STAG proteins promote cohesin ring loading at R-loops.

Hayley Porter[1+], Yang Li[1+], Wazeer Varsally[1], Maria Victoria Neguembor[2], Manuel Beltran[3], Dubravka Pezic[1], Laura Martin[2], Manuel Tavares Cornejo[3], 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

1 **ABSTRACT**

2

3 Most studies of cohesin function consider the Stromalin Antigen (STAG/SA) proteins

4 as core complex members given their ubiquitous interaction with the cohesin ring.

5 Here, we provide functional data to support the notion that the SA subunit is not a

6 mere passenger in this structure, but instead plays a key role in cohesins localization

7 to diverse biological processes and promotes loading of the complex at these sites.

8 We show that in cells acutely depleted for RAD21, SA proteins remain bound to

9 chromatin and interact with CTCF, as well as a wide range of RNA binding proteins

10 involved in multiple RNA processing mechanisms. Accordingly, SA proteins interact

11 with RNA and are localised to endogenous R-loops where they act to suppress R-loop

12 formation. Our results place SA proteins on chromatin upstream of the cohesin

13 complex and reveal a role for SA in cohesin loading at R-loops which is independent

14 of NIPBL, the canonical cohesin loader. We propose that SA takes advantage of this

15 structural R-loop platform to link cohesin loading and chromatin structure with diverse

16 genome functions. Since SA proteins are pan-cancer targets, and R-loops play an

17 increasingly prevalent role in cancer biology, our results have important implications

18 for the mechanistic understanding of SA proteins in cancer and disease.

19

20
21

22

23 **KEY WORDS**
24 Cohesin, STAG proteins, R-loops, genome.

INTRODUCTION

25

26 Cohesin complexes are master regulators of chromosome structure in interphase and

27 mitosis. Accordingly, mutations of cohesin subunits leads to changes in cellular

28 identity, both during development and aberrantly in cancer [1-3]. A prevailing model is

29 that cohesin contributes to cell identity changes in large part by dynamically regulating

30 genome organization and mediating communication between distal regulatory

31 elements [4-10]. Our understanding of how cohesin's component parts contribute to its

32 functions and where cohesin becomes associated to chromatin in order to perform its

33 critical roles in spatial genome organization is incomplete.

34 Most studies of cohesin function consider the Stromalin Antigen (STAG/SA)

35 proteins as core complex members given their ubiquitous interaction with the tripartite

36 cohesin ring (composed of SMC1, SMC3 and SCC1/RAD21). Rarely is the SA subunit

37 considered for its roles outside of cohesin, even though it is the subunit most

38 commonly mutated across a wide spectrum of cancers [1,11,12]. SA proteins play a role

39 in cohesin's association with DNA [13,14]. The yeast SA orthologue is critical for efficient

40 association of cohesin with DNA and its ATPase activation [13,14]. Recent crystallization

41 studies of cohesin in complex with its canonical loader NIPBL[15], suggest that NIPBL

42 and SA together wrap around both the cohesin ring and DNA to position and entrap

43 DNA [16-18], implying a role for SA in the initial recruitment of cohesin to DNA alongside

44 NIPBL. Further, SA proteins bridge the interaction between cohesin and CTCF [16,19,20],

45 and also bridge interactions with specific nucleic acid structures *in vitro*. SA1 binds to

46 AT-rich telomeric sequences [21,22] and SA2 displays sequence-independent affinity for

47 particular DNA structures commonly found at sites of repair, recombination, and

48 replication [23]. Consistent with this, results in yeast implicate non-canonical DNA

49 structures in cohesin loading in S-phase. *In vitro* experiments show that cohesin

50 captures the second strand of DNA via a single-strand intermediate [24], and chromatid

51 cohesion is impaired by de-stabilisation of single-strand DNA intermediates during

52 replication [25]. Together, this implicates SA proteins in playing a regulatory role in

53 guiding or stabilising cohesin localisations.

54 During transcription, the elongating nascent RNA can hybridise to the template

55 strand of the upstream DNA and form an R-loop, which is an intermediate RNA:DNA

56 hybrid conformation with a displaced single strand of DNA (Richardson, 1975; Roy

57   and Lieber, 2009; El Hage *et al.*, 2010; Roy *et al.*, 2010). A multitude of processes

58   have been linked to R-loop stability and metabolism. For example, proper co-

59   transcriptional RNA processing, splicing, and messenger ribonucleoprotein (mRNP)

60   assembly counteract R-loop formation (Li and Manley, 2005; Teloni *et al.*, 2019). R-

61   loop structures have also been shown to regulate transcription of both mRNA and

62   rRNAs by recruitment of transcription factors, displacement of nucleosomes, and

63   preservation of open chromatin (Dunn and Griffith, 1980; Powell *et al.*, 2013; Boque-

64   Sastre *et al.*, 2015). Hence, like at the replication fork, sites of active transcription

65   accumulate non-canonical nucleic acid structures.

66   We set out to investigate the nature of the association between SA proteins and

67   CTCF. We discovered that far from being 'passengers' in the cohesin complex, SA

68   proteins perform critical roles in their own right, directing cohesin's localization and

69   loading to chromatin. In cells acutely depleted of RAD21, SA proteins remain

70   associated with chromatin and CTCF where they are enriched at 3D clustered sites of

71   active chromatin. Moreover, we identify cohesin-independent binding of SA1 to

72   numerous proteins involved in RNA processing, ribosome biogenesis, and translation.

73   Consistent with this, SA1 and SA2 interact with RNA and non-canonical nucleic acid

74   structures in the form of R-loops where SA1 acts to suppress R-loop formation.

75   Importantly, SA proteins are required for loading of cohesin to chromatin in cells

76   deficient for NIPBL, and loading is enhanced by modulating the levels of R-loop

77   structures. Our results highlight a central role for SA proteins in cohesin biology.

78   Through their diverse interactions with proteins, RNA and DNA, SA proteins act as the

79   'seed' for cohesin loading to chromatin. Finally, the interaction of cohesin-independent

80   SA proteins with nucleolar and RNA processing factors, opens up a new

81   understanding of how cohesin mis-regulation can impact disease development that

82   moves us beyond its control of gene expression regulation.

83   **RESULTS**

84

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

86   To determine how CTCF and cohesin assemble on chromatin, we used human

87   HCT116 cells engineered to carry a miniAID tag (mAID) fused to monomeric Clover

88   (mClover) at the endogenous Rad21 locus and *OsTIR1* under the control of CMV

89   (herein RAD21^{mAC}) [26]. RAD21^{mAC} cells were cultured in control conditions (ethanol) or

90   in the presence of auxin (IAA) to induce rapid RAD21 degradation. We used

91   immunofluorescence (IF) to monitor the levels of mClover, SA1, SA2 and CTCF

92   (Fig.1a, b, S1a). While acute IAA treatment robustly reduced mClover levels by over

93   83% compared to control cells, SA2 levels were reduced by 63% (p=4.7E-76), and

94   SA1 was only reduced by 29% (p=7.9E-12) (Fig.1b). We also observed a variable

95   effect on CTCF in the absence of RAD21 (reduced mean value between 7-24%). The

96   retention of SA proteins on chromatin despite the degradation of RAD21 was

97   surprising given the fact that they are considered to be part of a stable biochemical

98   complex.

99   We sought to validate these observations using an orthogonal technique and

100   to establish whether the residual SA proteins retained the capacity to directly interact

101   with CTCF. We prepared chromatin extracts from RAD21^{mAC} cells treated with ethanol

102   or IAA and established a chromatin co-immunoprecipitation (coIP) protocol to probe

103   the interactions between SA proteins, cohesin core subunits and CTCF. Both SA1

104   and SA2 interacted with RAD21 and CTCF in control cells as expected [27,28], with

105   notable differences in their preferred interactions (Fig.1c). SA2 more strongly enriched

106   RAD21 while the SA1-CTCF interaction was significantly stronger than SA2-CTCF

107   (Fig.1c). Upon RAD21 degradation, we again observed a stronger effect on chromatin-

108   bound SA2 levels compared to SA1, implying stable binding of SA2 to chromatin is

109   more sensitive to cohesin loss than SA1. Not only did the residual SA proteins retain

110   their ability to interact with CTCF in the absence of Rad21, but the interactions

111   between SA1 and CTCF were further enhanced (Fig.1c). Reciprocal coIPs with CTCF

112   confirmed the CTCF-SA interactions in RAD21-depleted cells (Fig.1d). These results

113   were validated in a second cell line and upon siRNA-mediated knockdown of SMC3,

114   confirming the results (Fig S1b).

115   We performed two-color Stochastic Optical Reconstruction Microscopy
116   (STORM) to further assess the nuclear distribution and colocalization of SA1, SA2 and
117   CTCF with nanometric resolution in RAD21-degraded cells. Upon IAA treatment, we
118   observed a decreased density of detected SA1, SA2 and CTCF in two analyzed clones
119   (Fig.1e, f, S1c, d), suggesting that RAD21 degradation affects the stability of CTCF in
120   addition to SA proteins. As we observed by conventional confocal microscopy, SA2
121   localizations were more affected than SA1 (mean density reduction in SA1, 32% vs
122   SA2, 42%). Accordingly, SA1, SA2 and CTCF clusters were more sparsely distributed
123   across the nucleus upon RAD21 degradation as quantified by nearest neighbor
124   distance (NND) analysis of protein clusters (Fig.1g). This analysis also revealed a
125   higher density of SA1 and CTCF clusters compared to SA2, with shorter distances
126   between clusters, even in ethanol conditions (Fig.1g). To further confirm that SA and
127   CTCF were still co-localised in IAA conditions, we analyzed the relative distribution of
128   SA clusters to CTCF clusters by analyzing the NND distribution between SA1 and
129   CTCF, and SA2 and CTCF. NND showed that the association between SA1 and SA2
130   with CTCF is maintained upon RAD21 degradation as compared to both the control
131   cells and to a simulation of randomly-distributed protein clusters at the same density
132   (Fig. 1h). Interestingly, while the probability of SA1 at CTCF is only modestly affected
133   in IAA conditions, supporting their continued co-localization, SA2 at CTCF is more
134   affected in IAA treated cells, in line with results indicating that SA2 levels are more
135   affected than SA1 when cohesin is depleted. Together, our results confirm the
136   maintained interaction and localization patterns of SA proteins with CTCF and reveal
137   a difference in SA paralogue stability in the absence of the core cohesin trimer.

138

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

140   Previous analyses of the contribution of SA proteins to genome organization [7,9] were
141   performed in cells containing cohesin rings, possibly obscuring a functional role for SA
142   proteins themselves in genome organization. To determine if cohesin-independent SA
143   proteins may function at unique locations in the genome, we investigated whether the
144   residual SA-CTCF complexes (herein, SA-CTCF$^{\Delta Coh)}$ in IAA-treated RAD21^{mAC} cells
145   (Fig.1c, d) occupied the same chromatin locations as in control cells. Using chromatin
146   immunoprecipitation followed by sequencing (ChIP-seq), we determined the binding

147   profiles of CTCF, SA1, SA2, RAD21 and SMC3 in RAD21^{mAC} cells treated with ethanol

148   or IAA. Pairwise comparisons of CTCF ChIP-seq with RAD21 or SA in control

149   RAD21^{mAC} cells revealed the expected overlap in binding sites (Fig.1i, S1e).   In

150   contrast, both global and CTCF-overlapping RAD21 and SMC3 ChIP-seq signals were

151   dramatically lost in IAA-treated cells (Fig.1i, S1e). In agreement with our microscopy

152   and biochemistry results, we detected residual SA1 and SA2 binding sites in IAA-

153   treated cells which retained a substantial overlap with CTCF (Fig.1i, S1e). We

154   confirmed that the sites co-occupied by CTCF and SA proteins in RAD21-depleted

155   cells were previously bound in control conditions, indicating that CTCF and SA

156   maintain occupancy at their canonical binding sites in the absence of RAD21.   This

157   suggests that SA interaction with CTCF in the absence of the cohesin ring is a step in

158   normal cohesin activity.

