Identification of diverse cohesin-independent SA interactors that inform SA-specific functions.

Hayley Porter

A thesis submitted for the degree of
Doctor of Philosophy
UCL Cancer Institute
Department of Cancer Biology

Supervisors: Dr. Suzana Hadjur
Prof. Javier Herrero

31st March 2021

I, Hayley Porter confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

Cohesin complexes regulate genome organisation throughout the cell cycle. The molecular mechanisms by which cohesin governs this regulation are still not fully understood. SA1 and SA2 (SA) proteins are critical for cohesin function and are currently considered as core members of the complex due to their ubiquitous interaction with the ring protein members. This thesis investigates the role of the SA proteins in mediating interaction with CTCF. This work determines that following acute depletion of RAD21, SA proteins remain on chromatin and in complex with CTCF. The SA-CTCF interaction is dependent on the presence of nucleic acids and is localised at canonical cohesin binding sites in the genome. Mass spectrometry analysis further determines that cohesin-independent SA1, at least, does not just interact with CTCF, but also a range of additional proteins. The interactome of SA1 in the presence and absence of cohesin is identified. The SA1 interactome includes a wide variety of proteins spanning chromosome organisation, transcription, RNA processing, ribosome biogenesis, and translation. This thesis further reveals that cohesin-independent SA1 is enriched to proteins involved in RNA processing and ribosome biogenesis. R-loop proteins are highly enriched in the SA1 interactome and have previously been identified at sites encompassing all of these processes. Interaction of SA with R-loop structures and RNA itself is confirmed. A functional role for cohesin-independent SA is revealed in the association of cohesin with chromatin in the presence or absence of the NIPBL/MAU2 loader complex. While cohesin is found to load onto chromatin most efficiently in the presence of both SA and NIPBL/MAU2, this work reveals that SA alone can induce cohesin loading in a manner that is specifically linked to the abundance of R-loop structures present in the cell.

Impact Statement

DNA within a cell encodes the building blocks and regulatory factors, known as proteins, that determine the activities that cell can undertake. It is now well established that 3D organisation of the DNA impacts the ability of a cell to make specific proteins, thereby impacting its health and fate. Aberrations to proteins that mediate DNA organisation have been linked to developmental defects, such as Cornelia de Lange syndrome, and increasingly to cancer. Developmental disorders and cancer are important health and economic factors. This thesis investigates the mechanism by which a specific group of architectural proteins govern organisation of DNA in cells. Such understanding has implications in both academia and healthcare.

The research described in this thesis has two important academic impacts. Firstly, a methodology to efficiently isolate the protein interaction partners of the architectural proteins is described. Proteins rarely work alone in cells and so this is a key factor to understand their activity. This methodology will allow future work on the impact of each of these interaction partners on the activity of the architectural proteins. Variations of the methodology are also discussed that might allow this methodology to be used in altered conditions or to investigate alternative protein interaction partnerships. Hence, this work introduces a new technique to the academic community. Secondly, this thesis provides novel insight into the mechanism by which the architectural proteins drive DNA organisation. This is an important academic impact as it will enhance knowledge within the research community. This should help to drive even further discovery into the activity of these proteins.

The architectural proteins investigated in this thesis are also increasingly being recognised as drivers of cancer. In fact, one of the proteins, known as SA2, is one of only twelve factors that have been identified in four or more cancer types, underlying the importance of understanding its activity. Full comprehension of the basic mechanisms involved in structuring of DNA is important to determine

aetiology of disease caused by a change in the function of its regulatory proteins. Hence, the discovery of a novel method of their activity also has important implications for clinical work in these cancers.

The research described provides the groundwork understanding that can direct future cancer cell-/patient-specific studies to understand if this mechanism is altered with disease origin or evolution. If specifically altered in cancer, this mechanism could then be targeted as a treatment possibility to test if restoration of normal DNA organisation prevents cancer. Additional studies would need to be undertaken to move this work from bench to bedside, however, the described research represents the first steps in basic understanding required to achieve this. Therefore, this work has the potential to improve the health of many people in the future and to reduce the socio-economic impacts of related diseases.

Publications

A preprint of the work described in this thesis is available at:
doi: https://doi.org/10.1101/2021.02.20.432055

Submission for peer review is in progress.

Acknowledgements

First and foremost, I would like to thank my supervisor, Sue, for all her guidance, support, and mentorship. This thesis would not have been possible without her help. Thank you for helping me to develop as a scientist. Thank you also to my associate supervisor, Javier, for always being available to answer any of my bioinformatics queries. It very important that I thank CRUK for providing the funding that made this PhD research possible.

My greatest thanks to all past and present members of the Hadjur lab. It has been a pleasure to brainstorm, debate, and laugh with each of you. I would like to thank Stan for supporting all my questions at the beginning of my PhD, and for passing on chromatin and co-IP knowledge that helped to lay the foundation for this work. Sincere thank you to Chris and Waz for all the help with coding. Thank you to Yang for joining me in this work with such enthusiasm and great ideas and for helping to produce a great manuscript. Although we did not directly work on the same project, I thank Dubi and Sam for always engaging in my work and giving invaluable feedback and advice. It has been the greatest pleasure to work with you all. I would also like to extend sincere gratitude to Amandeep Bhamra and Silvia Surinova for all their guidance and help with the proteomics work in this study. And a special thanks to Amandeep for even answering my questions on the weekends.

For all the Friday drinks, I thank the pub crew, the stress release was essential to my PhD. I would like to especially thank Ellie, Sam, Maria, Jo, and Dhurva for being such great friends during this time. Last but by no means least I would like to thank my family and friends from back home in Ireland. To Ian, my constant supporter and sounding board – getting through this PhD would have been a lot more difficult without you. Stef, Rachel, Lucy, and Genevieve, thank you for all the visits and travelling fun through this time, I look forward to more in the future. And to my family, may you always be as loving, straight-up, and supportive of me as you are, you are the best. I appreciate everything you've done to enable me to pursue my interest in science.

Contents

Abstract....3
Impact Statement....4
Publications....6
Acknowledgements....7
List of figures....12
List of tables....17
List of abbreviations....18
1 Introduction....26
1.1 The cohesin complex....26
1.1.1 Structure of the cohesin complex....26
1.1.2 Regulators of cohesin....27
1.2 The canonical role of cohesin in sister chromatid cohesion....28
1.2.1 Association of cohesin with DNA....28
1.2.2 Stabilisation of cohesin-DNA interaction to established cohesion....30
1.2.3 Release of cohesin-DNA interaction to resolve cohesion....32
1.3 Non-canonical roles of cohesin in interphase....33
1.3.1 Cohesin and CTCF regulate chromatin architecture....35
1.4 Role of SA1 and SA2 in cohesin activity....43
1.4.1 Two distinct cohesin-SA complexes exist....43
1.4.2 The CES region of SA1 and SA2 promotes interaction with a variety of proteins....43
1.4.3 Interaction of SA1 and SA2 with CTCF....47
1.4.4 Distinct functions of SA1 and SA2....50
1.5 Function of the NIPBL/MAU2 loader complex....55
1.5.1 Topological loading of cohesin occurs in two steps....55
1.5.2 Mechanism of NIPBL-mediated loading....56

1.5.3 Role of NIPBL in cohesin activity....59
1.5.4 Where is cohesin loaded onto DNA?....61
1.5.5 Cohesin loading at sites of nucleic acid structure....63
1.5.6 Nucleic acid structure at sites of transcription....64
1.5.7 R-loop structures at sites of transcription....65
1.6 Research aims....70
2 Materials and methods....71
2.1 Cell culture....71
2.2 siRNA-mediated knockdowns....71
2.3 Plasmid DNA....72
2.4 Transient transfections of DNA plasmids....73
2.5 Chromatin Fractionation and co-immunoprecipitation....73
2.6 S9.6 IP and Dot Blot....75
2.7 Protein isolation and immunoprecipitation with the GFP-Trap....75
2.8 DNA/Protein isolation by ChIP protocol....76
2.9 Sodium Dodecyl Sulphate-Polyacrylamide gel electrophoresis (SDS-PAGE) and western blotting....78
2.10 ChIP-seq sample analysis....81
2.11 ChromHMM....81
2.12 Hi-C data and contact hotspots analysis....82
2.13 Mass spectrometry (MS) sample preparation and analysis....83
2.13.1 Full lane SA1 IP-MS....83
2.13.2 Banded SA and CTCF IP-MS....85
2.14 SLiMSearch analysis....85
3 CTCF and SA can interact independently of the cohesin ring....86
3.1 Introduction....86
3.2 Results....87
3.2.1 Characterisation of HCT116 RmAC OsTIR1 cells....87

3.2.2 Optimisation of co-immunoprecipitation protocol 92
3.2.3 CTCF and SA1 interact in the absence of RAD21 100
3.2.4 CTCF and SA1/SA2 colocalise on chromatin in the absence of cohesin 105
3.2.5 BiFC-ChIP, a new method to identify protein-protein interactions on chromatin 115
3.2.6 SA2 interacts with CTCF but not as robustly as SA1 128
3.2.7 Optimised nucleic acid digestion conditions for identification of CTCF and SA in complex 139
3.3 Discussion 147
4 SA1 interacts with a wide variety of proteins and with nucleic acids independently of RAD21 156
4.1 Introduction 156
4.2 Results 158
4.2.1 Banded mass spectrometry reveals a snapshot of SA1, SA2, and CTCF interactomes 158
4.2.2 Full lane SA1 mass spectrometry 165
4.2.3 SA1 interacts with CES-binding-motif-containing proteins in the presence and absence of RAD21 179
4.2.4 SA1 interacts with R-loop associated proteins 183
4.2.5 SA1 interacts with R-loops 186
4.2.6 SA1 interacts with RNA 200
4.3 Discussion 202
5 SA can load cohesin independently from the NIPBL-MAU2 loader complex 211
5.1 Introduction 211
5.2 Results 213
5.2.1 Optimising detection of NIPBL by western blot and investigating the role of NIPBL and MAU2 in cohesin loading 213
5.2.2 HCT116 RmAC OsTIR1 siRNA optimisation experiments 218

5.2.3 Interaction of SA1 with CTCF in the absence of RAD21 is not dependent on NIPBL 221
5.2.4 Reloading experiment optimisation 223
5.2.5 Influence of R-loop structures on RAD21 reloading 235
5.2.6 CTCF and SA colocalise at long-range contacts 246
5.3 Discussion 249
6 Conclusions & Future Perspectives 255
7 Supplemental Figures 262
8 References 272

List of figures

Figure 1: Schematic of the cohesin complex. 27
Figure 2: Structural comparison of cohesin regulators and SA. 44
Figure 3: Comparison of SA1 and SA2. 45
Figure 4: R-loop formation at sites of transcription.....66
Figure 5: Schematic of RAD21 depletion in HCT116 RmAC OsTIR1 cells. 88
Figure 6: Depletion of RAD21 in HCT116 RmAC OsTIR1 cells – polyclonal cells.....89
Figure 7: Depletion of RAD21 in HCT116 RmAC OsTIR1 cells – H2 and H11 clonal cells. 91
Figure 8: Schematic of chromatin fractionation protocol. 93
Figure 9: Optimisation of co-IP of cohesin complex members – Part 1.....94
Figure 10: Optimisation of co-IP of cohesin complex members – Part 2.....96
Figure 11: Increased RAD21 levels do not improve co-IP of with SA proteins..98
Figure 12: Optimisation of CTCF co-IP with SA proteins. 100
Figure 13: Co-IP of CTCF and SA in polyclonal HCT116 RmAC OsTIR1 cells treated with ethanol or auxin. 101
Figure 14: Co-IP of CTCF and SA in FACs-sorted clones of HCT116 RmAC OsTIR1 cells treated with ethanol or auxin. 104
Figure 15: Correlation analysis of cohesin and CTCF ChIP-seq replicates in ethanol- and auxin-treated samples. 107
Figure 16: Correlation analysis of the effect of auxin treatment on cohesin and CTCF ChIP-Seq samples. 109
Figure 17: Cohesin and CTCF peaks are correlated in ethanol conditions. 110
Figure 18: CTCF and SA are correlated in auxin conditions. 111
Figure 19: Colocalisation of CTCF and SA in ethanol and auxin conditions. ..112
Figure 20: ChIP-Seq analysis of cohesin and CTCF in ethanol and auxin conditions (6 hrs). 113
Figure 21: ChIP-Seq analysis of cohesin and CTCF in ethanol and auxin conditions (6 hrs) – SA heatmaps. 114
Figure 22: Schematic of BiFC assay obtained from Kodama and Hu (2012). .116
Figure 23: Validation of BiFC method. 118

Figure 24: Complemented Venus protein can be selectively immunoprecipitated.....119
Figure 25: Optimisation of BiFC-IP wash conditions – Part 1. 121
Figure 26: Optimisation of BiFC-IP wash conditions – Part 2. 122
Figure 27: BiFC-IP in ChIP buffer conditions. 124
Figure 28: BiFC-IP in ChIP conditions. 125
Figure 29: Optimising transfection concentration for BiFC-ChIP – Part 1. 126
Figure 30: Optimising transfection concentration for BiFC-ChIP – Part 2. 127
Figure 31: Effect of salt concentration and sonication conditions on co-IP of CTCF with SA2. 129
Figure 32: Analysis of naturally occurring variants of SA1 and SA2 and the contribution of a conserved C-terminal domain to interaction with CTCF. 130
Figure 33: Analysis of naturally occurring variants of SA1 and SA2 – mass spectrometry of IP bands. 134
Figure 34: Analysis of naturally occurring variants of SA1 and SA2 – mass spectrometry of CTCF IP bands.....136
Figure 35: DNA content of the SA1 IP. 139
Figure 36: Dependence of SA1-CTCF interaction on specific nucleic acid digestion – DNA analysis. 141
Figure 37: Dependence of SA1-CTCF interaction on specific nucleic acid digestion – RNA analysis. 142
Figure 38: Dependence of SA1-CTCF interaction on specific nucleic acid digestion – WB analysis. 143
Figure 39: Dependence of CTCF-SA2 interaction on specific nucleic acid digestion. 144
Figure 40: Co-IP of CTCF and SA in ethanol and auxin conditions (4 hrs) with further optimised nucleic acid digestion. 146
Figure 41: Panther protein classes enriched in SA1 banded MS1 and MS2…159
Figure 42: Panther protein classes enriched in overlap of SA1 banded MS1 & MS2. 161
Figure 43: Panther protein classes enriched in SA2 banded MS. 162
Figure 44: Panther protein classes enriched in CTCF banded MS. 164
Figure 45: Quality control analysis of SA1 full lane mass spectrometry replicates. 167
Figure 46: Correlation of SA1 IP-MS samples. 169

Figure 47: Subset view of the HCT116 SA1 interactome. 171
Figure 48: Subset view of SA1$^{\Delta Coh}$ interactome. 175
Figure 49: Effect of IAA treatment on SA1 interactors. 177
Figure 50: Validation of SA1 interactors in the presence and absence of RAD21. 178
Figure 51: FGF-like motif protein family members interact with SA1 in the presence or absence of cohesin. 180
Figure 52: FGF-like motif proteins interact with SA1 in the presence or absence of cohesin. 181
Figure 53: Solubilisation conditions for co-IP of FGF-like motif proteins with SA1. 183
Figure 54: Schematic of the structural similarity between the replication fork and an R-loop. 184
Figure 55: Enrichment of R-loop proteins in the SA1 interactome. 186
Figure 56: Overexpression of RNase H1 does not reduce enrichment of R-loop binding proteins with s9.6 – Intial test. 188
Figure 57: Overexpression of RNase H1 does not reduce enrichment of R-loop binding proteins with s9.6 – Optimisation of chromatin solubilisation conditions. 190
Figure 58: Overexpression of RNase H1 does not reduce enrichment of R-loop binding proteins with s9.6. 191
Figure 59: Optimisation of RNase A pre-treatment for s9.6 IP. 192
Figure 60: Overexpression of RNase H1 does not reduce enrichment of R-loop binding proteins with s9.6 following RNase A pre-treatment. 194
Figure 61: Optimisation of NEB RNase H digestion of R-loops 196
Figure 62: SA1 and SA2 co-IP with s9.6. 199
Figure 63: SA1 and SA2 interact with RNA in the presence and absence of RAD21. 201
Figure 64: Schematic of reloading experiment to test for role of SA proteins in association of cohesin ring with chromatin. 212
Figure 65: Detection of NIPBL knockdown by western blot analysis. 214
Figure 66: NIPBL and MAU2 knockdown differentially effect cohesin levels on chromatin. 216
Figure 67: Optimisation of siRNA transfection conditions. 219
Figure 68: Assessment of siNIPBL and auxin co-treatment. 221

Figure 69: SA1 can interact with chromatin and CTCF independently of NIPBL. 222
Figure 70: Re-assocation of RAD21 with chromatin in the absence of NIPBL in polyclonal HCT116 RmAC OsTIR1 cells. 225
Figure 71: Re-assocaition of RAD21 with chromatin in the absence of NIPBL in H11 HCT116 RmAC OsTIR1 cells. 227
Figure 72: The SA proteins contribute to cohesin loading in H11 HCT116 RmAC OsTIR1 cells. 229
Figure 73: MAU2 levels on chromatin are reduced with knockdown of SA1 and SA2. 230
Figure 74: Cohesin can re-associate with chromatin in the absence of NIPBL – representative WB. 232
Figure 75: Cohesin can re-associate with chromatin in the absence of NIPBL – densitometry. 234
Figure 76: Cohesin can re-associate with chromatin in the absence of NIPBL – alternative densitometry. 235
Figure 77: Optimisation of siAQR and siRNASH2A to increase R-loop levels. 238
Figure 78: RAD21 reassociation with chromatin is increased with knockdown of AQR – replicate 1. 240
Figure 79: RAD21 reassociation with chromatin is increased with knockdown of AQR – replicate 2. 243
Figure 80: RAD21 reassociation with chromatin is increased with increase in R-loops. 245
Figure 81: SA proteins localise to long-range chromatin contacts in the absence of cohesin. 247
Figure 82: SA proteins localise to long-range chromatin contacts at genic regions in the absence of cohesin - ChromHMM. 248

