Cohesin-independent STAG proteins interact with RNA and localise to R-loops to promote complex loading.

Hayley Porter[1+], Yang Li[1+], Maria Victoria Neguembor[2], Manuel Beltran[3], Wazeer Varsally[1], Laura Martin[2], Manuel Tavares Cornejo[3], Dubravka Pezic[1], Amandeep Bhamra[4], Silvia Surinova[4], Richard G. Jenner[3], Maria Pia Cosma[2,5,6], Suzana Hadjur[1]*

1 Research Department of Cancer Biology, Cancer Institute, University College London, 72 Huntley Street, London, United Kingdom

2 Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, 08003 Barcelona, Spain.

3 Regulatory Genomics Group, Cancer Institute, University College London, London WC1E 6BT, United Kingdom

4 Proteomics Research Translational Technology Platform, Cancer Institute, University College London, 72 Huntley Street, London, United Kingdom

5 Universitat Pompeu Fabra (UPF), Dr Aiguader 88, 08003 Barcelona, Spain

6 Institució Catalana de Recerca i Estudis Avançats (ICREA), Pg. Lluís Companys 23, 08010 Barcelona, Spain.

+ These authors contributed equally

* Correspondence: s.hadjur@ucl.ac.uk

ABSTRACT

Most studies of cohesin function consider the Stromalin Antigen (STAG/SA) proteins as core complex members given their ubiquitous interaction with the cohesin ring. Here, we provide functional data to support the notion that the SA subunit is not a mere passenger in this structure, but instead plays a key role in the localization of cohesin to diverse biological processes and promotes loading of the complex at these sites. We show that in cells acutely depleted for RAD21, SA proteins remain bound to chromatin, cluster in 3D and interact with CTCF, as well as with a wide range of RNA binding proteins involved in multiple RNA processing mechanisms. Accordingly, SA proteins interact with RNA and are localised to R-loops where they contribute to R-loop regulation. Our results place SA1 within R-loop domains upstream of the cohesin complex and reveal a role for SA1 in cohesin loading which is independent of NIPBL, the canonical cohesin loader. We propose that SA1 takes advantage of structural R-loop platforms to link cohesin loading and chromatin structure with diverse functions. Since SA proteins are pan-cancer targets, and R-loops play an increasingly prevalent role in cancer biology, our results have important implications for the mechanistic understanding of SA proteins in cancer and disease.

KEY WORDS

Cohesin, STAG proteins, R-loops, genome organization.

INTRODUCTION

Cohesin complexes are master regulators of chromosome structure in interphase and mitosis. Accordingly, mutations of cohesin subunits lead to changes in cellular identity, both during development and in cancer [1-3]. A prevailing model is that cohesin contributes to cell identity changes in large part by dynamically regulating 3D genome organization and mediating communication between distal regulatory elements [4-10]. Molecular insight into how and when cohesin subunits become associated with chromatin and contribute to this function *in vivo* in human cells is still lacking under this model.

Most studies of cohesin function consider the Stromalin Antigen (STAG/SA) proteins as core complex members given their ubiquitous interaction with the tripartite cohesin ring (composed of SMC1, SMC3 and SCC1/RAD21). Rarely is the SA subunit considered for its roles independent of the cohesin ring, even though it is the subunit most commonly mutated across a wide spectrum of cancers [1,11,12].

SA proteins contribute to cohesin's association with DNA [13,14]. The yeast SA orthologue is critical for efficient association of cohesin with DNA and its ATPase activation [13,14]. Separating interactions into SA-loader and cohesin ring-loader sub-complexes still impairs cohesin loading, indicating that SA functions as more than just a bridge protein [14]. Crystallisation studies reveal a striking similarity of SA with NIPBL (the canonical cohesin loader [15]), in that both are highly bent, HEAT-repeat containing proteins [16,17]. Of note, NIPBL and SA interact together and wrap around both the cohesin ring and DNA to position and entrap DNA [18-20], implying a potential role for SA in the initial recruitment of cohesin to DNA alongside NIPBL. Further, SA proteins bridge the interaction between cohesin and CTCF [18,21,22], and also bridge interactions with specific nucleic acid structures *in vitro* [23,24].

Mammalian cells express multiple SA paralogs. SA1 binds to AT-rich telomeric sequences [23,24] and SA2 displays sequence-independent affinity for particular DNA structures commonly found at sites of repair, recombination, and replication [25]. Consistent with this, results in yeast implicate non-canonical DNA structures in cohesin loading in S-phase. *In vitro* experiments show that cohesin captures the second strand of DNA via a single-strand intermediate [26], and chromatid cohesion is impaired by de-stabilisation of single-strand DNA intermediates during replication [27].

Together, these suggest that SA proteins and DNA structures may play a regulatory role in guiding or stabilising cohesin localisations.

During transcription, the elongating nascent RNA can hybridise to the template strand of the DNA and form an R-loop, which is an intermediate RNA:DNA hybrid conformation with a displaced single strand of DNA [28]. A multitude of processes have been linked to R-loop stability and metabolism. For example, co-transcriptional RNA processing, splicing, and messenger ribonucleoprotein (mRNP) assembly counteract R-loop formation [29,30]. R-loop structures have also been shown to regulate transcription of mRNA by recruitment of transcription factors, displacement of nucleosomes, and preservation of open chromatin [31,32]. Hence, like at the replication fork, sites of active transcription accumulate non-canonical nucleic acid structures.

We set out to investigate the nature of SA proteins and cohesin loading to DNA. We discovered independent functions of the SA proteins, providing critical insight into the importance they play in their own right to direct cohesin's localization and loading to chromatin. In cells acutely depleted of RAD21, SA proteins remain associated with chromatin and CTCF where they are enriched at chromatin sites clustered in 3D. Moreover, we identify numerous, diverse cohesin-independent SA1 interactors involved in RNA processing, ribosome biogenesis, and translation. Consistent with this, SA1 and SA2 interact with RNA and non-canonical nucleic acid structures in the form of R-loops where SA1 suppresses R-loop formation. Importantly, SA proteins are required for loading of cohesin to chromatin in cells deficient for NIPBL. Our results highlight a central role for SA proteins in cohesin biology and the cohesin-independent interaction of SA proteins with RNA processing factors opens up a new understanding of how SA dysregulation can impact disease development that moves us beyond the control of chromatin topology for gene expression regulation.

RESULTS

SA interacts with CTCF on chromatin in the absence of the cohesin trimer.

To determine how CTCF and cohesin assemble on chromatin, we used previously described [33] human HCT116 cells engineered to carry a miniAID tag (mAID) fused to monomeric Clover (mClover) at the endogenous RAD21 locus and *OsTIR1* under the control of CMV (herein RAD21^{mAC}). RAD21^{mAC} cells were cultured in control EtOH conditions (EtOH) or in the presence of auxin (IAA) to induce rapid RAD21 degradation (Fig. S1a). Immunofluorescence (IF) was used to monitor the levels of mClover, SA1, SA2 and CTCF (Fig.1a, b, S1b). As reported [33], acute IAA treatment dramatically reduced mClover levels compared to control cells (mean fluorescence intensity (MFI) reduction of 82%, p=2.9E-239). SA proteins and CTCF were also all significantly reduced, however the extent of the change was notably different. We observed a small but significant reduction in CTCF signal upon IAA treatment (mean reduction of 16%, p=1.6E-27). This was similar to the mean SA1 signal which was reduced by 22% compared to EtOH control (p=3.5E-21). However, SA2 levels mirrored more closely the effect on mClover, being reduced by 63% (p=1.9E-186), but not completely lost (Fig.1b). The retention of SA proteins despite the degradation of RAD21 was surprising given the fact that they are considered to be part of a stable biochemical complex.

We sought to validate these observations using an orthogonal technique and to establish whether the residual SA proteins retained the capacity to directly interact with CTCF. We prepared chromatin extracts from RAD21^{mAC} cells treated with EtOH or IAA and performed chromatin co-immunoprecipitation (coIP) to probe the interactions between SA proteins, RAD21 and CTCF. Both SA1 and SA2 interacted with RAD21 and CTCF in control cells as expected [34,35], with notable differences in their preferred interactions (Fig.1c). SA2 more strongly enriched RAD21 while the SA1-CTCF interaction was significantly stronger than SA2-CTCF (Fig.1c), in line with previous results [36]. Upon RAD21 degradation, we again observed a stronger effect on chromatin-bound SA2 levels compared to SA1, suggesting that SA2 is more sensitive to cohesin loss than SA1. Residual SA proteins retained their ability to interact with CTCF in the absence of RAD21 and additionally, the interaction between SA1 and

CTCF was further enhanced (Fig.1c). Reciprocal coIPs with CTCF confirmed the CTCF-SA interactions in RAD21-depleted cells, including the differences between SA1 and SA2 (Fig.1d). We validated these results in a second cell line and using siRNA-mediated knockdown (KD) of SMC3 (Fig. S1c). We also confirmed that approximately 20% of SA1 and SA2 were bound to chromatin without RAD21 even in unperturbed RAD21^{mAC} HCT116 cells (Fig S1d), reminiscent of reports of SA1 without cohesin at telomeres [23].

Next, we performed two-color Stochastic Optical Reconstruction Microscopy (STORM) to assess the nuclear distribution and co-localization of SA1, SA2 and CTCF with nanometric resolution in control and RAD21-degraded cells (Fig. 1e, S1e). In control RAD21^{mAC} cells, we observed clustering of CTCF, SA1 and SA2 localizations as quantified by cluster analysis and nearest neighbor distance (NND) analysis of protein clusters [37]. SA1 and CTCF exhibited higher densities compared to SA2, with shorter distances between clusters (mean NND of 68.9, 65.3 and 78.9nm for CTCF, SA1 and SA2, respectively) (Fig 1f, g). Furthermore, we analyzed the relative distribution of SA clusters to CTCF clusters by assessing the NND distribution between SA1 and CTCF, and SA2 and CTCF. Both SA1 and SA2 exhibited significant co-localization with CTCF at short distances in RAD21^{mAC} cells (Fig. 1h, S1f, g).

Upon IAA treatment, we observed a decreased density of detected SA1, SA2 and CTCF in two analyzed clones (Fig.1e, f, S1e), suggesting that RAD21 degradation affects the stability of SA proteins and CTCF. As observed by conventional confocal microscopy (Fig. 1b), SA2 localizations were more affected than SA1 (mean density reduction in SA1, 32% and SA2, 42%). Accordingly, SA1, SA2 and CTCF clusters were more sparsely distributed across the nucleus upon RAD21 degradation (mean NND 81.1, 109.7 and 117.5 nm, respectively) with SA2 significantly more affected than SA1 (mean NND increase 24.3%, 39% respectively) (Fig.1g). SA proteins and CTCF remained co-localised upon RAD21 degradation as compared to both the control cells and to a simulation of randomly-distributed protein clusters at the same density (Fig. 1h, S1f,g). Interestingly, while the probability of SA1 at CTCF is only modestly affected, SA2 at CTCF is more affected in IAA treated cells (Fig. 1h), in line with our previous observations. Together, our results confirm the maintained interaction and

spatial co-localization patterns of SA proteins with CTCF and reveal a difference in SA paralog stability in the absence of the core cohesin trimer.

Cohesin-independent SA proteins are localised at clustered regions in 3D.

Previous analyses of the contribution of SA proteins to genome organization [7,9] were performed in cells containing cohesin rings, possibly obscuring a functional role for SA proteins themselves in genome organization. To determine if cohesin-independent SA proteins may function at unique locations in the genome, we investigated whether the residual SA-CTCF complexes (herein, SA-CTCF$^{\Delta Coh)}$ in IAA-treated RAD21^{mAC} cells occupied the same chromatin locations as in control cells. Using chromatin immunopreciptation followed by sequencing (ChIP-seq), we determined the binding profiles of CTCF, SA1, SA2, RAD21 and SMC3 in RAD21^{mAC} cells treated with EtOH or IAA. Pairwise comparisons of CTCF ChIP-seq with RAD21 or SA in control RAD21^{mAC} cells revealed the expected overlap in binding sites (Fig.1i, S1h). In contrast, both global and CTCF-overlapping RAD21 and SMC3 ChIP-seq signals were dramatically lost in IAA-treated cells (Fig.1i, S1h). In agreement with our microscopy and biochemistry results, we detected residual SA1 and SA2 binding sites in IAA-treated cells which retained a substantial overlap with CTCF (Fig.1i, S1h). Furthermore, the sites co-occupied by CTCF and SA proteins in RAD21-depleted cells (SA-CTCF$^{\Delta Coh}$) were also bound in control conditions, were enriched at TSSs and were characterised by active chromatin marks including phospho-S5 POLR2A, H3K4me3 and H3K27ac (Fig.S1h). Thus, CTCF and SA maintain occupancy at their canonical binding sites in the absence of RAD21 (Fig 1i), suggesting that SA interaction with CTCF in the absence of the cohesin ring is a step in normal cohesin activity.

While depletion of cohesin results in a dramatic loss of Topologically Associated Domain (TAD) structure [8], the frequency of long-range inter-TAD, intra-compartment contacts (LRC) is increased [5,8], and enriched for CTCF [5] or active elements [8]. To determine whether residual, chromatin-bound SA could be associated with LRCs in the absence of RAD21, we re-analysed Hi-C data from control and IAA-treated RAD21^{mAC} cells [8]. We quantified all contacts within two different scales of genome organization; local TAD topology (100k-1Mb) and clustered LRCs (1-5Mb)

(Fig.1j). As previously shown [8], local TAD contacts are lost and clustered LRCs are enriched in IAA conditions (Fig.1j). We probed the Hi-C datasets for contacts containing the residual SA-CTCF$^{\Delta Coh}$ binding sites and observed a further enrichment in IAA conditions (Fig.1j, bottom right), indicating that SA-CTCF$^{\Delta Coh}$ are enriched at the clustered LRCs formed when cells are depleted of cohesin and thus implicating them in 3D structural configurations. Our results suggest that cohesin-independent SA, either with CTCF or alone, may itself contribute to higher-order arrangement of active chromatin and regulatory features in 3D space.

SA interacts with diverse 'CES-binding proteins' in cohesin-depleted cells.

SA proteins contain a highly conserved domain known as the 'stromalin conservative domain' [14,38], or the 'conserved essential surface' (CES). Structural analysis of CTCF-SA2-RAD21 has shown that a F/YXF motif in the N-terminus of CTCF engages with a composite binding surface containing amino acids from RAD21 and the CES of SA2, forming a tripartite interaction patch [18]. Further, the authors identified a similar F/YXF motif in other cohesin regulatory proteins, predicting interactions between the SA-RAD21 binding surface and additional chromatin proteins. To investigate whether SA proteins could associate with F/YXF-motif containing proteins beyond CTCF in the absence of RAD21 *in vivo*, we performed chromatin coIP with SA1 and SA2 in EtOH and IAA and probed for interaction with CTCF as before, and three additional F/YXF-motif containing proteins, CHD6, MCM3 and HNRNPUL2 (Fig. 2a, S2a). All of the proteins interacted with SA1 in RAD21-control cells and their interaction with SA1 was enriched upon RAD21-degradation, indicating that the association was not dependent on the amino acids contributed by RAD21 (Fig. 2a, S2b). Interestingly, despite SA2 also containing the conserved CES domain, the F/YXF-motif proteins showed little interaction with SA2 (Fig. 2a, S2b), pointing to the presence of additional features in SA1 that stabilise its interaction with F/YXF containing proteins *in vivo*. Together, our results confirm that SA, in particular SA1, can interact with proteins beyond just CTCF and reveal that RAD21 is not required for the interaction between SA and F/YXF proteins *in vivo.* These results prompted us to re-evaluate the role of SA in cohesin activity and consider possible novel functions for these proteins.

SA1 interacts with a diverse group of proteins in the absence of cohesin.