159   While depletion of cohesin results in a dramatic loss of Topologically

160   Associated Domain (TAD) structure [8], the frequency of long-range inter-TAD, intra-

161   compartment contacts (LRC) is increased [5,8], and enriched for CTCF [5] or active

162   enhancers [8]. To determine whether residual, chromatin-bound SA could be associated

163   with LRCs in the absence of RAD21, we re-analysed Hi-C data from control and IAA-

164   treated RAD21^{mAC} cells [8]. We quantified all contacts within two different scales of

165   genome organization; local TAD topology (100k-1Mb) and clustered LRCs (1-5Mb)

166   (Fig.1j). As previously shown, local TAD contacts are lost and clustered LRCs are

167   enriched in IAA conditions (Fig.1j). When we probed the Hi-C datasets for contacts

168   containing the residual SA-CTCF$^{\Delta Coh}$ binding sites, we observed a further enrichment

169   in IAA conditions (Fig.1j, bottom row), indicating that SA-CTCF$^{\Delta Coh}$ are enriched at the

170   clustered LRCs formed when cells are depleted of cohesin and thus implicating them

171   in 3D structural configurations. Finally, using ChromHMM, we discovered that SA-

172   CTCF$^{\Delta Coh}$ sites are characterised by active chromatin and enhancers (Fig.S1f). Our

173   results suggest that cohesin-independent SA, either with CTCF or alone, may itself

174   contribute to large-scale arrangement of active chromatin and regulatory features in

175   3D space.

176

177

178

SA interacts with diverse 'CES-binding proteins' in RAD21-depleted cells.

SA proteins contain a highly conserved domain termed the 'stromalin conservative domain' [14,29], or the 'conserved essential surface' (CES). Structural analysis of CTCF-SA2-SCC1(RAD21) has recently shown that FGF (F/YxF) motifs in the N-terminus of CTCF bind to the CES on the SA2-SCC1 sub-complex, forming a tripartite interaction patch [16]. Furthermore, the authors identified an FGF-like motif in additional cohesin regulators and showed that a consensus motif could be used to predict interaction with additional chromatin proteins. Thus, we investigated whether SA could associate with other FGF-motif containing proteins in native and IAA conditions in cells. We performed chromatin IP with SA1 and SA2 in ethanol and IAA and probed for interaction with CTCF, and three additional FGF-motif containing proteins, CHD6, MCM3 and HNRNPUL2 (Fig.2a, S2a). As with CTCF, all of the proteins directly interacted with SA1 in RAD21-control cells and furthermore their interaction with SA1 was enriched upon RAD21-degradation. Interestingly, despite SA2 also containing the conserved 'CES' domain, the FGF-motif proteins did not interact with SA2 as strongly (Fig.2a), pointing to an additional element which functions to stabilise SA with FGF-containing proteins *in vivo*. These results revealed that SA can interact with proteins beyond just CTCF in the absence of cohesin, indicating a need to re-evaluate the role of SA in cohesin activity and consider possible novel functions for SA 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 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 or RAD21-degraded RAD21^{mAC} cells. Three biological replicate sets 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 (Fig. S2b). Immunoprecipitated proteins were in-gel Trypsin digested, gel extracted, and identified by liquid chromatography tandem mass spectroscopy (LC-

8

210 MS/MS). SA1 peptides were robustly detected in all UT and siCON samples and never

211 detected in IgG controls, validating the specificity of the antibody.

212   We identified 1282 unique proteins that co-purified with SA1 with a False

213 Discovery Rate <1%. After filtering steps (methods), we used a pairwise analysis of

214 IAA vs UT samples to generate a fold-change value for each putative interactor. These

215 candidates were found in at least 2 of the 3 SA1 IP replicates, were changed by at

216 least 1.5-fold compared to UT controls, and sensitive to siSA1, yielding 134 high-

217 confidence cohesin-independent SA1 (SA1$^{\Delta Coh}$) interactors (Fig. 2b, Table 1). As

218 expected, core cohesin subunits SMC1A and SMC3 were strongly depleted while no

219 peptides were detected for RAD21 (Fig. 2b). SA1 itself was significantly depleted

220 compared to control cells, as were other cohesin regulators, known to directly interact

221 with SA1, such as PDS5B [30]. In line with the enrichment we observed for the CES-

222 binding proteins in IAA-conditions (Fig. 1c, 2a), the vast majority of the SA1$^{\Delta Coh}$

223 interactors were enriched for binding with SA1 in IAA conditions (117 of 134) (Fig. 2b).

224   We used STRING analysis to compute the associations between our SA1$^{\Delta Coh}$

225 interactome and to identify enriched biological processes and molecular functions.

226 This revealed that the SA1$^{\Delta Coh}$ PPI network included gene expression, chromatin,

227 cytoplasmic and RNA binding proteins representing a variety of functionally diverse

228 cellular processes. Among these are processes previously associated with cohesin

229 biology and identified in published cohesin mass-spec experiments [31], thus validating

230 our approach, such as MCM3 and SWI/SNF components INO80 and SMARCAL1

231 which are involved in DNA replication and chromatin remodelling, respectively.

232 Similarly, several transcriptional and epigenetic regulators were identified, such as

233 PRC2 component JARID2 and TAF15 and SPTY2D1.

234   In addition, we identified proteins associated with SA1 in IAA conditions that

235 were involved in functions that have not been previously associated with SA biology

236 (Fig. 2c, d). The most enriched category was RNA processing (p=3.62^{-39}), and

237 included proteins involved in RNA modification (YTHDC1, ADAR1, FTSJ3); mRNA

238 stabilization and export (SYNCRIP, FMR1); and several RNA splicing regulators

239 (SRSF1, SON). Accordingly, we found a significant enrichment for DNA and RNA

240 helicases (p=3.54^{-08}) (MCM3, DHX9, more) and RNA binding proteins (p=9.11^{-11})

9

241 within which were serveal hnRNP family members (hnRNPU or SAF-A). We also found

242 a highly significant enrichment of proteins associated with Ribosome biogenesis

243 ($p=2.20^{-30}$) including both large and small subunit components (RPL5, 17, 29, RPS9);

244 rRNA processing factors (BOP1, NOP56); and components of the snoRNA pathway.

245 Translation was significantly enriched as a biological process ($p=1.64^{-06}$), with several

246 cytoplasmic translation regulators also identified as SA1$^{\Delta Coh}$ interactors (Fig. 2c, d).

247 Among these are ESYT2 and EIF3B which we identify as FGF-containing proteins that

248 are primarily found in the cytoplasm (Fig. 2d). We validated 8 of the highest-ranking

249 proteins within the enriched functional categories described above by immunoblotting

250 in ethanol and IAA-treated RAD21^{mAC} cells (Fig. 2d). Overall, our results show that

251 SA1$^{\Delta Coh}$ PPIs contain not only for transcriptional and epigenetic regulators, but also

252 are predominantly enriched for proteins with roles in nuclear RNA processing and

253 modification, ribogenesis and translation pathways. Accordingly, this suggests that SA

254 may facilitate an aspect of cohesin regulation at a variety of functionally distinct cellular

255 locations through its association with these diverse proteins.

256

SA proteins directly bind RNA.

258 Since RNA binding and RNA processing were two of the most enriched categories in

259 the SA1$^{\Delta Coh}$ PPI network, we asked whether SA proteins could also bind RNA. We

260 performed CLIP (crosslinking and immunoprecipitation) to determine whether SA

261 proteins directly bind RNA in untreated RAD21^{mAC} cells. We found that both SA1 and

262 SA2 directly bound RNA (Fig.3a). This was evidenced by detection of RNPs of the

263 expected molecular weights, with a smear of trimmed RNA, which was stronger in the

264 +UV and +PNK conditions, which increased as the RNaseI concentration was reduced

265 and which was lost after siRNA-mediated KD (Fig.3a, S3a). We repeated the

266 experiment in IAA-treated RAD21^{mAC} cells to determine if the SA subunits can directly

267 bind RNA in the absence of cohesin. Although RAD21 depletion reduced SA1 and

268 SA2 stability, the amount of RNA crosslinked to the proteins remained proportional to

269 the amount of SA1 and SA2 protein, demonstrating that cohesin is not required for the

270 interaction of these proteins with RNA in cells (Fig.3b, S3b. Thus, cohesin-

271 independent SA proteins interact with a wide array of RNA binding proteins (RBP) as

272 well as with RNA itself.

10

273 **A variable exon in the C-terminus of SA tunes association w RNA binding**
274 **proteins.**

275 SA1 and SA2 express transcript variants in RAD21^{mAC} cells. One such prominent
276 variant arises from the alternative splicing of a single C-terminal exon, exon 31 in SA1
277 (SA1$^{e31\Delta}$) and exon 32 in SA2 (SA2$^{e32\Delta}$) (Fig.3c). This has been observed in many cell
278 types, however the significance of this variant is unknown. We re-analysed publicly
279 available RNA-seq datasets for gene expression and alternative splicing. Interestingly,
280 quantification of the splicing profiles using VAST-tools analysis [32] revealed that the
281 frequency of the e31 or e32 splicing events were dramatically different (Fig.3d). The
282 majority of SA1 mRNAs *include* e31 (average PSI 97.7%), while the majority of SA2
283 mRNAs *exclude* e32 (average PSI 20.4%). We confirmed this at the protein level by
284 designing custom esiRNAs to specifically target SA1 e31 or SA2 e32 (Methods).
285 Smartpool (SP) KD reduced the levels of SA1 and SA2 to similar extents compared to
286 scrambled controls (87% and 94% respectively compared to siRNA control) (Fig. 3e,
287 f). Specific targeting of SA1 e31 led to a reduction of 85% of SA1 compared to esiRNA
288 control (Fig.3e), while SA2 e32 targeting had a minimal effect on SA2 protein levels
289 compared to its esiRNA control (reduction of 2%) (Fig.3f), in line with the PSI data.

290 These results imply that cells 'tune' the availability of e31/32 domains in SA
291 proteins, prompting us to investigate the nature of these exons to shed light on their
292 potential function. Inspection of the amino acid (aa) sequence of the spliced exons
293 revealed that they encode a highly basic domain within an otherwise acidic C-terminus
294 (Fig.3c). Overall, the SA paralogs are highly homologous, however the N and C termini
295 diverge in their aa sequence. Despite this divergence, e31 and e32 have retained their
296 basic properties with a pI similar to histones (pI=10.4, 9.9 for e31 and e32,
297 respectively) (Fig.3c). Basic domains act as important regulatory cassettes and can
298 bind nucleic acids. Thus, we investigated whether the alternatively spliced basic exon
299 of SA proteins contributes to the differential association of SA with RNA (Fig.3a,b).
300 We cloned cDNAs from HCT116 cells representing full-length SA2 (SA2^{FL}) and the
301 variant lacking e32 (SA2$^{e32\Delta}$), tagged them with YFP and expressed them in HCT cells
302 (Fig.S3c). We used the GFP-TRAP system to specifically purify the YFP-tagged
303 isoforms from cells and compared their ability to interact with RNA. While CLIP
304 experiments revealed that the presence of the alternative e32 does not affect the
305 ability of SA2 to interact with RNA (Fig.3g, S3d), however it did reveal bands which

11

306 were enriched in the YFP-SA2^{FL} CLIP and not observed in the YFP-SA2$^{e32\Delta}$ samples,

307 revealing a role for the alternative exon in enhanced association of SA2 with RNA

308 binding proteins.