Supplemental Figure 1 262
Supplemental Figure 2 263
Supplemental Figure 3 265
Supplemental Figure 4 266
Supplemental Figure 5 268

Supplemental Figure 6 269
Supplemental Figure 7 270
Supplemental Figure 8 271

List of tables

Table 1: Details of siRNA used. 72
Table 2: Component make up of resolving and stacking gels used for SDS-PAGE. 78
Table 3: Details of antibodies used. 80
Table 4: Chip-seq datasets used. 82
Table 5: Alignment of SA1 exon31 and SA2 exon32. 132
Table 6: Detection of naturally occurring variants of SA1. 133
Table 7: Detection of naturally occurring variants of SA2. 135
Table 8: Analysis of naturally occurring variants of SA1 and SA2 – SA peptides identified in mass spectrometry of CTCF. 137
Table 9: Analysis of naturally occurring variants of SA1 and SA2 – SA2 peptides identified in IP-MS of SA2. 138
Table 10: Count of proteins detected in the SA1 full lane mass spectrometry replicates. 166
Table 11: Ontology of high confidence SA1 cohesin-independent interactors. 173
Table 12: FGF-like motif proteins identified in SA and CTCF MS experiments. 179
Table 13: Summary of technical differences between s9.6 IP experiments. 187

List of abbreviations

3D	Three-dimensional
AATF	Apoptosis-antagonizing transcription factor
AAVS1	Adeno-Associated Virus Integration Site 1
ADAR1	Adenosine deaminase acting on RNA 1-A
AID	Auxin-inducible degron
AML	Acute myeloid leukaemia
AQR	Aquarius
ATM	Ataxia telangiectasia mutated
ATP	Adenosine triphosphate
ATR	Ataxia telangiectasia and Rad3 related
B1R1	Biological replicate 1 technical replicate 1
B2R1	Biological replicate 2 technical replicate 1
BAMBI	BMP and activin membrane-bound inhibitor homolog
BiFC	bimolecular fluorescence complementation
BOP1	Block of proliferation 1
BORIS	Brother of the Regulator of Imprinted Sites
BRCA1	Breast cancer type 1 susceptibility protein
bZIP	Basic Leucine Zipper
CBX2	Chromobox homolog 2
CBX3	Chromobox homolog 3
CdLS	Cornelia de Lange syndrome
CES	Conserved Essential Surface
CHD	Chromodomain helicase DNA-binding
CHD1	Chromodomain helicase DNA-binding protein 1
CHD4	Chromodomain helicase DNA-binding protein 4
CHD6	Chromodomain helicase DNA binding protein 6
CHD8	Chromodomain helicase DNA-binding protein 8
CLIP	UV cross-linking and immunoprecipitation
CNC	Cohesin non-CTCF
COBALT	Constraint-based alignment tool
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
Cryo-EM	Cryogenic electron microscopy

CTCF	CCCTC-binding factor
CTCFL	CTCF like
DAPI	4′,6-diamidino-2-phenylindole
DDK	Dbf4-dependent kinase
DDX11	DEAD/H-BOX HELICASE 11
DHX9	DExH-Box helicase 9 (also known as ATP-dependent RNA helicase A)
DNA	Deoxyribonucleic acid
DNMT1	DNA methyltransferase 1
DPP3	Dipeptidyl-peptidase 3
DPP9	Dipeptidyl peptidase 9
DRIP	DNA-RNA immunoprecipitation
DSB	Double-strand break
EDTA	Ethylenediaminetetraacetic acid
EGTA	Ethylene glycol-bis(β-aminoethyl ether)-N,N,N′,N′-tetraacetic acid
eIF	eukaryotic initiation factor
EIF3I	Eukaryotic translation initiation factor 3 subunit I
ENCODE	Encyclopaedia of DNA Elements
ESCO1	Establishment of sister chromatid cohesion N-acetyltransferase 1
ESCO2	Establishment of sister chromatid cohesion N-acetyltransferase 2
ESYT2	Extended synaptotagmin-2
EtOH	Ethanol
ExPASy	Expert Protein Analysis System
F1	Fraction 1
F2	Fraction 2
FACS	Fluorescence-activated cell sorting
FACT	Facilitates chromatin transcription
FANCI	Fanconi anaemia, complementation group I
FANCM	Fanconi anaemia, complementation group M
FDR	False discovery rate
FISH	Fluorescence in situ hybridization

FMR1	Fragile X mental retardation 1
FP	Fluorescent protein
FRAP	Fluorescence recovery after photobleaching
FTSJ3	pre-rRNA 2'-O-ribose RNA methyltransferase FTSJ3
G1	Gap/growth 1 phase
G2	Gap/growth 2 phase
G4	G-quadruplex
G418	Geneticin
GADD45A	Growth arrest and DNA-damage-inducible protein GADD45A
GBM	Glioblastoma multiforme
GFP/eGFP	Green fluorescent protein/enhanced Green fluorescent protein
GO	Gene Ontology
GRCh38	Genome Reference Consortium Human Build 38
GST	Glutathione S-transferase
H1	Clonal HCT116 RmAC OsTIR1 cell line H1
H11	Clonal HCT116 RmAC OsTIR1 cell line H11
H2	Clonal HCT116 RmAC OsTIR1 cell line H2
H3	Histone H3
H3K27me3	Tri-methylation of lysine 27 on histone H3
H6	Clonal HCT116 RmAC OsTIR1 cell line H6
HA	Hemagglutinin
HCAEC	Human coronary artery endothelial cells
hCG_31253	Far upstream element-binding protein 3 (also known as FUBP3)
HEAT	Huntingtin-elongation factor 3-protein phosphatase 2A-TOR1
HM13	Histocompatibility Minor 13
HMEC	Human mammary epithelial cells
HNRN	Heterogeneous nuclear ribonucleoprotein
HNRNPD	Heterogeneous nuclear ribonucleoprotein D
HNRNPU	Heterogeneous nuclear ribonucleoprotein U-like protein (also known as SAF-A)
HNRNPUL2	Heterogeneous nuclear ribonucleoprotein U-like protein 2 (also known as SAF-A2)
HP1	Heterochromatin protein 1

HR	Homologous recombination
I152L	Isoleucine to Leucine mutation of amino acid 152
IAA	Indole-3-acetic acid
ICR	Imprinting control region
IF	Immunofluorescence
IFNG	Interferon gamma
IGF2	Insulin-like growth factor 2
IgG	Immunoglobulin G
IGH2	Immunoglobulin heavy chain 2
INO80	Chromatin-remodelling ATPase INO80 (also known as INOC1)
IP	Immunoprecipitation
KCl	Potassium chloride
KEGG	Kyoto Encyclopaedia of Genes and Genomes
LC	Liquid chromatography
LDS	Lithium dodecyl sulfate
log2FC	Log2-transformed fold change
LRC	Long-range contact
Mau2	MAU2 chromatid cohesion factor homolog (also known as Scc4)
MCM2	Minichromosome maintenance protein 2 homolog
MCM3	Minichromosome maintenance complex component 3
MCM6	Minichromosome maintenance complex component 6
MDC1	Mediator of DNA damage checkpoint protein 1
MEF	Mouse embryonic fibroblasts
mESCs	Mouse embryonic stem cells
MISA	MIcroSAtellite identification tool
MLH1	MutL protein homolog 1
MMR	Mismatch repair
mRNP	Messenger RNP (messenger ribonucleoprotein)
MS	Mass spectrometry
MS1	Banded mass spectrometry run 1
MS2	Banded mass spectrometry run 2
MSI/MIN	Microsatellite instability

NaCL	Sodium chloride
NB	Non-bound
NCAPD2	Non-SMC Condensin I Complex Subunit D2
NCBI	National Center for Biotechnology Information
NEB	New England Biolabs
NHEJ	Non-homologous end joining
NIPBL	Nipped-B-like protein (also known as Scc2)
NOP56	Nucleolar protein 56
NP-40	nonyl phenoxypolyethoxylethanol
NUMA1	Nuclear mitotic apparatus protein 1
OsTIR1	*Oryza sativa* Transport inhibitor response 1
PAGE	Polyacrylamide gel electrophoresis
PARP1	Poly [ADP-ribose] polymerase 1
PDB	Protein Data Bank
PBS	Phosphate-buffered saline
PCR	Polymerase chain reaction
PDS5A	Sister chromatid cohesion protein PDS5 homolog A
PDS5B	Sister chromatid cohesion protein PDS5 homolog B (also known as APRIN)
POGZ	Pogo transposable element derived with ZNF domain
Pol II	RNA polymerase II (also known as RNAP II)
POL2	DNA polymerase epsilon catalytic subunit A
PP2A	Protein phosphatase 2
PRC1	Polycomb repressive complex 1
PRC2	Polycomb repressive complex 2
Rad21	Double-strand-break repair protein rad21 homolog
RIF1	Rap1-interacting factor 1
RISC	RNA-induced silencing complex
RmAC	Rad21-mini auxin-inducible degon-mClover
RNA	Ribonucleic acid
RNASEH	Ribonuclease H
RNASEH1	Ribonuclease H1 (also known as RNH1)
RNASEH2A	Ribonuclease H2 subunit A (also known as RNH2A)
RPA	Replication protein A

RPA2	Replication protein A 32 kDa subunit
RPL	Ribosomal protein large subunit
RPL5	60S ribosomal protein L5
RPS	Ribosomal protein small subunit
RPS9	40S ribosomal protein S9
RT	Room temperature
SA	SA1 and SA2
SA1	Stromal antigen 1 (also known as STAG1)
SA2	Stromal antigen 2 (also known as STAG2)
SAF-A	Scaffold attachment factor A
SCF	S-phase kinase-associated protein 1–Cullin 1–F-box
SDS	Sodium dodecyl sulfate
SETX	Senataxin
SGO1	Shugoshin 1
SIN3A	Paired amphipathic helix protein Sin3a
SLiMSearch	Short, Linear Motif Search
SMARCA5	SWI/SNF-related matrix-associated actin-dependent regulator of chromatin subfamily A member 5
SMC1A	Structural maintenance of chromosomes protein 1A
SMC2	Structural maintenance of chromosomes protein 2
SMC3	Structural maintenance of chromosomes protein 3
SNF2H	Sucrose nonfermenting protein 2 homolog (also known as SMARCA5)
SON	SON DNA And RNA Binding Protein
SR	Serine–arginine
SRSF1	Serine/arginine-rich splicing factor 1
SSRP1	Structure specific recognition protein 1
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
SUZ12	SUZ12 Polycomb Repressive Complex 2 Subunit
SYNCRIP	Synaptotagmin-binding, cytoplasmic RNA-interacting protein
TAD	Topologically associated domain
TAF15	TATA-box binding protein-associated factor 5
TCF21	Transcription factor 21
TEAB	Tetraethylammonium Bromide

TES	Transcription end site
TET1	Ten-eleven translocation methylcytosine dioxygenase 1
TEV	Tobacco Etch Virus
TFIIH	Transcription factor II Human
TGX	Tris-Glycine eXtended
THOC1	THO Complex 1
TOP1	DNA Topoisomerase I
TOP2	DNA Topoisomerase 2
TOP3B	DNA Topoisomerase 3B
TRF1	Telomere repeat-binding factor 1
TSS	Transcription start site
U	Unit
UBAP2	Ubiquitin associated protein 2
UCSC	University of California, Santa Cruz
UTR	Untreated
UV	Ultraviolet
V	Volts
VC155	C-terminal fragment of Venus
VEH	Vehicle
VIM	Vimentin
VN155	Alternative N-terminal fragment of Venus
VN173	N-terminal fragment of Venus
WAPL	Wings apart-like protein homolog (also known as WAPAL)
WB	Western blot
WCE	Whole cell extract
WDHD1	WD repeat and HMG-box DNA-binding protein 1
WDR3	WD repeat-containing protein 3
YB1	Y box binding protein 1
YFP	Yellow fluorescent protein
YTHDC1	YTH domain-containing protein 1
YY1	Yin and yang 1
ZGPAT	Zinc finger CCCH-type with G patch domain-containing protein
ZMYM4	Zinc Finger MYM-Type Containing 4

ZNF326	Zinc Finger Protein 326

1

Introduction

3D organisation of the genome is critical for nuclear functions and cell fate. Organisation of chromosomes in the nucleus underlies gene expression, replication, and genome stability (Merkenschlager and Nora, 2016). As such, understanding chromatin organization and mechanisms of its regulation are pivotal biological questions. Cohesin and CTCF mediate important aspects of chromatin organization, however, the molecular mechanisms that govern cohesin and CTCF interaction and subsequent organization of chromosomes remain poorly understood. The focus of this thesis is to investigate interaction of cohesin with its regulatory proteins and how these proteins then shape cohesin activity, and ultimately, chromatin organization. In this thesis, the roles of SA1 and SA2 in cohesin biology are investigated. Here I discuss the background literature that exists regarding the roles of cohesin in genome organization, interaction of cohesin with CTCF, the role of SA1 and SA2 in cohesin activity, loading of cohesin on chromatin, and how the SA proteins may in fact function as regulators of the cohesin ring.

1.1 The cohesin complex

1.1.1 Structure of the cohesin complex

The cohesin complex, a member of the structural maintenance of chromosomes (SMC) family of ATPases, plays a central role in the structural changes that chromosomal DNA undergoes during the cell cycle. Cohesin is formed of three core proteins; SMC1, SMC3, and RAD21 (also known as Scc1), which come together to form a ring-shaped structure (Figure 1) (Sumara *et al.*, 2000; Haering

et al., 2002). A fourth cohesin subunit also associates with the core ring. In yeast this protein is known as Scc3 (Tóth *et al.*, 1999), while in higher eukaryotes, two Scc3 orthologs exist, termed SA1 and SA2 (Losada *et al.*, 2000). The Smc proteins each fold back on themselves in an antiparallel coiled-coil interaction, generating a long, flexible 'arm' section, with two globular 'head' domains at one end and a 'hinge' domain at the other (Melby *et al.*, 1998; Haering *et al.*, 2002; Hirano and Hirano, 2002). SMC1 and SMC3 come together at their hinge domains to form a V-shaped structure, with the gap between their head domains bridged by RAD21, as shown in Figure 1. Interaction of the SMC head domains (within one SMC protein or between both proteins) brings together an ATP binding site (walker A motif) from the N-terminal globular domain with a walker B motif from the C-terminal globular domain, forming the ATPase module of cohesin (Haering *et al.*, 2002).

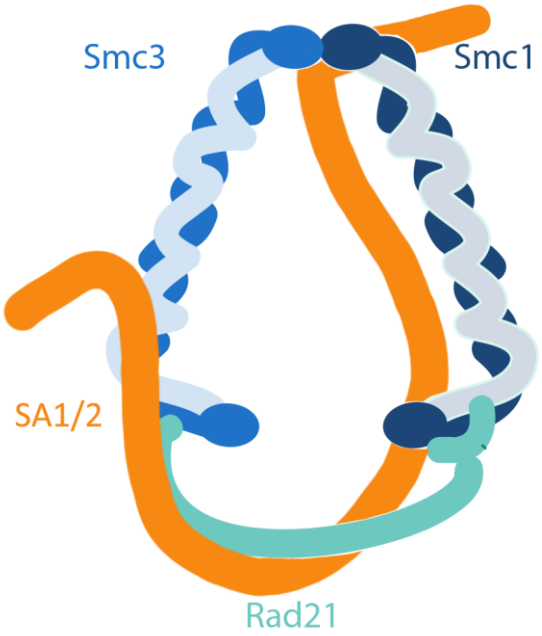


Figure 1: Schematic of the cohesin complex. The core cohesin ring is shown in blue and green and sitting within the 'U' surface of SA1/2, shown in orange.

1.1.2 Regulators of cohesin

A number of HEAT (Huntingtin-elongation factor 3-protein phosphatase 2A-TOR1) repeat proteins interact with the cohesin complex and influence its activity. Work in *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* has

shown that while cohesin has an intrinsic ability to bind DNA, Nipped-B-like protein (NIPBL or Scc2) and its co-factor MAU2 (Scc4) are required for efficient loading of cohesin onto chromatin, via activation of cohesins ATPase activity (Ciosk *et al.*, 2000; Murayama and Uhlmann, 2013). Disassociation of cohesin from chromatin is mediated by Wings apart-like protein homolog (Wapl) and the Pds5 protein (exists as two isoforms; Pds5A and Pds5B)(Shintomi and Hirano, 2009; Sutani *et al.*, 2009). Finally, the transcription factor CCCTC-binding factor (CTCF) is thought to determine the distribution of cohesin on chromatin (Parelho *et al.*, 2008; Rubio *et al.*, 2008; Wendt *et al.*, 2008). RAD21 and SA2 may act as a hub for binding of these regulators to cohesin (Hara *et al.*, 2014; Li *et al.*, 2020). More details on each of the cohesin regulators is discussed below.