To delineate novel protein binding partners and putative biological functions of cohesin-indepenent SA1, we optimised our chromatin-bound, endogenous SA1 co-IP protocol to be compatible with mass-spectrometry (IP-MS) and used this to comprehensively characterize the SA1 protein–protein interaction (PPI) network in control and RAD21-degraded RAD21^{mAC} cells. Three biological replicates were prepared from RAD21^{mAC} cells that were either untreated (UT) or treated with IAA (IAA) and processed for IP with both SA1 and IgG antibodies. In parallel, RAD21^{mAC} cells were also treated with scrambled siRNAs or with siRNA to SA1 to confirm the specificity of putative interactors. Immunoprecipitated proteins were identified by liquid chromatography tandem mass spectroscopy (LC-MS/MS). SA1 peptides were robustly detected in all UT and siCON samples and never detected in IgG controls, validating the specificity of the antibody. We used a pairwise analysis of IAA vs UT SA1 samples to generate a fold-change value for each putative interactor. These candidates were changed by at least 1.5-fold compared to UT controls, and sensitive to siSA1, yielding 136 high-confidence interactors whose abundance was significantly altered with RAD21 loss (SA1$^{\Delta Coh}$) (Fig. 2b, Table S1). As expected, core cohesin subunits SMC1A and SMC3 were strongly depleted while no peptides were detected for RAD21 (Fig. 2b). SA1 itself was significantly depleted compared to control cells, as were other cohesin regulators, known to directly interact with SA1, such as PDS5B [39]. In line with the enrichment we observed for the CES-binding proteins in IAA-conditions (Fig. 1c, 2a), the vast majority of the SA1$^{\Delta Coh}$ interactors were enriched for binding with SA1 in IAA conditions (117 of 136), and represent the cohesin-independent SA1$^{\Delta Coh}$ interactors (Fig. 2b).

We used STRING to analyse associations in the SA1$^{\Delta Coh}$ interactome and to identify enriched biological processes and molecular functions. We calculated enrichment relative to either the whole genome or to the SA1 interactome, which was determined in parallel from control cells (Fig. S2c), revealing similar enrichment terms to previously published cohesin interactomes [40,41], and validating our approach. Compared to the whole genome background, processes enriched in the presence of cohesin were still enriched in the SA1$^{\Delta Coh}$ PPI network and include a variety of functionally diverse cellular processes such as chromosome organization,

transcription, RNA processing, ribosome biogenesis, and translation (Fig. 2c, d). Within this group there are chromatin remodeling proteins (INO80 and SMARCAL1) and several transcriptional and epigenetic regulators such as JARID2 and TAF15. Similar to our ChIP and coIP results, this suggests that SA1 maintains interaction with proteins that localize with it in the presence of cohesin, albeit at different abundances.

RNA processing was the most enriched category in the SA1$^{\Delta Coh}$ PPI network (FDR=3.62x10^{-39}) and included proteins involved in RNA modification (YTHDC1, ADAR1, FTSJ3), mRNA stabilization and export (SYNCRIP, FMR1), and RNA splicing regulators (SRSF1, SON). We also found a significant enrichment for DNA and RNA helicases (FDR=3.54x10^{-08}) as well as RNA binding proteins (FDR=9.11x10^{-11}) within which were many HNRNP family members (HNRNPU, aka SAF-A). We also found a highly significant enrichment of proteins associated with ribosome biogenesis (FDR=2.20x10^{-30}) including both large and small subunit components; rRNA processing factors and components of the snoRNA pathway (FDR=4.39x10^{-05}). Finally, translation was significantly enriched as a biological process (p=1.64x10^{-06}), with several cytoplasmic translation regulators identified as SA1$^{\Delta Coh}$ interactors (DHX29, GCN1L1) (Fig. 2c, d, S2d, e). Among these is ESYT2 which is primarily found in the cytoplasm and contains a F/YXF-motif (Fig. 2c,d). We validated 8 of the highest-ranking proteins within the enriched functional categories described above in EtOH and IAA-treated RAD21^{mAC} cells (Fig. 2d). Importantly, the enrichment of these proteins with SA1 in the IAA condition suggests that SA may have a role in these processes independently of the core cohesin complex.

Comparison of the SA1$^{\Delta Coh}$ interactome with the SA1 interactome revealed that the proteins involved in RNA processing (FDR= 0.0298), ribosome biogenesis (0.0197), ribonucleoprotein complex biogenesis (0.0298) and rRNA processing (0.0409) were enriched with SA1 following IAA treatment compared to SA1 in the presence of RAD21 (Fig 2c, dotted lines). Overall, our results show that SA1$^{\Delta Coh}$ PPIs contain not only transcriptional and epigenetic regulators, but are in fact predominantly enriched for proteins with roles in RNA processing and modification, ribosome biogenesis and translation pathways. Thus, SA1 is involved in several biological

processes and may facilitate an aspect of cohesin regulation at a variety of functionally distinct locations.

SA proteins bind RNA independently of cohesin.

Since RNA binding and RNA processing were among the most enriched categories in the SA1$^{\Delta Coh}$ PPI network, we hypothesized that SA proteins may also bind RNA. We performed SA-crosslinking and immunoprecipitation (CLIP) in untreated RAD21^{mAC} cells and found that both SA1 and SA2 directly bound RNA (Fig. 3a, b). This was evidenced by detection of RNPs of the expected molecular weights, with a smear of trimmed RNA, which was stronger in the +UV and +PNK conditions, increased as the RNaseI concentration was reduced, and which was lost after siRNA-mediated SA KD (Fig. S3a-c). We repeated the experiment in EtOH- and IAA-treated RAD21^{mAC} cells to determine if the SA subunits can bind RNA in the absence of cohesin. As before, RAD21 depletion reduced SA1 and SA2 (Fig. S3d) and the amount of RNA crosslinked remained proportional to the amount of residual SA1 and SA2 protein (Fig. 3c, S3e), demonstrating that cohesin is not required for the interaction of these proteins with RNA in cells. Thus, cohesin-independent SA proteins interact with a wide array of RNA binding proteins (RBPs) as well as with RNA itself.

SA proteins localise to endogenous R-loops in the absence of cohesin.

Proteins involved in RNA processing, such as splicing, modification and export, act as regulators of R-loops [42]. Furthermore, R-loops accumulate at sites of multiple biological processes including transcription, DNA replication and DNA repair [42]. As many of these processes were enriched in the SA1 interactome, we reasoned that the diversity of biological processes represented in the SA1$^{\Delta Coh}$ PPI network may be reflective of a role for SA proteins in R-loop biology.

We performed a number of experiments to investigate the localization of SA proteins at endogenous R-loops. First, we found a correlation between global SA and R-loop levels. We depleted endogenous R-loops by overexpressing ppyCAG-RNaseH-V5 in HCT116 cells. IF using the R-loop specific antibody S9.6 revealed that nuclear S9.6 levels were significantly reduced in cells which expressed V5 (38% of controls, p=0.04) and that mean SA1 signal was significantly reduced by 29% in the same cells (Fig. 3d). Furthermore, RAD21^{mAC} cells treated with scramble control

siRNAs or Smartpool (SP) siRNAs to AQR (a known suppressor of R-loops [43]), SA1 or SA2 revealed that S9.6 IF signal was significantly increased in siAQR and siSA1 but not siSA2 cells compared to the siScr control (mean S9.6 signal increased by 28%, p=0.0004; 32%, p=3.90E-8; reduced by 10%, p=0.17, respectively) (Fig. S3f). Although S9.6 signal was reduced by IF in RAD21^{mAC} cells treated with IAA, this did not represent a significant change using this method (Fig. S3g).

We also performed STORM imaging on EtOH and IAA-treated RAD21^{mAC} cells to assess the nuclear distribution of SA1 in the context of R-loops with and without RAD21. We measured the ratio of the SA1 signal inside and outside of the S9.6 signal mask. A ratio of 1 indicates a random distribution of SA1 with respect to S9.6 domains while a ratio above 1 reflects enrichment within S9.6 domains. In EtOH conditions, we did not detect enrichment of SA1 localizations, in fact SA1 was modestly depleted (mean ratio 0.93). However, upon IAA treatment, we observed a significant enrichment of SA1 localizations within S9.6 domains (mean ratio 1.24, $p<0.0001$) (Fig. 3e), strongly suggesting that SA1 proteins are localized within R-loop domains independently of cohesin.

In addition, we returned to our IP-MS experiment to analyse enrichment of R-loop-associated proteins in our SA1$^{\Delta Coh}$ interactome. We overlapped the proteins identified in two independent IP-MS experiments for R-loop interactors [44,45] to create a high-confidence 'R-loop interactome' and then used a hypergeometric distribution to determine the significance of this category in the SA1$^{\Delta Coh}$ interactome (Methods). Both the custom R-loop interactome as well as proteins from the individual studies were highly over-enriched in the SA1$^{\Delta Coh}$ interactome (FDR=1.1x10^{-15}, 1.4x10^{-47}, 7.7x10^{-19}, respectively) (Fig. 3f). To directly measure this, we optimised a coIP method using the S9.6 antibody in RAD21^{mAC} cells (Fig. 3g, S3h). In agreement with published results, we found that S9.6 precipitated the known R-loop helicases AQR, DHX9, RNase H2 [43,44] as well as MCM3 and RNA Pol II (POLR2)[46]. Both SA1 and SA2 precipitated with S9.6 and treatment with RNase H (RNH) revealed the specificity of the S9.6-SA interactions since the reduction of R-loop signal was proportional to the observed reduction in coIP of SA1 by S9.6 (Fig. 3g, S3h, i).

Finally, we used a high resolution, genome-wide method to detect R-loops in HCT116 cells. RAD21^{mAC} cells were treated with RNH to confirm the specificity of our

335 method and with EtOH or IAA to assess the impact of cohesin loss on R-loops and
336 subjected to DNA-RNA Immunoprecipitation coupled with sequencing (DRIP-seq)
337 using the S9.6 antibody. We combined these datasets with our ChIP-seq for SA
338 proteins, RAD21 and CTCF in EtOH or IAA conditions to confirm the associations
339 described above. We detected 50,338 RNH-sensitive R-loop sites which were also
340 sensitive to acute degradation of RAD21, albeit not to the same extent as RNH
341 treatment (average S9.6 signal was reduced by 31.4% in RNH and 16.8% in IAA) (Fig.
342 3h, i). Among the RNH-sensitive R-loop sites, we detected two regimes of SA-R-loop
343 biology. A small proportion of R-loop sites directly overlapped with SA1/2, RAD21 and
344 CTCF in control EtOH conditions. These sites were enriched at genes and both the
345 SA1 and SA2 read density was sensitive to RAD21 loss (Fig. 3h, i, S3j, k). On the
346 other hand, a larger proportion of R-loops had SA signals adjacent (bound within 2kb
347 of the R-loop peak). Interestingly, these SA sites were enriched in repressed chromatin
348 and were not sensitive to RAD21 loss, in fact their read density was enriched
349 compared to EtOH conditions (Fig. 3h, i, S3j, k), reminiscent of the enrichment
350 observed previously by STORM imaging (Fig. 3e).
351
352 **NIPBL-independent cohesin loading mediated by SA proteins.**
353 Our results thus far revealed that $SA^{\Delta Coh}$ is localised to clustered regions, engages
354 with RNA and various RBPs and is localised to R-loops hybrids. Several lines of
355 evidence suggest that alongside the canonical NIPBL/Mau2 loading complex, SA
356 proteins contribute to cohesin's association with chromatin [13,14] and that its functions
357 may go beyond simply acting as a bridging protein [14]. Thus, we hypothesized that SA
358 proteins support genome organization in their own right and herein facilitate cohesin's
359 association with chromatin.
360 The $RAD21^{mAC}$ system has the advantage that when IAA is washed-off cells,
361 the RAD21 protein is no longer degraded and can become 're-loaded' onto chromatin.
362 We assessed this by measuring mClover signal intensity using IF and observed that it
363 was robustly lost in IAA conditions and was partially restored to EtOH levels within 4hr
364 of IAA withdrawal (Fig. 4a,b, S4a). We note the spatial distribution of RAD21 was itself
365 variable, ranging between highly compartmentalised and randomly distributed (Fig.

4a, c). This provided a unique opportunity to assess how SA influences cohesin reloading *in vivo* and the potential role for RNA and R-loops in this process.

We assessed reloading using both single-cell and bulk methods, coupled with siRNA-mediated KD to determine how specific proteins affected cohesin reloading *in vivo*. We first measured the impact of the canonical cohesin loader, NIPBL. RAD21^{mAC} cells were treated with scramble or NIPBL siRNAs and subsequently grown in EtOH or IAA. The '0h' and '4h' post EtOH/IAA wash-off samples represent the extent of cohesin *degradation* or *reloading*, respectively (Fig.S4b). Chromatin fractionation in high-salt conditions followed by immunoblot analysis confirmed the loss of the loader complex, NIPBL and MAU2 (known to become destablised upon NIPBL loss[47]). As expected, in NIPBL KD conditions, mean RAD21 re-loading efficiency was reduced, although surprisingly, this was incomplete (41% of the siRNA controls; mean re-loading siNIPBL, 2.1 vs siCon, 3.6), and did not represent a statistically significant difference (p=0.33) (Fig. 4d, e, S4c). This result was reproduced using IF, where mean mClover signal in siNIPBL-treated cells was 45.1% of siRNA control (MFI siCON, 6563 vs siNIPBL, 3602) (Fig 4g), indicating that cells can still load cohesin in the absence of NIPBL.

We reasoned that SA proteins may be contributing to the observed NIPBL-independent reloading. Thus, we repeated the experiments to include siRNA to SA1 and SA2 together (siSA), and a siNIPBL+siSA condition. In both population and single cell analysis of reloading, SA KD had a more dramatic effect on cohesin re-loading efficiency than NIPBL KD, reducing RAD21 on chromatin to 51% of scramble controls (mean siSA, 1.9 vs siCon, 5.1, p=0.002 for Fig. 4f, S4d and MFI siSA,2303 p<0.0001 for Fig. 4g). In the absence of both SA and NIPBL, cohesin reloading was reduced further (mean siNIPBL+siSA, 1.4 vs siCon, 5.1, p=0.001 for Fig. 4f, S4d and MFI siNIPBL+SA,1925 p<0.001 for Fig. 4g), indicating that SA performs an important and complementary step to NIPBL during normal reloading. Given the differences between SA1 and SA2 reported herein, we also performed the reloading experiment to separate the effects of SA1 and SA2. As expected from our co-IP results (Fig. 1c), RAD21 levels in RAD21^{mAC} cells were more affected by siSA2 than siSA1 (Fig. S4e). We observed that cohesin reloading was more efficient in siSA2 (where SA1 is present) than in siSA1 (where only SA2 is present), and that siSA1 was similar in reloading to siSA

(Fig. S4e). Together these observations suggest that the bulk of the reloading in IAA conditions is supported by SA1.

SA proteins stabilize nascent RNA in the absence of cohesin.

Given the association of SA proteins with RNA and RBPs and the dependence of cohesin reloading on SA1, we tested the requirement for RNA in cohesin reloading. Cells were treated as above with a pulse of 5 ethynyl uridine (EU) prior to collection. EU becomes actively incorporated into nascent RNA and can be measured by IF alongside the change in RAD21-mClover. While a significant reduction in nascent RNA signal was detected upon treatment with Triptolide (TRP), mClover signal was not significantly changed compared to IAA washoff conditions, indicating that RNA is not a key determinant of cohesin reloading *per se* (Fig 4h, left panel). However, we did observe an increase in nascent RNA upon acute RAD21 degradation which returned to EtOH levels when cohesin became reloaded onto chromatin (Fig 4h, right panel). These results pointed to the stabilization of nascent RNA in the absence of cohesin. Given the association of SA1$^{\Delta Coh}$ with RNA and RBP, we repeated the experiment in siSA1 KD conditions. Again, we observed an increase in nascent RNA in IAA conditions compared to EtOH control, but this was no longer detected in cells treated with siRNA to SA1 (Fig. 4i). Our results point to a role for SA1 in the stabilization of nascent RNA in the absence of cohesin and a possible competition between SA-RNA and SA-cohesin associations.