309

310 **SA proteins bind to endogenous R-loops.**

311 Regulators of RNA processing, such as splicing, modification and export factors, act

312 as regulators of R-loops [33]. In addition, R-loops are found at sites of multiple biological

313 processes including transcription (of both mRNAs and rRNA), DNA replication and

314 DNA repair [33]. Given the fact that many of these processes were enriched in the SA1

315 interactome and our observations that SA proteins can interact with RNA, we reasoned

316 that the diversity of biological processes represented in the SA1$^{\Delta Coh}$ PPI network may

317 be reflective of a role for SA proteins in R-loop biology.

318   To address this, we returned to our IP-MS experiment to analyse enrichment of

319 R-loop-associated proteins in our SA1$^{\Delta Coh}$ interactome. We overlapped the proteins

320 identified in two independent IP-MS experiments using the R-loop specific antibody,

321 S9.6 [34,35] to create a custom high-confidence 'R-loop interactome' and then used a

322 hygrometric distribution to determine the significance of this category in the SA1$^{\Delta Coh}$

323 interactome (methods). Both the custom R-loop interactome as well as the S9.6

324 interactomes from Cristini et al., and Wang et al., were highly over-enriched in the

325 SA1$^{\Delta Coh}$ interactome (FDR=1.10×10^{-15}, 1.38×10^{-47}, respectively) (Fig.4a). As an

326 independent validation of these observations, we optimised an S9.6 coIP method in

327 RAD21^{mAC} cells (Fig.4b, methods). In agreement with published results, we found that

328 S9.6 precipitated with the known R-loop helicases AQR, DHX9, RNase H2 [34,36] as well

329 as MCM3 and RNA Pol II [37]. Both SA1 and SA2 precipitated with S9.6 (Fig.4b, S3x),

330 indicating a function at R-loops and supporting the observed enrichment of R-loop

331 proteins in the SA1 interactome.

332   To understand the causal relationship between R-loops and SA proteins, and

333 to determine the specificity of S9.6-SA interactions, we used RNase H1 to selectively

334 degrade the RNA component of RNA:DNA hybrids [38]. We were able to achieve a

335 ~30% reduction in R-loops upon treatment of chromatin lysates with RNase H (Fig.4b,

336 S4a). This reduction was proportional to the observed reduction in coIP of SA1 by

337 S9.6 (Fig.4b, S4b). In parallel, we assessed the effect of R-loop degradation on

12

338 chromatin-bound SA levels in single cells using confocal microscopy. Treatment of
339 RAD21^{mAC} cells with RNase H1 reduces S9.6 staining by >50% of untreated controls
340 (Fig.4c, d). In agreement with the S9.6 coIP results, mean levels of SA1 and SA2 were
341 significantly reduced by 35% and 18.5%, respectively compared to control cells in
342 response to RNase H treatment (Fig. 4c, d). Finally, we also depleted R-loops *in vivo*
343 by overexpressing ppyCAG-v5-RNaseH1 in cells. IF revealed that nuclear S9.6 levels
344 were significantly reduced in cells which expressed v5 (to 38% of controls) and that
345 mean levels of chromatin-associated SA1 were similarly reduced by 29.4% (p=4.05E-
346 8) (Fig.S4c), further confirming the causal relationship between R-loops and SA
347 proteins.

348

SA1 proteins act as suppressors of R-loops.

350 Proteins that act to suppress R-loops *in vivo*, such as AQR [36], have an inverse
351 correlation with S9.6 levels. From our IF results (Fig. 4c, d), we noticed that SA1 had
352 a similar negative relationship with S9.6 (Fig. S4d), prompting us to investigate
353 whether SA proteins could act as suppressors of R-loop formation. To this end, we
354 treated RAD21^{mAC} cells with scramble control siRNAs or siRNA to SA1, SA2 or AQR
355 and used IF to assess the impact on nuclear S9.6 levels in KD cells (Fig.4e, f). As
356 previously reported, AQR KD resulted in a 30.1% increase of mean nuclear S9.6 levels
357 (p=0.0004). Compared to control siRNA-treated cells, mean SA1 levels were reduced
358 by 56.2% (p=4.1E-40), while mean nuclear S9.6 staining was significantly increased
359 in the same cells by 55.3% (p=3.90E-08) (Fig.3f, g). We note that perturbing SA1
360 levels increased nuclear S9.6 staining to a similar extent as what was observed upon
361 AQR KD, a bonafide R-loop regulator. When we treated cells with the custom siRNA
362 to SA1 e31 (Fig 3e), we also observed an increase in S9.6 signals (Fig S4e),
363 suggesting that this basic exon plays a role in R-loop stability. Surprisingly, despite
364 efficient KD of SA2 (68% reduction), there was no significant change in nuclear S9.6
365 staining (mean S9.6 reduced by 10% compared to control, p=0.17), indicating that
366 although SA2 is localised to R-loops (Fig S4b), it does not seem to contribute to their
367 regulation. Taken together, our results confirm the presence of SA proteins at
368 endogenous R-loops *in vivo* and reveal a role for SA1 in R-loop suppresion.

369

370 **SA contributes to cohesin loading independently of NIPBL.**

371 Our results thus far support a hypothesis whereby $SA^{\Delta Coh}$ engages with RNA and

372 various RNA binding proteins at clustered regulatory regions (possibly R-loops) to

373 structurally support them and/or facilitate cohesin's association with chromatin.

374 Indeed, several lines of evidence suggest that alongside the canonical NIPBL/Mau2

375 loading complex, SA proteins contribute to cohesin's association with chromatin. In

376 yeast, interaction of the SA orthologue with the loader complex is required for efficient

377 association of the cohesin ring with DNA and subsequent ATPase activation [13,14].

378 Separating interactions into SA-loader and cohesin ring-loader subcomplexes still

379 impairs cohesin loading, indicating that SA functions as more than just a bridge protein

380 [14]. Crystallisation studies reveal a striking similarity of NIPBL and SA, in that both are

381 highly bent, HEAT-repeat proteins [39,40]. Indeed, NIPBL and SA1 interact together in

382 an antiparallel arrangement and wrap around DNA and the cohesin ring via similar

383 interactions in their respective 'U' surfaces, implying that SA1 has a role in the initial

384 recruitment of cohesin to DNA alongside NIPBL (Shi 2020, Higashi 2020).

385 The $RAD21^{mAC}$ system has the unique advantage that when IAA is washed off

386 cells, RAD21 proteins are no longer degraded and can be 're-loaded' back onto

387 chromatin (Fig.5a). We coupled this to an siRNA-mediated KD of NIPBL to investigate

388 whether cohesin re-loading onto chromatin is influenced by SA proteins in human cells

389 in native conditions. $RAD21^{mAC}$ cells were treated with scramble or NIPBL siRNAs and

390 subsequently grown in ethanol or IAA. The '0h post IAA wash-off' sample represents

391 the extent of cohesin *degradation* in the IAA-treated cells. In parallel, IAA was washed

392 out and the cells were left for 4h to recover. This sample, '4h post IAA wash-off'

393 represents the extent of cohesin *re-loading* in the respective genetic background

394 (Fig.5a, b). We confirmed loss of the loader complex by immunoblot for both NIPBL

395 and MAU2 as it is known that MAU2 is de-stablised upon NIPBL loss [41]. We note that

396 re-loading was not fully restored to the levels observed in ethanol-treated cells and

397 varied between experiments (Fig.5c), which may reflect differences in the initial

398 amounts of RAD21 or NIPBL (see methods). Despite this variation, we observed a

399 consistent effect on RAD21 re-loading across 8 independent experiments. As

400 expected, in NIPBL KD conditions, mean RAD21 re-loading efficiency was reduced to

401 40.9% of the siRNA controls (mean re-loading siNIPBL, 2.1 vs siCon, 3.6), however

14

402   this did not represent a statistically significant difference (p=0.33) and accordingly, a

403   large fraction of chromatin-associated RAD21 could still be detected in NIPBL KD cells

404   (Fig. 5c), indicating that cells have a NIPBL-independent cohesin re-loading

405   mechanism.

406   We performed the same experiment and this time, in addition to treating cells

407   with siRNAs to NIPBL, we also included siRNA to SA1 and SA2 together (siSA), and

408   a siNIPBL+ siSA condition to ask if SA proteins contribute to the observed NIPBL-

409   independent reloading. Across 5 independent experiments, SA KD had a more

410   dramatic effect on cohesin re-loading efficiency than NIPBL KD, reducing RAD21 re-

411   loading on chromatin to 51% of scramble controls (mean re-loading siSA, 1.9 vs siCon,

412   5.1, p=0.061), (Fig. 5d, e). However, only when SA and NIPBL were both reduced in

413   cells, was there a statistically significant change to cohesin re-loading, reducing

414   RAD21 on chromatin to 64.9% of scramble control cells (mean re-loading

415   siNIPBL+siSA, 1.42 vs siCon, 5.1, p=0.001), indicating that SA performs an important

416   and complementary step to NIPBL and MAU2 during normal reloading (Fig.5d, e).

417   Finally, given that SA localises to R-loops and these can be localized to many

418   places across the genome, we reasoned that SA could use this structural platform to

419   link the loading of cohesin to diverse biological processes. Therefore, we repeated the

420   cohesin re-loading experiments in the presence of siRNAs to AQR, which we had

421   previously shown act as suppressors of R-loops (Fig. 4e, f). AQR KD alone had little

422   effect on cohesin re-loading efficiency (Fig. 5f, g), however when R-loops were

423   increased in the context of reduced NIPBL, we observed an increase in the efficiency

424   of cohesin re-loading compared to control cells (Fig. 5f, g). This increase in re-loading

425   efficiency corresponded with a 2.08-fold increase in SA1 levels and a 1.46-fold

426   increase in s9.6 levels, relative to siCon (Fig. S5a) while MAU2 and AQR showed a

427   corresponding fold-change of 0.48 and 0.69, respectively, indicating the specificity of

428   SA1 and R-loop increase. Our results support a role for R-loops in SA-mediated

429   cohesin loading.

430

DISCUSSION

431 Whether SA proteins function in their own right outside of the cohesin complex is rarely

432 considered. Consequently, our understanding of how these proteins contribute to

433 cohesin function and disease is incomplete. In this study, we shed light on this

434 question by uncovering a diverse repertoire of SA interactors in cells acutely depleted

435 for the cohesin trimer. These range from proteins associated with translation and

436 ribogenesis to RNA processing factors and regulators of the epitranscriptome. These

437 observations suggest that SA proteins have a previously unappreciated role in post-

438 transcriptional regulation of gene expression which offers much-needed new insight

439 into its roles in disease and cancer.

440

441 Acute depletion of the cohesin ring has allowed us to capture a moment in the

442 normal life cycle of cohesin – DNA associations and unveiled a previously

443 unappreciated step for SA proteins. We show that cohesin-independent SA proteins

444 bind to DNA and RNA, in the context of non-canonical RNA:DNA hybrid structures as

445 we have shown here, or sequentially, and use this platform for the loading of cohesin

446 to chromatin. Our results are supportive of biophysical observations of SA proteins

447 and R-loops [42] and *in vitro* assessment of cohesin loading at DNA intermediates [24].