1.2 The canonical role of cohesin in sister chromatid cohesion

The cohesin complex can topologically encircle and entrap DNA, thus physically holding replicated sister chromatids together in a process known as sister chromatin cohesion. This binding is coordinated from S phase until the onset of anaphase allowing for the synchronous segregation of the two chromatids and ensuring faithful replication and separation of homologous chromosomes (Michaelis, Ciosk and Nasmyth, 1997; Sonoda *et al.*, 2001; Nasmyth and Haering, 2005). The role of cohesin in sister chromatid cohesion has been studied for over 20 years, revealing valuable insight into molecular mechanisms of cohesin activity and important regulators. Key discoveries are discussed below.

1.2.1 Association of cohesin with DNA

Initial association of cohesin with chromatin occurs prior to DNA replication, in late G1 in yeast cells (Michaelis, Ciosk and Nasmyth, 1997) and in telophase of the previous cell cycle in mammalian cells (Darwiche, Freeman and Strunnikov, 1999; Sumara *et al.*, 2000). This difference is thought to be mediated by availability of free cohesin and is discussed more in section 1.2.3. Scc2 was found to be essential for proper sister chromatid cohesion, although it did not complex

with cohesin (Michaelis, Ciosk and Nasmyth, 1997; Furuya, Takahashi and Yanagida, 1998). Subsequently, Scc2 was found to complex with Scc4 and be required for association of cohesin with DNA (Tóth *et al.*, 1999; Ciosk *et al.*, 2000). Ciosk *et al.* (2000) collected chromatin pellets from yeast cells synchronized from G1 to anaphase with wild-type or mutant *scc2* and *scc4* (*scc2-4)*. In wild-type cells, Scc1 and Scc3 were detectable in the chromatin pellet following release from G1 arrest and decreased at anaphase. Smc1 was detected in the chromatin pellet at all of the timepoints tested. In contrast, in *scc2-4* mutants Scc1 and Scc3 were not detectable in the chromatin pellet and Smc3 levels were considerably reduced. In whole cell extracts the protein levels were similar in wild-type and *scc2-4* mutant samples, so the authors asked whether *scc2-4* mutation was affecting complex assembly. Smc1 and Scc1 could be detected in Scc3 immunoprecipitates from wild-type and *scc2-4* mutant extracts. Hence, it was determined that Scc2-4 was required for association of cohesin with chromatin. Following establishment of cohesion, Scc2-4 was found to be dispensable for maintenance of cohesion, confirming specificity of Scc2-4 for cohesin loading onto chromatin (Ciosk *et al.*, 2000; Lengronne *et al.*, 2006). The role of Scc2-4 in loading cohesin onto chromatin was similarly confirmed for homologous proteins in Xenopus and human cells (Gillespie and Hirano, 2004; Watrin *et al.*, 2006). Further details of the role of Scc2 in cohesin activity and the molecular mechanism of cohesin loading is discussed in section 1.5.

Two key characteristics of the cohesin complex led to the hypothesis that cohesin may topologically entrap DNA inside the SMC and RAD21 protein ring. Firstly, all four of the complex subunits are required for efficient association of each other with chromatin (Tóth *et al.*, 1999; Ciosk *et al.*, 2000). Although it is worth noting that this reliance is not complete as Smc1 may be able to bind to chromatin in early G1 in the absence of Scc1 (Ciosk *et al.*, 2000). Secondly, cleavage of the cohesin protein backbone releases cohesin from yeast chromosomes (Uhlmann, Lottspelch and Nasmyth, 1999). Evidence of topological loading arose from the fact that artificial cleavage of Scc1 or Smc3 releases cohesin from mini-chromosome *in vitro* and in chromosomal spreads and cleavage of mini-chromosomes releases bound cohesin (Gruber, Haering and Nasmyth, 2003; Ivanov and Nasmyth, 2005). Characterization of DNA release by opening of cohesin ring shape also supports the hypothesis of topological binding to DNA

(Chan *et al.*, 2012; Buheitel and Stemmann, 2013; Eichinger *et al.*, 2013; discussed in section 1.2.3).

Two models of cohesin-DNA topological interaction have been suggested in the literature. Firstly, the “embrace” model predicts that two DNA strands are held in the ring of a single cohesin complex (Haering *et al.*, 2008). This model was described following the finding that chemical fusion of all cohesin subunit interfaces allowed denaturation and identification of DNA dimers with single cohesin rings (Haering *et al.*, 2008; Gligoris *et al.*, 2014). Alternatively, the “handcuff” model predicts that cohesin rings each hold one DNA strand and two cohesin complexes interact together to hold the DNA stands together (Zhang *et al.*, 2008). This model was described following the finding that differentially tagged RAD21 or SMC proteins can all co-IP themselves but differentially tagged SA1 and SA2 do not. Thus, it was proposed that cohesin rings can interact via their SA subunit (Zhang *et al.*, 2008). Sufficient evidence to prove or disprove either model has not yet been achieved and it is possible that cohesin can function via a mix of the two mechanisms.

1.2.2 Stabilisation of cohesin-DNA interaction to established cohesion

During sister chromatid cohesion cohesin transforms to a more stable “cohesed” state. Fluorescence recovery after photo-bleaching (FRAP) of EGFP-tagged SMC3 and SA1 and Halo-tagged RAD21 in fly and mammalian cells has revealed two modes of cohesin binding to chromatin – a more transient mode and a more stable mode (Gerlich *et al.*, 2006; Kueng *et al.*, 2006; Gause *et al.*, 2010; Hansen *et al.*, 2017). The stable mode of cohesin binding is specific to the G2 phase of the cells cycle and in mammalian cells increases cohesin residence time from ~25mins to > 6hrs (Gerlich *et al.*, 2006; Kueng *et al.*, 2006; Hansen *et al.*, 2017). Single molecule tracking has further identified a third pool of chromatin-bound cohesin whose dissociation constant suggests non-topological association (Hansen *et al.*, 2017). It is not known if this third pool represents cohesin bound to chromatin independently of NIPBL or is actively loaded.

Stabilisation of cohesin is achieved by acetylation of conserved lysine residues in Smc3 by Eco1 acetyltransferases (Ben-Shahar *et al.*, 2008; J. Zhang *et al.*, 2008; Chan *et al.*, 2012). Two Eco1 orthologs exist in mammalian cells, ESCO1 and ESCO2. While both ESCO1 and ESCO2 contribute to proper sister chromatid cohesion, ESCO1 is active throughout the cell cycle and ESCO2 activity is specific to S phase at the establishment of sister chromatid cohesion (Hou and Zou, 2005; Alomer *et al.*, 2017). The effect of SMC3 acetylation was found to stabilise cohesin by shifting the dynamics of antagonistic cohesin regulators sororin and the interaction partners WAPL and PDS5 (Nishiyama *et al.*, 2010).

Sororin was identified as a cell cycle-regulated protein that accumulates in S phase and is degraded as cells exit mitosis, and is required for proper sister chromatid cohesion (Rankin, Ayad and Kirschner, 2005). Sororin is not required for association of cohesin with chromatin, however, RNAi-mediated knockdown of sororin decreases the stabilized pool of G2 cohesin by ~1/2 in Hela cells, indicating a role in the stabilization of chromatin-bound cohesin (Schmitz *et al.*, 2007; Ladurner *et al.*, 2016). Nishiyama *et al.* (2010) determined that SMC3 acetylation facilitates increased interaction of sororin with cohesin and sororin promotes cohesion in a manner dependent on a conserved C-terminal FGF motif. IF of Xenopus chromosomes further revealed that sororin and WAPL codepletion induces cohesion defects that mimic WAPL depletion alone, indicating that sororin antagonizes WAPL function in cohesion. WAPL interacts with the cohesin regulator PDS5 via multiple FGF motifs to induce dissociation of cohesin from chromatin (Shintomi and Hirano, 2009). Given that sororin and WAPL both contain FGF motifs, Nishiyama *et al.* (2020) proposed that sororin-mediated stabilization of cohesin may occur by displacement of WAPL from cohesin-bound PDS5. Displacement of WAPL from PDS5 could be observed in solution, however, sororin did not displace WAPL from Xenopus chromatin extracts, suggesting continued interaction with cohesin (Nishiyama *et al.*, 2010). Therefore, the molecular mechanism of sororin-mediated stabilization remains unclear.

1.2.3 Release of cohesin-DNA interaction to resolve cohesion

Coordinated release of cohesin binding is required for proper segregation of replicated sister chromatids. In budding yeast, the cysteine protease separase proteolytically cleaves Scc1 triggering sister separation and metaphase to anaphase transition (Uhlmann, Lottspelch and Nasmyth, 1999; Uhlmann *et al.*, 2000). In higher eukaryotes, cohesin release occurs in two stages. The “prophase pathway” first removes the majority of cohesin from chromosome arms during prophase (Waizenegger *et al.*, 2000). Release is dependent on WAPL, which together with PDS5 is proposed to release cohesin by transiently opening the SMC3-RAD21 interface. Evidence for this proposal comes from the fact that fusion of the SMC3 and klesin interface reduces cohesin turnover in budding yeast, Drosophila, and human cells (Chan *et al.*, 2012; Buheitel and Stemmann, 2013; Eichinger *et al.*, 2013). Reciprocally, mutation of four residues in the RAD21 portion of the interface abolished the G2 stabilized pool of cohesin and prevented stabilization by depletion of WAPL, indicating a state of constant destabilization (Huis In’t Veld *et al.*, 2014). Finally, Wapl-Pds5- and Wapl-mediated release of an N-terminal TEV-cleaved fragment of Scc1 from Smc3 has been shown *in vitro* and in yeast, demonstrating specific opening of the Scc1-Smc3 interface (Murayama and Uhlmann, 2014; Beckouët *et al.*, 2016). Phosphorylation of sororin and SA2 mediates the switch from sororin-mediated stabilization to Wapl-Pds5-mediated release, potentially by weakening the sororin-Pds5 interaction (Hauf *et al.*, 2005; Nishiyama *et al.*, 2013).

Cohesin at centromeric chromatin is protected from release during the prophase pathway by a complex of the phosphatase PP2A and shugoshin. PP2A dephosphorylates SA2 and sororin preventing Wapl-mediated release (Kitajima *et al.*, 2006; Riedel *et al.*, 2006; Liu, Rankin and Yu, 2013; Nishiyama *et al.*, 2013). Wapl and shugoshin peptides compete for binding to an SA2-Scc1 subcomplex *in vitro* suggesting that shugoshin can also protect centromeric cohesin by sterically hindering Wapl binding (Hara *et al.*, 2014). An FGF-like motif in shugoshin mediates interaction with the same sites in SA2-Scc1 as Wapl, accounting for this competitive binding (Hara *et al.*, 2014; Li *et al.*, 2020; discussed in more detail below). Mutations to the shared Wapl and shugoshin interaction site in SA2 and overexpression of shugoshin does not completely

abolish Wapl binding, indicating that Wapl may bind to SA2-Scc1 via multiple interaction surfaces (Hara *et al.*, 2014). Wapl also contains multiple FGF motifs, raising the question of whether this prevents easy antagonism of Wapl-mediated release and maintains the majority of cohesin in dynamic association with chromatin.

Centromeric cohesin is released from the sister chromatids at the onset of anaphase by separase-mediated cleavage of Scc1, thus allowing final segregation of the replicated sister and cytokinesis (Waizenegger *et al.*, 2000; Hauf, Waizenegger and Peters, 2001). RAD21 and SA2 dissociate from chromatin with similar kinetics, as shown in human cells by IF analysis of RAD21 and SA2 distribution during the cell cycle (Tóth *et al.*, 1999; Prieto *et al.*, 2002). In contrast, select reports of retained DNA interaction for the Smc3 protein until late anaphase have been made in yeast and the parasite *Trypanosoma brucei* (Tanaka *et al.*, 1999; Bessat and Ersfeld, 2009). In *T. brucei*, this Smc3 signal was assessed by IF and found to be detergent-sensitive, indicating that Smc3 may only interact with DNA weakly (Bessat and Ersfeld, 2009). The authors propose that this may represent a soluble pool of Smc3 that is ready to reassociate with chromatin upon interaction with newly translated Scc1. The existence of the prophase pathway in higher eukaryotes has similarly been suggested to have been selected through evolution as it produces a large pool of soluble cohesin that facilitates proper cohesin occupancy in subsequent telophase and G1 phases (Tedeschi *et al.*, 2013).

1.3 Non-canonical roles of cohesin in interphase

Cohesin's ability to topologically bind chromosomal DNA has also implicated it as an important regulatory factor during interphase for DNA repair (Birkenbihl and Subramani, 1992; Bauerschmidt *et al.*, 2009) and structural organization of chromosome, including regulation of enhancer-promoter contacts (Wendt *et al.*, 2008; Hadjur *et al.*, 2009; Kagey *et al.*, 2010) and demarcation of chromosomes into TADs (Seitan *et al.*, 2013; Sofueva *et al.*, 2013; Zuin *et al.*, 2014). Mutations of cohesin ring components (Seitan *et al.*, 2013; Sofueva *et al.*, 2013; Zuin *et al.*,

2014; Rao *et al.*, 2017a) and regulator proteins (Haarhuis *et al.*, 2017; Schwarzer *et al.*, 2017; Wutz *et al.*, 2017) leads to disruption of these structures, highlighting its functional importance in genome organization (discussed in detail below).

Evidence for a role of cohesin in gene regulation was originally observed in *S. cerevisiae* when the Smc proteins where found to be required for insulation of the transcriptionally repressed *HMR* locus (Donze *et al.*, 1999). At the same time, in drosophila, the cohesin loader ortholog Nipped-B/delangin was identified as an architectural factor that facilitated expression between the *cut* homeobox gene and an upstream enhancer (Rollins, Morcillo and Dorsett, 1999). Cohesin was subsequently found to repress expression of *cut* as decreasing cohesin via mutating the *smc1* or *pds5* genes or RNAi-mediated depletion of SA2 increased *cut* expression (Rollins *et al.*, 2004; Dorsett *et al.*, 2005). These papers led to the idea that cohesin could interfere with enhancer-promoter interactions and Nipped-B could regulate the levels of cohesin on chromosomes mediating the interference.

Evidence for a role of cohesin in gene regulation widely comes from study of the developmental disorder Cornelia de Lange syndrome (CdLS). CdLS is predominantly characterized by loss-of-function mutations in NIPBL, as well as mutations in SMC and RAD21 cohesin subunits (Krantz *et al.*, 2004; Tonkin *et al.*, 2004; Deardorff *et al.*, 2007; Minor *et al.*, 2014). Importantly, CdLS derived mutations do not invariably induce cohesion defects, indicating a secondary critical role for cohesin (Kaur *et al.*, 2005; Castronovo *et al.*, 2009). Analysis of patient-derived cell lines or cDNA has revealed broad transcriptional deregulation, indicating a functional link between cohesin and gene expression (Liu *et al.*, 2009; Yuan *et al.*, 2015). Experiments in model organisms confirmed transcriptional deregulation at example genes following cohesin and NIPBL mutation (Horsfield *et al.*, 2007; Rhodes *et al.*, 2010; Remeseiro *et al.*, 2013). Key insight into the molecular mechanism by which cohesin regulates gene expression was derived from identification and investigation of interaction with CTCF.

1.3.1 Cohesin and CTCF regulate chromatin architecture

1.3.1.1 CTCF is a DNA binding protein with insulator function

CTCF is an 11 zinc finger protein that is highly conserved in higher eukaryotes, but absent in yeast and *Caenorhabditis elegans* model organisms (Filippova *et al.*, 1996; Moon *et al.*, 2005; Heger *et al.*, 2012). CTCF was identified as a repressor of *myc* via binding to CCCTC repeats in the *myc* promoter (Lobanenkov *et al.*, 1990). Subsequently, CTCF was characterised as an insulator protein due to its ability to disrupt enhancer-promoter interaction (Bell, West and Felsenfeld, 1999; Bell and Felsenfeld, 2000; Hark *et al.*, 2000; Kanduri *et al.*, 2000; Szabó *et al.*, 2000). Insulator proteins can also buffer spread of active and repressive histone marks at the border of heterochromatin and euchromatin – a phenomenon known as positive effect variegation (Eissenberg *et al.*, 1992; Recillas-Targa *et al.*, 2002). Genome-wide ChIP-seq analyses of CTCF binding identified significant overlap with H3K27me3 at the boarder of heterochromatin domains and insulation of the H3K27me3 signal by CTCF (Bartkuhn *et al.*, 2009; Cuddapah *et al.*, 2009).

Genome-wide ChIP-seq and *in vitro* analyses established a consensus motif of ~20bp for CTCF binding (Filippova *et al.*, 1996; Ohlsson, Renkawitz and Lobanenkov, 2001; Kim *et al.*, 2007; Parelho *et al.*, 2008; Wendt *et al.*, 2008; Nakahashi *et al.*, 2013). Mutation of individual zinc fingers and structural studies indicate that CTCF binds to its consensus motif via combined binding of multiple zinc fingers across the sequence, thereby allowing variation in the consensus motif without loss of CTCF binding (Filippova *et al.*, 1996; Nakahashi *et al.*, 2013; Hashimoto *et al.*, 2017). As well as establishing the basis of CTCF distribution on DNA, ChIP-seq analyses have revealed the global characteristics of CTCF binding sites. Approximately half of all CTCF binding sites are located in intergenic regions, with the remain CTCF variably distributed in promoter regions (12-20%), introns (22-30%), and exons (5-12%) (Kim *et al.*, 2007; Chen *et al.*, 2012).