Our results thus far showed that SA proteins remain chromatin associated in the absence of cohesin (Fig. 1), when they bind RBP (Fig. 2) and RNA and are localized to R-loops (Fig. 3). We also report that SA1 proteins contribute to cohesin's re-association with chromatin and that this involves nascent RNA (Fig. 4). Thus, we reasoned that SA may facilitate cohesin reloading at R-loops. It was technically challenging to measure reloading upon over-expression of ppyCAG-RNaseH. As an alternative, we used STORM imaging to assess the nuclear distribution of the reloaded cohesin in the context of R-loop clusters by comparing EtOH- and IAA-treated to IAA-washoff RAD21^{mAC} cells (Fig. 4j). As before, we measured the ratio of signal (this time RAD21-mClover) inside and outside of the S9.6 mask. Interestingly, in EtOH

conditions, RAD21 localizations were depleted from the S9.6 domain (mean ratio 0.95) (Fig. 4j) similar to what we observed for SA1 (Fig 3h). Since STORM is such a sensitive approach, trace localizations of mClover will always be detected, even in IAA conditions when the bulk of the signal is lost. The few localizations we observed were indeed modestly enriched within the S9.6 mask, although these were not significantly different from EtOH (mean ratio 1.08, p=0.10). These localizations may represent either extremely stable or freshly loaded cohesin. Upon IAA washoff, new RAD21-mClover molecules are readily detected, became significantly enriched within S9.6 domains compared to EtOH treated cells (mean ratio 1.19, p=0.029) and were sensitive to treatment with RNase H (mean ratio 0.98) (Fig 4j). Overall, our results point to a role for SA1 proteins in mediating reloading of cohesin at R-loops.

A basic exon in the C-terminus of SA2 tunes interactions with RBPs.

While both SA1 and SA2 played a role in cohesin's reloading, SA1 was the dominant paralog (Fig S4e). In addition, SA2 was not able to compensate for SA1 in R-loop stability (Fig S3f), despite its interaction with RNA (Fig. 3a, b) and R-loops (Fig. S3h). Previous publications have described association of RBP from SA2 MS-IP in HCT cells [40]. Indeed, several of these RBPs overlap with the proteins described here as SA1 interactors (Fig. 2b, c) and are enriched in SA1 IP in IAA conditions (Fig. 2a, d). However, we did not observe robust enrichment of RBPs compared to input in SA2 IP, in either EtOH or IAA conditions. This was reminiscent of the differential interactions between SA1 and SA2 with F/YXF containing proteins (Fig. 2a). These results thus raised the question of whether additional features in SA2 may be required to stabilize these interactions and functions.

SA1 and SA2 express transcript variants in RAD21^{mAC} cells. We re-analysed publicly available RNA-seq datasets and quantified alternative splicing profiles using VAST-tools analysis [48]. One prominent variant which is conserved between human and mouse (Fig. S5a, b), arises from the alternative splicing of a single C-terminal exon, exon 31 in SA1 (SA1$^{e31\Delta}$) and exon 32 in SA2 (SA2$^{e32\Delta}$) (Fig. 5a), The significance of this is unknown. We found that in human HCT cells, the majority of SA1 mRNAs *include* e31 (average 'percent spliced in' (PSI) 97.7%), while the majority of SA2 mRNAs *exclude* e32 (average PSI 20.4%) (Fig. 5b, S5a, b). We confirmed this at the protein level by designing custom esiRNAs to specifically target SA1 e31 or

SA2 e32 (Methods). Smartpool (SP) KD reduced the levels of SA1 and SA2 to similar extents compared to scrambled controls (87% and 94%, respectively) (Fig. 5c). Specific targeting of SA1 e31 led to a reduction of 85% of SA1 compared to esiRNA control (which was comparable to SP KD). In contrast, SA2 e32 targeting had a minimal effect on SA2 protein levels compared to its esiRNA control (reduction of 2%) (Fig. 5c), in line with the PSI data (Fig. 5b) and indicating that the dominant SA2 isoform does not contain e32.

These results imply that cells 'tune' the availability of e31/32 in SA proteins, prompting us to investigate the nature of these exons. Interestingly, the amino acid (aa) sequence of the spliced SA exons encode a highly basic domain within an otherwise acidic C-terminus (Fig. 5a, zoom-in). Overall, the SA paralogs are highly homologous, however the N- and C-termini diverge in their aa sequence. Despite this, e31 and e32 have retained their basic properties (pI=10.4 and 9.9, respectively) (Fig.5a, zoom-in). Basic patches can act as regulatory domains and bind nucleic acids prompting us to ask whether these alternatively spliced basic exons contribute to the association of SA proteins with RNA (Fig 3a). We cloned cDNAs from HCT116 cells representing the exon32-containing SA2 ($SA2^{e32+}$) and the canonical exon32-lacking SA2 ($SA2^{e32\Delta}$), tagged them with YFP, expressed them in HCT116 cells and purified the tagged isoforms to compare their ability to interact with RNA (Fig. 5d) using CLIP. While the presence of e32 did not change the ability of SA2 to interact with RNA (Fig. 5d, blue arrows), cells expressing the alternative exon routinely enriched RBPs with molecular weights ~110-140kDa (Fig 5d, black arrow), strongly suggesting that the e32 domain may act to stabilize the association of SA2 with RBPs.

To identify the proteins stabilized by the presence of e32, we coupled YFP-SA2 isoform CLIP with Mass Spectrometry. Three biological replicate IPs were prepared from $RAD21^{mAC}$ cells that were transfected with either YFP-$SA2^{e32+}$ or YFP-$SA2^{e32\Delta}$. YFP IP efficiency for $SA2^{e32+}$ or $SA2^{e32\Delta}$ was similar and both isoforms interacted with core cohesin subunits (Fig S5c). We identified a total of 238 proteins, the majority of which overlap in the two SA IPs and with a previously published SA2 IP [40] (Fig S5c, d). We used a pairwise analysis of $SA2^{e32+}$ vs $SA2^{e32\Delta}$ samples to generate a fold-change value for each putative interactor (Fig 5e). GO analysis of proteins changed by at least 1.5-fold, and absent in Mock IP revealed a mild enrichment for post-

translational modification category from the SA2$^{e32\Delta}$ IP (FDR=0.0234, p=1.35e-06), and conversely an enrichment of the RNA Binding category from SA2^{e32+} (FDR= 3.43E-05, p=6.56E-09). Interestingly, the enriched proteins included YTHDC1 and YTHDF3 (previously identified in the SA1$^{\Delta Coh}$ interactome, Fig 2b), DIS3 and POLR2B, all known to play key roles in RNA-protein complexes and stability, have molecular weights ~110-140kDa and thus likely represent the specifically enriched band in the CLIP experiments (Fig 5d, black arrow). Finally, the observation that a basic exon 32 domain in SA2 supports the stability of RNA-RBP interactions led us to investigate if exon 32 also stabilized SA2 at R-loops. We repeated the S9.6 IP in RAD21^{mAC} cells expressing either YFP-SA2^{e32+} or YFP-SA2$^{e32\Delta}$. As before, AQR and MCM3 were enriched by S9.6 IP (Fig. 5f) and we found that SA2^{e32+} was more enriched in the S9.6 IP compared to SA2$^{e32\Delta}$ (enrichment of 1.8x and 1.24x respectively, relative to endogenous SA2) (Fig 5f, g). Taken together, our results support a role for the alternatively spliced C-terminal basic domain of SA in stabilizing interactions with RBPs and R-loops.

DISCUSSION

Whether SA proteins function in their own right outside of the cohesin complex is rarely considered. Consequently, our understanding of how these proteins contribute to cohesin function and disease is incomplete. In this study, we shed light on this question by uncovering a diverse repertoire of SA1 interactors in cells acutely depleted for the cohesin ring. This ranges from proteins associated with translation and ribosome biogenesis to RNA processing factors and regulators of the epitranscriptome. These observations suggest that SA1 has a previously unappreciated role in post-transcriptional regulation of gene expression which offers much-needed new insight into its roles in disease and cancer.

Acute depletion of the cohesin ring has allowed us to capture a moment in the normal life cycle of cohesin–DNA associations and unveiled a previously unknown step for SA proteins herein. We show that SA proteins, independent of the cohesin trimer, bind to DNA and RNA, either in the context of RNA:DNA hybrid structures, as we have shown here, or perhaps sequentially, and use this platform for the loading of

cohesin to chromatin. Our results are supportive of biophysical observations of SA proteins and R-loops [49] and *in vitro* assessment of cohesin loading at DNA intermediates [26]. These results point to the importance of DNA *structure,* as opposed to sequence, in the targeting of cohesin to chromatin. Furthermore, structural studies suggest that NIPBL and SA1 together bend DNA and cohesin to guide DNA entering into the cohesin ring [19,20,50]. Our work shows that in cells lacking either the canonical NIPBL/MAU2 loader complex or the SA proteins, cohesin can still associate with chromatin, suggesting that loading can occur with either component alone, albeit most effectively together.

Our results represent a new view onto the role of SA proteins in cohesin biology. Since SA paralogs have distinct terminal ends and nucleic acid targeting mechanisms [24,25], their recruitment to chromatin may be specified by unique DNA, RNA or protein-interactions, or indeed all three. Such diversification of loading platforms would be important in large mammalian genomes to ensure sufficient amounts of cohesin on chromatin or to stabilize cell-type specific chromatin structures [51]. Indeed, SA1 and SA2 show clear differences in interaction with F/YXF-motif containing proteins, despite the fact that both paralogs contain a CES domain [52], underscoring the importance of investigating interactions *in vivo* and arguing that additional features play an important role in complex stabilization. In this context, RNA-associated protein interaction has previously been shown to support cohesin stabilisation at CTCF at the *IGF2/H19* locus [53]. These results are in line with our findings that a basic domain in the unstructured C-terminal portion of SA supports RNA-associated protein interactions.

This study also reveals SA1 as a novel regulator of R-loop homeostasis. It is noteworthy that other suppressors of R-loop formation include RNA processing factors, chromatin remodellers and DNA repair proteins [28] which all function in the context of nuclear bodies [54]. We find that SA1 proteins are enriched at very distal chromatin interactions in cohesin-depleted Hi-C data, interact with numerous RBPs known to condense in 3D [55,56] and are enriched in S9.6 domains in cells where we find cohesin becomes associated with chromatin. Harnessing such condensates would provide an efficient loading platform for cohesin at sites of similar biological function. Yeast cohesin has been shown to mediate phase separated condensate structures [57]. Our results support this view and further suggest that it is SA (and possibly

559 predominantly SA1 in HCT116 cells), with its propensity for intrinsically disordered
560 domains [51] that contribute to this formation, thereby linking cohesin loading to
561 biological functions. We note that if SA paralogs or isoforms direct different localization
562 of cohesin loading or stability of its association, this could have important implications
563 in our understanding of disease where SA proteins are commonly mutated, such as in
564 cancers.

ACKNOWLEDGMENTS

This work was supported by a Senior Research Fellowship from the Wellcome Trust awarded to S.H. (106985/Z/15/Z) and a CRUK PhD studentship awarded to H.P. The Proteomics work was supported by the CRUK–UCL Centre Award [C416/A25145]. We are grateful to Jernej Ule for his support with DRIP-sequencing and to Julian Zagalak and the CRICK sequencing facility for reagents, advice and assistance. We thank Stanimir Dulev for his contributions at the early stages of the project and Jiten Manji for his support with microscopy. We also thank Konstantina Skourti-Stathaki for advice about S9.6 antibody, IFs and R-loops. We are grateful to the members of the Hadjur lab for critical discussions and reading of the manuscript.

AUTHOR CONTRIBUTIONS

H.P. and S.H. conceived the project. H.P. designed and performed the coIP, Mass spectrometry, DRIP-sequencing and cohesin re-loading experiments, analysed the ChIP, DRIP and Hi-C data and performed the statistical analysis for mass spectrometry with the support of A.B. and S.S. Y.L. performed and analysed all imaging experiments (apart from STORM), derived clonal lines of RAD21^{mAC} cells, cloned YFP-tagged SA2 cDNAs and performed CLIP and CLIP-MS together with M.T.C. W.V. performed Hi-C, ChIP-seq and splicing analyses. M.V.N., L.M. and M.P.C. performed and analysed STORM imaging. D.P. discovered splicing features of the SA isoforms. H.P., Y.L. and M.B. prepared cellular materials for CLIP, which was carried out by M.B. and M.T.C. and supervised by R.G.J. A.B. and S.S. performed mass spectrometric and proteomic analysis. H.P., Y.L and S.H. formatted all figures and wrote the manuscript with input from all authors.

Accession Numbers

Genomic data generated in this study (ChIP-seq) was submitted to GEO with the Accession GSE167887. The mass spectrometry proteomics data was deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD024354.

Declaration of Interests

597 The authors declare no competing interests.

FIGURE LEGENDS

Figure 1. SA interacts with CTCF in the absence of cohesin.

a) Representative confocal images of SA1 and CTCF IF in RAD21^{mAC} cells treated with EtOH (EtOH) as a control or Auxin (IAA) for 4hrs. Nuclei were counterstained with DAPI.

b) Imaris quantification of the relative mean fluorescence intensity (MFI) of mClover, CTCF, SA1 and SA2 in EtOH and IAA-treated RAD21^{mAC} cells. Whiskers and boxes indicate all and 50% of values, respectively. Central line represents the median. Asterisks indicate a statistically significant difference as assessed using two-tailed t-test. **** $p<0.0001$. $n>50$ cells/condition from 3 biological replicates.

Chromatin coIP of (c) SA1, SA2 and IgG with RAD21 and CTCF or (d) CTCF and IgG with RAD21, SA1 and SA2 in RAD21^{mAC} cells treated with EtOH or IAA for 4hrs. Input represents (c) 2.5% and (d) 1.25% of the material used for immunoprecipitation.

e) Dual-color STORM images of SA1 (green) and CTCF (magenta) in EtOH and IAA-treated RAD21^{mAC} cells. Representative full nuclei and zoomed nuclear areas are shown. Line denotes 2 microns and 200nm for full nuclei and zoomed areas respectively. See Supplementary Figures for SA2 STORM images.

f) Mean CTCF, SA1 and SA2 localization densities (localizations normalized with nuclear area) in EtOH and IAA-treated RAD21^{mAC} cells ($n = >30, >17$ and >15 nuclei for CTCF, SA1 and SA2 respectively). Mean and SD are plotted, Mann Whitney test. ** $p<0.005$, *** $p<0.0005$, **** $p<0.0001$.

g) Mean Nearest Neighbor Distance (NND) of CTCF, SA1 and SA2 clusters in nanometers in EtOH and IAA-treated cells ($n = >38, >14$ and >23 nuclei for CTCF, SA1 and SA2 respectively). Mean and SD are plotted, Mann Whitney test. **** $p<0.0001$.

h) NND distribution plot of the distance between CTCF and SA1 (left panel) or SA2 (right panel) clusters in EtOH and IAA-treated cells. Experimental data are shown as continuous lines, random simulated data are displayed as dotted lines.

i) ChIP–seq deepTools heat map of CTCF, SA1, SA2, Rad21 and SMC3 binding profiles in control (EtOH) and IAA-treated RAD21^{mAC} cells. Selected regions are bound by CTCF in control conditions.

j) Analysis of contact frequency hotspots from Hi-C libraries generated from EtOH-treated (top row) and IAA-treated (bottom row) RAD21^{mAC} cells. Contact frequencies were calculated in two distance ranges of 100kb – 1Mb and 1-5Mb. The last column includes contact frequencies specifically at SA-CTCF$^{\Delta Coh}$ binding sites.