448 Structural studies suggest that NIPBL and SA1 together bend DNA and cohesin to

449 guide DNA entering into the cohesin ring [17,18,43]. Our work shows that in cells lacking

450 either the canonical NIPBL/MAU2 loader complex or the SA proteins, cohesin can still

451 associate with chromatin, suggesting that loading can occur with either component

452 alone, albeit most effectively together.

453 Since SA paralogues have distinct terminal ends and nucleic acid targeting

454 mechanisms [22,23], their initial recruitment to chromatin may be specified by unique

455 DNA, RNA or protein-interactions, or indeed all three. Such diversification of loading

456 platforms would be important in large mammalian genomes to ensure sufficient

457 cohesin was chromatin associated or to direct stabilization of particular biological

458 processes for a given cell fate [44]. Indeed, SA1 and SA2 show clear differences in

459 interaction with FGF-motif containing proteins, despite the fact that both paralogs

460 contain a CES domain [45], underscoring the importance of *in vivo* studies and arguing

461 that additional factors play an important role in complex stabilization. In this context,

462 RNA-associated protein interaction has previously been shown to support cohesin

16

463 stabilisation at CTCF at the *IGF2/H19* locus [46]. These results are in line with our

464 findings that a basic domain in the unstructured C-terminal portion of SA supports

465 RNA-associated protein interactions and R-loop stability.

466   This study identifies SA proteins as novel regulators of RNA:DNA hybrid

467 homeostasis. It is noteworthy that other suppressors of R-loop formation include

468 mRNA processing factors, chromatin remodellers and DNA repair proteins [47] which all

469 function in the context of nuclear bodies [48]. We find that SA proteins are enriched at

470 very distal chromatin interactions in cohesin-depleted Hi-C data and they interact with

471 numerous RNA binding proteins known to condense in 3D [49,50]. Harnessing such

472 condensates would provide an efficient loading platform for cohesin at sites of similar

473 biological function. If SA paralogs direct different localization of cohesin loading or

474 stability of its association, this could have important implications in our understanding

475 of disease and cancer.

ACKNOWLEDGMENTS

AUTHOR CONTRIBUTIONS

H.P. and S.H. conceived the project. H.P. designed and performed all the coIP, Mass spectrometry and cohesin re-loading experiments, analysed the ChIP 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 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. and Y.L. prepared cellular materials for CLIP, which was carried out by M.B., M.T.C., and R.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.

Declaration of Interests

The authors declare no competing interests.

18

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 ethanol (EtOH) as a control or Auxin (IAA) for 4hrs. Nuclei were counterstained with DAPI.

B) Imaris quantification of the fluorescence intensity of mClover, SA1 (top panel), SA2 (bottom panel) and CTCF 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<0.05, ** p<0.005, *** p<0.0005, n.s., not significant. 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<0.0001.

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) Nearest Neighbor Distance (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 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 residual SA-CTCF binding sites.

19

548   **Figure 2. Characterization of SA1 protein-protein interaction network in RAD21-**
549   **depleted cells.**
550   A) Chromatin coIP of SA1, SA2 and IgG with 4 predicted CES-binding proteins in
551   RAD21^{mAC} cells treated with EtOH or IAA for 4hrs. Input represents 1.25% of the
552   material used for immunoprecipitation.

553
554   B) Volcano plot displaying the statistical significance (-log10 p-value) versus
555   magnitude of change (log2 fold change) from SA1 IP-MS data produced
556   from RAD21^{mAC} cells treated with ethanol or IAA (n=3). Vertical dashed lines represent
557   changes of 1.5-fold. Horizontal dashed line represents a pvalue of 0.1.
558   Cohesin complex members and validated high-confidence proteins have been
559   highlighted.

560
561   C) SA1$^{\Delta Coh}$ interaction network of protein–protein interactions identified in
562   RAD21^{mAC} cells using STRING. Node colours describe the major
563   enriched categories, with squares denoting helicases and polygons
564   denoting hnRNP family members. Proteins within each enrichment category were
565   subset based on p-value change in B). See supplemental figures for full network.

566
567   D) Chromatin IP of SA1 and IgG in RAD21^{mAC} cells treated with EtOH or IAA and
568   immunoblotted with antibodies to validate the proteins identified by IP-MS. Input
569   represents 1.25% of the material used for immunoprecipitation. * We note that ESYT2
570   is a FGF-containing protein.

571
572
573
574   **Figure 3. SA proteins bind to RNA.**
575   A) CLIP for SA1, SA2 and non-specific IgG controls. Autoradiograms of crosslinked
576   ^{32}P-labelled RNA are shown at the top and the corresponding immunoblots, below.
577   CLIP was performed with and without UV crosslinking and polynucleotide kinase
578   (PNK) and with high (H; 1/50 dilution) or low (L; 1/500 dilution) concentrations of
579   RNase I.

580
581   B) CLIP for SA1, SA2 and non-specific IgG control in ethanol (-) or IAA-treated
582   Rad21^{mAC} cells. ^{32}P-labelled RNA and the corresponding immunoblots are shown as
583   above.

584
585   C) Schematic of the SA1 and SA2 proteins showing the SA1-specific AT-hook, the
586   conserved CES domain (blue) and the acidic C-terminus (green) which contains
587   the basic alternatively spliced exon (red). Right-hand zoom-in indicates the spliced
588   exons for SA1 (top) and SA2 (bottom) and the pI for each. The conservation scores for
589   the divergent N- and C-termini and the middle portion of the proteins which contains
590   the CES domain are shown.

591
592   (D) Percent Spliced In (PSI) calculations for SA1 exon 31 (black) and SA2 exon 32
593   (grey) based on VAST-Tools analysis of RNA-seq from multiple datasets (see
594   Methods).

595

596 (E) Western blot analysis of SA1 levels in whole cell lysates after treatment
597 with scrambled siRNAs (siCon), SmartPool SA1 siRNAs (siSA1 SP),
598 control esiRNAs (esiCon) and esiRNA designed to target SA1 exon
599 31 for 48hrs in RAD21^{mAC} cells. H3 serves as a loading control. The percentage of
600 knockdown (KD) after SA1 signal is normalised to H3 is shown.

602 (F) Western blot analysis of SA2 levels in whole cell lysates after treatment with
603 scrambled siRNAs (siCon), SmartPool SA2 siRNAs (siSA2 SP),
604 control esiRNAs (esiCon) and siRNA designed to target SA2 exon 32
605 for 48hrs in RAD21^{mAC} cells. H3 serves as a loading control. . The percentage of
606 knockdown (KD) after SA2 signal is normalised to H3 is shown.

608 (G) CLIP with endogenous SA2 (as in A, B above), IgG control and cells where
609 either YFP-tagged full-length SA2 or YFP-tagged SA2 lacking exon 32 are expressed
610 for 48hrs. CLIP reveals proteins which specifically associate with exon-32 containing
611 SA2 (arrow).

614 **Figure 4. SA proteins act as suppressors of endogenous R-loops.**
615 (A) Adjusted p-value (FDR) for enrichment of S9.6 interactome data from Cristini et al.
616 and Wang et al., with the SA1$^{\Delta Coh}$ interactome. Overlap indicates the proteins
617 identified in both S9.6 interactome datasets, representing a high confidence R-loop
618 interactome list.

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

625 (C) Representative confocal images of S9.6, SA1 and SA2 IF in RAD21^{mAC} cells
626 treated with control buffer or RNase H enzyme. Nuclear outlines (white) are derived
627 from DAPI counterstain.

629 (D) Fluorescence Intensity of S9.6, SA1 and SA2 protein assessed by
630 Immunofluorescence in (C). Data are from three biological replicates with >50 cells
631 counted/condition). Quantifications and statistical analysis were done as above.

633 (E) Representative confocal images of S9.6 and SA1, SA2 or AQR IF in
634 RAD21^{mAC} cells treated with scramble control siRNA (si scr) or siRNA to the protein of
635 interest. Nuclear outlines (white) are derived from DAPI counterstain.

637 (F) Fluorescence Intensity of S9.6, SA1, SA2 and AQR proteins assessed by
638 Immunofluorescence in (E). Data are from three biological replicates with >50 cells
639 counted/condition). Quantifications and statistical analysis were done as previously
640 stated.

21

Figure 5. SA proteins contribute to cohesin loading at R-loops.

(A) Schematic of experimental set-up. RAD21^{mAC} cells expressing mClover (green cells in dishes) were treated with scramble siRNAs or siRNA to NIPBL. Prior to collection, cells were cultured in ethanol or IAA for 4hrs to degrade RAD21 (0 h timepoints). Each sample was then split into those cultured in ethanol or IAA for 4hrs to degrade RAD21 (0h timepoints). The IAA treatment was washed-off and the cells were left to recover for 4hrs (4h timepoints). Chromatin fractions were prepared from all samples and used in immunoblot analysis.

(B) Representative western blot analysis of chromatin-bound RAD21, MAU2 and NIPBL levels in RAD21^{mAC} cells treated according to the schematic shown in (A). H3 was used as a loading control. *NB* The full blots can be seen in the supplement.

(C) RAD21 fold change relative to siCon samples at the 0h timepoint in siCon 4h (light 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.

(D) Representative western blot analysis of chromatin-bound RAD21, SA1, SA2, MAU2 and NIPBL levels in RAD21mAC cells treated according to the schematic shown in (A) and including samples treated with siRNA to SA1 and SA2 together (siSA) and siRNA to NIPBL + siSA. H3 was used as a loading control.

(E) RAD21 fold change relative to siCon samples at the 0h timepoint in siCon, siNIPBL, siSA and siNIPBL+siSA. Asterisks indicate a statistically significant difference as assessed using 2-tailed T-test. ** $p<0.005$. Data is from 5 biological replicates.

(F) Representative western blot analysis of chromatin-bound RAD21, AQR, MAU2 and NIPBL levels in RAD21mAC cells treated according to the schematic shown in (A) and including samples treated with siRNA to AQR and siRNA to NIPBL + AQR. H3 was used as a loading control. RAD21* denotes increased exposure. NIPBL levels could not be determined due to incompatibility with assessing RNA:DNA hybrids. *NB* siCon samples were cropped here and the full blots can be seen in the supplement.

(G) RAD21 fold change relative to siCon samples at the 0h timepoint in siCon, siNIPBL, siAQR and siNIPBL+siAQR. Data is from 2 biological replicates and is represented as mean +/- SEM.

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. Li, Y. *et al.* Structural basis for Scc3-dependent cohesin recruitment to chromatin. *eLIFE* **7,** 352 (2018).

17. Shi, Z., Gao, H., Bai, X.-C. & Yu, H. Cryo-EM structure of the human cohesin-NIPBL-DNA complex. *Science* **368,** 1454–1459 (2020).

18. Higashi, T. L. *et al.* A Structure-Based Mechanism for DNA Entry into the Cohesin Ring. *Molecular Cell* **79,** 917–933.e9 (2020).

19. 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).

20. 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).

21. 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).

22. 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).

23. 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).

24. Murayama, Y., Samora, C. P., Kurokawa, Y., Iwasaki, H. & Uhlmann, F. Establishment of DNA-DNA Interactions by the Cohesin Ring. *Cell* 1–29 (2018).