1.3.1.1.1 Regulation of CTCF-DNA interaction

Binding of CTCF to DNA is regulated at multiple levels. Positioning of nucleosomes within a CTCF consensus motif disrupts CTCF binding *in vitro*

(Kanduri *et al.*, 2002). *In vivo* bound CTCF is flanked by nucleosomes with specific spacing and sites that lose CTCF binding following differentiation of mESCs show nucleosome rearrangement (Clarkson *et al.*, 2019). Whether nucleosome remodelling or CTCF binding occurs upstream of each is unclear. Owens *et al.* (2019) suggest that CTCF drives nucleosome positioning as nucleosomes reposition into CTCF sites following depletion in mESCs, however, it is not clear if the loss of CTCF has any downstream effect on nucleosome remodelling complexes. Nucleosomes flanking CTCF have also been found to be enriched for the histone variant H2A.Z (Fu *et al.*, 2008).

CTCF occupancy is also regulated by methylation. Methylation of cytosines within its consensus motif impairs CTCF binding (Bell and Felsenfeld, 2000; Hark *et al.*, 2000; Wang *et al.*, 2012). *In vitro* structural analysis further suggests that methylation at different positions within the CTCF consensus binding sequence differentially impairs or enhances CTCF binding, allowing plasticity in CTCF-DNA interaction (Hashimoto *et al.*, 2017).

A variety of interaction partners have also been identified that influence CTCF behaviour and DNA binding. For example, CTCF can interact with and activate PARP1, thereby inducing inactivation of DNMT1 and preserving the unmethylated status of its binding site (Guastafierro *et al.*, 2008; Zampieri *et al.*, 2012). Protein interactions of CTCF have widely been identified with DNA binding proteins (YY1, YB1, and Kasio) chromatin proteins (Suz12, Taf1), and helicases (CHD8, p68) (Ziatanova and Caiafa, 2009; Ghirlando and Felsenfeld, 2016). Cohesin has been identified as one of the most important interaction partners of CTCF (discussed in detail below).

Finally, transcription and interaction with RNA has recently been identified as an important regulator of CTCF binding to DNA. Deletion of RNA-binding domains in CTCF zinc fingers 1 and 10 reduces binding at promoters, intronic, and intergenic regions (Saldaña-Meyer *et al.*, 2019). Differential loss was observed at these regions depending on the specific RNA-binding domain deleted, further suggesting variability in RNA-mediated stabilisation at distinct DNA loci. At the *IGF2/H19* locus p68 interacts with CTCF alongside RNA to stabilise interaction with cohesin (Yao *et al.*, 2010). This raises the question of whether CTCF-RNA

interaction stabilises CTCF binding to DNA via stabilisation of protein interaction partners. Deletion of RNA binding domains in CTCF did not decrease the bulk levels of cohesin interacting with CTCF, however, potential effect on stabilisation of the interaction was not assessed (Saldaña-Meyer *et al.*, 2019).

1.3.1.2 Role of cohesin and CTCF in chromatin loops

Genome-wide ChIP-seq analyses revealed extensive overlap of CTCF and cohesin (Parelho *et al.*, 2008; Rubio *et al.*, 2008; Wendt *et al.*, 2008). The importance of this co-localisation was revealed as CTCF was shown to recruit cohesin to specific sites. This was shown as siRNA-mediated knockdown of CTCF induced loss of cohesin from CTCF sites assessed by ChIP-qPCR, despite not affecting overall levels of cohesin on chromatin (Parelho *et al.*, 2008; Rubio *et al.*, 2008; Wendt *et al.*, 2008). Reciprocal knockdown of RAD21 had differential effect on CTCF occupancy in two studies, despite investigation of Hela cells in both cases (Parelho *et al.*, 2008; Wendt *et al.*, 2008). Together these studies indicated that CTCF interaction mediates the distribution of cohesin on chromatin rather than the loading of cohesin and that cohesin may contribute to CTCF localisation. To the best of my knowledge, FRAP of cohesin in CTCF depleted cells has not been reported in the literature so stability of the cohesin on chromatin in the absence of CTCF is not clear. This raised the question of whether recruitment of cohesin to CTCF binding sites contributed to the insulator function attributed to both proteins.

Chromosome conformation capture (3C) analyses linked CTCF insulator function to looping of chromatin *in cis* (Kurukuti *et al.*, 2006; Splinter *et al.*, 2006; Majumder *et al.*, 2008). At the *Igf2/H19* imprinting locus, the paternally expressed *Igf2* and maternally expressed *H19* genes share enhancers and the imprinting control region (ICR) located downstream of *H19* and upstream of *Igf2* is specifically methylated in the paternal allele (Leighton *et al.*, 1995; Thorvaldsen, Duran and Bartolomei, 1998). Hence, CTCF specifically binds to the ICR in the non-methylated maternal allele and restricts access of *Igf2* and the shared enhancer (Bell and Felsenfeld, 2000). Kurukuti *et al.* (2006) identified chromatin loops specific to the paternal and maternal alleles – at the paternal allele looping between the enhancers and *Igf2* promoter was observed, whereas, at the maternal allele a tight loop excluding *Igf2* from the enhancer was observed.

Importantly, mutation of CTCF binding sites in the maternal allele depleted CTCF binding and abolished chromosomal contacts within the tight loop. Similarly, depletion of CTCF in mouse cells was found to reduce chromosomal contacts in the *β-globin* locus (Splinter *et al.*, 2006). Thus, CTCF was implicated as a regulator of chromatin architecture.

Knockdown of RAD21 reduced insulation at the chicken *β-globin* locus and the *Igf2/H19* imprinting locus similarly to CTCF knockdown in two separate studies, indicating that cohesin may contribute to CTCF-mediated insulator function (Parelho *et al.*, 2008; Wendt *et al.*, 2008). This led to the hypothesis that cohesin may anchor chromatin contacts *in cis* in interphase cells at sites defined by CTCF binding. Direct evidence confirming this hypothesis came from 3C of the *IFNG* locus in human T-cells. Hadjur *et al.* (2009) differentiated T-cells and identified that cell-type specific induction of *IFNG* expression corresponded with cell-type specific chromosomal contracts at CTCF/cohesin co-bound sites. Importantly, siRNA-mediated knockdown of RAD21 strongly reduced these contacts and reduced *IFNG* expression, without loss of CTCF occupancy (Hadjur *et al.*, 2009). Hence, this work determined that CTCF alone cannot maintain chromosome loops at the human *IFNG* locus and provided evidence of cohesin-mediated loops *in cis*. Similarly, cohesin was shown to play a key role in maternal *Igf2* suppression and the chromatin loops mediated by CTCF binding at the *Igf2/H19* imprinting locus and at the mammalian *β-globin* locus (Nativio *et al.*, 2009; Chien *et al.*, 2011). Thus, cohesin was identified as an architectural protein capable of mediating chromosomal contacts *in cis* and thereby regulating expression of bound genes. The importance of cohesin in anchoring chromatin loops was also indicated as similar 3C results were observed for chromatin loops at developmentally important genes at cohesin non-CTCF sites in mESCs (Kagey *et al.*, 2010). Cohesin at non-CTCF sites was found to overlap with tissue-specific transcription factors and the mediator complex, indicating a role for cohesin in cell fate specification (Kagey *et al.*, 2010; Schmidt *et al.*, 2010).

1.3.1.3 Role of cohesin and CTCF in chromatin domains

Hi-C is a 3C derivative that pairs proximity ligation of chromatin fragments with massively parallel sequencing to generate a genome-wide map of contact frequencies (Lieberman-Aiden *et al.*, 2009). Hi-C experiments have revealed that

chromosomes are segregated into topologically associated domains (TADs) or regions of chromosomes within which chromatin contacts are enriched, but between which, contacts are depleted (Dixon *et al.*, 2012; Nora *et al.*, 2012; Sexton *et al.*, 2012). This spatial partitioning has been associated with specific histone modifications (Sexton *et al.*, 2012), coordinated gene expression (Nora *et al.*, 2012), and DNA replication timing (Pope *et al.*, 2014). TAD borders show a high degree of overlap across cell types despite changes in transcriptional programmes and are conserved across species (Dixon *et al.*, 2012; Rao *et al.*, 2014; Vietri Rudan *et al.*, 2015), indicating biological importance.

Mammalian TADs encompass genomic regions ~1Mb in size and characteristically have borders that engage in frequent interaction in the cell population, generating a characteristic peak in the Hi-C map (Dixon *et al.*, 2012; Nora *et al.*, 2012). CTCF and cohesin binding was found to correlate highly with these strong borders, suggesting a role in segregation of genomes into interaction domains (Dixon *et al.*, 2012; Phillips-Cremins *et al.*, 2013; Rao *et al.*, 2014; Vietri Rudan *et al.*, 2015). Genetic partial depletion of cohesin in non-cycling cells produced global architectural changes, including loss of contacts between CTCF sites and loss of interactions within TADs, indication a role of cohesin and CTCF in mediating contacts at multiple scales (Seitan *et al.*, 2013; Sofueva *et al.*, 2013). Inducible cleavage of RAD21 produced similar loss of interaction within TADs, suggesting that cohesin mediates these contacts by topological binding to chromatin (Zuin *et al.*, 2014). Increased interaction between TADs were observed with CTCF depletion and variably with cohesin depletion, despite evidence of retained domain borders, suggesting that the boundaries of domains are maintained by the loops within (Sofueva *et al.*, 2013; Zuin *et al.*, 2014).

More recent studies utilizing auxin-inducible degradation of RAD21 or CTCF have achieved more rapid and robust depletion of the proteins. Rao *et al.* (2017) robustly degraded RAD21 in HCT116 cells and observed loss of contacts within and at the border of TADs. Whether loss of all border signal was due to increased depletion of RAD21, change to the timescale of depletion, or investigating in dividing cells is not clear. As previously, CTCF binding was relatively unaffected, although its average ChIP-seq signal intensity may have been slightly reduced. Auxin-mediated depletion of CTCF also resulted in loss of contacts within and at

the border of TADs, although ~18% of borders did remain (Nora *et al.*, 2017). Transcriptional changes were observed in all of the cohesin and CTCF depletion samples, albeit to varying degrees, indicating a functional link between the chromatin contacts observed and gene regulation (Seitan *et al.*, 2013; Sofueva *et al.*, 2013; Nora *et al.*, 2017; Rao *et al.*, 2017a). In cohesin-deficient cells, gene expression was decreased for highly expressed genes and increased for lowly expressed genes, suggesting that cohesin impacts gene regulation by equalising expression of genes across the genome (Seitan *et al.*, 2013; Sofueva *et al.*, 2013). The importance of these chromatin interaction can be seen as even in cells with modest changes to global transcription levels, expression of cell fate genes such as *myc* can be linked cell-type specific CTCF-mediated architecture (Hyle *et al.*, 2019).

1.3.1.4 Role of cohesin and CTCF in chromatin compartments

Imaging and Hi-C analysis have also determined that genomes are organised into megabase-scale territories of the chromatin that preferentially interact (Lieberman-Aiden *et al.*, 2009; Rao *et al.*, 2014; Wang *et al.*, 2016). These territories are termed compartments and were originally subdivided into A and B types based on gene-rich and -poor characterisation, respectively (Lieberman-Aiden *et al.*, 2009). Higher resolution Hi-C maps have identified further subcompartments that correspond with distinct replication timing and histone marks (Yaffe and Tanay, 2011; Rao *et al.*, 2014). Compartmentalisation of genomes has been shown to shift between differentiated cell types, indicating correlation with the gene expression programme of the cell (Dixon *et al.*, 2012). However, this shift was not completely pervasive, indicating that some gene expression changes are not sufficient to induce a switch in compartment or vice versa. The role of cohesin and CTCF in formation and maintenance of compartments is not clear. Historically mild enhancement of compartmentalisation or no change has been reported with cohesin or CTCF depletion (Seitan *et al.*, 2013; Sofueva *et al.*, 2013; Zuin *et al.*, 2014; Nora *et al.*, 2017; Rao *et al.*, 2017a). Whereas, more recently, increased compartmentalisation has been observed with NIPBL knockdown or a short timecourse of acute cohesin depletion (Schwarzer *et al.*, 2017; Wutz *et al.*, 2017).

1.3.1.5 Role of cohesin in phase separation

Finally, many different forms of condensates or foci of concentrated proteins and nucleic acids have been discovered in the nucleus, primarily by imaging techniques. Examples include cajal bodies, nuclear speckles and polycomb bodies (Mao, Zhang and Spector, 2011; Banani *et al.*, 2017). Importantly, the condensates differ from canonical protein complexes as they show the physical properties of phase separation, or liquid-liquid demixing, from the surrounding nucleus. These properties include; i) the spherical shape of the condensate, ii) fusion of condensates that touch together, with reformation of the spherical shape, and iii) molecules within the condensate are highly mobile and rearrange following photobleaching (Banani *et al.*, 2017). Condensate in the nucleus have also been linked to chromatin organisation. For example, Scaffold attachment factor A (SAF-A, also known as HNRNPU), oligomerises with chromatin associated RNAs to form condensates that are required for open chromatin conformation at active genes (Nozawa *et al.*, 2017; Fan *et al.*, 2018). A functional role for cohesin in this process was suggested as SAF-A, RAD21, and CTCF were identified in reciprocal co-IPs and depletion of SAF-A reduced RAD21 binding at 3816 sites and increased RAD21 binding at 535 sites by ChIP-seq (Fan *et al.*, 2018). Indeed comparison of the number of cohesin binding sites in the genome in a population of yeast cells and quantification of total chromatin-bound cohesin suggest that 5-20 cohesin molecules bind chromatin per site and cohesin may exist in cluster on chromatin (Weitzer, Lehane and Uhlmann, 2003). In addition, low levels of recombinant cohesin-SA1 incubated with DNA *in vitro* have been observed to frequently form foci, suggesting a molecular mechanism for aggregation of cohesin (Davidson *et al.*, 2019).

Ryu *et al.* (2021) assessed *in vitro* DNA organisation by yeast cohesin at higher, physiologically relevant concentrations of cohesin and observed condensation of DNA clusters. These clusters contained many cohesin complexes and characteristics of liquid droplets, including spherical shape that reformed when two clusters merged. Hence, cohesin may be involved in phase separation. DNA over 3 kbp in length was required for clustering to occur and simulations predict that when many cohesin binding sites are found on the DNA this phase separation behaviour produces contact maps reminiscent of compartments found in mammalian Hi-C maps. Finally, Ryu *et al.* (2021) observed bridging of DNA by

single cohesin complexes (plus its loader) and propose that part of cohesin may topologically bind to DNA while a second part of cohesin interacts with DNA as part of a condensate. A recent pre-print from our lab determined that overexpression of SA1 induces condensation of heterochromatin in mESCs (Pežić *et al.*, 2021). Intrinsically disordered regions and multivalency are characteristics of proteins involved in phase separation (Larson *et al.*, 2017). Intrinsically disordered regions were identified in the N- and C-terminal ends of SA1, suggesting it may be capable of mediating phase-separation of the condensates formed. Multivalency is required for the bridging-induced phase separation observed for cohesin *in vitro,* however, the molecular basis of this remains to be formally determined (Ryu *et al.*, 2021). Structural analysis of cohesin, its loader complex and DNA suggest that DNA may be contacted and bent by both SA1 and NIPBL (Higashi *et al.*, 2020; Shi *et al.*, 2020). This raises the question of whether multivalency in cohesin is achieve across multiple protein components and regulators.

In summary, cohesin has been identified as an architectural protein that regulates chromatin organisation at multiple levels in interphase cells and interaction with CTCF is important for this function. As such, understanding the molecular mechanism of cohesin-CTCF interaction is key to understanding cohesin functions at CTCF and non-CTCF sites. Interaction with CTCF is thought to be mediated by the SA proteins, hence, the role of the SA proteins in cohesin interactions and specific localisation, and the molecular mechanism of SA-CTCF interaction are discussed below.

1.4 Role of SA1 and SA2 in cohesin activity

1.4.1 Two distinct cohesin-SA complexes exist

Immunoprecipitation of SA1 and SA2 in Xenopus and Hela cells determined mutually exclusive interaction with SMC and RAD21 proteins, suggesting that two distinct forms of cohesin exist in cells; cohesin–SA1 and cohesin–SA2 (Losada *et al.*, 2000; Sumara *et al.*, 2000). Quantitative mass spectrometry of cohesin subunits in G1 Hela cells has also shown that the combined total of SA1 and SA2 molecules per cell is approximately equal to the individual amounts of the three ring subunits, suggesting that each cohesin ring interacts with one of SA1 or SA2 (Holzmann *et al.*, 2019). ChIP-seq experiments have documented overlap of SA1 and SA2 localisation in the genome (Kojic *et al.*, 2018; Cuadrado *et al.*, 2019; Casa *et al.*, 2020), suggesting that the two cohesin-SA proteins have the ability to localise to the same regions. It is not clear from bulk ChIP-seq experiments if this co-localisation occurs in the form of two cohesin molecules, one SA1-bound and one SA2-bound, or if it represents differential binding of cohesin-SA1 and cohesin-SA2 in different cells in the population. Casa *et al.* (2020) performed sequential ChIP, or “Re-ChIP”, to determine if SA1 and SA2 could co-occupy specific fragments of DNA together. No SA1 and SA2 co-occupancy was observed, whereas both SA1 and SA2 were detected following sequential ChIP from SMC3 (Casa *et al.*, 2020). Together these papers indicate that SA proteins interact with cohesin in a mutually exclusive manner.

1.4.2 The CES region of SA1 and SA2 promotes interaction with a variety of proteins

Yeast Scc3 and human SA are HEAT-repeat containing proteins (Hara *et al.*, 2014; Roig *et al.*, 2014). Human SA2 contains 17 HEAT repeats and bends significantly to form a ‘dragon’ shape with a snout and head at its N-terminal end and a sharp bend in its centre (Hara *et al.*, 2014). Interestingly, SA2, PDS5, and NIPBL/Scc2 all contain bent HEAT repeat regions that give them a similar, highly curved structure (Figure 2, Hara *et al.* 2014; Lee *et al.* 2016; Muir *et al.* 2016).