Figure 2. Characterization of SA1 protein-protein interaction network in RAD21-depleted cells.

a) Chromatin coIP of SA1, SA2 and IgG with 4 predicted CES-binding proteins in RAD21^{mAC} cells treated with EtOH or IAA for 4hrs. Input represents 1.25% of the material used for immunoprecipitation.

b) Volcano plot displaying the statistical significance (-log10 p-value) versus magnitude of change (log2 fold change) from SA1 IP-MS data produced from untreated or IAA-treated RAD21^{mAC} cells (n=3). Vertical dashed lines represent changes of 1.5-fold. Horizontal dashed line represents a pvalue of 0.1. Cohesin complex members and validated high-confidence proteins have been highlighted.

c) SA1$^{\Delta Coh}$ interaction network of protein–protein interactions identified in RAD21^{mAC} cells using STRING. Node colours describe the major enriched categories, with squares denoting helicases. Proteins within each enrichment category were subset based on p-value change in B). See supplemental figures for full network. Dashed boxes indicate the proteins and categories which were specifically enriched in IAA-treatment compared to the SA1 interactome.

d) Chromatin IP of SA1 and IgG in RAD21^{mAC} cells treated with EtOH or IAA and immunoblotted with antibodies to validate the proteins identified by IP-MS. Input represents 1.25% of the material used for immunoprecipitation. * We note that ESYT2 is a F/YXF-motif containing protein.

Figure 3. SA proteins bind to RNA and localise to R-loops.

a) CLIP for SA1, b) SA2 and IgG controls. Autoradiograms of crosslinked ^{32}P-labelled RNA are shown at the top and the corresponding immunoblots, below. CLIP was performed with and without UV crosslinking and polynucleotide kinase (PNK) and with high (H; 1/50 dilution) or low (L; 1/500 dilution) concentrations of RNase I.

c) CLIP for SA1, SA2 and IgG control in EtOH (-) or IAA-treated (+) Rad21^{mAC} cells. ^{32}P-labelled RNA and the corresponding immunoblots are shown as above.

d) Top, Representative confocal images of S9.6 and SA1 IF in RAD21^{mAC} cells untreated or overexpressing ppyCAG-RNaseH-v5. Expressing cells were identified with v5 staining. Nuclear outlines (white) are derived from DAPI counterstain. Bottom, Imaris quantification of the relative mean fluorescence Intensity (MFI) of S9.6 and SA1. Data are from two biological replicates with >75 cells counted/condition. Quantifications and statistical analysis were done as above.

e) STORM analysis of localization density for SA1 in S9.6 masks in EtOH and IAA. Ratio of SA1 localizations inside and outside S9.6 masks is shown. Ratio of above 1 represents an enrichment within the S9.6 domain. Mean and SD are plotted, statistics based on One-Way Anova test. Data are from two biological replicates.

f) -log10 transformed adjusted p-value (FDR) for enrichment of S9.6 interactome data from Cristini et al. and Wang et al., with the SA1$^{\Delta Coh}$ interactome. Overlap indicates the proteins identified in both of the S9.6 interactome datasets, representing a high confidence R-loop interactome list.

g) Chromatin coIP of S9.6 and IgG in RAD21^{mAC} cells treated with RNase H and immunoblotted with antibodies representing known R-loop proteins, as well as SA1. Input represents 1.25% of the material used for immunoprecipitation. Bottom, S9.6 dot blot of lysates used in coIP.

h) deepTools heatmap of DRIP-seq and ChIP-seq from RAD21^{mAC} cells. DRIP-seq was carried out in control (EtOH), RNase H (RNH) and IAA-treated cells. ChIP-seq was carried out for SA1, SA2, RAD21, CTCF and IgG in EtOH and IAA treated cells. Regions were selected based on DRIP-seq sensitivity to RNH and proximity with SA1 ChIP-seq. BEDTools identified regions of overlap or adjacent SA1 co-binding.

i) Summary plots showing mean DRIP-seq (top) or ChIP-seq (bottom) read density across the regions from h), including sites of R-loop and SA 'overlap' (Left) or 'adjacent' (right) regions. Input samples are indicated with dotted lines.

Figure 4. SA proteins contribute to cohesin loading.

a) Representative confocal images of immunofluorescence for mClover in EtOH, IAA and IAA washoff conditions. White lines denote nuclei based on DAPI staining.

b) Imaris quantification of the mean fluorescence intensity (MFI) of mClover in EtOH, IAA-treated and IAA washoff RAD21^{mAC} cells. Analysis and statistics as before. n>50 cells/condition from 2 biological replicates.

c) Examples of individual cells 4hr post IAA washoff showing different distributions of mClover signal within the nucleus. White lines denote nuclei based on DAPI staining.

d) Representative immunoblot analysis of chromatin-bound RAD21, MAU2 and NIPBL levels in RAD21^{mAC} cells treated with scramble control siRNA (si Con) or siRNA to NIPBL followed by EtOH or IAA treatment. 0h and 4h represent no wash-off of IAA or a sample taken 4 hours after washoff of IAA (the 'reloading timepoint). H3 was used as a loading control. *NB* The full blots are in Fig. S4c.

e) Western blot densitometry quantification. RAD21 fold change relative to siCon samples at the 0h timepoint in siCon 4h (grey), siNIPBL 0hr (light blue) and siNIPBL 4hr (dark blue). Whiskers and boxes indicate all and 50% of values respectively. Central line represents the median. Statistical analysis as assessed using a two-tailed t-test. Data is from 8 biological replicates.

f) Representative immunoblot analysis of chromatin-bound RAD21, SA1, SA2, MAU2 and NIPBL levels in RAD21^{mAC} cells treated according to the schematic described in Fig. S4b and including samples treated with siRNA to SA1 and SA2 together (siSA)

and siRNA to NIPBL + siSA. H3 was used as a loading control. Quantifications can be seen in in Fig. S4c.

g) Imaris quantification of the mClover MFI in EtOH, IAA-treated and IAA washoff RAD21^{mAC} cells treated with siRNA to NIPBL, SA1/2 and siRNA to NIPBL + siSA. Asterisks indicate a statistically significant difference as assessed using 2-tailed T-test. Data is from 2 biological replicates with >50 cells per experiment.

h) Imaris quantification of the mClover MFI (left) and RNA (based on EU incorporation (right) in EtOH and IAA-treated RAD21^{mAC} cells. Whiskers and boxes indicate all and 50% of values, respectively. Central line represents the median. Asterisks indicate a statistically significant difference as assessed using one-way ANOVA. n>50 cells/condition from 2 biological replicates.

i) Analysis of mClover MFI as in h) above, this time treated with siRNA to SA1/2 or scrambled controls. Analysis and statistics as above. n>50 cells/condition from 2 biological replicates.

j) STORM analysis of localization density for RAD21-mClover in S9.6 masks in EtOH (black), IAA (grey) and IAA washoff (green) conditions. Ratio of mClover localizations inside and outside S9.6 masks is shown. Ratio of above 1 represents an enrichment within the S9.6 domain. Mean and SD are plotted, statistics based on One-Way Anova test. Data are from two biological replicates.

Figure 5. A basic exon in SA2 influences RBP stability.

a) Schematic of the SA1 and SA2 proteins showing the SA1-specific AT-hook, the conserved CES domain (blue) and the acidic C-terminus (green) which contains the basic alternatively spliced exon (red). Right-hand zoom-in indicates the spliced exons for SA1 (top) and SA2 (bottom) and the pI for each. The conservation scores for the divergent N- and C-termini and the middle portion of the proteins which contains the CES domain are shown.

b) Percent Spliced In (PSI) calculations for SA1 exon 31 (black) and SA2 exon 32 (grey) based on VAST-Tools analysis of RNA-seq from multiple datasets (see Methods).

c) (top) Immunoblot analysis of SA1 levels in chromatin lysates after treatment with scrambled siRNAs (siCon), SmartPool SA1 siRNAs (siSA1 SP), control esiRNAs (esiCon) and esiRNA designed to target SA1 exon 31 for 48hrs in RAD21^{mAC} cells. (bottom) Immunoblot analysis of SA2 levels in chromatin lysates after treatment with scrambled siRNAs (siCon), SmartPool SA2 siRNAs (siSA2 SP), control esiRNAs (esiCon) and siRNA designed to target SA2 exon 32 for 48hrs in RAD21^{mAC} cells. H3 serves as a loading control. The percentage of knockdown (KD) after SA signal is normalised to H3 is shown.

d) CLIP with endogenous SA2, IgG control and cells where either YFP-tagged SA2 containing exon 32 (e32+) or YFP-tagged SA2 lacking exon 32 are expressed for

48hrs. CLIP reveals RNA associated with SA2 (blue arrows) and RBPs which specifically associate with exon-32 containing SA2 (black arrow).

e) Volcano plot displaying the statistical significance (-log10 p-value) versus magnitude of change (log2 fold change) from IP-MS of HCT116 cells expressing either YFP-SA2^{e32+} or YFP-SA2$^{e32\Delta}$ (n=3 biological replicate IP). Cohesin complex members are highlighted in green and the two most enriched functional categories of RNA-binding proteins in blue or Post-translational modification in red.

f) Chromatin coIP of S9.6 and IgG in RAD21^{mAC} cells expressing the YFP-SA2 isoforms and immunoblotted with antibodies representing known R-loop proteins, as well as endogenous SA2 (eSA2). Input represents 1.25% of the material used for immunoprecipitation. Bottom, S9.6 dot blot of lysates used in coIP. *NB* the shift in SA2 signal representing overexpressed protein (SA2-YFP).

g) Quantification of the immunoblot signal from f) of SA2 in the YFP-SA2 isoform band relative to input and to eSA2 signal. YFP-SA2^{e32+} is more enriched by S9.6 IP compared to YFP-SA2$^{e32\Delta}$.

REFERENCES

1. Leiserson, M. D. M. *et al.* Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nat. Genet.* **47,** 106–114 (2015).
2. Horsfield, J. A. *et al.* Cohesin-dependent regulation of Runx genes. *Development* **134,** 2639–2649 (2007).
3. Viny, A. D. & Levine, R. L. Cohesin mutations in myeloid malignancies made simple. *Curr. Opin. Hematol.* **25,** 61–66 (2018).
4. Hadjur, S. *et al.* Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. *Nature* **460,** 410–413 (2009).
5. Sofueva, S. *et al.* Cohesin-mediated interactions organize chromosomal domain architecture. *EMBO J.* **32,** 3119–3129 (2013).
6. Zuin, J. *et al.* Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc. Natl. Acad. Sci. U.S.A.* **111,** 996–1001 (2014).
7. Kojic, A. *et al.* Distinct roles of cohesin-SA1 and cohesin-SA2 in 3D chromosome organization. *Nature Publishing Group* **25,** 496–504 (2018).
8. Rao, S. S. P. *et al.* Cohesin Loss Eliminates All Loop Domains. *Cell* **171,** 305–309.e24 (2017).
9. Wutz, G. *et al.* Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J.* **36,** 3573–3599 (2017).
10. Fudenberg, G., Abdennur, N., Imakaev, M., Goloborodko, A. & Mirny, L. Emerging Evidence of Chromosome Folding by Loop Extrusion. *bioRxiv* 1–24 (2018). doi:10.1101/264648
11. Balbás-Martínez, C. *et al.* Recurrent inactivation of STAG2 in bladder cancer is not associated with aneuploidy. *Nat. Genet.* **45,** 1464–1469 (2013).
12. Solomon, D. A. *et al.* Frequent truncating mutations of STAG2 in bladder cancer. *Nat. Genet.* **45,** 1428–1430 (2013).
13. Murayama, Y. & Uhlmann, F. Biochemical reconstitution of topological DNA binding by the cohesin ring. *Nature* **505,** 367–371 (2014).
14. Orgil, O. *et al.* A conserved domain in the scc3 subunit of cohesin mediates the interaction with both mcd1 and the cohesin loader complex. *PLoS Genet.* **11,** e1005036 (2015).
15. Ciosk, R. *et al.* Cohesin's Binding to Chromosomes Depends on a Separate Complex Consisting of Scc2 and Scc4 Proteins. *Molecular Cell* **5,** 243–254 (2000).
16. Kikuchi, S., Borek, D. M., Otwinowski, Z., Tomchick, D. R. & Yu, H. Crystal structure of the cohesin loader Scc2 and insight into cohesinopathy. *Proc. Natl. Acad. Sci. U.S.A.* **113,** 12444–12449 (2016).
17. Hara, K. *et al.* Structure of cohesin subcomplex pinpoints direct shugoshin-Wapl antagonism in centromeric cohesion. *Nature Publishing Group* **21,** 864–870 (2014).
18. Li, Y. *et al.* Structural basis for Scc3-dependent cohesin recruitment to chromatin. *eLIFE* **7,** 352 (2018).
19. Shi, Z., Gao, H., Bai, X.-C. & Yu, H. Cryo-EM structure of the human cohesin-NIPBL-DNA complex. *Science* **368,** 1454–1459 (2020).
20. Higashi, T. L. *et al.* A Structure-Based Mechanism for DNA Entry into the Cohesin Ring. *Molecular Cell* **79,** 917–933.e9 (2020).
21. Xiao, T., Wallace, J. & Felsenfeld, G. Specific sites in the C terminus of CTCF interact with the SA2 subunit of the cohesin complex and are required for cohesin-dependent insulation activity. *Mol. Cell. Biol.* **31,** 2174–2183 (2011).
22. Saldaña-Meyer, R. *et al.* CTCF regulates the human p53 gene through direct interaction with its natural antisense transcript, Wrap53. *Genes Dev.* **28,** 723–734 (2014).
23. Bisht, K. K., Daniloski, Z. & Smith, S. SA1 binds directly to DNA through its unique AT-hook to promote sister chromatid cohesion at telomeres. *J. Cell. Sci.* **126,** 3493–3503 (2013).
24. Lin, J. *et al.* Functional interplay between SA1 and TRF1 in telomeric DNA binding and DNA-DNA pairing. *Nucleic Acids Research* **44,** 6363–6376 (2016).