742		doi:10.1016/j.cell.2017.12.021
743	25.	Zheng, G., Kanchwala, M. & Xing, C. MCM2–7-dependent cohesin loading during S
744		phase promotes sister-chromatid cohesion. *eLIFE* 1–25 (2018).
745		doi:10.7554/eLife.33920.001
746	26.	Natsume, T., Kiyomitsu, T., Saga, Y. & Kanemaki, M. T. Rapid Protein Depletion in
747		Human Cells by Auxin-Inducible Degron Tagging with Short Homology Donors. *Cell Rep*
748		**15,** 210–218 (2016).
749	27.	Parelho, V. *et al.* Cohesins functionally associate with CTCF on mammalian
750		chromosome arms. *Cell* **132,** 422–433 (2008).
751	28.	Wendt, K. S. *et al.* Cohesin mediates transcriptional insulation by CCCTC-binding factor.
752		*Nature* **451,** 796–801 (2008).
753	29.	Roig, M. B. *et al.* Structure and function of cohesin's Scc3/SA regulatory subunit.
754		*FEBS Letters* **588,** 3692–3702 (2014).
755	30.	Hons, M. T. *et al.* Topology and structure of an engineered human cohesin complex
756		bound to Pds5B. *Nature Communications* **7,** 1–11 (2017).
757	31.	Kim, J.-S. *et al.* Systematic proteomics of endogenous human cohesin reveals an
758		interaction with diverse splicing factors and RNA-binding proteins required for mitotic
759		progression. *J. Biol. Chem.* **294,** 8760–8772 (2019).
760	32.	Irimia, M. *et al.* A Highly Conserved Program of Neuronal Microexons Is Misregulated in
761		Autistic Brains. *Cell* **159,** 1511–1523 (2014).
762	33.	Santos-Pereira, J. M. & Aguilera, A. R loops: new modulators of genome dynamics and
763		function. *Nat. Rev. Genet.* **16,** 583–597 (2015).
764	34.	Cristini, A., Groh, M., Kristiansen, M. S. & Gromak, N. RNA/DNA Hybrid Interactome
765		Identifies DXH9 as a Molecular Player in Transcriptional Termination and R-Loop-
766		Associated DNA Damage. *Cell Rep* **23,** 1891–1905 (2018).
767	35.	Wang, I. X. *et al.* Human proteins that interact with RNA/DNA hybrids. *Genome Res.* **28,**
768		1405–1414 (2018).
769	36.	Sollier, J. *et al.* Transcription-coupled nucleotide excision repair factors promote R-loop-
770		induced genome instability. *Molecular Cell* **56,** 777–785 (2014).
771	37.	Skourti-Stathaki, K., Kamieniarz-Gdula, K. & Proudfoot, N. J. R-loops induce repressive
772		chromatin marks over mammalian gene terminators. *Nature* **516,** 436–439 (2014).
773	38.	Cerritelli, S. M. & Crouch, R. J. Ribonuclease H: the enzymes in eukaryotes. *FEBS J.*
774		**276,** 1494–1505 (2009).
775	39.	Kikuchi, S., Borek, D. M., Otwinowski, Z., Tomchick, D. R. & Yu, H. Crystal structure of
776		the cohesin loader Scc2 and insight into cohesinopathy. *Proc. Natl. Acad. Sci. U.S.A.*
777		**113,** 12444–12449 (2016).
778	40.	Hara, K. *et al.* Structure of cohesin subcomplex pinpoints direct shugoshin-Wapl
779		antagonism in centromeric cohesion. *Nature Publishing Group* **21,** 864–870 (2014).
780	41.	Watrin, E. *et al.* Human Scc4 is required for cohesin binding to chromatin, sister-
781		chromatid cohesion, and mitotic progression. *Curr. Biol.* **16,** 863–874 (2006).
782	42.	Pan, H. *et al.* Cohesin SA1 and SA2 are RNA binding proteins that localize to RNA
783		containing regions on DNA. *Nucleic Acids Research* **24,** 105–17 (2020).
784	43.	Chao, W. C. H. *et al.* Structural Studies Reveal the Functional Modularity of the Scc2-
785		Scc4 Cohesin Loader. *Cell Rep* **12,** 719–725 (2015).
786	44.	Pezic, D. *et al.* The cohesin regulator Stag1 promotes cell plasticity through
787		heterochromatin regulation. *bioRxiv* 1–60 (2021). doi:10.1101/2021.02.14.429938
788	45.	Li, Y. *et al.* The structural basis for cohesin-CTCF-anchored loops. *Nature* **578,** 472–476
789		(2020).
790	46.	Yao, H. *et al.* Mediation of CTCF transcriptional insulation by DEAD-box RNA-binding
791		protein p68 and steroid receptor RNA activator SRA. *Genes Dev.* **24,** 2543–2555
792		(2010).
793	47.	García-Muse, T. & Aguilera, A. R Loops: From Physiological to Pathological Roles. *Cell*
794		1–15 (2019). doi:10.1016/j.cell.2019.08.055
795	48.	Misteli, T. Beyond the Sequence: Cellular Organization of Genome Function. *Cell* **128,**
796		787–800 (2007).
797	49.	Nozawa, R.-S. *et al.* SAF-A Regulates Interphase Chromosome Structure through
798		Oligomerization with Chromatin- Associated RNAs. *Cell* **169,** 1214–1227.e18 (2017).
799	50.	Huo, X. *et al.* The Nuclear Matrix Protein SAFB Cooperates with Major Satellite RNAs to

800   Stabilize Heterochromatin Architecture Partially through Phase Separation. *Molecular*
801   *Cell* **77,** 368–383.e7 (2020).
802   51.   Barrington, C., Georgopoulou, D., Nature, D. P.2019. Enhancer accessibility and CTCF
803   occupancy underlie asymmetric TAD architecture and cell type specific genome
804   topology. *Nat Communications.* doi:10.1038/s41467-019-10725-9
805   52.   Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized
806   p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat Biotechnol*
807   **26,** 1367–1372 (2008).
808   53.   Choi, M. *et al.* MSstats: an R package for statistical analysis of quantitative mass
809   spectrometry-based proteomic experiments. *Bioinformatics* **30,** 2524–2526 (2014).
810   54.   Beltran, M. *et al.* The interaction of PRC2 with RNA or chromatin is mutually
811   antagonistic. *Genome Res.* **26,** 896–907 (2016).
812

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. 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 in (Natsume). 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 ethanol) 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 48 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

26

854 another 40h until harvest. Efficiency of KD was assessed by WB. siRNA information
855 can be found in Table 1.

856

857 **Immunofluorescence**

858 Cells were adhered onto poly L-lysine coated glass coverslips in 6 well
859 culture dishes and were washed twice with ice-cold PBS before IF procedures. For
860 RAD21-depletion analysis, cells were fixed for 10 mins at room temperature with 3.7%
861 paraformaldehyde (Alfa Aesar) in PBS, washed 3 times with PBS and
862 then permeabilized at room temperature for 10 mins with 0.25% Triton X-100 in PBS
863 (Sigma Aldrich). For R-loop imaging, cells were fixed and permeabilised with ice-cold
864 ultra-pure methanol (Sigma Aldrich) for 10 mins at -20°C. After 3 washes with PBS,
865 cells were blocked for 45 mins at room temperature with 10% FCS-PBS. For
866 RNASEH1 enzyme treatment, cells were incubated with blocking solution
867 supplemented with 1x RNASEH1 reaction buffer alone (50 mM Tris-HCl, 75 mM KCl,
868 3 mM $MgCl_2$, 10 mM DTT) or 5 units of RNASEH1 enzyme (M0297, New England
869 Biolabs) for 30 mins at 37 °C, PBS-washed twice, before blocking. Cells were washed
870 twice with PBS before incubation with primary antibodies diluted in 5% FCS-PBS at 4
871 °C overnight. Anti-SA1, anti-SA2 and anti-AQR were used at 1:3000 dilutions; anti-
872 CTCF was used at 1:2500 dilution; anti-s9.6 was used at 1:1000 dilution; anti-V5 was
873 used at 1:1000.   After 4 washes with PBS, cells were incubated with secondary
874 antibodies (donkey anti-Goat AF555 or AF647 for SA1/2 used at 1:3000; donkey anti-
875 Rabbit AF647 for CTCF used at 1:2500; donkey anti-Mouse AF555 for s9.6 used at
876 1:2000; donkey anti-Rabbit AF647 for AQR used at 1:3000; donkey anti-Rabbit AF488
877 for V5 used at 1:2000)) in 5% FCS-PBS for 1 hour at room temperature, and washed
878 4 times with PBS before being mounted onto glass slides with ProLong[TM] Diamond
879 Antifade Mountant with DAPI (Thermo Fisher Scientific) to stabilise overnight in dark
880 before imaging. See Table 2 for details of where antibodies were purchased.

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

895 Significance is denoted as p>0.05 = not significant (ns), p≤0.05 = *, p≤0.01 = **,

896 p≤0.001 = *** and p≤0.0001 = ****.

897

898 **Chromatin Fractionation and coImmunoprecipitation.**

899 Cells were washed twice with ice-cold PBS (Sigma Aldrich) and lysed in Buffer A (10

900 mM HEPES, 10mM KCl, 1.5 mM MgCl2, 0.34 M Sucrose, 10% Glycerol, 1mM DTT,

901 1mM PMSF/Pefabloc, protease inhibitor), supplemented with 0.1% T-X100, for 10 min

902 on ice. Lysed cells were collected by scraping. Nuclei and cytoplasmic

903 material was separated by centrifugation for 4 min at 1300 g at 4oC. The supernatant

904 was collected as the cytoplasmic fraction and cleared of any insoluble material with

905 further centrifugation for 15 min at 20,000 g at 4oC. The nuclear pellet was washed

906 once with buffer A before lysis in buffer B (3mM EDTA, 0.2mM EGTA, 1mM DTT, 1mM

907 PMSF/Pefabloc, protease inhibitor) with rotation for 30 min at 4oC. Insoluble nuclear

908 material was spun down for 4 min at 1700 g at 4oC and the supernatant taken as

909 nuclear soluble fraction. The insoluble material was wash once with buffer B and then

910 resuspended in high-salt chromatin solubilization buffer (50mM Tris-HCl pH 7.5, 1.5

911 mM MgCl2, 500mM KCl, 1mM EDTA, 20% Glycerol, 0.1% NP-40, 1mM

912 PMSF/Pefabloc, protease inhibitor). The lysate was vortexed for 2 min to aid

913 solubilization. Nucleic acids were digested with 85U benzonase (Sigma-Aldrich) per

914 100 x 106 cells, with incubation for 10 min at 37oC and 20 min at 4oC. Chromatin was

915 further solubilized with ultra-sonication for 3 x 10 sec at an amplitude of 30. The lysate

916 was diluted to 200 mM KCl and insoluble material was removed by centrifugation at

917 15,000 RPM for 30 min at 4oC.