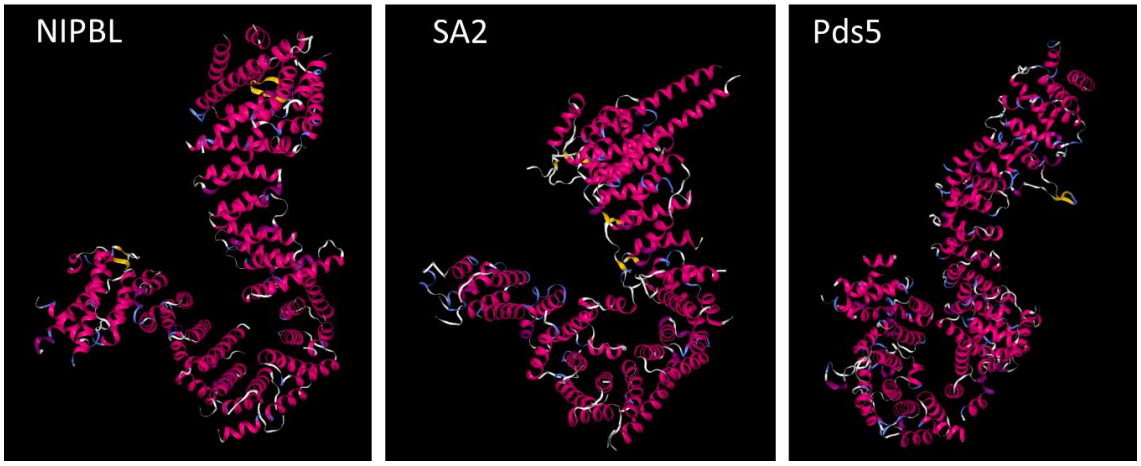


Figure 2: Structural comparison of cohesin regulators and SA. Crystal structure of NIPBL (PDB ID: 5ME3 (Chao et al., 2017)), SA2 (PDB ID: 4PJU (Hara et al., 2014)), and Pds5 (PDB ID: 5F0N (Lee et al., 2016)), coloured by secondary structure. All three proteins form hook-shaped structures with anti-parallel HEAT repeats forming their highly bent architecture.

SA1 and SA2 have 70% sequence homology (Figure 3A; Carramolino *et al.*, 1997; Losada *et al.*, 2000). This homology suggests that the SA proteins may overlap in functions mediated by these conserved domains. The crystal structure of SA2 in complex with a portion of RAD21 and the cryo-EM structure of SA1 in complex with cohesin, NIPBL, and DNA have been reported (Hara *et al.*, 2014; Shi *et al.*, 2020). The structure of the N- and C-terminal ends of the SA proteins have not been reported due to their disordered nature. The reported portions of SA1 and SA2 structurally align well, with a root mean square deviation of 2.44, and are topologically very similar, with a template modelling score of 0.91 (Figure 3B). Template modelling scores are reported between 0 and 1, with 1 representing a perfect match, indicating the high similarity between SA1 and SA2. This similarity in folding further suggests that the SA proteins may overlap in functions mediated by this central region.

Comparison of SA orthologs between species has identified a conserved N-terminal domain that is essential for cell survival, known as the conserved essential surface (CES) or stromalin conserved domain (Roig *et al.*, 2014; Orgil *et al.*, 2015). As well as showing sequence conservation across species, the CES regions of human SA1 and SA2 structurally align well, with a root mean square deviation of 1.26, and are topically similar, with a template modelling score of 0.81 (Figure 3C). This similarity in folding may suggest that the CES can function similarly in SA1 and SA2. The CES is conserved even in organisms lacking Pds5,

Wapl, and Scc2/4, highlighting the functional importance of this domain. The exact role the CES plays in cohesin biology is still unclear however amino acids in the CES have been shown to mediate interaction with RAD21, NIPBL, and CTCF.

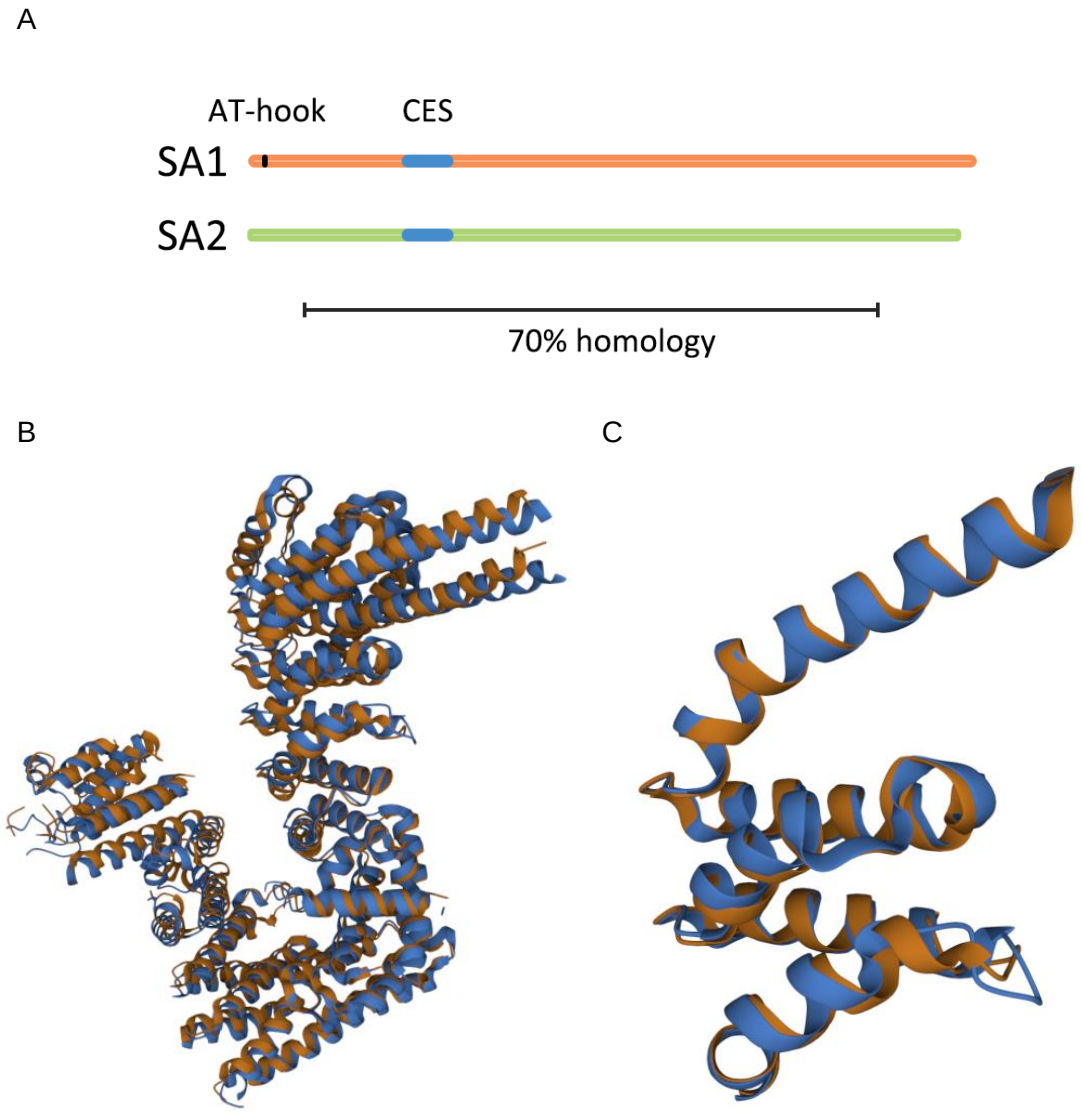


Figure 3: Comparison of SA1 and SA2. (A) Schematic of SA1 and SA2 comparison. A central conserved region with 70% homology is indicated. The conserved essential surface (CES) or stromalin conserved domain is indicated in blue (Roig et al., 2014; Orgil et al., 2015). An AT-hook domain in the N-terminus of SA1 is indicated in black (Bisht, Daniloski and Smith, 2013; Lin et al., 2016). (B) Snapshot of rigid pairwise structural alignment of SA2 (PDB ID: 4PJU (Hara et al., 2014)) and SA1 (PDB ID: 6WG3 (Shi *et al.*, 2020)). (C) Snapshot of rigid pairwise structural alignment of the CES of SA2 (PDB ID: 4PJU (Hara et al., 2014)) and SA1 (PDB ID: 6WG3 (Shi *et al.*, 2020)).

Mutational mapping of regions of recombinant human SA2 and RAD21 required for interaction *in vitro* suggested that the CES is required for interaction of SA proteins with RAD21 (Zhang *et al.*, 2013). This work was validated in yeast, where

random insertion of mutations in the CES region of the SA ortholog Scc3 abolished co-IP of the yeast RAD21 ortholog Mcd1 (Orgil *et al.*, 2015). However, mutation of regions upstream of the CES also abolished interaction with Mcd1, leaving the question of why the CES in particular is required for cell survival unanswered.

Peptide array and immunoprecipitation experiments determined that in yeast amino acids within CES are also involved in interaction with the cohesin loader Scc2. In addition, Scc3 was found to be essential for loading of cohesin onto DNA and simulation of cohesin ATPase activity (Murayama and Uhlmann, 2014). This suggested that the CES is required for loading of cohesin onto DNA via interaction with the loader complex. However, this interaction may not fully explain the essential function of the CES *in vivo* as the CES is also conserved in organisms that lack Scc2 and depletion of Scc3 in yeast cells does not completely abolish co-IP of Mdc1 with Scc2 or *in vitro* binding of the cohesin ring to DNA (Murayama and Uhlmann, 2014; Roig *et al.*, 2014; Orgil *et al.*, 2015). It is not clear if the cohesin loaded by Scc2 in these conditions can maintain proper chromosome organisation in the absence of Scc3. Indeed, despite binding to DNA, specific amino acids in the CES were required for Scc3 localisation at centromeric and ribosomal DNA (rDNA) loci in yeast and mutation greatly reduced condensation of chromatin at the rDNA locus (Orgil *et al.*, 2015). Hence, the CES may be required for efficient loading of cohesin, perhaps at a step distinct from interaction of the loader complex with the cohesin ring, and is involved in localisation of Scc3 and condensation of the bound chromatin.

Li et al. (2020) crystallised an N-terminal portion of CTCF in association with a SA2-RAD21 complex. The authors determined that the CES of SA2 formed the binding pocket for CTCF along with RAD21 amino acids bound within the CES. Two regulators of cohesin, WAPL and shugoshin, have also been found to interact with SA2-RAD21 via SA2's CES (Hara *et al.*, 2014). Repeating FGF motifs in WAPL have been shown to be involved in its interaction with SA2-RAD21 (Shintomi and Hirano, 2009), so, Li et al. (2020) devised a motif for binding to the CES by alignment of cohesin regulators containing FGF-like motifs (CTCF, WAPL, sororin, shugoshin, and NIPBL). The presence of this CES-binding motif across the human proteome identified known cohesin regulators as

well as a number of novel potential binding factors that were confirmed to interact with SA2-RAD21 by peptide array (Li *et al.*, 2020). This suggests that SA2 may localise cohesin across the genome by interaction with a range of proteins through its CES. The authors did not determine if SA1 also interacts with the FGF-like motif proteins via its CES.

1.4.3 Interaction of SA1 and SA2 with CTCF

SA1 and SA2 have been proposed to bridge the interaction between cohesin and CTCF. This was first shown by Xiao et al. (2011) who incubated GST-tagged CTCF with *in vitro*-translated SA1, SA2, SMC1, SMC3, or RAD21. Only SA1 and SA2 were pulled down with CTCF, a finding that was interpreted to mean that the SA proteins bridge interaction between the cohesin ring and CTCF (Xiao, Wallace and Felsenfeld, 2011). Immunoprecipitation of tagged fragments of CTCF and SA2 suggested that the C-terminus of CTCF and an N-terminal/central fragment of SA2 from amino acids 162 to 290 mediate the interaction. Further studies suggest that the C-terminus of CTCF may not be solely responsible for interaction with cohesin as its deletion does not prevent co-IP of SA1 (Saldaña-Meyer *et al.*, 2014) or the cohesin ring (RAD21 and SMC1A) (Hansen *et al.*, 2019).

CRISPR-Cas9-mediated deletion of CTCF zinc fingers 9 – 11 depletes CTCF and cohesin from ~5,000 sites in mouse B cell lymphoma cells (Vian *et al.*, 2018; Pugacheva *et al.*, 2020). Expression of the N-terminal domain of CTCF fused to zinc fingers 1 -11 was able to restore both CTCF and RAD21 to the lost sites in these cells, whereas the C-terminal domain of CTCF fused to zinc fingers 1 – 11 was only able to restore CTCF binding (Pugacheva *et al.*, 2020). This demonstrated that the N-terminal domain of CTCF is required for cohesin localisation at a subset of CTCF sites, at least. Similarly, investigation of a natural variant of CTCF determined that the N-terminus of CTCF confers ability to interact with RAD21. A natural variant of CTCF expressed in germ cells, known as CTCF like (CTCFL) or BORIS (Brother of the Regulator of Imprinted Sites), does not interact with cohesin (Nishana *et al.*, 2020; Pugacheva *et al.*, 2020). CTCF and BORIS share 74% identify across their central zinc finger domains but have variable N- and C-terminal domains (Loukinov *et al.*, 2002). CRISPR-Cas9-mediated replacement of BORIS N-terminus with CTCF N-terminus induced

interaction with RAD21 by co-IP (Nishana *et al.*, 2020) and co-localisation of BORIS and RAD21 by ChIP-seq (Loukinov *et al.*, 2002). Whereas, replacement of BORIS C-terminus with CTCF C-terminus did not induce RAD21 co-IP or co-localisation with BORIS, indicating that the N-terminus of CTCF specifically confers ability to interact with RAD21 (Nishana *et al.*, 2020; Pugacheva *et al.*, 2020).

The CTCF N-terminus was not sufficient to target RAD21 binding to sites not bound by CTCF in control cells, indicating that additional factors contribute to cohesin-CTCF interaction (Pugacheva *et al.*, 2020). Pugacheva *et al.* (2020) further determined that the first two zinc fingers in CTCF are required for efficient RAD21 co-localisation, suggesting that specific nucleic acid binding may also be important to the cohesin-CTCF interaction. The first two zinc fingers and the terminal ends of CTCF were still not sufficient to target RAD21 binding (Pugacheva *et al.*, 2020). RAD21 recruitment was only considered at a subset of sites in this study, however, these results suggest that tertiary structure of CTCF or interactions across the entirety of the protein (potentially with nucleic acids and/or proteins) are required for robust cohesin-CTCF interaction.

Finally, Li *et al.* (2020) used GST-tagged CTCF fragments to pull down a complex of SA2 and RAD21 together and identified amino acids 222–231 of CTCF as the exact portion of the N-terminus required for interaction. While a GST pull down method similar to that used by Xiao et al. (2011) was employed in this paper, incubation with a complex of SA2 and RAD21 together may account for the different fragment of CTCF identified as the cohesin interactor domain. Therefore, it is not definitively clear what regions of CTCF mediate interaction with cohesin, although it is possible that both N- and C-terminal regions of CTCF interact with members of the cohesin ring and the different methodologies used allowed differential detection of the multiple interactions. In addition, binding to two or more CTCF proteins at the base of chromatin loop structures *in vivo* could alter how cohesin complex members and CTCF interact.

These papers indicate that the SA proteins likely play an important role in the interaction of cohesin with CTCF and highlight the importance of this interaction for proper localisation of cohesin. However, these experiments utilised modified

proteins or *in vitro* techniques that may not tell the full story of interaction *in vivo*. Hansen *et al.* (2017) in fact suggest that cohesin and CTCF do not interact in a stable complex. The authors Halo-tagged CTCF in mESCs and utilised single-molecule tracking to estimate an average residence time of ~1min on chromatin. In contrast, FRAP of RAD21 estimated a residence time of ~22min (Hansen *et al.*, 2017). This suggests that CTCF-SA binding is a dynamic event and CTCF does not form a stable complex with cohesin. Co-IP of wild-type and Halo-tagged CTCF with RAD21, SMC1, and SMC3 was unchanged in the mESCs suggesting that interaction was not affected by Halo-tagging of CTCF (Hansen *et al.*, 2017). However, co-IP with SA proteins was not assessed, leaving it unclear if CTCF-SA interaction was disrupted by the Halo-tag. In contrast, *in vitro* assessment of CTCF residence on the *amyloid precursor protein* promoter sequence revealed a half-life of ~22hrs (Quitschke *et al.*, 2000). Hence, stable binding of CTCF can occur under specific conditions.