25. Countryman, P. *et al.* Cohesin SA2 is a sequence-independent DNA-binding protein that recognizes DNA replication and repair intermediates. *J. Biol. Chem.* **293,** 1054–1069 (2018).
26. Murayama, Y., Samora, C. P., Kurokawa, Y., Iwasaki, H. & Uhlmann, F. Establishment of DNA-DNA Interactions by the Cohesin Ring. *Cell* 1–29 (2018). doi:10.1016/j.cell.2017.12.021
27. Zheng, G., Kanchwala, M. & Xing, C. MCM2–7-dependent cohesin loading during S phase promotes sister-chromatid cohesion. *eLIFE* 1–25 (2018). doi:10.7554/eLife.33920.001
28. García-Muse, T. & Aguilera, A. R Loops: From Physiological to Pathological Roles. *Cell* 1–15 (2019). doi:10.1016/j.cell.2019.08.055
29. Crossley, M. P., Bocek, M. & Cimprich, K. A. R-Loops as Cellular Regulators and Genomic Threats. *Molecular Cell* **73,** 398–411 (2019).
30. Li, X. & Manley, J. L. Inactivation of the SR protein splicing factor ASF/SF2 results in genomic instability. *Cell* **122,** 365–378 (2005).
31. Boque-Sastre, R. *et al.* Head-to-head antisense transcription and R-loop formation promotes transcriptional activation. *Proc. Natl. Acad. Sci. U.S.A.* **112,** 5785–5790 (2015).
32. Powell, W. T. *et al.* R-loop formation at Snord116 mediates topotecan inhibition of Ube3a-antisense and allele-specific chromatin decondensation. *Proc. Natl. Acad. Sci. U.S.A.* **110,** 13938–13943 (2013).
33. Natsume, T., Kiyomitsu, T., Saga, Y. & Kanemaki, M. T. Rapid Protein Depletion in Human Cells by Auxin-Inducible Degron Tagging with Short Homology Donors. *Cell Rep* **15,** 210–218 (2016).
34. Parelho, V. *et al.* Cohesins functionally associate with CTCF on mammalian chromosome arms. *Cell* **132,** 422–433 (2008).
35. Wendt, K. S. *et al.* Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature* **451,** 796–801 (2008).
36. Wutz, G. *et al.* ESCO1 and CTCF enable formation of long chromatin loops by protecting cohesinSTAG1 from WAPL. *eLIFE* **9,** (2020).
37. Ricci, M. A., Manzo, C., García-Parajo, M. F., Lakadamyali, M. & Cosma, M. P. Chromatin fibers are formed by heterogeneous groups of nucleosomes in vivo. *Cell* **160,** 1145–1158 (2015).
38. Roig, M. B. *et al.* Structure and function of cohesinâ€™s Scc3/SA regulatory subunit. *FEBS Letters* **588,** 3692–3702 (2014).
39. Hons, M. T. *et al.* Topology and structure of an engineered human cohesin complex bound to Pds5B. *Nature Communications* **7,** 1–11 (2017).
40. Kim, J.-S. *et al.* Systematic proteomics of endogenous human cohesin reveals an interaction with diverse splicing factors and RNA-binding proteins required for mitotic progression. *J. Biol. Chem.* **294,** 8760–8772 (2019).
41. Panigrahi, A. K., Zhang, N., Otta, S. K., Journal, D. P. B.2012. A cohesin–RAD21 interactome. doi:10.1042/BJ20111745
42. Santos-Pereira, J. M. & Aguilera, A. R loops: new modulators of genome dynamics and function. *Nat. Rev. Genet.* **16,** 583–597 (2015).
43. Sollier, J. *et al.* Transcription-coupled nucleotide excision repair factors promote R-loop-induced genome instability. *Molecular Cell* **56,** 777–785 (2014).
44. Cristini, A., Groh, M., Kristiansen, M. S. & Gromak, N. RNA/DNA Hybrid Interactome Identifies DXH9 as a Molecular Player in Transcriptional Termination and R-Loop-Associated DNA Damage. *Cell Rep* **23,** 1891–1905 (2018).
45. Wang, I. X. *et al.* Human proteins that interact with RNA/DNA hybrids. *Genome Res.* **28,** 1405–1414 (2018).
46. Skourti-Stathaki, K., Kamieniarz-Gdula, K. & Proudfoot, N. J. R-loops induce repressive chromatin marks over mammalian gene terminators. *Nature* **516,** 436–439 (2014).
47. Watrin, E. *et al.* Human Scc4 is required for cohesin binding to chromatin, sister-chromatid cohesion, and mitotic progression. *Curr. Biol.* **16,** 863–874 (2006).
48. Irimia, M. *et al.* A Highly Conserved Program of Neuronal Microexons Is Misregulated in Autistic Brains. *Cell* **159,** 1511–1523 (2014).
49. Pan, H. *et al.* Cohesin SA1 and SA2 are RNA binding proteins that localize to RNA

containing regions on DNA. *Nucleic Acids Research* **24,** 105–17 (2020).
50. Chao, W. C. H. *et al.* Structural Studies Reveal the Functional Modularity of the Scc2-Scc4 Cohesin Loader. *Cell Rep* **12,** 719–725 (2015).
51. Pezic, D. *et al.* The cohesin regulator Stag1 promotes cell plasticity through heterochromatin regulation. *bioRxiv* 1–60 (2021). doi:10.1101/2021.02.14.429938
52. Li, Y. *et al.* The structural basis for cohesin-CTCF-anchored loops. *Nature* **578,** 472–476 (2020).
53. Yao, H. *et al.* Mediation of CTCF transcriptional insulation by DEAD-box RNA-binding protein p68 and steroid receptor RNA activator SRA. *Genes Dev.* **24,** 2543–2555 (2010).
54. Misteli, T. Beyond the Sequence: Cellular Organization of Genome Function. *Cell* **128,** 787–800 (2007).
55. Nozawa, R.-S. *et al.* SAF-A Regulates Interphase Chromosome Structure through Oligomerization with Chromatin- Associated RNAs. *Cell* **169,** 1214–1227.e18 (2017).
56. Huo, X. *et al.* The Nuclear Matrix Protein SAFB Cooperates with Major Satellite RNAs to Stabilize Heterochromatin Architecture Partially through Phase Separation. *Molecular Cell* **77,** 368–383.e7 (2020).
57. Ryu, J.-K. *et al.* Phase separation induced by cohesin SMC protein complexes. *bioRxiv* **58,** 142–40 (2020).
58. Skourti-Stathaki, K. *et al.* R-Loops Enhance Polycomb Repression at a Subset of Developmental Regulator Genes. *Molecular Cell* **73,** 930–945.e4 (2019).
59. Nadel, J. *et al.* RNA:DNA hybrids in the human genome have distinctive nucleotide characteristics, chromatin composition, and transcriptional relationships. *Epigenetics & Chromatin* **8,** 46–19 (2015).
60. Barrington, C., Georgopoulou, D., Nature, D. P.2019. Enhancer accessibility and CTCF occupancy underlie asymmetric TAD architecture and cell type specific genome topology. *nature.com* doi:10.1038/s41467-019-10725-9
61. 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).
62. Choi, M. *et al.* MSstats: an R package for statistical analysis of quantitative mass spectrometry-based proteomic experiments. *Bioinformatics* **30,** 2524–2526 (2014).
63. Beltran, M. *et al.* The interaction of PRC2 with RNA or chromatin is mutually antagonistic. *Genome Res.* **26,** 896–907 (2016).

METHODS

Cell culture and IAA-mediated degradation of Rad21.

HCT116 cells with engineered RAD21-miniAID-mClover (RAD21mAC), or OsTIR1-only, or both (RAD21mAC-OsTIR) were obtained from Masato T. Kanemaki. Throughout this study we used RAD21mAC-OsTIR cells, and for simplicity we refer to them in the text as RAD21^{mAC}. The cells were maintained in McCoy's 5A medium with Glutamax (Thermo Fisher Scientific) supplemented with 10% Heat-inactivated FBS (Gibco), 700μg/ml Geneticin, 100μg/ml Hygromycin B Gold and 100μg/ml Puromycin as described. We clonally selected the RAD21mAC-OsTIR cells by sorting green fluorescence positive single cells on a FACS Aria Fusion cell sorter (BD Bioscience). Single cells were individually seeded into one well of a 96-well plate, expanded for 10 days into 6cm culture dishes and selected with Geneticin, Hygromycin B Gold and Puromycin as indicated above in McCoy's medium for another 10 days. Each clone was assessed for efficiency of Rad21 degradation using FACS analysis and western blotting (WB) using mClover, mAID and OsTIR antibodies. Two clones (H2 and H11) were taken forward and used throughout this study. To deplete RAD21, RAD21mAC-OsTIR cells were grown in adherent conditions for 3 days and treated with 500 μM Indole-3-acetic acid (IAA, Auxin, diluted in EtOH) for 4 hours. For IAA withdrawal, IAA treated cells were washed with PBS and replaced with fresh supplemented McCoy's medium for another 4 hours. Cells were washed twice with ice-cold PBS before being harvested for later experimental procedures.

siRNA-mediated knockdowns.

For siRNA transfections, RAD21^{mAC}-OsTIR cells were reverse transfected with scramble siRNA (siCon) or siRNAs targeting SA1, SA2, NIPBL, or AQR (Dharmacon, Horizon Discovery). A final concentration of 10 nM of siSA1, siSA2, or siNIPBL or 5 nM of siAQR was reverse transfected into the cells using Lipofectamine RNAiMAX reagent (Invitrogen), as per the manufacturer's instructions. Cells were plated at a density of 1 – 1.25 x 10^6 cells per 10 cm dish and harvested 72hrs post-transfection, at a confluency of ~70%. The Lipofectamine-containing media was replaced with fresh media 12-16 hrs post-transfection to avoid toxicity. For Figure 5f/g, incubation time was reduced to 40 hrs. To account for the reduced growth time, cells were plated at a density of 2-3 x 10^6 cells per 10 cm dish. Here siCon- and siNIPBL-transfected cells were plated at a lower cell number than siAQR-transfected cells to ensure equalised confluence (~70%) at the time of collection. When IAA-treatment was combined with siRNA mediated KD, the IAA was added at the end of the normal KD condition so that total KD time was not changed compared to UT cells. For esiRNA treatment, RAD21^{mAC}-OsTIR cells were reverse-transfected with 20 μM FLUC control esiRNA or esiRNA custom designed to SA1 exon31 or SA2 exon32 (MISSION® siRNA, Sigma Aldrich) using RNAiMAX (Invitrogen). Cells were incubated in transfection mixture for 7-8

hours before being replaced with fresh supplemented McCoy's medium and left for another 40h until harvest. Efficiency of KD was assessed by WB. siRNA information can be found in Table 1.

Immunofluorescence

Cells were adhered onto poly L-lysine coated glass coverslips in 6 well culture dishes and were washed twice with ice-cold PBS before IF procedures. For RAD21-depletion analysis, cells were fixed for 10 mins at room temperature with 3.7% paraformaldehyde (Alfa Aesar) in PBS, washed 3 times with PBS and then permeabilized at room temperature for 10 mins with 0.25% Triton X-100 in PBS (Sigma Aldrich). For R-loop imaging, cells were fixed and permeabilised with ice-cold ultra-pure mEtOH (Sigma Aldrich) for 10 mins at -20°C. After 3 washes with PBS, cells were blocked for 45 mins at room temperature with 10% FCS-PBS. For RNASEH1 enzyme treatment, cells were incubated with blocking solution supplemented with 1x RNASEH1 reaction buffer alone (50 mM Tris-HCl, 75 mM KCl, 3 mM $MgCl_2$, 10 mM DTT) or 5 units of RNASEH1 enzyme (M0297, New England Biolabs) for 30 mins at 37 °C, PBS-washed twice, before blocking. Cells were washed twice with PBS before incubation with primary antibodies diluted in 5% FCS-PBS at 4 °C overnight. Anti-SA1, anti-SA2 and anti-AQR were used at 1:3000 dilutions; anti-CTCF was used at 1:2500 dilution; anti-s9.6 was used at 1:1000 dilution; anti-V5 was used at 1:1000. After 4 washes with PBS, cells were incubated with secondary antibodies (donkey anti-Goat AF555 or AF647 for SA1/2 used at 1:3000; donkey anti-Rabbit AF647 for CTCF used at 1:2500; donkey anti-Mouse AF555 for s9.6 used at 1:2000; donkey anti-Rabbit AF647 for AQR used at 1:3000; donkey anti-Rabbit AF488 for V5 used at 1:2000)) in 5% FCS-PBS for 1 hour at room temperature, and washed 4 times with PBS before being mounted onto glass slides with ProLong™ Diamond Antifade Mountant with DAPI (Thermo Fisher Scientific) to stabilise overnight in dark before imaging. See Table 2 for details of where antibodies were purchased.

Imaging was performed on Zeiss LSM confocal microscopes using 63x/1.40 NA Oil Plan-Apochromat objective lens (Carl Zeiss, Inc.). Images were captured as z-stacks and under consistent digital gain, laser intensity and resolution for each experiment. Numerical analysis was carried out using Imaris software (Oxford Instruments, version 9.5.1) and representative images are shown as maximum z-projected views generated using Fiji Image J. In brief, z-stack images were imported into Imaris, cells were identified using DAPI and only those located 1 μm away from image boundary and sized between 120-800 μm^3 were selected. A seed-split function of 7.5um was used to separate closely situated cells. Fluorescence intensities of individual DAPI-selections in each channel were determined by Imaris and exported into Excel for further analysis. Distribution plots were generated from >50 cells of each replicate with 3 biological replicates per experiment. Student's *t*-test was performed between control and experimental conditions and statistical significance was determined by detecting the difference between means (unequal variance, two-tailed).

Significance is denoted as p>0.05 = not significant (ns), p≤0.05 = *, p≤0.005 = **, p≤0.0005 = *** and p≤0.0001 = ****.

Chromatin Fractionation and coImmunoprecipitation.

Cells were washed twice with ice-cold PBS (Sigma Aldrich) and lysed in Buffer A (10 mM HEPES, 10mM KCl, 1.5 mM MgCl2, 0.34 M Sucrose, 10% Glycerol, 1mM DTT, 1mM PMSF/Pefabloc, protease inhibitor), supplemented with 0.1% T-X100, for 10 min on ice. Lysed cells were collected by scraping. Nuclei and cytoplasmic material was separated by centrifugation for 4 min at 1300 g at 4oC. The supernatant was collected as the cytoplasmic fraction and cleared of any insoluble material with further centrifugation for 15 min at 20,000 g at 4oC. The nuclear pellet was washed once with buffer A before lysis in buffer B (3mM EDTA, 0.2mM EGTA, 1mM DTT, 1mM PMSF/Pefabloc, protease inhibitor) with rotation for 30 min at 4oC. Insoluble nuclear material was spun down for 4 min at 1700 g at 4oC and the supernatant taken as nuclear soluble fraction. The insoluble material was wash once with buffer B and then resuspended in high-salt chromatin solubilization buffer (50mM Tris-HCl pH 7.5, 1.5 mM MgCl2, 500mM KCl, 1mM EDTA, 20% Glycerol, 0.1% NP-40, 1mM PMSF/Pefabloc, protease inhibitor). The lysate was vortexed for 2 min to aid solubilization. Nucleic acids were digested with 85U benzonase (Sigma-Aldrich) per 100 x 106 cells, with incubation for 10 min at 37oC and 20 min at 4oC. Chromatin was further solubilized with ultra-sonication for 3 x 10 sec at an amplitude of 30. The lysate was diluted to 200 mM KCl and insoluble material was removed by centrifugation at 15,000 RPM for 30 min at 4oC.

For coIP, antibodies were bound to Dynabead Protein A/G beads (ThermoFisher Scientific) for 10 min at room temperature and ~ 5 hr at 4oC. For mock IgG IPs, beads were incubated with serum from the same host type as the antibody of interest. 1mg of chromatin extract was incubated with the antibody-bead conjugate per IP for approximately 16 hr at 4oC. IPs were washed x5 with IP buffer (200mM chromatin solubilization buffer) and eluted by boiling in either 2x Laemmeli sample buffer (BioRad) or 4x NuPAGE LDS sample buffer (ThermoFisher Scientific). Proteins ≤ 250 kDa were separated by SDS-PAGE electrophoresis using 4–20% Mini-PROTEAN® TGX™ Precast Protein Gels (BioRad) and transferred to Immobilon-P PVDF Membrane (Merck Millipore) for detection. Proteins ≥ 250 kDa were separated by SDS-PAGE electrophoresis using Invitrogen NuPAGE 3-8% Tris-Acetate precast protein gels. Transfer was extended to overnight with low voltage (20V) to aid in transfer of the high-molecular weight proteins. Membranes were incubated in primary antibody solution overnight at 4oC and images were detected using chemiluminescent fluorescence. Densitometry was carried out using ImageStudio Lite software with statistical significance calculated by unpaired t test, unless otherwise specified. Fold enrichment quantifications were performed by first normalising the raw densitometry value to its corresponding Histone H3 quantification and the comparing between the samples indicated. See Table 2 for details of antibodies.

S9.6 IP and Dot Blot.

Cells were fractioned and processed for S9.6 IP as described above, with the following modifications. To avoid digestion of RNA:DNA hybrids, samples were not treated with benzonase during chromatin solubilization and sonication was carried out for 10 min (Diagenode Biorupter) as in [44]. Where indicated, chromatin samples were treated with Ribonuclease H enzyme (NEB) overnight at 37°C to digest RNA:DNA hybrids in the extract. To avoid detection of single-stranded RNA by the S9.6 antibody, all S9.6 IP samples were pre-treated with Purelink RNase A (Thermo Fisher Scientific) at 0.25ug/1mg chromatin extract for 1 hr 30 min at 4°C. The reaction was stopped with addition of 143U Invitrogen SUPERase•In RNase Inhibitor (Thermo Fisher Scientific). RNA:DNA hybrid levels were assessed in chromatin samples by dot blot. Specifically, the chromatin lysate was directly wicked onto Amersham Protran nitrocellulose membrane (Merck) by pipetting small volumes above the membrane. Membranes were blocked in 5% (w/v) non-fat dry milk in PBS-0.1% Tween and incubated with S9.6 antibody overnight as for standard western blot. As above, detection was carried out using chemiluminescent fluorescence. RNase A-mediated digestion of RNA:DNA hybrids was performed using a non-ssRNA-specific enzyme (Thermo Scientific) at 1.5ug/25ug chromatin extract at 37°C.

