots indicate "differentially expressed" genes, defined by FDR < 0.05 and $|\log_2 FC| > 1.5$. Orange dots indicate HERVs belonging to cluster 1 (see Fig. 6d). **b.** Hierarchical clustering of HERV data using correlation distance. **c.** Time profiles ($\log_2$ ratio between a given time point and DMSO normalised counts) of 8 selected HERV elements. Each panel shows expression from all detected loci containing this element. Black lines indicate HERVs belonging to cluster 1 (see **Fig. 6d**). **d.** Expression from five selected HERV loci. Colours indicate replicates.

Supplementary Figure 10. (S10)



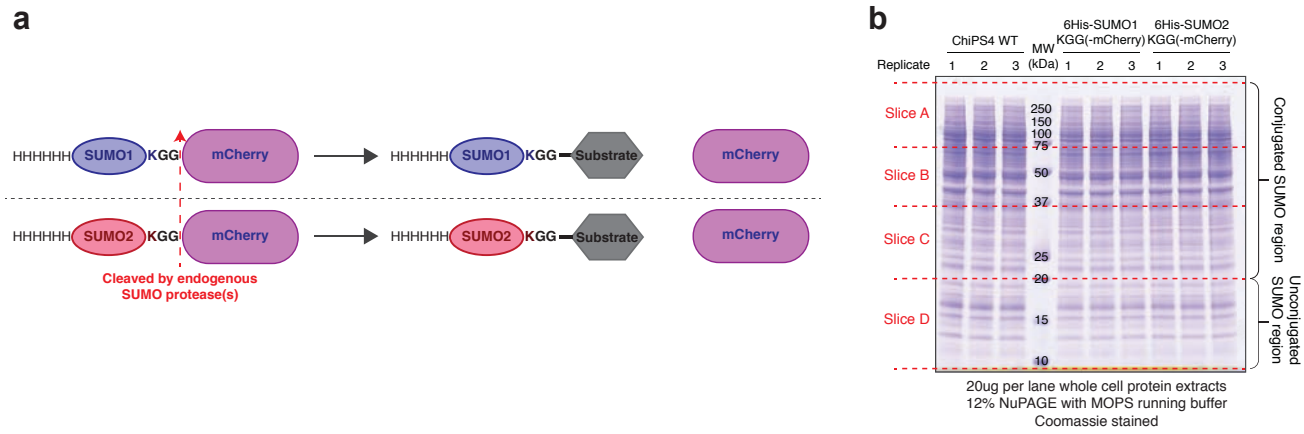
Supplementary Figure 10. Generation of 6xHis-SUMO1/2-mCherry ChiPS4 cell lines. a. Design of the piggyBac constructs used for generation of ChiPS4 cell lines expressing SUMO1 or SUMO2. Single cell clones were selected based on the expression levels of mCherry using flow cytometry with a view of having a similar level of His-tagged SUMO1 or SUMO2 for SUMO site proteomics experiments. **b.** Selected single cell clonal lines were further validated using High Content Screening microscopy. Cells were fixed and stained using DAPI, Cy5 Cell Mask as well as anti-NANOG and anti-PML antibodies and further assessed for mCherry expression. The result of a one-way ANOVA adjusted for multiple comparisons using Holm-Sidak's method is shown. ****P < 0.001 **c.** ChiPS4 WT, SUMO1-KGG and SUMO2-KGG expressing cell lines were exposed to heat shock for 15 minutes at 42°C and total protein lysates were analysed by Western blot using anti-SUMO1, anti-SUMO2/3, anti-His and anti-tubulin (loading control) antibodies.

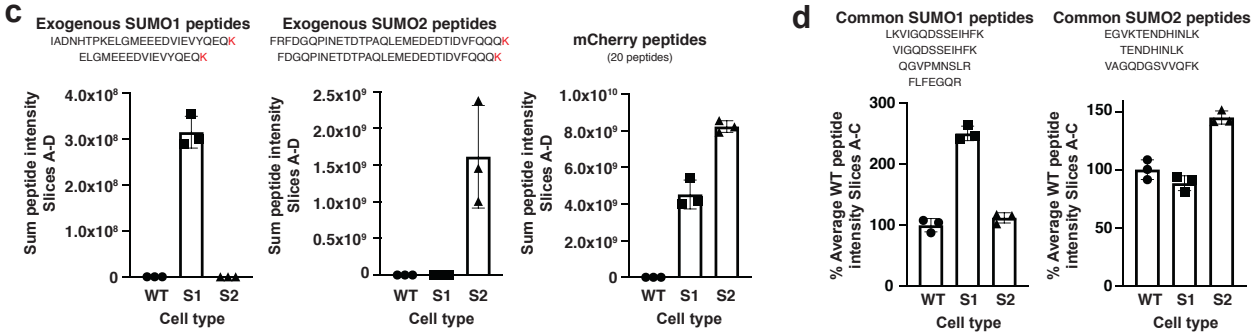
Supplementary Figure 11. (S11)