SMC3 is acetylated in interphase cells by ESCO1 (Alomer *et al.*, 2017). This suggests that stabilised cohesin-mediated loops may exist in interphase similar to that seen during sister chromatid cohesion. Indeed, in the absence of ATP, DNA replication, or transcription DNA loops have been observed for hours (Vian *et al.*, 2018). This suggested that stationary cohesin may be stabilised on chromatin and raised the question of whether cohesin may be protected from WAPL-mediated release in these situations. Wutz *et al.* (2020) identified preferential ESCO1-mediated acetylation of SMC3 in complex with SA1 compared to SA2 in Hela cells. CTCF was required for SMC3 acetylation and together ESCO1 and CTCF were required for a pool of cohesin-SA1 with an increased average residence time of ~4 hrs. Co-depletion of CTCF and WAPL prevented the loss of cohesin acetylation observed with knockdown of CTCF alone, suggesting that ESCO1 and CTCF interaction with cohesin-SA1 protect it from WAPL-mediated release from chromatin (Wutz *et al.*, 2020). Whether CTCF is also stabilised on chromatin at these sites or can mediate this activity while dynamically associating and dissociating is not clear. The authors suggest that cohesin-SA1 mediates long-range chromatin loops as SA2 knockdown increased SA1 levels on chromatin and increased the highest contact frequency from 200kb to 800kb, and SA1 knockdown does not increase the highest contact frequency compared to control (Wutz *et al.*, 2020). The dynamics of CTCF-SA interaction

under endogenous, unmodified conditions are not yet clear, however, this paper suggests that CTCF interaction with SA1 and SA2 has distinct functional outcomes.

1.4.4 Distinct functions of SA1 and SA2

Mutation of SA2 has been identified in numerous cancers, including, bladder cancer, Ewing sarcoma, glioblastoma multiform (GBM), and acute myeloid leukaemia (Rocquain *et al.*, 2010; Solomon *et al.*, 2011; Balbás-Martínez *et al.*, 2013; Guo *et al.*, 2013). In fact, SA2 is one of only twelve genes that contains statistically significant somatic point mutations in four or more cancer types (Lawrence *et al.*, 2014). SA1 mutation has also been linked to cancer, albeit to a lesser extent; Romero-Pérez et al. (2019) preformed a meta-analysis of 53,691 patient datasets from cBioPortal and identified that somatic mutation of SA1 and SA2 occur at a frequency of 0.9 and 2%, respectively (Romero-Pérez *et al.*, 2019). As such, understanding SA function is key to understanding the aetiology of these cancers and whether SA2-specific functions contribute to its prevalence in cancer.

The N- and C-terminal domains of SA1 and SA2 share the lowest homology and likely play a role in the functional specificities of each protein. For example, during cohesion, both SA proteins are involved in holding together sister chromatids, however, SA1 is specifically important for cohesion at telomeres and SA2 is specifically important for cohesion at centromeres (Canudas and Smith, 2009; Remeseiro, Cuadrado, Carretero, *et al.*, 2012). SA1 contains an AT-hook domain in its N-terminus that allows it to bind to AT-rich DNA, such as that found at telomeres (Bisht, Daniloski and Smith, 2013). SA1 may also be able to recognise telomeric DNA sequences, as it binds more strongly to telomeric repeats than scrambled control DNA of the same length *in vitro* (Lin *et al.*, 2016). Interestingly, telomere-specific SA1-mediated cohesion is more reliant on interaction with DNA and with the shelterin protein TRF1, but less so with the cohesin ring subunits SMC3 or RAD21, suggesting a cohesin-independent role for SA1 at telomeres (Bisht, Daniloski and Smith, 2013; Lin *et al.*, 2016). In contrast, SA2-mediated cohesion at centromeres requires the cohesin ring for DNA-DNA pairing (Bisht, Daniloski and Smith, 2013). SA2 does not contain an AT-hook domain and has

been shown to bind DNA in a sequence-independent manner. DNA recognition by SA2 may instead be mediated by structure; SA2 shows a high affinity for DNA ends, single-stranded gap DNA, and flap and fork DNA intermediate structures (Countryman *et al.*, 2018). As such, cohesin-SA2 could be enriched at sites of DNA repair, recombination, and replication. Both SA1 and SA2 diffuse on dsDNA, which is thought to represent a searching mode of cohesin before it stably binds to the sites, such as by the mechanisms discussed above (Lin *et al.*, 2016; Countryman *et al.*, 2018). Therefore, specific nucleic acid binding capabilities of SA1 and SA2 may impart distinct regulatory roles for the SA proteins in guiding cohesin activity and even an ability for SA1 to act independently of the cohesin ring.

Similarly, SA-specific responses to DNA damage has been reported. Preferential clustering of cohesin-SA2 was recorded at induced double-strand breaks (DSBs) in the DNA of Hela cells (Kong *et al.*, 2014). Recruitment of SMC1 and NIPBL to the DSB was dependent on the presence of SA2 in the cells. In contrast, cohesin-SA1 was not recruited to the sites of damage. Interestingly, generation of an SA1-SA2 chimera in which the C-terminus of SA1 was replaced with the C-terminus of SA2 induced recruitment of SA1 to sites of DSBs (Kong *et al.*, 2014). This again demonstrates the importance of the terminal ends of the SA proteins for specific functions. Furthermore, the C-terminus of SA2 alone was not sufficient to induce recruitment to DSBs, indicating that multiple domains in SA are required (Kong *et al.*, 2014).

1.4.4.1 Localisation of SA1 and SA2 at sites of transcription

As described above, SA2, at least, may localise cohesin to a range of FGF-like motif protein via its CES (Hara *et al.*, 2014; Li *et al.*, 2020). In line with these findings, ChIP-seq experiments have revealed that cohesin non-CTCF (CNC) sites exist across the genome, albeit at a reduced number compared to colocalised cohesin–CTCF sites (Kagey *et al.*, 2010; Schmidt *et al.*, 2010; Faure *et al.*, 2012). By overlap with ChIP tracks for transcription factors, enhancer marks, histone modifications, and RNA Polymerase II, CNC sites were found to coincide with transcription factors specific to the tissue type analysed, enhancer marks, and active histone modifications (Faure *et al.*, 2012). The authors thus concluded that a second mode of cohesin activity exists at sites of active

transcription. Interestingly, although not discussed, k-means clustering of the transcription factor ChIP peaks showed that at sites absent for any of the 11 factors considered, CTCF was present, alongside RAD21, SA1, and SA2. Whereas, at sites with the highest number of transcription factors present, only RAD21 and SA2 were present, with SA1 showing signal similar to CTCF (Faure *et al.*, 2012). This paper suggests that cohesin mediates 3D organisation at two types sites, CTCF-bound sites and sites of transcription. How cohesin is targeted to the transcription sites is not clear, however, the authors suggest differential enrichment of SA1 and SA2 at the different site types.

Similarly, in human mammary epithelial cells (HMECs) and human cardiac endothelial cells (HCAECs) the majority of CNC sites are occupied by cohesin-SA2 (Kojic *et al.*, 2018). These SA2 CNC sites were similarly found at enhancer regions and were co-occupied by transcription factors. siRNA-mediated loss of SA2 resulted in deregulation of 630 genes, including genes involved in cell identity in mammary cells and the majority of which were not affected by CTCF loss. Cohesin-SA1 did not relocate to SA2 CNC sites upon SA2 loss, indicating that the two proteins bind to distinct subsets of sites within the genome. Corresponding Hi-C samples determined that siRNA-mediated loss of SA1 or SA2 had distinct effects on chromatin contacts. For example, intra-TAD contacts were increased with SA2 loss. This suggests that distinct binding of the SA proteins may contribute to different aspects of chromatin organisation (Kojic *et al.*, 2018). These papers suggest that SA1s main function during interphase involves interaction with CTCF and maintenance of TADs. However, during embryonic development, at least, cohesin-SA1 is enriched at promoter sites and loss of SA1 results in large-scale disruption of gene expression, a significant subset of which overlap with genes deregulated upon loss of NIPBL (Cuadrado *et al.*, 2012; Remeseiro, Cuadrado, Gómez-Lãpez, *et al.*, 2012). Similar to incomplete rescue of cohesin-SA2 occupancy by cohesin-SA1 in SA2 knockdown cells, SA2 was not re-located to all cohesin-SA1 sites upon SA1 depletion in these cells. Hence, SA1 and SA2 have important function roles that likely extend beyond bridging interaction of cohesin with CTCF and the mechanism underlying which are not fully understood.

Three recent papers also support these findings and suggest specific functions of SA1 and SA2. In mESCs cohesin-SA1 has also been shown to be important for maintenance of TAD borders, whereas cohesin-SA2 mediates looping between superenhancers and Polycomb domains (Cuadrado *et al.*, 2019). ChIP-seq in the mESCs revealed overlap of SA1 and SA2 binding sites with CTCF and a distinct set of SA2 sites that did not overlap with SA1 or CTCF. These 'SA2 only' sites were found to overlap with either Polycomb group members or active enhancer marks. TAD borders were affected by knockdown of SA1 and SA2, however, there was a more significant effect following loss of SA1, again suggesting that cohesin-SA1 mediates TAD organisation. SA1 and SA2 knockdown also had differential effect on chromatin contacts at Hox loci, which are commonly mediated by Polycomb binding. SA1 knockdown strengthened contacts between Hox loci, whereas SA2 knockdown reduced contacts between the Hox loci, perhaps due to a loss of SA2-mediated recruitment of PRC1, although specific interaction of SA2 and PRC1 was not determined (Cuadrado *et al.*, 2019).

AID-tagging of SA1 and SA2 in two separate HCT116 cell lines has also revealed specific functions of the SA proteins (Casa *et al.*, 2020). SA1 and SA2 could both be depleted rapidly without loss of SMC1A or the other SA protein from chromatin. In these cells 60% of cohesin sites were bound by SA1 and SA2, with 34% bound by SA1 only and 4% bound by SA2. Co-occupancy of CTCF was observed at all of the sites analysed. SA1 was re-distributed to SA2 sites upon SA2 depletion, whereas SA2 showed minimal re-distribution to SA1 sites upon SA1 depletion. Active promoters and enhancers were enriched in the SA overlapping sites and were more highly enriched at SA2 only sites compared to SA1 only sites. This suggesting that the discrepancy in redistribution may occur as SA can only redistribute to active promoter and enhancer sites. SA1 depletion more strongly affected TAD level contacts compared to SA2, although TAD borders were relatively unaffected in both SA depletion samples. Short-range contacts were decreased with SA2 depletion and longer-range contacts (>2 Mb) were slightly decreased with SA1 loss.

Biological relevance of SA functional distinction has been shown in haematopoiesis in mice. Viny *et al.* (2019) observed increased self-renewal and

decreased differential capacity in hematopoietic stem cells in mice with deletion of SA2, but not SA1. Double deletion of SA1 and SA2 was lethal, indicating that, at some level, SA1 and SA2 have redundant roles that can prevent lethality with the loss of just one SA protein. However, as in the studies described above, SA1 did not relocate to all SA2 sites following deletion of SA2, including some critical hematopoietic regulators. SA2-specific loci contained lineage-specific transcription factor binding motifs and showed a loss of insulation and downregulation following SA2 deletion. Hence, the function of SA proteins is important for cell fate and SA2 specifically regulates differentiation potential in hematopoietic cells, loss-of-function of which can result in transformation.

Altogether these papers indicate that i) SA1 may contribute to TAD and long-range chromatin interactions more strongly than SA2, ii) SA2 may contribute more strongly to local chromatin interactions than SA1, and perhaps is more important for tissue-specific promoter-enhancer interactions, iii) SA1 and SA2 share a number of CTCF and non-CTCF binding sites in the genome but do not specifically co-localise at these sites in single cells, iii) both SA proteins can redistribute to each other's bindings sites following knockdown of the other, but not completely, and iv) an unknown factor means that SA1 can redistribute to SA2-bound sites more efficiently that SA2 can redistribute to SA1-bound sites. As discussed in section 1.4.3, cohesin-SA1 may be preferentially protected from WAPL-mediated release from chromatin (Wutz *et al.*, 2020). It is possible that different stabilities on chromatin influence our ability to detect each SA at specific sites in the genome in bulk experiments, hence, contributing to the perceived differences in their localisations.

Although SA2 was more strongly enriched to promoter and enhancer sites, both SA proteins were identified at sites of transcription in these papers. Cohesin loading is proposed to occur at sites of transcription and is discussed below.

1.5 Function of the NIPBL/MAU2 loader complex

1.5.1 Topological loading of cohesin occurs in two steps

As discussed in section 1.2.1, yeast Scc2 and its co-factor Scc4 are required for loading of cohesin onto chromatin (Tóth *et al.*, 1999; Ciosk *et al.*, 2000). Orthologs of Scc2 exist in eukaryotes, including fission yeast (Mis4), Drosophila (Nipped-B), and humans (NIPBL) (Furuya, Takahashi and Yanagida, 1998; Rollins, Morcillo and Dorsett, 1999; Krantz *et al.*, 2004; Tonkin *et al.*, 2004). ATP hydrolysis was identified as a key factor in loading of cohesin in yeast. Loss of ATP hydrolysis capacity in either Smc3 or Smc1 abolished association of cohesin with chromatin without preventing assembly of the cohesin ring (Arumugam *et al.*, 2003; Weitzer, Lehane and Uhlmann, 2003). ChIP-seq of ATP hydrolysis mutants identified accumulation of cohesin with Scc2 at centromeric and chromosome arm sequences (Hu *et al.*, 2011). Association of SMC3 ATPase mutants with chromatin has also been observed in human cells (Ladurner *et al.*, 2014). In both yeast and human cells, FRAP determined that average residence time of this cohesin was reduced to seconds and depletion of Scc2/NIPBL abolished interaction completely (Hu *et al.*, 2011; Ladurner *et al.*, 2014). Altogether these papers suggest that ATP hydrolysis occurs as a subsequent step in loading following interaction of cohesin and Scc2 on chromatin.

Scc2-4 mediated loading of cohesin onto chromatin produces an interaction that is salt resistant (Ciosk *et al.*, 2000). *In vitro* recapitulation of cohesin loading on DNA molecules occurs via a salt sensitive intermediate, again indicating two steps/modes in the loading reaction (Onn and Koshland, 2011; Murayama and Uhlmann, 2014). As the second mode of DNA binding is salt resistant, the ATP hydrolysis-dependent second step of cohesin loading likely represents topological loading of cohesin (Murayama and Uhlmann, 2014). Linearization of circular DNA IP'd by cohesin was used as a test to confirm topological loading and determined that cohesin has an intrinsic ability to topologically bind to DNA, however Scc2-4 is required for efficient topological loading (Murayama and Uhlmann, 2014). Finally, single molecule tracking has identified two modes of cohesin diffusion on chromatin. In both G1 and G2/S phase cells, ~1/2 of all cohesin molecules identified fit a model of specific chromatin binding while 13%

fit a model of non-specific chromatin binding (remaining ~40% of cohesin in 3D diffusion, not bound to chromatin) (Hansen *et al.*, 2017). The topological nature of this non-specific fraction remains to be determined.

1.5.2 Mechanism of NIPBL-mediated loading

The crystal structure of Scc2 has been solved in the fungi *Ashyba gossypii* ($Scc2^{378-1479}$; Chao *et al.*, 2017) and *Chaetomium thermophilum* ($Scc2^{385-1840}$; Kikuchi *et al.*, 2016). Both fungal Scc2 proteins contain an N-terminal disordered region that binds to Scc4, globular N- and C-terminal ends, and a central HEAT repeat domain that bends to form a hook-shaped structure. Deletion of different regions of NIPBL suggest that the N-terminus of the protein is required for its stabilisation, via interaction with MAU2, while the C-terminal end of the protein is involved in regulation of cohesin's ATPase activity (Hinshaw *et al.*, 2015; Haarhuis *et al.*, 2017). A similar bent architecture has also been reported for Scc2 in *S. cerevisiae* by electron microscopy (EM), suggesting that this represents a conserved feature of the cohesin loader (Hinshaw *et al.*, 2015). It is worth noting that human NIPBL has a predicted molecular weight that is over twice the size of the Scc2 crystal structures solved. In addition, the N-terminal domain of human NIPBL does not align well with yeast Scc2. Thus, human NIPBL may contain additional structure that are not yet understood. However, mutations identified in patients with CdLS map to conserved regions of NIPBL, suggesting that important domains in human NIPBL are also found in yeast Scc2 (Chao *et al.*, 2017).

While a similar structure for Scc2 has been reported in different studies using different species, how Scc2 interacts with cohesin has not been definitively established. *In vitro* binding to cohesin peptide fragments suggest that in *S. pombe*, Scc2-Scc4 interacts with cohesin through SMC1, SMC3, Scc1, and Scc3 (Murayama and Uhlmann, 2014). Multiple contacts between *S. cerevisiae* Scc2-Scc4 and cohesin subunits were also observed in an independent study using amine cross-linking coupled with mass spectrometry (Chao *et al.*, 2017). Alongside their finding that *A. gossypii* Scc2 is highly flexible, Chao et al. (2015) suggest that these multiple contacts may allow Scc2 to contact cohesin at multiple points and induce loading by regulating large-scale conformational changes in the complex. In particular, they observed that the hook domain of

Scc2 is able to compact and extend from ~170 Å to ~280 Å (Chao *et al.*, 2015). The authors propose a model for loading in which Scc2-Scc4 first localises to chromatin and adopts its extended conformation, thereby allowing capture of the cohesin ring and transition to its compacted form, which may then induce a conformational change in cohesin that allows the complex to open and entrap the DNA. In contrast, GST pull-down of *C. thermophilum* GST-Scc2 found that Scc2 only binds to an N-terminal portion of Scc1, but not SMC1, SMC3, or Scc3 (Kikuchi *et al.*, 2016). Scc3 and Pds5 interact with cohesin through Scc1; so, the common structure of cohesin regulators may promote interaction via Scc1 and suggests that cohesins association with chromatin is regulated via Scc1. Alternatively, Scc2 may only interact with multiple cohesin subunits in certain conformations and thus, these interactions are not captured by single cohesin component GST pull-downs. Recent structural studies of cohesin in complex with DNA and NIPBL indicate that NIPBL and SA interact together in an antiparallel arrangement and wrap around both the cohesin ring and DNA to position and entrap DNA (Higashi *et al.*, 2020; Shi *et al.*, 2020).