ChIP-sequencing, library preparation and analysis.

ChIP lysates were prepared from RAD21^{mAC} cells treated with EtOH or IAA for 6hrs in two biological replicates. Formaldehyde (1%) was added to the culture medium for 10min at room temperature. Fixation was blocked with 0.125M glycine and cells were washed in cold PBS. Nuclear extracts were prepared by douncing (20 strokes, medium pestle) in swelling buffer (25 mM HEPES pH8, 1.5 mM MgCl2, 10mM KCL, 0.1% NP40, 1 mM DTT and protease inhibitors) and centrifuged for 5min at 2000rpm at 4C. Nuclear pellets were resuspended in Sucrose buffer I (15mM Hepes pH 8, 340 mM Sucrose, 60mM KCL, 2mM EDTA, 0.5 mM EGTA, 0.5% BSA, 0.5 mM DTT and protease inhibitors) and dounced again with 20 strokes. The lysate was carefully laid on top of an equal volume of Sucrose buffer II (15mM Hepes pH 8, 30% Sucrose, 60mM KCL, 2mM EDTA, 0.5 mM EGTA, 0.5 mM DTT and protease inhibitors) and centrifuged for 15min at 4000rpm at 4C. Nuclei were washed twice to remove cytoplasmic proteins, centrifuged and the pellet was resuspended in Sonication/RIPA buffer (50mM Tris, pH 8.0, 140 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% Na-deoxycholate, 0.1% SDS and protease inhibitors) at a concentration of 5 x10^6 nuclei in 130µl buffer. This was transferred to a sonication tube (AFA Fiber Pre-Slit Snap-Cap 6x16mm) and sonicated in a Covaris S2 (settings; 4 cycles of 60 seconds, 10% duty cycle, intensity: 5, 200 cycles per burst). Soluable chromatin was in the range of 200 - 400 bp. Triton X100 was added (final concentration 1%) to the sonicated chromatin and moved to a low-retention tube (Eppendorf) before

centrifugation at 14,000 rpm for 15min at 4C and pellets were discarded. 1/100th of the chromatin lysate was retained as the Input sample.

For Immunoprecipitation, 200ug chromatin aliquots/IP were precleared with a slurry of Protein A/G (50:50) (Dynabeads) an incubated for 4hr at 4C. Meanwhile, washed Protein A/G beads (40ul per IP) were mixed with primary antibodies and incubated for 4hrs at 4C. The following amounts of antibodies were used: anti-CTCF, 5ug/ChIP; anti-SA1, 15ug/ChIP; anti-SA2, 10ug of the mixed antibody pack/ChIP; anti-Smc3, 5ug/ChIP and anti-IgG, 10ug/ChIP. See Table 2 for information about the antibodies. Washed, pre-bound Protein A/G beads+antibody were mixed with pre-cleared chromatin lysates and incubated overnight with rotation at 4C. The next day, the supernatant was removed and the beads were washed 9 times with increasing salt concentrations. Protein-DNA crosslinks were reversed in ChIP elution buffer (1% SDS, 5 mM EDTA, 10 mM Tris HCl pH 8) + 2.5 ul of Proteinse K and incubated for 1 hour at 55°C and overnight at 65°C. Samples were phenol–chloroform extracted, resuspended in TE buffer and assessed by qPCR as a quality control. Libraries were prepared from 5-10ng of purified DNA, depending on availability of material, using NEBNext Ultra II DNA Library Prep Kit for Illumina kit and using NEBNext Multiplex Oligos for Illumina (Index Primers Set 2) according to manufacturer's instructions using 6-8 cycles of PCR. ChIP-seq libraries from one biological set (all ChIP libraries for both EtOH and IAA) were multiplexed and sequenced on the Illumina HiSeq2500 platform, 80bp single-end reads. Each biological set was sequenced on a separate run.

Quality control of reads was preformed using FASTQC. Reads were aligned to the hg19 reference genome using Bowtie with 3 mismatches. PCR duplicates were detected and removed using SAMTOOLS. Bam files were imported into MISHA (v 3.5.6) and peaks were identified using a 0.995 percentile. Peaks that overlapped in both replicates were retained. Only replicate 1 of the SA1 library was used. Correlation plots of peaks across the genome from different ChIP libraries were compared with log-transformed percentiles plotted as a smoothed scatter plot. Comparison of peaks at regions of interest were carried out using deepTools (Version 3.1.0-2). For input into deepTools, peak data was converted to bigwig format, with a bin size of 500, using the UCSC bedGraphtoBigWig package. The signal matrix was calculated for a window 2,000 bp up- and down-stream of the region of interest, missing data was treated as zero, and all other parameters were as default. Heatmaps were generated within deepTools, with parameters as default. Read density profile plots were plotted in ggplot using deepTools profilePlot -perGroup data and smoothed using geom_smooth default 'gam' settings.

DRIP-sequencing.

DRIP lysates were prepared from chromatin. Chromatin was fractionated as described for ChIP samples above, with the following changes. Samples were not fixed and were collected from the plate by scraping in ice-cold PBS. Sonication was performed to

solubilise the chromatin using a picorupter with 10 cycles of 30 sec on, 30 sec off. Following sonication, 60U of RNase I (Ambion) was incubated with each sample for 1.5hrs at 37°C to reduce off-target RNA binding by S9.6. Protein was digested with 184ug of proteinase K incubated with each sample for 2.5hrs at 45°C and 3.5hrs at 55°C. Samples were spun briefly and the supernatant taken. Nucleic acid material was isolated from the sample by phenol-chloroform extraction and isopropanol precipitation. Purified nucleic acid material was IP'd as in [58], including resuspension nuclear lysis buffer and fragmentation by sonication using a picorupter for 4 cycles of 30sec on, 30 sec off. Following elution, reverse crosslinking, and proteinase K treatment the immunoprecipitated nucleic acid was purified by phenol-chloroform extraction and isopropanol precipitation. Second strand synthesis was carried out according to [59] with minor changes. Namely, the reaction was set up in a PCR tube as: ~28ul eluted DRIP sample, ~27ul nuclease-free water (depending on DRIP sample, for a total of 75ul), 15ul 5X ss buffer, 2ul 10mM dNTP, 0.5ul DNA ligase (E. coli, NEB cat. M0205S), 2ul DNA polymerase I (E.coli, NEB cat. M0209S), 0.5ul RNase H (2.5 units). This was incubated for 2hrs at 16°C. dsDNA was purified using a 1.8X ratio of SPRI beads and eluted in ultrapure water. Libraries were prepared using NEB Ultra II DNA library prep kit according to manufacturer's instructions and sequenced on the Illumina NovaSeq platform, 100bp paired-end reads (30M per sample). Reads were aligned and processed as for the ChIP-seq samples above with two changes; i) reads were quality trimmed using trimfq -b 14 -e 16, and ii) a maximum insert size of 1000 was set for bowtie. DeepTools analysis was carried out as for ChIP-seq samples and with binSize set to 2bp.

ChromHMM.

ChIP-seq data for YY1, CBX3, SIN3A, POLR2A, POLR2AphosphoS5, H3K27ac, H3K4me3, H3K4me1, H3K27me3, EZH2, and H3K9me3 from HCT116 cells were obtained from ENCODE and processed as above. ChIP-seq, DRIP-seq, and ENCODE ChIP-seq. BAM files were binarized in ChromHMM using a bin size of 200 bp and a shift of 150 bp. Where input files were available on ENCODE they were used in ChromHMM to determine the binarization threshold, otherwise the ChromHMM default of a uniform background was assumed. The chromatin state model was generated for 15 states and compared to the hg19 genome assembly. All other parameters were as default.

Hi-C data and contact hotspots analysis.

Generating hotspots - Previously published Hi-C datasets derived from RAD21^{mAC} cells treated with EtOH or IAA [8] were analyzed as previously described [60]. Custom R scripts were written to identify Hi-C hotspots, i.e. regions of Hi-C maps with high contact frequency. To begin, for each chromosome, all contacts were extracted and subsetted for only high scoring (>=60) contacts between a band of 10e3 – 70e6. Using KNN, for each high scoring contact, the 250 nearest neighbour contacts

were identified and subset for only the high-scoring neighbours. This created a list of high scoring neighbours for each high scoring contact, where the first neighbour is the contact itself with a distance of 0. This allowed the neighbour information to be converted into edge information, thereby allowing high score fend contacts to be grouped into cluster hotspots using the R package 'igraph'. Hotspots that contained less than the minimum number of high scoring fends (<100) were removed. The output list of hotspots were represented as 2D intervals which contained high scoring contacts. In total, 5539 hotspots were identified in EtOH and 759 in IAA Hi-C data.

Creating aggregate plots - To calculate and visualise the contact enrichment at hotspots in the EtOH and IAA Hi-C, we used the R package 'shaman'. Firstly, we used the function 'shaman_generate_feature_grid' to calculate the enrichment profile at EtOH and IAA hotspots. Using the weighted centre for each hotspot, represented as a 2D interval we used the function to build grids for the EtOH and IAA hotspots in the HiC data at 3 specific bands, 100k – 1MB, 1MB – 5MB, 5MB – 10MB. A range of 250kb was visualised around the weighted centre. The grid was built by taking all combinations interval1 and interval2 of the EtOH and IAA hotspot centres, with each combination termed a 'window'. Hotspots were not filtered for size or shape. A score threshold of 60 was used to focus on enriched pairs, those windows that did not contain at least one point with a score of 60 were discarded. Each window was then split into 1000nt bins and the windows were summed together to generate a grid containing the observed and expected contacts. We visualised the grid using 'shaman_plot_feature_grid' using 'enrichment' mode and a plot_resolution value of 6000, due to the large range being visualised.

STORM – Immunolabelling and imaging.

Two clones of RAD21^{mAC}-OsTIR cells were seeded at a density of 30,000 cells per well per 400ul) onto poly-L-lysine coated 8-well chamber slides (Lab-Tek™ 155411) overnight. Each clone was treated with EtOH, IAA or IAA washoff and then fixed with PFA 4% (Alfa Aesar) for 10 min at room temperature and rinsed with PBS three times for 5 min each. The cells were shipped to the Cosma Lab after fixation for STORM processing and imaging. Cells were permeabilized with 0.3% Triton X-100 in PBS and blocked in blocking buffer (10% BSA – 0.01 % Triton X-100 in PBS) for one hour at room temperature. Cells were incubated with primary antibodies (see Table 2) in blocking buffer at 1:50 dilution. For combined S9.6/STAG1 and S9.6/RAD21-GFP imaging, cells were incubated with primary antibodies in blocking buffer (dilutions 1:100 for S9.6 and STAG1, 1:250 for GFP). Cells were washed three times for 5 min each with wash buffer (2% BSA – 0.01 % Triton X-100 in PBS) and incubated in secondary antibody. For STORM imaging, home-made (Bates et al., 2007) dye pair labeled secondary antibodies were added at a 1:50 dilution in blocking buffer and were incubated for 45 min at room temperature or single fluorophore labeled commercial antibodies were added at a 1:250 dilution in blocking buffer and were incubated for 45

min at room temperature (see Table 2). Cells were washed three times for 5 min each with wash buffer.

STORM imaging was performed on an N-STORM 4.0 microscope (Nikon) equipped with a CFI HP Apochromat TIRF 100x 1.49 oil objective and an iXon Ultra 897 camera (Andor) and using Highly Inclined and Laminated Optical sheet illumination (HILO). Dual color STORM imaging was performed with a double activator and single reporter strategy by combining AF405_AF647 anti-Goat secondary with Cy3_AF647 anti-Rabbit secondary antibodies. Sequential imaging acquisition was performed (1 frame of 405 nm activation followed by 3 frames of 647 nm reporter and 1 frame of 560 nm activation followed by 3 frames of 647 nm reporter) with 10 ms exposure time for 120000 frames. 647 nm laser was used at constant ~2 kW/cm^2 power density and 405 nm and 560 nm laser powers were gradually increased over the imaging. For S9.6 experiments, before STORM imaging, conventional images were taken for s9.6 signal (AF568 labeled) with a TRITC filter, for endogenous mClover signal with a FITC filter and for STAG1 or GFP (AF647 labeled) with a Quadband filter. STORM imaging for STAG1 or GFP was performed with continuous imaging acquisition (i.e. simultaneous stimulation with 405 and 647 nm lasers) with 10 ms exposure time for 60000 frames. 647 nm laser was used at constant ~2 kW/cm^2 power density and 405 nm was gradually increased over the imaging. Imaging buffer composition for STORM imaging was 100 mM Cysteamine MEA (Sigma-Aldrich, #30070) - 5% Glucose (Sigma-Aldrich, #G8270) – 1% Glox Solution (0.5 mg/ml glucose oxidase, 40 mg/ml catalase (Sigma-Aldrich, #G2133 and #C100)) in PBS.

STORM imaging analysis and quantifications.

STORM images were analyzed and rendered with Insight3 software (kind gift of Bo Huang, UCSF, Huang et al., 2008) as previously described (Bates et al., 2007; Rust et al., 2006). Localizations were identified based on a threshold and fit to a simple Gaussian to determine the x and y positions. Cluster analysis of CTCF, SA1 and SA2 STORM signal was performed as previously described (Ricci et al., 2015) to obtain cluster size and positions and to measure Nearest Neighbour distributions (NND) between clusters of the same protein in individual nuclei. NND between clusters' centroids of two different proteins (i.e. CTCF-SA1 and CTCF-SA2) was calculated by knnsearch.m Matlab function and the NND histogram of experimental data was obtained by considering all the NNDs of individual nuclei (histogram bin, from 0 to 500 nm, 5 nm steps). Simulated NNDs recapitulating random spatial distribution of cluster centroids were first obtained for each nucleus separately and then merged to calculate the simulated NND histogram (histogram bin, from 0 to 500 nm, 5 nm steps). The difference plot reports the difference between experimental NND and simulated NND. Quantification and analysis of STORM images was performed in Matlab and statistical analysis was performed in Graphpad Prism (v7.0e). The type of statistical test is specified in each case. Statistical significance is represented as indicated above.

Analysis for S9.6 experiments was performed in the following way. After generating localization lists for each STORM image, nuclear masks and S9.6 were generated to segment the obtained localizations belonging to nuclear areas inside or outside s9.6 enriched regions. Masks generation, quantification of masks' areas and segmentation of STORM localizations were performed in Fiji/ImageJ. Nuclear masks were manually designed based on STAG1/GFP signal. S9.6 masks were generated by applying an automatic threshold on s9.6 images based on nuclear intensity signal. Masks were visually inspected individually and adjusted manually in cases where dim signal or noise from cytosolic signal compromised the identification of the mask. The area of all masks was calculated. Masks were applied to the STORM localization lists to generate segmented lists with the localizations belonging to the entire nucleus and to the s9.6 enriched areas. Finally, the density of localizations (number of localizations/area of the mask) for the areas inside and outside s9.6 masks was calculated and the ratio between both values was plotted.

Graphpad Prism software used for statistical analysis can be found at: https://www.graphpad.com/scientific-software/prism/ MatLab software used for imaging data analysis can be found at: https://www.mathworks.com/products/matlab.html

Mass spectrometry sample preparation and running.