918 For coIP, antibodies were bound to Dynabead Protein A/G beads

919 (ThermoFisher Scientific) for 10 min at room temperature and ~ 5 hr at 4oC. For mock

920 IgG IPs, beads were incubated with serum from the same host type as the antibody of

921 interest. 1mg of chromatin extract was incubated with the antibody-bead conjugate per

922 IP for approximately 16 hr at 4oC. IPs were washed x5 with IP buffer (200mM

923 chromatin solubilization buffer) and eluted by boiling in either 2x Laemmeli sample

924 buffer (BioRad) or 4x NuPAGE LDS sample buffer (ThermoFisher Scientific). Proteins

925 ≤ 250 kDa were separated by SDS-PAGE electrophoresis using 4–20% Mini-

926 PROTEAN® TGX™ Precast Protein Gels (BioRad) and transferred to Immobilon-P

927 PVDF Membrane (Merck Millipore) for detection. Proteins ≥ 250 kDa were separated

928 by SDS-PAGE electrophoresis using Invitrogen NuPAGE 3-8% Tris-Acetate precast

929 protein gels. Transfer was extended to overnight with low voltage (20V) to aid in

930 transfer of the high-molecular weight proteins. Membranes were incubated in primary

931 antibody solution overnight at 4oC and images were detected using chemiluminescent

932 fluorescence. Densitometry was carried out using ImageStudio Lite software with

933 statistical significance calculated by unpaired t test, unless otherwise specified. Fold

934 enrichment quantifications were performed by first normalising the raw densitometry

935 value to its corresponding Histone H3 quantification and the comparing between the

936 samples indicated. See Table 2 for details of antibodies.

937

S9.6 IP and Dot Blot.

938 **S9.6 IP and Dot Blot.**

939 Cells were fractioned and processed for S9.6 IP as described above, with the following
940 modifications. To avoid digestion of RNA:DNA hybrids, samples were not treated
941 with benzonase during chromatin solubilization and sonication was carried out for 10
942 min (Diagenode Biorupter) as in [34]. Where indicated, chromatin samples were treated
943 with Ribonuclease H enzyme (NEB) overnight at 37°C to digest RNA:DNA hybrids in
944 the extract. To avoid detection of single-stranded RNA by the S9.6 antibody, all S9.6
945 IP samples were pre-treated with Purelink RNase A (Thermo Fisher Scientific) at
946 0.25ug/1mg chromatin extract for 1 hr 30 min at 4°C. The reaction was stopped with
947 addition of 143U Invitrogen SUPERase•In RNase Inhibitor (Thermo Fisher
948 Scientific). RNA:DNA hybrid levels were assessed in chromatin samples by dot blot.
949 Specifically, the chromatin lysate was directly wicked
950 onto Amersham Protran nitrocellulose membrane (Merck) by pipetting small volumes
951 above the membrane. Membranes were blocked in 5% (w/v) non-fat dry milk in PBS-
952 0.1% Tween and incubated with S9.6 antibody overnight as for standard western blot.
953 As above, detection was carried out using chemiluminescent fluorescence. RNase A-
954 mediated digestion of RNA:DNA hybrids was performed using a non-ssRNA-specific
955 enzyme (Thermo Scientific) at 1.5ug/25ug chromatin extract at 37°C.

956

957 **ChIP-sequencing, library preparation and analysis.**

958 ChIP lysates were prepared from RAD21^{mAC} cells treated with ethanol or IAA for
959 4hrs in two biological replicates. Formaldehyde (1%) was added to the culture
960 medium for 10min at room temperature. Fixation was blocked with 0.125M glycine and
961 cells were washed in cold PBS. Nuclear extracts were prepared by douncing (20
962 strokes, medium pestle) in swelling buffer (25 mM HEPES pH8, 1.5 mM MgCl2, 10mM
963 KCL, 0.1% NP40, 1 mM DTT and protease inhibitors) and centrifuged for 5min at
964 2000rpm at 4C. Nuclear pellets were resuspended in Sucrose buffer I
965 (15mM Hepes pH 8, 340 mM Sucrose, 60mM KCL, 2mM EDTA, 0.5 mM EGTA, 0.5%
966 BSA, 0.5 mM DTT and protease inhibitors) and dounced again with 20 strokes. The
967 lysate was carefully laid on top of an equal volume of Sucrose buffer
968 II (15mM Hepes pH 8, 30% Sucrose, 60mM KCL, 2mM EDTA, 0.5 mM EGTA, 0.5 mM
969 DTT and protease inhibitors) and centrifuged for 15min at 4000rpm at 4C. Nuclei were
970 washed twice to remove cytoplasmic proteins, centrifuged and the pellet was
971 resuspended in Sonication/RIPA buffer (50mM Tris, pH 8.0, 140 mM NaCl, 1 mM
972 EDTA, 1% Triton X-100, 0.1% Na-deoxycholate, 0.1% SDS and protease inhibitors) at
973 a concentration of 5 x10^6 nuclei in 130µl buffer. This was transferred to a sonication
974 tube (AFA Fiber Pre-Slit Snap-Cap 6x16mm) and sonicated in a Covaris S2 (settings;
975 4 cycles of 60 seconds, 10% duty cycle, intensity: 5, 200 cycles per
976 burst). Soluable chromatin was in the range of 200 - 400 bp. Triton X100 was added
977 (final concentration 1%) to the sonicated chromatin and moved to a low-retention

978   tube (Eppendorf) before centrifugation at 14,000 rpm for 15min at 4C and pellets were

979   discarded. 1/100th of the chromatin lysate was retained as the Input sample.

980   For Immunoprecipitation, 200ug chromatin aliquots/IP were precleared with a

981   slurry of Protein A/G (50:50) (Dynabeads) an incubated for 4hr at 4C. Meanwhile,

982   washed Protein A/G beads (40ul per IP) were mixed with primary antibodies and

983   incubated for 4hrs at 4C. The following amounts of antibodies were used: anti-

984   CTCF, 5ug/ChIP; anti-SA1, 15ug/ChIP; anti-SA2, 10ug of the mixed antibody

985   pack/ChIP; anti-Smc3, 5ug/ChIP and anti-IgG, 10ug/ChIP. See Table 2 for information

986   about the antibodies. Washed, pre-bound Protein A/G beads+antibody were mixed

987   with pre-cleared chromatin lysates and incubated overnight with rotation at 4C. The

988   next day, the supernatant was removed and the beads were washed 9 times with

989   increasing salt concentrations. Protein-DNA crosslinks were reversed in ChIP elution

990   buffer (1% SDS, 5 mM EDTA, 10 mM Tris HCl pH 8) + 2.5 ul of Proteinse K and

991   incubated for 1 hour at 55°C and overnight at 65°C. Samples were phenol–chloroform

992   extracted, resuspended in TE buffer and assessed by qPCR as a quality

993   control. Libraries were prepared from 5-10ng of purified DNA, depending on

994   availability of material, using NEBNext Ultra II DNA Library Prep Kit for Illumina kit and

995   using NEBNext Multiplex Oligos for Illumina (Index Primers Set 2) according to

996   manufacturer's instructions using 6-8 cycles of PCR. ChIP-seq libraries from one

997   biological   set   (all   ChIP   libraries   for   both   ethanol   and

998   IAA) were multiplexed and sequenced on the Illumina HiSeq2500 platform, 80bp

999   single-end reads. Each biological set was sequenced on a separate run.

1000   Quality control of reads was preformed using FASTQC. Reads were aligned to

1001   the hg19 reference genome using Bowtie with 3 mismatches. Only replicate 1 of the

1002   SA1 librraye was used. PCR duplicates were detected and removed using

1003   SAMTOOLS. Bam files were imported into MISHA (v 3.5.6) and peaks were identified

1004   using a 0.995 percentile. Peaks that overlapped in both replicates were

1005   retained. Correlation plots of peaks across the genome from different ChIP libraries

1006   were compared with log-transformed percentiles plotted as a smoothed scatter plot.

1007   Comparison of peaks at regions of interest were carried out using deepTools (Version

1008   3.1.0-2 REF). For input into deepTools, peak data was converted to bigwig format,

1009   with a bin size of 500, using the UCSC bedGraphtoBigWig package. The signal matrix

1010   was calculated for a window 2,000 bp up- and down-stream of the region of interest,

1011   missing data was treated as zero, and all other parameters were as default. Heatmaps

1012   were generated within deepTools, with parameters as default.

1013

1014   **Hi-C data and contact hotspots analysis.**

1015   Generating   hotspots   -   Previously   published   Hi-C   datasets   derived   from

1016   RAD21[mAC] cells treated with ethanol or IAA [8] were analyzed as previously described

1017   [51]. Custom R scripts were written to identify Hi-C hotspots, i.e. regions of Hi-C maps

1018   with high contact frequency. To begin, for each chromosome, all contacts were

1019   extracted and subsetted for only high scoring (>=60) contacts between a band of 10e3

1020 – 70e6. Using KNN, for each high scoring contact, the 250 nearest neighbour contacts
1021 were identified and subset for only the high-scoring neighbours. This created a list of
1022 high scoring neighbours for each high scoring contact, where the first neighbour is the
1023 contact itself with a distance of 0. This allowed the neighbour information to be
1024 converted into edge information, thereby allowing high score fend contacts to be
1025 grouped into cluster hotspots using the R package 'igraph'. Hotspots that contained
1026 less than the minimum number of high scoring fends (<100) were removed. The output
1027 list of hotspots were represented as 2D intervals which contained high scoring
1028 contacts. In total, 5539 hotspots were identified in EtOH and 759 in IAA Hi-C data.

1029   Creating aggregate plots - To calculate and visualise the contact enrichment at
1030 hotspots in the EtOH and IAA Hi-C, we used the R package 'shaman'. Firstly, we used
1031 the function 'shaman_generate_feature_grid' to calculate the enrichment profile at
1032 EtOH and IAA hotspots. Using the weighted centre for each hotspot, represented as
1033 a 2D interval we used the function to build grids for the EtOH and IAA hotspots in the
1034 HiC data at 3 specific bands, 100k – 1MB, 1MB – 5MB, 5MB – 10MB. A range of
1035 250kb was visualised around the weighted centre. The grid was built by taking all
1036 combinations interval1 and interval2 of the EtOH and IAA hotspot centres, with each
1037 combination termed a 'window'. Hotspots were not filtered for size or shape. A score
1038 threshold of 60 was used to focus on enriched pairs, those windows that did not
1039 contain at least one point with a score of 60 were discarded. Each window was then
1040 split into 1000nt bins and the windows were summed together to generate a grid
1041 containing the observed and expected contacts. We visualised the grid using
1042 'shaman_plot_feature_grid' using 'enrichment' mode and a plot_resolution value of
1043 6000, due to the large range being visualised.

1044
1045

1046 **STORM – Immunolabelling and imaging.**
1047 Two clones of RAD21^{mAC}-OsTIR cells were seeded at a density of 30,000 cells per
1048 well per 400ul) onto poly-L-lysine coated 8-well chamber slides (Lab-Tek™ 155411)
1049 overnight. Each clone was treated with either ethanol (EtOH) or Auxin (IAA) for
1050 4hr and then fixed with PFA 4% (Alfa Aesar) for 10 min at room temperature and rinsed
1051 with PBS three times for 5 min each. The cells were shipped to the Cosma Lab after
1052 fixation for STORM processing and imaging. Cells were permeabilized with 0.3%
1053 Triton X-100 in PBS and blocked in blocking buffer (10% BSA – 0.01 % Triton X-100
1054 in PBS) for one hour at room temperature. Cells were incubated with primary
1055 antibodies (see Table 2) in blocking buffer at 1:50 dilution. Cells were washed three
1056 times for 5 min each with wash buffer (2% BSA – 0.01 % Triton X-100 in PBS) and
1057 incubated in secondary antibody. For STORM imaging, home-made (Bates et al.,
1058 2007) dye pair labeled secondary antibodies were added at a 1:50 dilution in blocking
1059 buffer and were incubated for 45 min at room temperature. Cells were washed three
1060 times for 5 min each with wash buffer.