The exact nature the ATP hydrolysis-dependent second step of cohesin loading that mediates topological loading is not yet clear. It has been proposed that ATP hydrolysis catalyzes opening of an interface in the cohesin complex to allow DNA entry into the ring-shaped structure. Gruber *et al* (2006) substituted the Smc hinge domain such that it was locked shut and found that yeast cohesin could not load onto DNA, whereas locking of Smc3-Scc1 or Scc1-Smc1 interfaces did not affect loading. Similarly, in human cells knockdown of endogenous SMC proteins and expression of SMC proteins that lock their hinge closed prevented loading of RAD21 onto DNA (Buheitel and Stemmann, 2013). These papers suggest that opening of the SMC1-SMC3 dimer is required for loading of cohesin onto DNA. Maintained interaction with the loader complex in the hinge mutants was not assessed and is required to conform specificity of the loss of loading. *In vitro* analysis of fission yeast cohesin suggests that the Rad21-Psm3(Smc3) gate opens instead in a reaction mediated by Pds5-Wapl. Evidence for this was that co-IP of DNA with the cohesin complex was increased with addition of Pds5-Wapl and the N-terminal end of artificially cleaved Rad21 is lost from an Smc3 IP with addition of Pds5-Wapl (Murayama and Uhlmann, 2015). The authors propose that the cohesin loader may first induce a conformation change in cohesin that

allows the Smc head domains to open in an ATP-hydrolysis-dependent manner and to expose the Rad21-Smc3 gate to Pds5-Wapl. It is not clear if this mechanism occurs *in vivo* as depletion of Pds5 or mutation of its interaction with Scc1 do not effect cohesin levels on chromatin (Kulemzina *et al.*, 2012; Chan *et al.*, 2013).

In order to mediate sister chromatid cohesion or looping of interphase chromatin cohesin needs to tether two chromatin stands. Murayama *et al.* (2018) assayed capture of ds- and ss-DNA by cohesin. Multiple combinatorial protocols were tested, and it was determined that in order to tether two DNA strands, cohesin must first load onto dsDNA prior to addition of the second DNA. The cohesin-dsDNA complex is then able to capture ssDNA but not dsDNA. Capture of the second strand was strictly dependent on the presence of the cohesin loader in the solution, despite the fact that as a first capture event cohesin could IP ssDNA equally efficiently in the presence of absence of the loader complex. This suggests that the second DNA capture assayed here is distinct from cohesins ability to bind to DNA non-topologically. Finally, conversion of the second DNA from ssDNA to dsDNA by DNA polymerase increased DNA retention with salt wash from 10 to 70%, indicating a stabilisation of the second DNA strand capture. Topological loading on the stabilised second strand was confirmed as linearisation of second strand or artificial cleavage of Rad21 released DNA from the IP material (Murayama *et al.*, 2018). In summary, this work indicates that cohesin captures the second strand of DNA via a single-strand intermediate.

MAU2 is well conserved from yeast to humans, suggesting conserved function (Watrin *et al.*, 2006; Hinshaw *et al.*, 2015). In human cells, MAU2 may not be directly required for loading of cohesin onto DNA as CRISPR-Cas9-mediated deletion of the first 10 exons in NIPBL reduced total MAU2 levels, increased total NIPBL levels, and resulted in normal levels of RAD21 on chromatin (Haarhuis *et al.*, 2017). Yeast Scc4 has been suggest to targeting NIPBL-mediated loading of cohesin (Hinshaw *et al.*, 2015). Evidence includes targeted localisation of yeast cohesin to centromeres via interaction of Scc4 with Ctf19 kinetochore protein (Hinshaw *et al.*, 2017). In addition, while both Scc2 and Scc4 have been linked to localisation with chromatin remodelers, Scc4 loss can be compensated for by fusion of a C-terminal portion of Scc2 to the chromatin remodelers proteins

(Muñoz *et al.*, 2019). This suggests that as well as stabilising Scc2 the importance of Scc4 may lie in targeting Scc2 binding.

1.5.3 Role of NIPBL in cohesin activity

Heterozygous loss of NIPBL in MEFs did not significantly reduce bulk levels of cohesin on chromatin, induce cohesion defects, or alter sensitivity to DNA damage (Remeseiro *et al.*, 2013). However, reduced cohesin levels were observed at the promoters of a subset of genes and altered gene expression was recorded, indicating the importance of NIPBL function. The authors also recorded significant upregulation of SA1 expression, but not SA2 or RAD21 (Remeseiro *et al.*, 2013). This raises the question of functional interplay between SA1 and NIPBL.

Deletion of *Nipbl* in non-dividing mouse liver cells reduced cohesin levels on chromatin ~5-fold, although SA1 could still be detected on chromatin by western blot (Schwarzer *et al.*, 2017). Chromatin loops and TADs were lost with *Nipbl* deletion; however, compartmentalization was enhanced, with small B-like regions observed in A-type compartments. These B-like regions correlated well with repressed genes. This suggests that cohesin prevents compartmentalization of repressed genes in active regions under normal conditions, or that compartmentalization factors were enhanced with NIPBL loss (Schwarzer *et al.*, 2017).

1.5.3.1 Role of Scc2 in cohesin translocation and potential loop extrusion

As chromatin loops may form the fundamental units of chromatin organisation, understanding the molecular mechanisms of chromatin loop formation and maintenance are key to our understanding of chromatin biology. An influential model termed the “loop extrusion model” postulates that the cohesin complex acts an extrusion machine that expels a loop in the chromatin until it dissociates or is stabilised at a boundary element, such as CTCF (Alipour and Marko, 2012; Sanborn *et al.*, 2015; Fudenberg *et al.*, 2016a). Under this model, TADs are formed from progressive, dynamic extrusion of loops within domains constrained by boundary elements and cohesin would be loaded at sites distinct from loop/TAD anchor sites (Alipour and Marko, 2012; Fudenberg *et al.*, 2016a).

Evidence for the loop extrusion model has emerged in recent years. Most notably, extrusion of loops in DNA molecules has been observed *in vitro* in the presence of ATP and the NIPBL-MAU2 loader complex (Davidson *et al.*, 2019; Y. Kim *et al.*, 2019). Interestingly, continued presence of NIPBL-MAU2 and ATP was required for loop maintenance, suggesting that cohesin and the loader may interact outside of the loading reaction (Davidson *et al.*, 2019; discussed further below). Additionally, extrusion was observed without topological loading of cohesin onto the DNA, but only in the presence of SA1, indicating a key role for NIPBL-MAU2 and SA1 in the molecular mechanism of loop extrusion (Davidson *et al.*, 2019).

Loop extrusion has not been observed *in vivo*, however translocation of cohesin along DNA has been observed (Lengronne, Katou, Mori, Yokabayashi, *et al.*, 2004; Busslinger *et al.*, 2017). It is not clear if this translocation is involved in active extrusion of a loop in the chromatin or movement of cohesin along DNA by transcription machinery.

To investigate cohesin-NIPBL interaction over time, Rhodes et al. (2017) utilized FRAP and single-molecule tracking (Rhodes *et al.*, 2017). The authors inactivated WAPL in their cells, leading to stabilization of cohesin on chromatin and compaction of chromosomes into ‘vermicelli’ structures. They reported that upon WAPL inactivation, increased binding of NIPBL to chromosomes occurs, despite a reduced availability of free cohesin. Importantly, the diffusion co-efficient of unbound NIPBL molecules was not altered by WAPL depletion or cohesin degradation in wild-type cells, indicating that NIPBL dynamics were not affected by the formation of compacted chromatin or changes in cohesin levels. In addition, NIPBL recovery after photobleaching was seen to process gradually from vermicelli close to the unbleached zone across the cell. The authors argue that this occurs due to ‘hopping’ of NIPBL from one cohesin molecule to the next. The authors suggest that these findings indicate that NIPBL associates with cohesin for some function besides just loading. It is not clear from this analysis whether such hopping behaviour occurs when cohesin is not stabilised by WAPL depletion. SA2-dependent recruitment of NIPBL to DSBs has been observed in Hela cells following laser-induced DNA damage, validating the idea that cohesin may be involved in association of NIPBL with chromatin (Kong *et al.*, 2014).

Continued association beyond the initial recruitment was not investigated in this paper.

Depletion of WAPL increases cohesin residence time and amount on chromatin (Kueng *et al.*, 2006; Tedeschi *et al.*, 2013; Wutz *et al.*, 2017). In agreement with the loop extrusion model, Hi-C analysis has shown that loops weakly identified in control cells were elongated and increased in frequency with WAPL depletion (Haarhuis *et al.*, 2017; Wutz *et al.*, 2017). Interactions within TADs were decreased while contacts at and beyond control TAD boundaries were increased, indicating that dynamic cohesin binding is required for TAD boundary maintenance and suggesting that TADs represent dynamic loops within (Haarhuis *et al.*, 2017; Wutz *et al.*, 2017). Finally, deletion of MAU2 and subsequent destabilization of NIPBL was observed to result in reduced loop sizes in both a wild-type or WAPL mutant background (Haarhuis *et al.*, 2017). This suggests that NIPBL-MAU2 is involved in the extension of DNA loops, however, no direct link to the cohesin activity was shown for this result – it may just be that loops are now governed by another protein, rather than decreased cohesin translocation due to loss of NIPBL-MAU2. Hence, continued association of NIPBL with translocating cohesin may occur in cells but has not been observed under wild-type conditions.

1.5.4 Where is cohesin loaded onto DNA?

As discussed above, NIPBL's ability to stimulate cohesin's ATPase activity has implicated it in the loading of cohesin onto DNA (Ciosk *et al.*, 2000; Murayama and Uhlmann, 2013), cohesin's ability to translocate along DNA efficiently (Kanke *et al.*, 2016; Rhodes *et al.*, 2017) and the extension of loops formed by cohesin (Haarhuis *et al.*, 2017). Given these findings, NIPBL might be expected to co-localise with cohesin along the genome, however, mixed findings have been reported, clouding our understanding of how cohesin interacts with and shapes chromosomes.

ChIP-seq analysis in *S. cerevisiae* characterized Scc2-Scc4 at sites largely distinct from cohesin, suggesting that cohesin moves away from its initial loading spot following stable association with the DNA (Lengronne, Katou, Mori,

Yokobayashi, *et al.*, 2004). In a contrasting study carried out in mouse embryonic stem cells (ESCs), cohesin was found to overlap NIPBL binding sites at the enhancers and core promoter sites of actively transcribed genes (Kagey *et al.*, 2010). Kagey *et al.* (2010) also identified the mediator complex at a high percentage of the cohesin-NIPBL co-bound sites. As mediator and cohesin co-occupy different promoters in different cells, cell-type-specific DNA loops were linked to the gene expression program of each cell in this study (Kagey *et al.*, 2010).

By characterising the available NIPBL antibodies and using the one that performed best in human cells, Zuin *et al.* (2014) reported that there are two different types of NIPBL binding sites; NIPBL major sites, as detected by NIPBL#1 antibodies, which are localized at promoters and do not overlap with cohesin, and NIPBL minor sites, as detected by NIPBL#1&#6 antibodies, which do overlap with cohesin binding sites (Zuin *et al.*, 2014). This work suggests that NIPBL may carry out some function distinct from cohesin loading and use of different NIPBL antibodies can affect the NIPBL signal detected.

van den Berg et al. (2016), however, suggested that this difference is not simply an artifact of the antibody used. Their ChIP-Seq analysis of neural stem cells detected a very small percentage of Scc2 sites that were co-bound by cohesin, despite using the same antibody as Kagey *et al.* (2010). Given that Scc2 was still predominantly located at the enhancers and core promoters of actively transcribed genes, the authors suggest that co-occupancy may be missed due to differences in cell cycle length or cohesin dynamics between the cells, which results in almost all cohesin having translocated along the DNA to the boundary element CTCF (van den Berg *et al.*, 2017). ChIP-Seq analysis of CTCF depleted cells supported this interpretation, as loss of CTCF was found to result in an increase in co-localisation of cohesin at Scc2 sites from 20% to 50%, without significant changes in gene expression (Busslinger *et al.*, 2017). In this case, cohesin may not translocate away from its loading sites in the absence of CTCF or it may progress beyond its normal sites of activity to adjacent NIPBL binding sites, which would then act as boundary elements.

In summary, robust observation of when and where cohesin loading occurs has not been achieved, limiting our ability to understand the mechanisms that underlie chromosomal organisation and its impact on gene regulation. A number of factors may account for the lack of consensus regarding cohesin-Scc2 interaction; including that, a) the interaction may be transient, which would make it difficult to detect by conventional ChIP-seq methods; b) the localisation of Scc2-mediated loading may be cell-type specific, c) Scc2 may also be carrying out cell-type dependent activities distinct from loading, and these activities may or may not involve interaction with cohesin, d) antibodies with varying efficiencies are used in different studies, and e) cell cycle dynamics of different cell types may determine where cohesin can be detected. Furthermore, ChIP experiments are carried out on a population of cells, hence, 'colocalised peaks' do not directly indicate that the same fragment of chromatin is bound by the two proteins in any given cell. However, as discussed above, multiple studies suggest that loading of cohesin may occur at the enhancer and core promoter sites of actively transcribed genes.

1.5.5 Cohesin loading at sites of nucleic acid structure

As discussed above, some evidence suggests that NIPBL-mediated loading of cohesin during interphase may occur at enhancers and core promoter sites of actively transcribed genes. Yet, how the NIPBL-MAU2 loading complex is recruited to these sites remains unknown. During S phase, NIPBL-MAU2 and cohesin interact with DDK-phosphorylated MCM2-7 (Zheng *et al.*, 2018). Hence, cohesin can be deposited at sites of replication and mediate cohesion of sister chromatids. In yeast, cohesin captures the second strand of DNA via a single-strand intermediate, at least *in vitro* (Murayama *et al.*, 2018). Accordingly, cohesion establishment in these cells seemed to be upstream of Okasaki fragment processing and histone deposition (Zheng *et al.*, 2018). Defects in cohesion and reduced interaction with MCM2 were also observed following depletion of the replisome components WDHD1, TIMELESS, and DDX11, and depletion of RPA2, a subunit of the replication protein A complex that binds to and stabilises single-stranded DNA intermediates at sites of replication and repair (Zheng *et al.*, 2018). Thus, compound protein interactions and DNA structure are required for cohesin loading during S phase.

In yeast, the cohesin loader complex interacts with chromatin remodelers to localise loading along chromosome arms and at centromeres (Lopez-Serra *et al.*, 2014; Muñoz *et al.*, 2019). ATPase activity of the remodelers is required for cohesin loading and nucleosome-free DNA was required for *in vitro* reconstitution of the loading reaction (Muñoz *et al.*, 2019). Histone modifications, chromatin remodelers, and histone chaperones regulate nucleosome turnover at regions of transcription, opening chromatin structure and allowing access to transcription factors and RNA polymerase (Thurman *et al.*, 2012; Venkatesh and Workman, 2015). Hence, in human cells NIPBL-MAU2 may simply take advantage of open chromatin at sites of transcription to load cohesin at accessible regions. However, given the involvement of multiple proteins and DNA structure to cohesin loading during S phase, a more complex mechanism may be at play.

1.5.6 Nucleic acid structure at sites of transcription

Similar to the replication fork, transcription bubbles represent a nexus of protein complexes and nucleic acid molecules in various structures. Transcription initiation is a multi-step process involving binding of DNA by numerous proteins and conformational changes to the promoter DNA. RNA polymerase II and a number of general transcription factors sequentially bind to promoter elements and form the pre-initiation complex (PIC) (Buratowski *et al.*, 1989). The general transcription factor IIH (TFIIH) then acts as a DNA helicase that induces ATP-dependent duplex DNA melting downstream of the PIC (Kim, Ebright and Reinberg, 2000). The template strand of the resulting single-stranded DNA “bubble” can then bind the active site of the PIC forming the open promoter complex, from which, RNA synthesis can commence (He *et al.*, 2016). Early transcription is unstable and will go through multiple abortive cycles until the RNA transcript reaches around 15 nucleotides, after which, ‘promoter escape’ and elongation can commence (Kugel and Goodrich, 1998). During elongation, RNA polymerase is bound to the transcription bubble as part of the elongation complex and transcribed RNA will hybridise to the template DNA as it exits the polymerase active site. The RNA-DNA hybrid will extend for ~9 base pairs before the RNA and DNA strands separate. The Rpb1 and Rpb2 subunits of the elongation complex create an exit tunnel for the transcribed RNA and may be involved in

resolution of the RNA-DNA hybrid and maintenance of the upstream end of the transcription bubble (Gnatt *et al.*, 2001).

1.5.7 R-loop structures at sites of transcription

During transcription, synthesised RNA is processed (such as 5' capping and splicing) as elongation occurs (Bentley, 2014). Fully synthesised and processed mRNA is then exported from the nucleus to the cytoplasm for translation to its corresponding protein. Under certain circumstances, the elongating RNA can instead hybridise to the template strand of the upstream DNA, forming an R-loop - an intermediate RNA:DNA conformation and a displaced single strand of DNA (Figure 4). As discussed above, the crystal structure of RNA polymerase indicates that RNA and DNA exit the elongation complex through different exit channels. Thus, a thread back model of R-loop formation is favoured, however, it is possible that under the circumstances that induce R-loop formation the RNA-DNA hybrid formed within the transcription bubble is not separated during exit and instead continues to be extended.