SA1 immunoprecipitation samples were analysed by liquid chromatography–tandem mass spectrometry (LC-MS/MS). Three biological replicate experiments were carried out for MS and each included four samples, untreated (UT), treated with IAA for 4hrs, siCon, or siSA1, generated as described above. Cells were fractionated to purify chromatin-bound proteins as above and immunoprecipitated with IgG- or SA1-bead conjugates. To maximise IP material for the MS, the antibody amount was increased to 15ug and the chromatin amount was increased to 2mg.

The IP eluates were loaded into a pre-cast SDS-PAGE gel (4–20% Mini-PROTEAN® TGX™ Precast Protein Gel, 10-well, 50 μL) and proteins were run approximately 1 cm to prevent protein separation. Protein bands were excised and diced, and proteins were reduced with 5 mM TCEP in 50 mM triethylammonium bicarbonate (TEAB) at 37°C for 20 min, alkylated with 10 mM 2-chloroacetamide in 50 mM TEAB at ambient temperature for 20 min in the dark. Proteins were then digested with 150ng trypsin, at 37°C for 4 h followed by a second trypsin addition for 4 h, then overnight at room temperature. After digestion, peptides were extracted with acetonitrile and 50 mM TEAB washes. Samples were evaporated to dryness at 30°C and resolubilised in 0.1% formic acid.

nLC-MS/MS was performed on a Q Exactive Orbitrap Plus interfaced to a NANOSPRAY FLEX ion source and coupled to an Easy-nLC 1200 (Thermo Scientific). 25% (first, second and fourth biological replicate) or 50% (third biological replicate) of each sample was loaded as 5 or 10 μL injections. Peptides were separated on a 27cm fused silica emitter, 75 µm diameter, packed in-house with Reprosil-Pur 200 C18-AQ,

2.4 µm resin (Dr. Maisch) using a linear gradient from 5% to 30% acetonitrile/ 0.1% formic acid over 60 min, at a flow rate of 250 nL/min. Peptides were ionised by electrospray ionisation using 1.8 kV applied immediately prior to the analytical column via a microtee built into the nanospray source with the ion transfer tube heated to 320°C and the S-lens set to 60%. Precursor ions were measured in a data-dependent mode in the orbitrap analyser at a resolution of 70,000 and a target value of 3e6 ions. The ten most intense ions from each MS1 scan were isolated, fragmented in the HCD cell, and measured in the orbitrap at a resolution of 17,500.

Mass spectrometry analysis

Raw data was analysed with MaxQuant [61] version 1.5.5.1 where they were searched against the human UniProtKB database using default settings (http://www.uniprot.org/). Carbamidomethylation of cysteines was set as fixed modification, and oxidation of methionines and acetylation at protein N-termini were set as variable modifications. Enzyme specificity was set to trypsin with maximally 2 missed cleavages allowed. To ensure high confidence identifications, PSMs, peptides, and proteins were filtered at a less than 1% false discovery rate (FDR). Label-free quantification in MaxQuant was used with LFQ minimum ratio count set to 2 with 'FastLFQ' (LFQ minimum number of neighbours = 3, and LFQ average number of neighbours = 6) and 'Skip normalisation' selected. In Advanced identifications, 'Second peptides' was selected and the 'match between runs' feature was not selected. Statistical protein quantification analysis was done in MSstats [62](version 3.14.0) run through RStudio. Contaminants and reverse sequences were removed and data was log2 transformed. To find differential abundant proteins across conditions, paired significance analysis consisting of fitting a statistical model and performing model-based comparison of conditions. The group comparison function was employed to test for differential abundance between conditions. Unadjusted p-values were used to rank the testing results and to define regulated proteins between groups.

Proteins with peptides discovered in the IgG samples were disregarded from downstream analyses. Significantly depleted/enriched proteins were considered with an absolute log2foldchange > 0.58 (1.5-fold change) and a p-value < 0.1. SA1 interactome analysis was performed in STRING. The network was generated as a full STRING network with a minimum interaction score of 0.7 required. Over-enrichment of GO biological process and molecular function terms was calculated with the human genome as background. Network analysis of the SA1 interactome in IAA-treated samples was generated from the significantly depleted/enriched proteins, with a minimum interaction score of 0.4 required. Two conditions for functional enrichments were considered; i) enrichment was calculated with the human genome as background to determine the full SA1 interactome in the absence of cohesin, and ii) enrichment was calculated with the untreated SA1 interactome as background, to determine the statistical effect of cohesin loss of the SA1 interactome itself. The network developed

in i) was manually rearranged in Cytoscape for visual clarity, enriched categories were visualized using the STRING pie chart function and half of the proteins within each category were subset from the network based on pvalue change between UTR and IAA samples.

Over-enrichment of the s9.6 interactome was calculated separately using the hypergeometric distribution for comparison with [44,45]. Significance was calculated using the dhyper function in R and multiple testing was corrected for using the p.adjust Benjamini & Hochberg method. To compare with a minimal background protein list, http://www.humanproteomemap.org was analysed on the Expression Atlas database to determine a list of proteins expressed in one or more of three tissue types corresponding to the cell types used across the different studies.

SLiMSearch analysis

The SLiMSearch tool http://slim.icr.ac.uk/slimsearch/, with default parameters was used to search the human proteome for additional proteins that contained the FGF-like motif determined in [18] to predict binding to SA proteins. The motif was input as [PFCAVIYL][FY][GDEN]F.{0,1}[DANE].{0,1}[DE]. Along with CTCF, four proteins found to contain the FGF-like motif, CHD6, MCM3, HNRNPUL2 and ESYT2 were validated for interaction with SA.

CLIP

Crosslinking immunoprecipitation (CLIP) was performed as previously described[63]. Briefly, mESC or HCT116 cells were irradiated with 0.2 J/cm2 of 254 nm UV light in a Stratalinker 2400 (Stratagene). Cells were lysed in 1 ml of lysis buffer with Complete protease inhibitor (Roche). Lysates were passed through a 27 G needle, 1.6 U DNase Turbo (Thermofisher) per 10^6 cells and 0.8 (low) or 8 U (high) U RNase I (Ambion) per 10^6 cells added, and incubated in a thermomixer at 37°C and 1100 rpm for 3 minutes. Lysates were then cleared by centrifugation and using Proteus clarification spin column, according to the manufacturer's instructions. Endogenous SA1 and SA2 were immunoprecipitated with 10 μg SA1 and SA2 antibodies or non-specific IgG control (Sigma) conjugated to protein G dynabeads (Dynal) for 4 hrs at 4°C. Tagged SA2 proteins were immunoprecipitated from HCT116 cells 40 hours after transfection with 30 μ l GFP-Trap beads. IPs were washed three times with high salt buffer (containing 1M NaCl and 1M urea) and once with PNK buffer and RNA labelled with 8 μl radioactive ^{32}P-gamma-ATP (Hartmann Analytic) for 5 mins at 37°C. RNPs were eluted in LDS loading buffer (Invitrogen) and resolved on a 4-12% gradient NuPAGE Bis-Tris gel (Invitrogen) and transferred onto 0.2 μm diameter pore nitrocellulose membrane. After blocking with PBST+milk, membranes were washed and exposed overnight to phosphorimager screen (Fuji) and RNA-^{32}P visualized using a Typhoon phosphorimager (GE) and ImageQuant TL (GE). Membranes were then immunoblotted for SA1, SA2, and RAD21 and visualized using an ImageQuantLAS 4000 imager (GE). See Table 2 for details on antibodies.

GFP-TRAP + Cloning of STAG2 isoforms and YFP constructs.

SA2 cDNAs were cloned directly from HCT116 cells by PCR using KAPA HiFi HotStart PCR kit (Roche) (Fwd: ATGATAGCAGCTCCAGAAAACCAACTG; Rev: TTAAAACATTGACACTCCAAGAACTGATTCATCC). Two major isoforms were detected, $SA2^{\Delta ex32}$ where exon32 has been spliced out and $SA2^{+ex32}$ where exon 32 has been spliced in. Both SA2 cDNAs were cloned into pENTR/D vector (Invitrogen) and then into an N-terminal YFP-tagged Gateway cloning vector (a kind gift from Endre Kiss-Toth, University of Sheffield). Sequences were confirmed by restriction enzyme digestion and Sanger sequencing. Recombinant YFP-$SA2^{\Delta ex32}$ or YFP-$SA2^{+ex32}$ were transfected into adherent HCT116 cells for 40 hours before being harvested. Cells were lysed and fractionated as indicated for CLIP. One third of the whole cell lysate was pre-cleared with a 50:50 mixture of protein A/G magnetic beads and GFP-Trap (Chromotek, gtd-20) was pre-blocked with 1mg/mL ultra-pure BSA (AM2616, Invitrogen) for 2h at 4°C. After blocking, GFP-Trap was washed twice with CLIP lysis buffer and added to pre-cleared lysates to immunoprecipitate proteins for 1h at 4°C. Samples were washed in high salt buffer (50mM Tris-HCL pH 7.4, 1M NaCl, 1mM EDTA, 1% NP-40, 0.1% SDS, 0.5% sodium deoxycholate, 1M Urea) and low salt PNK buffer (20mM Tris-HCL pH 7.4, 10mM $MgCl_2$, 0.2% Tween-20), and eluted in 2x Laemmli buffer (Bio-Rad). Proteins were separated by SDS-PAGE on a 4-20% gradient mini-PROTEAN® Precast Gel (Bio-Rad) and transferred onto PVDF membrane for visualization.

VAST-TOOLS

VAST-TOOLS was used to generate Percent Spliced In (PSI) scores, a statistic which represents how often a particular exon is spliced into a transcript using the ratio between reads which include and exclude said exon. Paired-end RNA-seq datasets were submitted to VAST-TOOLS (v2.1.3) using the Mmu genome (Tapial J et al, Gen Res 2017). Briefly, reads are split into 50nt words with a 25nt sliding window. The 50nt words are aligned to a reference genome using Bowtie to obtain unmapped reads. These unmapped reads are then aligned to a set of predefined exon-exon junction (EJJ) libraries allowing for the quantification of alternative exon events. The output was further interrogated using a script which searches all hypothetical EEJ combinations between potential donors and acceptors within Stag1. PSI scores could be obtained providing there was at least a single read within the RNAseq data that supported the event, although we only considered events supported by a minimum of 50 reads. Calculated PSI values for each alternatively spliced exon as well as the average PSI reported in the text are shown below. See Supp Fig 5 for names of published datasets used in this analysis.

1453 **Table 1. siRNAs used in this study.**

siRNA name	Company	Target	Catalogue no.	custom siRNA sequence
si scramble control	Dharmacon	Smartpool	D-001810-10-05	
siSA1	Dharmacon	Smartpool	L-010638-01-0010	
siSA2	Dharmacon	Smartpool	L-021351-00-0010	
siNIPBL	Dharmacon	Smartpool	L-012980-00-0010	
siAQR	Dharmacon	Smartpool	L-022214-01-0005	
esi control	Sigma	Luciferase	EHUFLUC	
esi SA1	Sigma	exon 31	custom esiRNA	TCCTCAGATGCAGATCTCTTGGTTAGGCC AGCCGAAGTTAGAAGACTTAAATCGGAAG GACAGAACAGGAATGAACTACATGAAAGTG AGAACTGGAGTGAGGCATGCTGT
esi SA2	Sigma	exon 32	custom esiRNA	CACGCAGGTAACATGGATGTTAGCTCAAAG ACAACAAGAGGAAGCAAGGCAACAGCAGG AGAGAGCAGCAATGAGCTATGTTAAACTG CGAACTAATCTTCAGCATGCCAT

1454
1455
1456 **Table 2. Antibodies used in this study.**

Protein	Company	Catalogue No.	Species	Figure Reference
SA1	Abcam	ab4455	mouse	1a, c, d, e, 2a, b, d, 3a, b, e, 4g, 5b, f, h, i
SA1	Abcam	ab4457	mouse	1i
SA2	Bethyl	A300-159	goat	1b, c, d, e, 2a, 3a, b, f, 4g, e, 5c, g
SA2	Bethyl, AbVantage Pack	A310-941A	goat	1i
CTCF	Diagenode	C15410210	rabbit	1c, d, i, 2a
CTCF	Cell signalling	2899s	rabbit	1a, e
RAD21	Abcam	ab992	rabbit	1c, d, i, 2a, d, 4e, g, j
GFP-TRAP	ChromoTek	gtd-20		1i, 3g
GFP	Invitrogen	A11122	rabbit	1a, e
mAID	MBL	M214-3	mouse	S1a
OsTIR	MBL	PD048	rabbit	S1a
SMC3	Abcam	ab9263	rabbit	1i
CHD6	Bethyl	A301-221A	rabbit	2a
MCM3	Bethyl	A300-124A	goat	2a, 5b
HNRNPUL2	Abcam	ab195338	rabbit	2a
YTHDC1	Abcam	ab122340	rabbit	2d
FTSJ3	Bethyl	A304-199A-M	rabbit	2d
FANCI	Bethyl	A301-254A-M	rabbit	2d
TAF15	Abcam	ab134916	rabbit	2d
DHX9	Abcam	Ab26271	rabbit	2d, 5b
SSRP1	Abcam	ab26212	mouse	2d
INO80	Proteintech	18810-1-AP	rabbit	2d
ESYT2	Sigma-Aldrich	HPA002132	rabbit	2d
S9.6	Kerafast	ENH001	mouse	5b-j
RNASE H2	Novus	NBP1-76981	rabbit	5b
AQR	Bethyl	A302-547A	rabbit	5b, e
POLR2	Covance	MMS-1289	mouse	5b
MAU2	Abcam	ab183033	rabbit	4e, g, f
NIPBL	Abbiotec	250133	rat	4e, g, f, S5
H3	Abcam	ab1791	rabbit	3e, f, 4e, g, f

Name (SecondaryAbs)	Fluorophore	Company	Catalogue No.	Figure Reference
Donkey anti-Rabbit	Cy3_AF647	Home made from Jackson Immunoresearch IgG	Home made from 711-005-152	1e, S1
Donkey anti-Goat	AF405_AF647	Home made from Jackson Immunoresearch IgG	Home made from 705-005-147	1e, S1
Donkey anti-mouse	AF647	Invitrogen	A31570	1a, d, e
Donkey anti-rabbit	AF488	Invitrogen	A21206	1a, d, e
Donkey anti-rabbit	AF647	Invitrogen	A31573	1a, d, e
Donkey anti-goat	AF555	Invitrogen	A21432	1a, d, e
Donkey anti-goat	AF647	Invitrogen	A21447	1a, d, e
Goat anti-Mouse	AF568	ThermoFisher Scientific	A-11031	
Goat anti-Rabbit	AF647	ThermoFisher Scientific	A-21244	
Rabbit anti- Goat	AF647	ThermoFisher Scientific	A-21446	

Table 3. Published datasets used in this study.

Accession no.	Analysis description	Publication DOI or Ref	Figure Reference
GSE104334	Long-range contact analysis of Hi-C datasets	10.1016/j.cell.2017.09.026	1i
GSE89729	Percent Spliced In (PSI) analysis of RNA-seq datasets	10.1172/jci.insight.91419	4d, “HCT Zuo”
GSM958749	Percent Spliced In (PSI) analysis of RNA-seq datasets	ENCODE HCT116 RNAseq	4d, “HCT ENCODE”
GSM958735	Percent Spliced In (PSI) analysis of RNA-seq datasets	ENCODE HeLa RNAseq	4d, “HeLa”

Table 4. Published ChIP-seq datasets used for ChromHMM.