1061   STORM imaging was performed on an N-STORM 4.0 microscope (Nikon)
1062 equipped with a CFI HP Apochromat TIRF 100x 1.49 oil objective and an iXon Ultra
1063 897 camera (Andor) and using Highly Inclined and Laminated Optical sheet
1064 illumination (HILO). Dual color STORM imaging was performed with a double activator
1065 and single reporter strategy by combining AF405_AF647 anti-Goat secondary with
1066 Cy3_AF647 anti-Rabbit secondary antibodies. Sequential imaging acquisition was
1067 performed (1 frame of 405 nm activation followed by 3 frames of 647 nm reporter and
1068 1 frame of 560 nm activation followed by 3 frames of 647 nm reporter) with
1069 10 ms exposure time for 120000 frames. 647 nm laser was used at constant ~2
1070 kW/cm^2 power density and 405 nm and 560 nm laser powers were gradually increased
1071 over the imaging. Imaging buffer composition for STORM imaging was 100 mM
1072 Cysteamine MEA (Sigma-Aldrich, #30070) - 5% Glucose (Sigma-Aldrich, #G8270) –
1073 1% Glox Solution (0.5 mg/ml glucose oxidase, 40 mg/ml catalase (Sigma-Aldrich,
1074 #G2133 and #C100)) in PBS.

1075

1076 **STORM imaging analysis and quantifications.**
1077 STORM images were analyzed and rendered in Insight3 (kind gift of Bo Huang, UCSF)
1078 as previously described (Bates et al., 2007; Rust et al., 2006). Localizations were
1079 identified based on a threshold and fit to a simple Gaussian to determine the x and y
1080 positions. Cluster analysis of CTCF, SA1 and SA2 STORM signal was performed as
1081 previously described (Ricci et al., 2015) to obtain cluster size and positions and to
1082 measure Nearest Neighbour distributions (NND) between clusters of the same protein
1083 in individual nuclei. NND between clusters' centroids of two different proteins
1084 (i.e. CTCF-SA1 and CTCF-SA2) was calculated by knnsearch.m Matlab function and
1085 the NND histogram of experimental data was obtained by considering all the NNDs of
1086 individual nuclei (histogram bin, from 0 to 500 nm, 5 nm steps). Simulated NNDs
1087 recapitulating random spatial distribution of cluster centroids were first obtained for
1088 each nucleus separately and then merged to calculate the simulated NND histogram
1089 (histogram bin, from 0 to 500 nm, 5 nm steps). The difference plot reports the
1090 difference between experimental NND and simulated NND. Quantification and
1091 analysis of STORM images was performed in Matlab and statistical
1092 analysis was performed in Graphpad Prism (v7.0e). The type of statistical test is
1093 specified in each case. Statistical significance is represented as indicated above.
1094 Insight3 Software used for STORM image processing has been generated (Huang et
1095 al., 2008) and kindly provided by Dr Bo Huang (UCSF). Graphpad Prism software
1096 used for statistical analysis can be found at: https://www.graphpad.com/scientific-
1097 software/prism/ MatLab software used for imaging data analysis can be found
1098 at: https://www.mathworks.com/products/matlab.html

1099

1100 **Mass spectrometry sample preparation and running**.
1101 SA1 immunoprecipitation samples were analysed by liquid chromatography–tandem
1102 mass spectrometry (LC-MS/MS). Three biological replicate experiments were carried

1103   out for MS and each included four samples, untreated (UT), treated with IAA for
1104   4hrs, siCon, or siSA1, generated as described above. A fourth technical replicate was
1105   also included for the UTR samples. Cells were fractionated to purify chromatin-bound
1106   proteins   as   above   and   immunoprecipitated   with   IgG-   or   SA1-bead
1107   conjugates. To maximise IP material for the MS, the antibody amount was increased
1108   to 15ug and the chromatin amount was increased to 2mg.

1109   The IP eluates were loaded into a pre-cast SDS-PAGE gel (4–20% Mini-
1110   PROTEAN® TGX™ Precast Protein Gel, 10-well, 50 μL) and proteins were run
1111   approximately 1 cm to prevent protein separation. Protein bands were excised and
1112   diced, and proteins were reduced with 5 mM TCEP in 50 mM triethylammonium
1113   bicarbonate (TEAB) at 37°C for 20 min, alkylated with 10 mM 2-chloroacetamide in 50
1114   mM TEAB at ambient temperature for 20 min in the dark. Proteins were then digested
1115   with 150ng trypsin, at 37°C for 3 h followed by a second trypsin addition for 4 h, then
1116   overnight at room temperature. After digestion, peptides were extracted with
1117   acetonitrile and 50 mM TEAB washes. Samples were evaporated to dryness at 30°C
1118   and resolubilised in 0.1% formic acid.

1119   nLC-MS/MS was performed on a Q Exactive Orbitrap Plus interfaced to a
1120   NANOSPRAY FLEX ion source and coupled to an Easy-nLC 1200 (Thermo Scientific).
1121   25% (first, second and fourth biological replicate) or 50% (third biological replicate) of
1122   each sample was loaded as 5 or 10 μL injections. Peptides were separated on a 27cm
1123   fused silica emitter, 75 µm diameter, packed in-house with Reprosil-Pur 200 C18-AQ,
1124   2.4 µm resin (Dr. Maisch) using a linear gradient from 5% to 30% acetonitrile/ 0.1%
1125   formic acid over 60 min, at a flow rate of 250 nL/min. Peptides were ionised by
1126   electrospray ionisation using 1.8 kV applied immediately prior to the analytical column
1127   via a microtee built into the nanospray source with the ion transfer tube heated to
1128   320°C and the S-lens set to 60%. Precursor ions were measured in a data-dependent
1129   mode in the orbitrap analyser at a resolution of 70,000 and a target value of 3e6 ions.
1130   The ten most intense ions from each MS1 scan were isolated, fragmented in the HCD
1131   cell, and measured in the orbitrap at a resolution of 17,500.

1132

1133   **Mass spectrometry analysis**
1134   Raw data was analysed with MaxQuant [52] version 1.5.5.1 where they were searched
1135   against   the   human   UniProtKB   database   using   default   settings
1136   (http://www.uniprot.org/). Carbamidomethylation of cysteines was set as fixed
1137   modification, and oxidation of methionines and acetylation at protein N-termini were
1138   set as variable modifications. Enzyme specificity was set to trypsin with maximally 2
1139   missed cleavages allowed. To ensure high confidence identifications, PSMs, peptides,
1140   and proteins were filtered at a less than 1% false discovery rate (FDR). Label-free
1141   quantification in MaxQuant was used with LFQ minimum ratio count set to 2 with
1142   'FastLFQ' (LFQ minimum number of neighbours = 3, and LFQ average number of
1143   neighbours = 6) and 'Skip normalisation' selected. In Advanced identifications,
1144   'Second peptides' was selected and the 'match between runs' feature was not

selected. Statistical protein quantification analysis was done in MSstats [53] (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, compared to the genome, 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 [34,35]. 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 [16] 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 and iCLIP

Crosslinking immunoprecipitation (CLIP) was performed as previously described[54]. 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 ɡ 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 ɭl 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. For RNaseH1 treatment, YFP-SA2 samples were split and either treated with PNK buffer alone or PNK buffer containing 50 U RNaseH1 for 15 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 s/l 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, fractionated and sonicated following the same protocol for chromatin fractionation with the variation of chromatin solubilisation in NaCl IP buffer (50mM Tris-HCL pH 7.5, 150mM NaCl, 1mM EDTA, 0.1% NP-40, 20% Glycerol, 1mM DTT). 1mg chromatin 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

1229   washed twice with NaCl IP buffer and added to pre-cleared lysates to
1230   immunoprecipitate proteins for 1h at 4°C. Samples were washed in NaCl IP buffer and
1231   eluted in 2x Laemmli buffer (Bio-Rad). Proteins were separated by SDS-PAGE on a
1232   4-20% gradient mini-PROTEAN® Precast Gel (Bio-Rad) and transferred onto PVDF
1233   membrane for visualization.

1234

1235   **VAST-TOOLS**
1236   VAST-TOOLS was used to generate Percent Spliced In (PSI) scores, a statistic which
1237   represents how often a particular exon is spliced into a transcript using the ratio
1238   between reads which include and exclude said exon. Paired-end RNA-seq datasets
1239   were submitted to VAST-TOOLS (v2.1.3) using the Mmu genome (Tapial J et al, Gen
1240   Res 2017). Briefly, reads are split into 50nt words with a 25nt sliding window. The 50nt
1241   words are aligned to a reference genome using Bowtie to obtain unmapped reads.
1242   These unmapped reads are then aligned to a set of predefined exon-exon junction
1243   (EJJ) libraries allowing for the quantification of alternative exon events. The output
1244   was further interrogated using a script which searches all hypothetical EEJ
1245   combinations between potential donors and acceptors within Stag1. PSI scores could
1246   be obtained providing there was at least a single read within the RNAseq data that
1247   supported the event, although we only considered events supported by a minimum
1248   of 50 reads. Calculated PSI values for each alternatively spliced exon (and shown in
1249   Fig 3d), as well as the average PSI reported in the text are shown below. See
1250   Table 3 for names of published datasets used in this analysis.

1251

Dataset	SA2 e32 PSI	Reads across junction	SA1 e31 PSI	Reads across junction
ENCODE HCT	21.92	202.98	98.99	588
Zuo HCT	19.18	94.75	97.91	278.06
ENCODE HeLa	19.95	90.57	96.36	156.1
AVG	**20.35**		**97.75**	

1252
1253

1254 **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	
siControl (scramble)	Dharmacon	Smartpool	D-001810-10-20	
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

1255

1256

1257 **Table 2. Antibodies used in this study.**

Protein	Company	Catalogue No.	Species	Figure References
SA1	Abcam	ab4455	mouse	1a, c, d, e, 2a, b, d, 3a, b, e, 4b, c, e, 5d
SA1	Abcam	ab4457	mouse	1i
SA2	Bethyl	A300-159	goat	1b, c, d, e, 2a, 3a, b, f, 4c, e, 5d
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, 5b, d, f
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, 4b
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, 4b
SSRP1	Abcam	ab26212	mouse	2d
INO80	Proteintech	18810-1-AP	rabbit	2d
ESYT2	Sigma-Aldrich	HPA002132	rabbit	2d
S9.6	Kerafast	ENH001	mouse	4b, c, e, 5f
RNase H2	Novus	NBP1-76981	rabbit	4b
AQR	Bethyl	A302-547A	rabbit	4b, e, 5f
Pol2	Covance	MMS-1289	mouse	4b
Mau2	Abcam	ab183033	rabbit	5b, d, f
NIPBL	Abbiotec	250133	rat	5b, d, f, S5
H3	Abcam	ab1791	rabbit	3e, f, 5b, d, f

37

Name (Secondary Abs)	Fluorophore	Company	Catalogue No.	Figure References
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