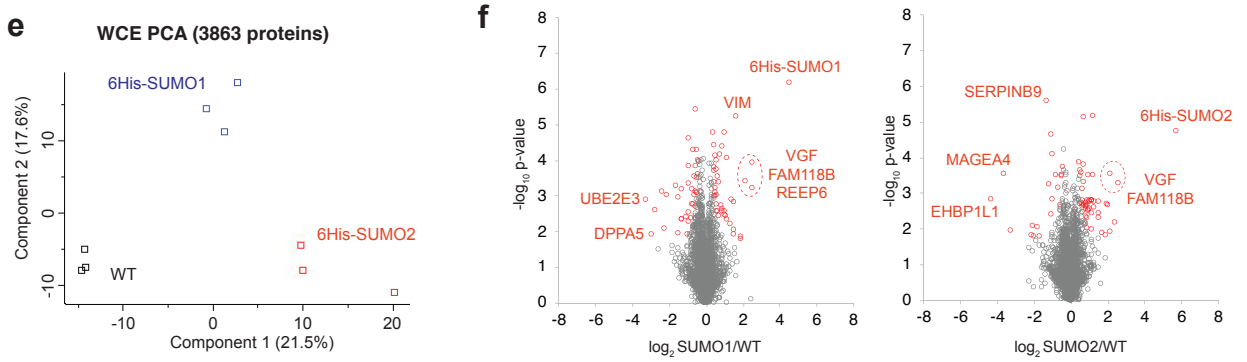


Supplementary Figure 11. Validation of pluripotency status of 6xHis-SUMO1/2-mCherry ChiPS4 cell lines. a. Flow cytometry analysis of cell cycle and **b.** NANOG expression. **c.** IF analysis of the expression of pluripotency associated markers (NANOG, SOX2, OCT4, TRA-1-60) in ChiPS4 WT, SUMO1-KGG and SUMO2-KGG cell lines. **d.** *In vitro* differentiation potential of ChiPS4 SUMO1-KGG and SUMO2-KGG cell lines was assessed by IF staining with DAPI and specific antibodies against CYTOKERATIN 17 (Endoderm), NESTIN (Ectoderm) and SMA (Mesoderm). IF images were obtained using a Leica DM-IRB microscope equipped with a Hamamatsu CCD camera and 20x 0.3C-Plan lens. All images contain 100 μm scale bar.

Supplementary Figure 12. (S12)