Protein	Accession	Publication	Matched input
NIPBL (EtOH- and IAA-treated)	GSE104334	(Rao *et al.*, 2017)	-
CBX1	GSM1010758	(Gertz *et al.*, 2013)	
EZH2	GSM3498250	(Dunham *et al.*,	GSM2308475;GSM2308476
POLR2A	GSM935426	(Dunham *et al.*,	GSM2308422
POLR2AphosphoS5	GSM803474	(Gertz *et al.*, 2013)	GSM803475
SIN3A	GSM1010905	(Gertz *et al.*, 2013)	
YY1	GSM803354	(Gertz *et al.*, 2013)	GSM803475
H3K4me1	GSM945858		GSM2308475; GSM2308476
H3K4me1	GSM2527549	(Dunham *et al.*,	GSM2308422
H3K4me3	GSM2533929	(Dunham *et al.*,	GSM2308475; GSM2308476
H3K4me3	GSM945304	(Thurman *et al.*,	GSM945287
H3K9me3	GSM2527565	(Dunham *et al.*,	

H3K9me3	GSM2308431	(Dunham *et al.*,	
H3K27ac	GSM2534277	(Dunham *et al.*,	GSM2308422
H3K27me3	GSM2308612	(Dunham *et al.*,	

1462

Figure 1.

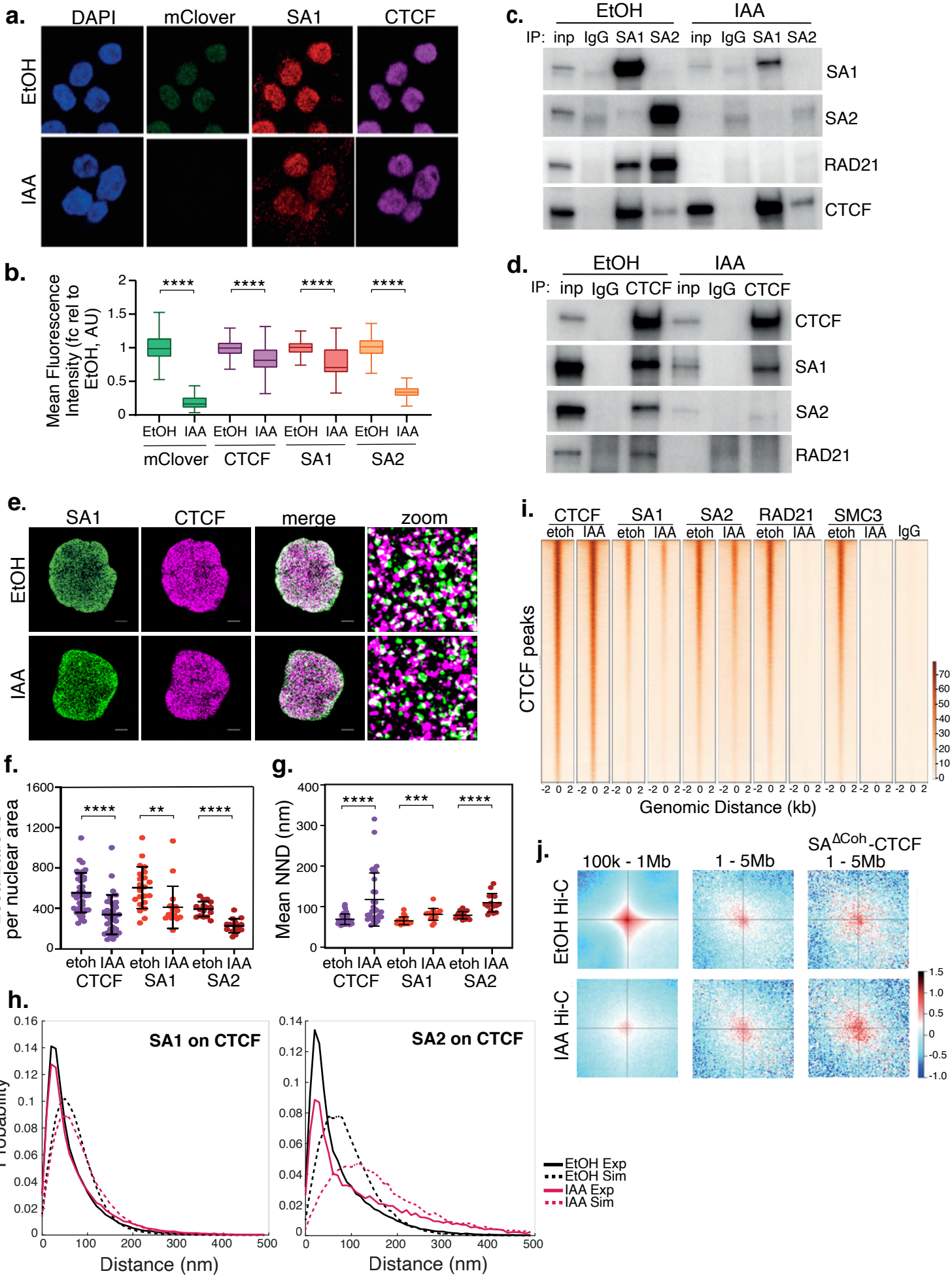

Figure 2.

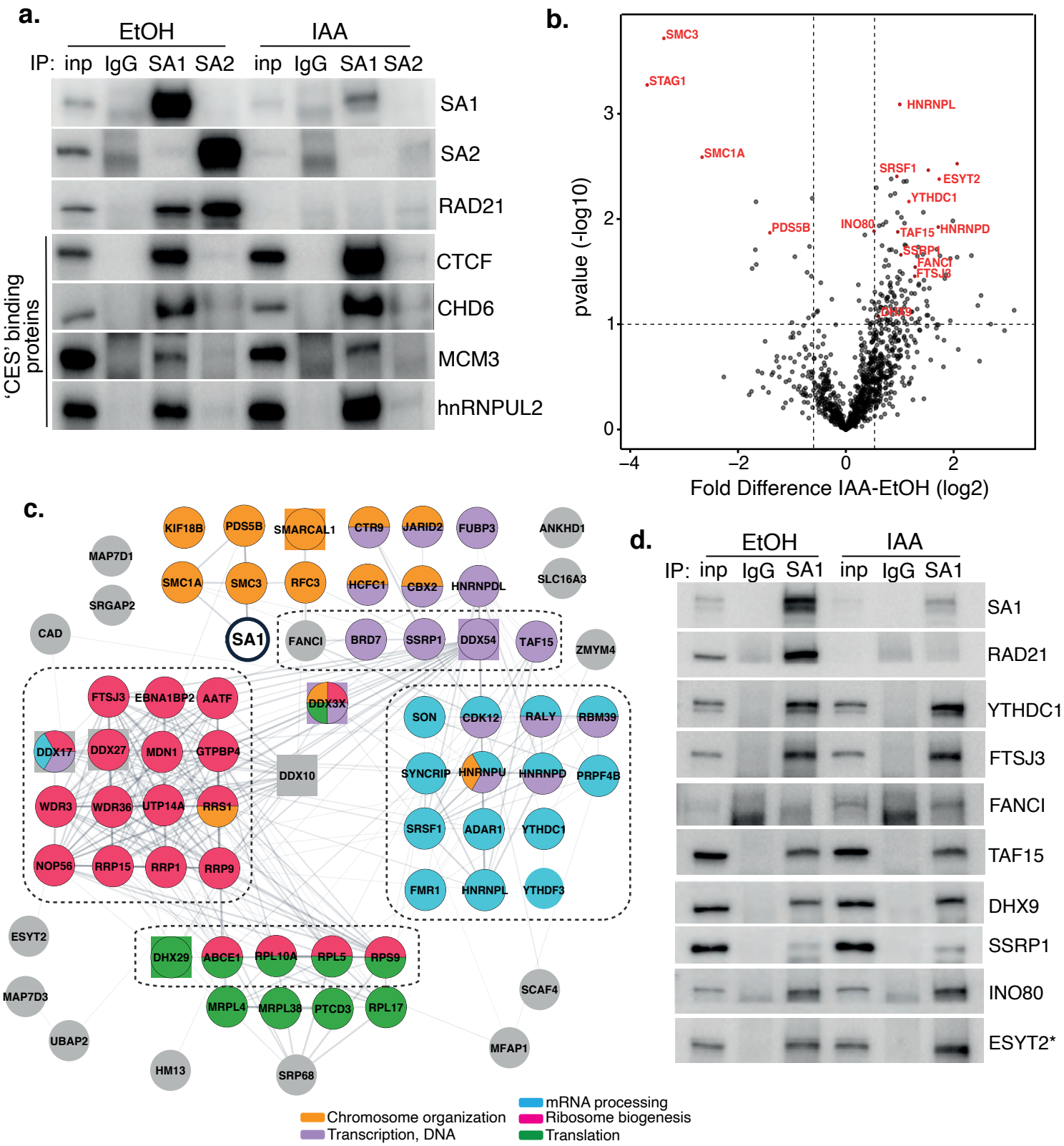

Figure 3.

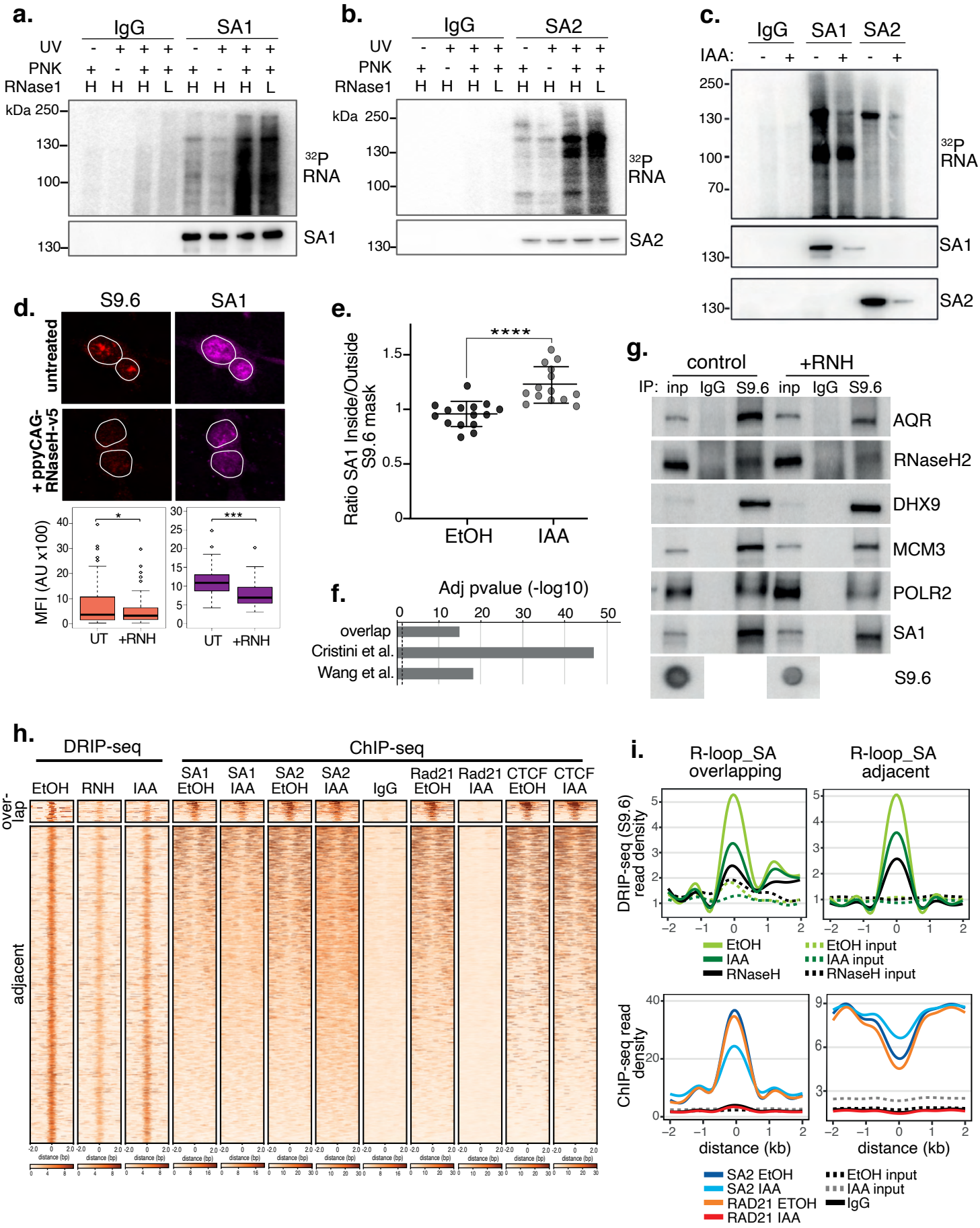

Figure 4

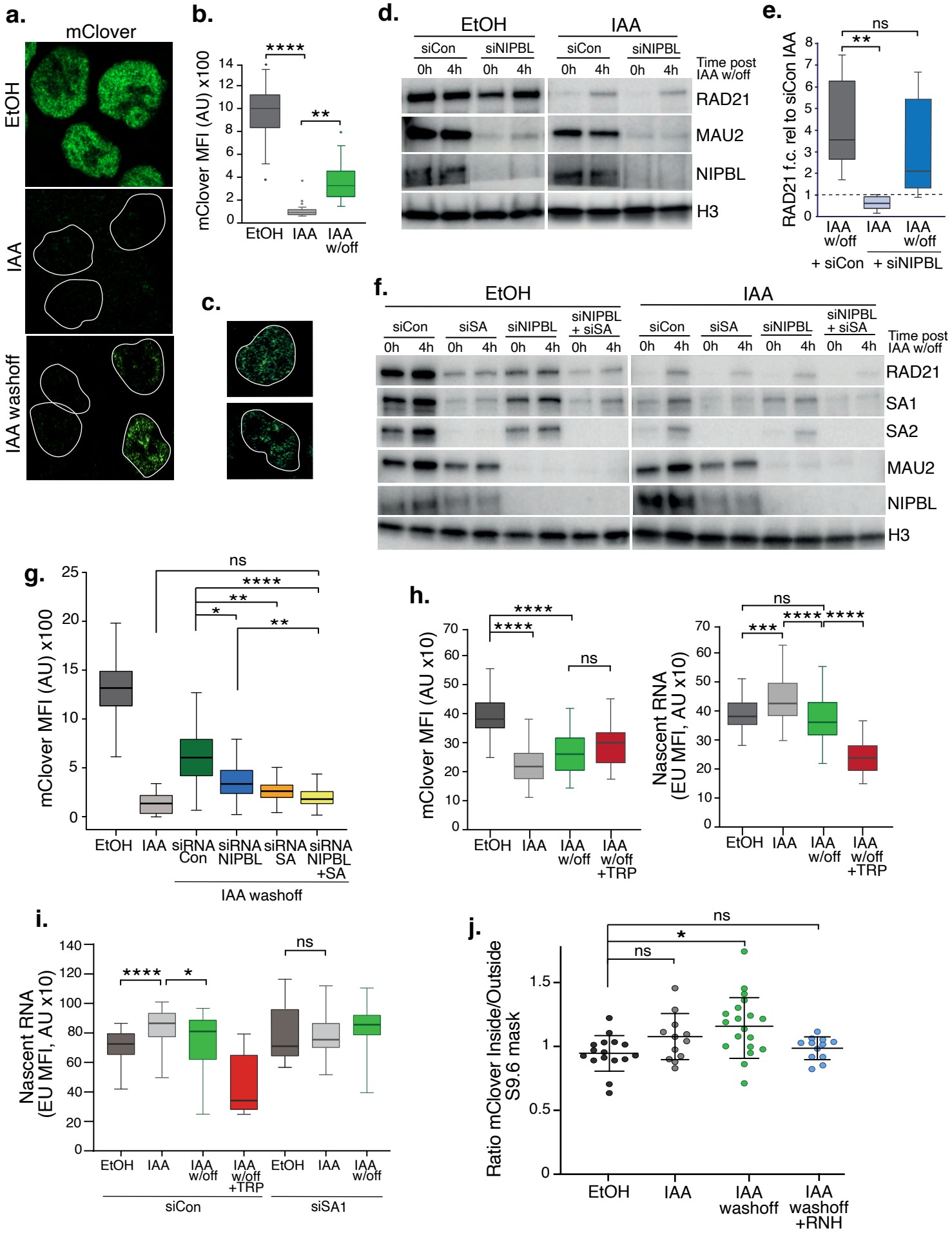

Figure 5.