1258

1259 **Table 3. Published datasets used in this study.**

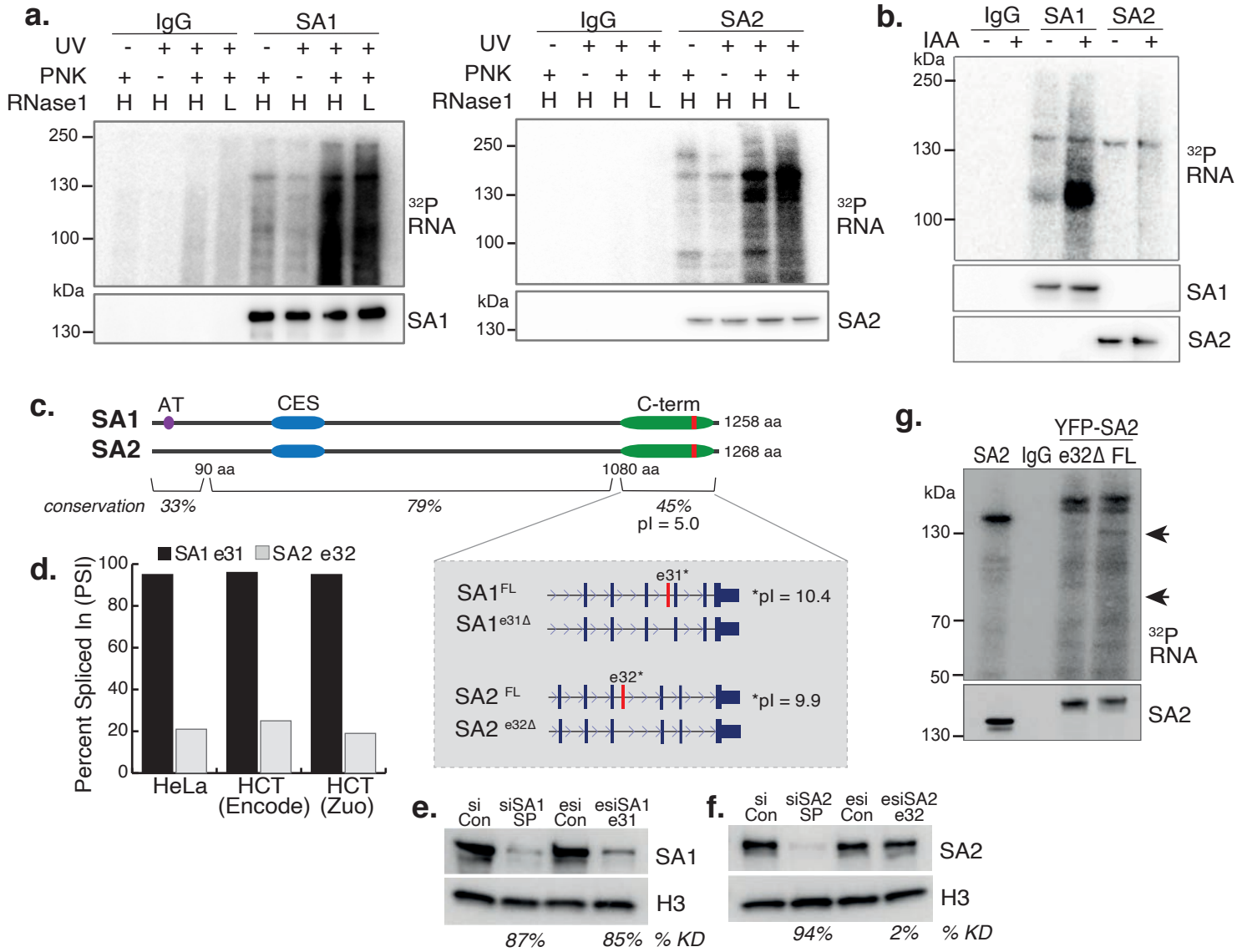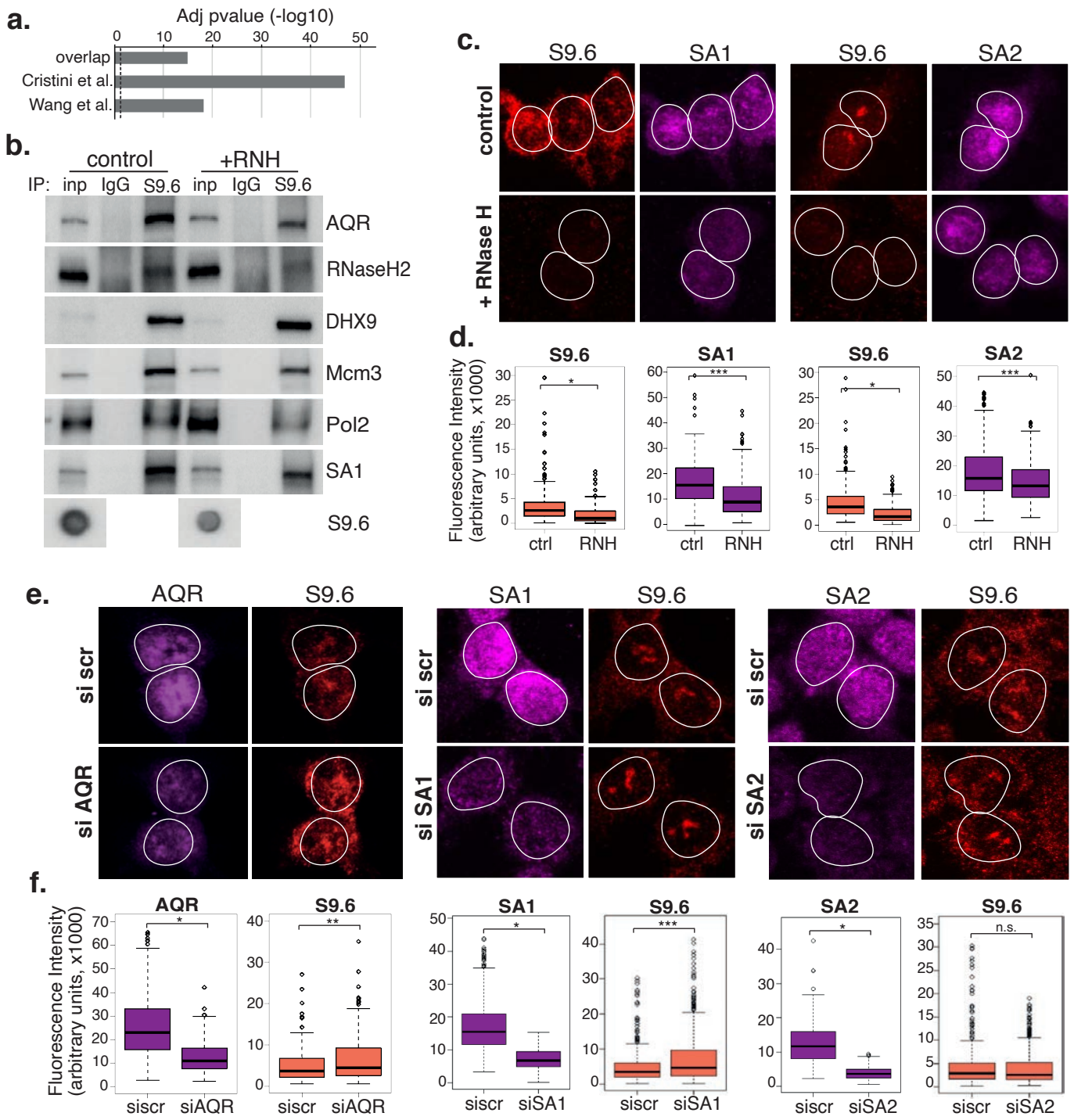| Accession no.	Analysis description	Publication DOI or Ref	Figure Reference
GSE104334	Long-range contact analysis of Hi-C datasets	10.1016/j.cell.2017.09.026	1i
GSE89729	Percent Spliced In (PSI) analysis of RNA-seq datasets	10.1172/jci.insight.91419	4d, "HCT Zuo"
GSM958749	Percent Spliced In (PSI) analysis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

1260

Figure 1.
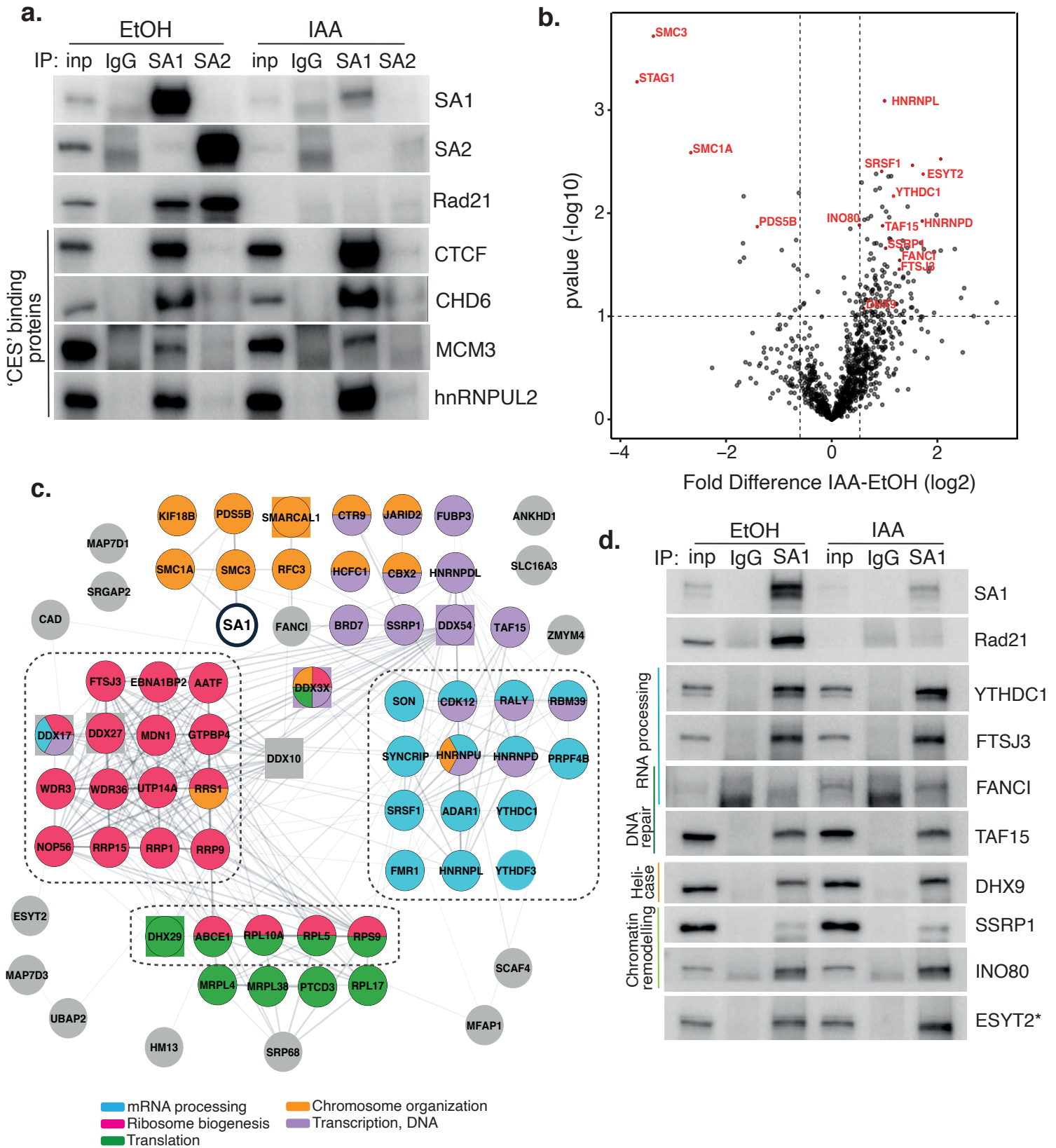
Figure 2.

Figure 3.

Figure 4.

Figure 5.