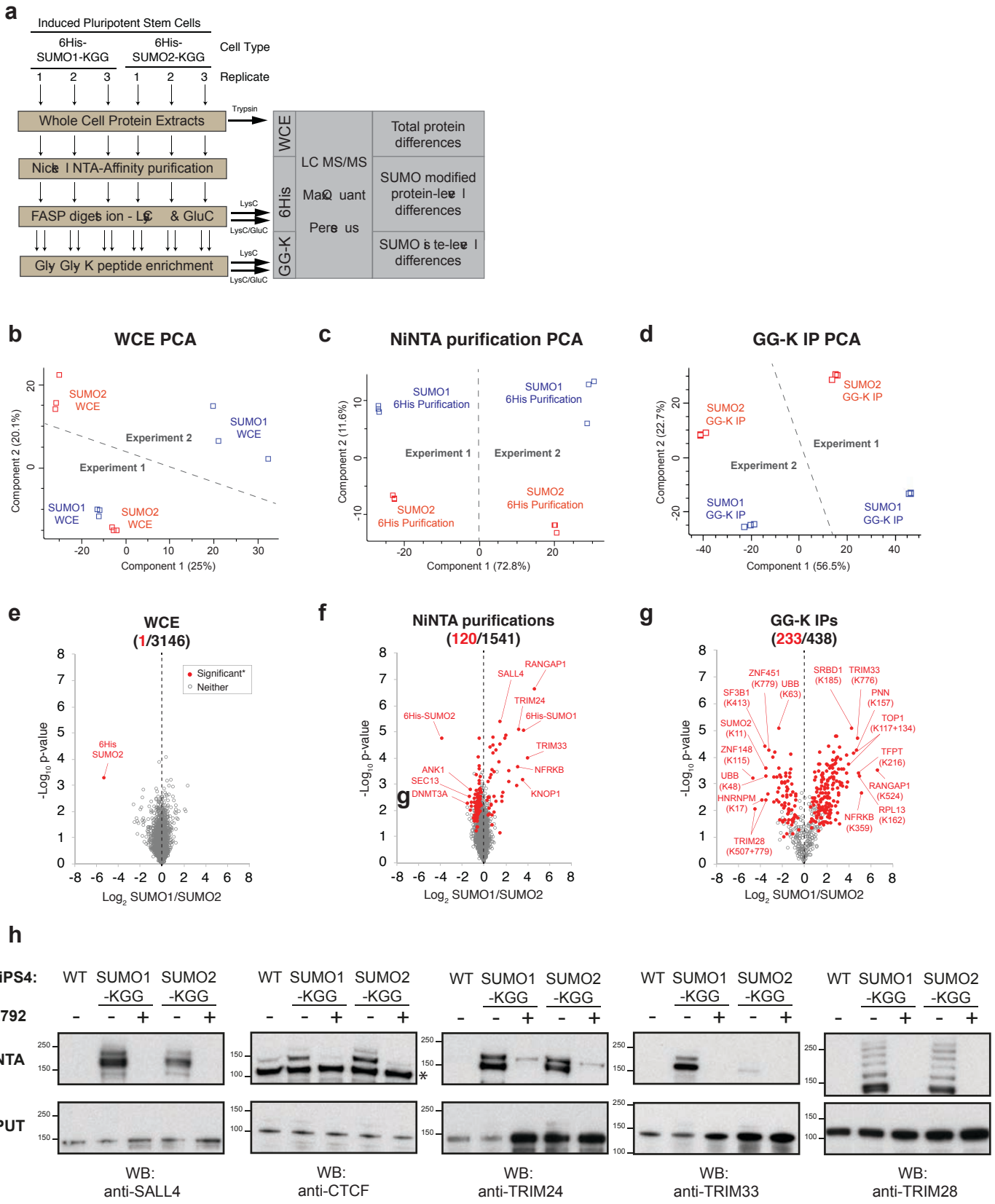






Supplementary Figure 12. Mass spectrometry-based validation of 6xHis-SUMO1/2-mCherry ChiPS4 cell lines. a. Schematic overview of the 6His-SUMO-KGG-mCherry overexpression constructs stably expressed in ChiPS4 cells. The C-terminal mCherry protein used for cell selection is cleaved from 6His-SUMO by endogenous SUMO proteases. **b.** Coomassie stained SDS-PAGE gel fractionating whole cell protein extracts from parental ChiPS4 cells (WT) and the selected 6His-SUMO-KGG-mCherry clones. Samples were prepared in triplicate. Each lane was excised into 4 sections allowing differentiation between conjugated (slices A-C) and unconjugated (Slice D) SUMO forms. Tryptic peptides from each slice were analysed by LC-MS/MS and data processed by MaxQuant. **c.** Two peptides each from 6His-SUMO1-KGG (left) and 6His-SUMO2-KGG (centre) are specific to the exogenous construct and not the endogenous proteins. The sum of the MaxQaunt LFQ peptide intensity is shown for each replicate in each cell type. Data for 20 mCherry peptides is also shown (right). **d.** Four peptides from SUMO1 (left) and three from SUMO2 (right) are common to both the endogenous and exogenous forms of the proteins. These intensities can be used to assess over-expression levels of the 6His-SUMO-KGG constructs relative to their endogenous counterparts and are presented relative to parental (WT) cell intensity. Data from slice D was omitted to allow comparisons in context of the conjugated forms of the proteins. **e.** Quantitative data from 3863 proteins identified from the gel shown in **b.** were compared by principal component analysis. **f.** Numerical ratio and unpaired student's t-test results comparing WT parental cells with 6His-SUMO1-KGG-mCherry cells (left), and WT with 6His-SUMO2-KGG-mCherry cells (right). Outliers (red markers) were defined in Perseus by 5% FDR with an S0 value of 0.1 (79 outliers from WT vs SUMO1 cells and 73 from WT vs SUMO2 cells - 22 common). Gene names from extreme outliers are indicated. *MaxQuant assigned all mCherry peptides to the 6His-SUMO1-KGG-mCherry protein group, so mCherry peptides derived from 6His-SUMO2-KGG-mCherry falsely shows enrichment of the SUMO1 construct in SUMO2 cells.

Supplementary Figure 13. (S13)



Supplementary Figure 13. Design and quality control of SUMO site proteomics experiments. **a.** Overview of a proteomics experiment to identify IPS-specific SUMO1 and SUMO2 substrates. Two experimental runs were performed with two different hiPSC lines (expressing 6His-SUMO1-KGG or 6His-SUMO2-KGG), each one was performed in triplicate. Three protein fractions were analysed; whole cell extracts (WCE), NiNTA column elutions (6HIS), GlyGly-K immunoprecipitated peptide elutions (GG-K IP). All peptides were analysed by LC-MS/MS and data processed by MaxQuant. **b.-d.** Principal component analysis of MS data from the three different cell fractions. **e.-g.** Scatter plots of $\log_2$ SUMO1/SUMO2 ratio and $-\log_{10}$ t-test p-value for proteins or peptides detected in different cellular fractions as indicated in **a.**: **e.** WCE - Whole cell extract (measuring total protein abundance difference), **f.** NiNTA elutions (difference in proteins abundance in 6His-SUMO purifications), **g.** GGK-IP (site-level SUMO preference). *Red markers were found to be significantly different in both experimental runs. Selected outliers are indicated. Numbers of significantly differing proteins or peptides compared to the entire set of proteins or peptides quantified are shown. **h.** NiNTA purification of His-SUMO modified proteins was performed using WT, 6His-SUMO1-KGG or 6His-SUMO2-KGG ChiPS4 cell lines that were treated with DMSO vehicle or 400 nM ML792 for 48h. Input and NiNTA elutions were analysed by Western blot using specific antibodies directed against following protein targets: SALL4, TRIM24, TRIM28, TRIM33, and CTCF. Bands for unmodified proteins that are non-specifically pulled down on NiNTA resin are labelled with an *.