Formation and stabilisation of an R-loop occurs in circumstances that make binding to the nascent RNA favourable over the non-template strand. Such circumstances include the presence of guanine-rich clusters in the 5' end of the nascent RNA, further runs of guanines in the extending sequence, negative supercoiling of the trailing fork, and nicks in the non-template strand (Richardson, 1975; Roy and Lieber, 2009; El Hage *et al.*, 2010; Roy *et al.*, 2010). In mammalian cells, R-loops are predominately detected at the promoter and termination sites of active genes (Ginno *et al.*, 2012; Sanz *et al.*, 2016).

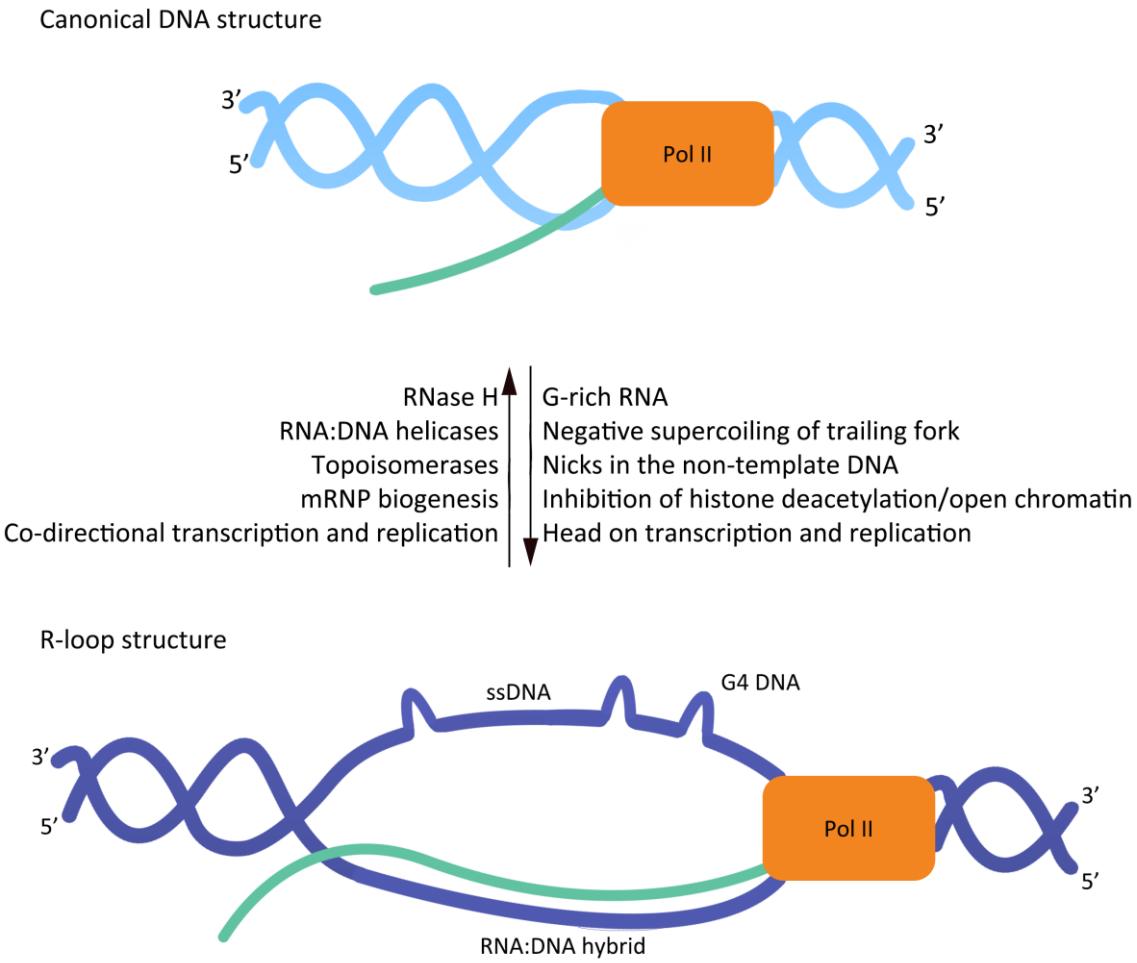


Figure 4: R-loop formation at sites of transcription. A schematic of transcription with downstream canonical DNA helix reformation (top, blue) or formation of an R-loop (bottom, purple). The transcribed RNA that hybridises to the DNA is shown in green. Factors that repress or induce R-loop formation are indicated beside direction arrors. Details of these factors are in the main text below. G4 = G-quadruplex; PolII = RNA Polymerase II; ssDNA = single-stranded DNA.

Once hybridised, the RNA:DNA hybrid is more stable than duplex DNA and is not likely to spontaneously resolve (Roberts and Crothers, 1992). Accordingly, numerous proteins have been shown to counteract R-loop formation. The RNase H nucleases (RNase H1 and RNase H2) digest RNA present in RNA-DNA lesions, removing it from the R-loop (Stein and Hausen, 1969; Wahba *et al.*, 2011). RNA-DNA helicases, such as SETX, AQR, DHX9, and FANCM, unwind RNA-DNA hybrids and prevent their extension (Chakraborty and Grosse, 2011; Skourti-Stathaki, Proudfoot and Gromak, 2011; Sollier *et al.*, 2014; Schwab *et al.*, 2015). Three topoisomerase enzymes, TOP1, TOP2, and TOP3B, have been shown to reduce R-loop levels by decreasing negative supercoiling behind the RNA polymerase, thus relieving stress that might allow local unwinding of DNA

and preferential RNA binding (Tuduri *et al.*, 2009; El Hage *et al.*, 2010; Yang *et al.*, 2014).

Proper co-transcriptional RNA processing and messenger ribonucleoprotein (mRNP) assembly also counteract R-loop formation. The THO/TREX complex (THO) plays an important role in mediating RNA processing from transcription of the nascent RNA to export of mRNP to the cytoplasm, including by recruitment of splicing factors, heterogeneous nuclear ribonucleoproteins (HNRNPs), and export machinery onto the RNA (Rappsilber *et al.*, 2002; Masuda *et al.*, 2005; Cheng *et al.*, 2006). THO also interacts with the histone deacetylase Sin3A (Salas-Armenteros *et al.*, 2017). Depletion of the THO subunit THOC1 in complex with Sin3A promotes RNA:DNA hybrid accumulation, a phenomenon that was also observed with chemical inhibition of histone deacetylase activity. Therefore, the THO/TREX complex plays an important role in R-loop suppression via mRNP biogenesis and chromatin modification. Promotion of R-loop formation at open chromatin during transcription and reduction of R-loop levels with topoisomerase activity behind RNA polymerase suggests that tight regulation of chromatin organisation around R-loops is important for their regulation.

R-loops have been implicated as a source of genomic instability, for example, depleting cells of the RNA splicing factor serine/arginine-rich splicing factor 1 (SRSF1) results in increased R-loops and subsequently, the rapid appearance of double-strand DNA breaks (Li and Manley, 2005). The most supported mechanism for R-loop mediated genomic instability postulates that R-loops induce pausing of RNA polymerase, and subsequently, collision of transcription and replication machinery. Evidence for this mechanism includes the findings that i) genes that are more likely to form R-loops show increased frequency of paused replication forks, ii) resolution of R-loops reduces enrichment of helicase involved in fixing stalled replication forks at these genes, and iii) resolution of R-loops decreases replication fork stalling and chromosome breaks (Azvolinsky *et al.*, 2009; Tuduri *et al.*, 2009; Gómez-González *et al.*, 2011).

In vitro study indicates that co-directional transcription and replication collision reduces R-loop levels and induces ATM autophosphorylation, whereas, head-on transcription and replication collision increases R-loop levels and induces ATR

phosphorylation (Hamperl *et al.*, 2017). Downstream ATM and ATR signalling molecules were also specifically phosphorylated by the specific collision orientations. Importantly, collision of transcription and replication machinery in the absence of a co-transcription R-loops did not induce either of these pathways. ATM functions in response to double-strand breaks, which may be formed by RNA polymerase dissociating from the DNA and the replisome resuming leading strand synthesis using the R-loop RNA as a primer, ultimately leaving a break in the DNA which will convert to a DSB in the next round of replication (Cimprich and Cortez, 2008; Pomerantz and O'Donnell, 2008). In contrast, ATR responds to single-stranded DNA (Cimprich and Cortez, 2008). Uncoupling of MCM2-7 from the replisome has been recorded following replication stalling *in vitro*, allowing MCM2-7 to continue to unwind DNA in the vicinity (Byun *et al.*, 2005). Large amounts of ssDNA generated by both the R-loop and uncoupled helicase may then induce ATR DNA damage response signalling. Head-on collisions induce R-loop structure, perhaps helping to generate even more ssDNA. As cohesin can interact with MCM2-7 and ssDNA, this raises the question of whether MCM2-7 and the presence of ssDNA may help target cohesin to the region to help correct the DNA damage.

While R-loops can induce DNA damage, they may also facilitate DNA repair. Yasuhara et al. (2018) discovered that active transcription and R-loop levels were important for Rad52 recruitment to DSBs. Rad52 then recruits XPG and BRCA1 to resolve the R-loop and expose ssDNA for RPA binding and to repress binding of the NHEJ factors RIF1 and 53BP1, respectively. Together, these events stimulate end resection and ATM-signalling-mediated homologous repair (HR) (Yasuhara *et al.*, 2018). Thus, R-loop stability and levels may fine tune the choice between HR and NHEJ as the presence of R-loops blocks end resection of breaks and binding of repair proteins, yet, the act of resolving R-loops promotes end resection and HR (Ohle *et al.*, 2016).

Studies also indicate a multifaceted contribution of R-loops to gene expression regulation. Non-methylated cytosine nucleotides followed by a guanine characterise regions of the genome termed CpG islands and demarcate the promoters of most active mammalian genes. The G-rich characteristic of CpG islands contributes to the preferential formation of R-loops at these active genes,

wherein the RNA:DNA hybrid can then preserve unmethylation by protecting from DNA methyltransferases (Ginno *et al.*, 2012; Grunseich *et al.*, 2018). This preservation may be direct, as DNA methyltransferases preferentially bind to DNA:DNA structures over RNA:DNA hybrid structures, or indirect, as at the TCF21 tumour suppressor gene; here GADD45A binds to an R-loop generated by transcription at TCF21 and recruits TET1, a methylcytosine dioxygenase that can demethylate the local region (Ginno *et al.*, 2012; Grunseich *et al.*, 2018; Arab *et al.*, 2019). Hence, R-loops form part of a feedback loop to ensure further transcription. As in the indirect method of methylation regulation discussed above, R-loops of different genes have been shown to dynamically regulate the binding or repulsion of transcription factors. For example, antisense transcription at the promoter of the vimentin (VIM) gene generates a long non-coding RNA, VIM-AS1, which binds to the DNA forming an R-loop structure. Knockdown of the VIM-AS1 RNA or digestion of RNA:DNA hybrids diminished binding of the transcriptional activator p65 and reduced VIM expression levels (Boque-Sastre *et al.*, 2015). This and other studies also demonstrated the ability of R-loops to displace nucleosomes and safeguard open chromatin (Dunn and Griffith, 1980; Powell *et al.*, 2013). Thus, R-loops can regulate transcription of proximal genes by modulating transcription factor binding and by modulating the epigenetic landscape of the region.

Together with H3K9me3, R-loops are important to achieve efficient pause-mediated RNA polymerase II termination. Loss of R-loops at pause regions of transcription demonstrates their requirement for in situ antisense transcription, formation of double-stranded RNA, and recruitment of the RNA-induced silencing complex (RISC). Recruitment of RISC leads to deposition of H3K9me3, binding of HP1γ, and pausing of Pol II (Skourti-Stathaki, Kamieniarz-Gdula and Proudfoot, 2014). In yeast, HP1- (known as Swi6-) mediated transcription termination is dependent on cohesin recruitment, which is then thought to facilitate formation of the 3’ end of the mRNA and transcription termination (Gullerova and Proudfoot, 2008). In contrast, in mammalian cells, senataxin has been proposed to resolve the RNA/DNA hybrids at pause sites, allowing access of the 5’ to 3’ exonuclease Xrn2 to the elongating transcript, whereby it can digest up to the Poll II and promote termination (West, Gromak and Proudfoot, 2004; Skourti-Stathaki, Proudfoot and Gromak, 2011).

1.6 Research aims

The main goal of this thesis was to investigate the role of the SA proteins in cohesin biology. The first aim was to assess CTCF-SA interaction in human cells under endogenous conditions and to determine the role of RAD21 in this interaction. The second aim was to characterise non-CTCF interaction partners of the SA proteins. Given the ability of SA2, at least, to recognise distinct DNA structures, the intrinsic ability of cohesin to load onto chromatin, and the striking similarity of NIPBL and SA crystal structures, I hypothesized that the SA proteins may be capable of inducing loading of cohesin onto chromatin. Hence, the third aim was to assess the role of the SA proteins in loading of cohesin onto chromatin. R-loops represent the intersection of many protein interactions and show complex nucleic acid structure. Hence, I hypothesized that R-loops may represent a molecular alternative to the replication fork for cohesin loading during interphase and so assessed the impact of R-loop levels on cohesin loading.

2

Materials and methods

2.1 Cell culture

Human HCT116, Hela, and U2OS cells were cultured at 37ºC in a humidified incubator maintaining a ratio of 5% CO_2, 95% air. HCT116 and Hela cells were cultured in Gibco™ McCoy's 5A (Modified) GlutaMAX™ Medium (ThermoFisher) supplemented with 10% (v/v) fetal bovine serum (ThermoFisher) as a basal cell culture medium. U2OS cells were cultured in Gibco™ DMEM, high glucose, GlutaMAX™ medium (ThermoFisher) supplemented with 10% (v/v) fetal bovine serum (ThermoFisher) as a basal cell culture medium. HCT116 Rad21-mAID-mClover (RmAC) OsTIR1, HCT116 RmAC, and HCT116 OsTIR1 cell lines were obtained from Natsume *et al.*, (2016). Cells were maintained as for standard HCT116 cells, except for supplementation of the media with 700µg/ml Geneticin and 100µg/ml Hygromycin B Gold for cell with tagged Rad21 and 100µg/ml Puromycin for cells with integrated OsTIR1. Rad21 depletion was achieved by addition of 500 µM Indole-3-acetic acid (IAA/Auxin) diluted in ethanol to the cell media.

2.2 siRNA-mediated knockdowns

For siRNA transfections, HCT116 or Hela cells were reverse transfected with scramble siRNA (siCon) or siRNAs targeting SA1, SA2, NIPBL, MAU2, SMC3, AQR, or RNASEH2A (Table 1). Transfection concentrations are indicated or described in figure legends. siRNAs were 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 siAQR and siRNASEH2A, incubation time was reduced to 40 hrs. To account for the reduced growth time, cells were plated at a density of 2-3 x 10^6 cells per 10 cm dish. Here siCon- and siNIPBL-transfected cells were plated at a lower cell number than siAQR-/siRNASEH2A-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.

siRNA name	Company	Target	Catalogue no.
siControl (scramble)	Dharmacon	Smartpool	D-001810-10-20
siSA1	Dharmacon	Smartpool	L-010638-01-0010
siSA2	Dharmacon	Smartpool	L-021351-00-0010
siSMC3	Dharmacon	Smartpool	L-006834-00-0010
siNIPBL	Dharmacon	Smartpool	L-012980-00-0010
siMAU2	Dharmacon	Smartpool	L-031981-01-0010
siAQR	Dharmacon	Smartpool	L-022214-01-0005
siRNASEH2A	Dharmacon	Smartpool	L-003535-01-0005

Table 1: Details of siRNA used.

2.3 Plasmid DNA

The Fos and Jun BiFC plasmids, pBiFC-bFosVC155, pBiFC-bJunVN173, pBiFC-bJunVN155(I152L), and pBiFC-bFosDeltaZIPVC155 were a gift from Chang-Deng Hu (Addgene plasmid # 22013, 22012, 27098, 22014, respectively)(Shyu *et al.*, 2006; Kodama and Hu, 2010). NIPBL_A-GFP was obtained from Lena Strom(Bot *et al.*, 2017) and transformed into Invitrogen™ One Shot™ Stbl3™ Chemically Competent *E. coli* (ThermoFisher), according to the manufacturer's instructions. Purification of plasmid DNA was performed using Qiagen midi or maxi-prep kits, according to the manufacturer's instructions. NIPBL_A-GFP preps were incubated at 30ºC to allow for the growth of the large plasmid. pEGFP-RNASEH1 was a gift from Andrew Jackson & Martin Reijns (Addgene plasmid # 108699)(Bubeck *et al.*, 2011).

2.4 Transient transfections of DNA plasmids

U2OS cells were electroporated with the Fos and Jun BiFC plasmids using the Neon® Transfection System (ThermoFisher), as per the manufacturer's instructions. Cells were grown to 60-80% confluency before collection with Trypsin-EDTA (0.25%), phenol red (ThermoFisher). Unless otherwise stated, 5 – 8ug of purified plasmid DNA was electroporated per 1 x 10^6 cells, with 2 x 10^6 cells plated per 10cm plate. U2OS cells electroporated in the presence of no plasmid DNA were included as a 'Mock' control. Cells were incubated for 16 hrs before lysis and IP with the GFP-Trap. HCT116 RmAC OsTIR1 cells were transfected with pEGFP-RNASEH1 using Lipofectamine 3000, according to the manufacturer's instructions. Cells were plated 22 hrs prior to transfection and unless otherwise stated 2ug of plasmid DNA was added per 1 x 10^6 cells. Unless otherwise indicated, cells were collected 40 hrs post-transfection. Hela cells were transfected with NIPBL_A-GFP by the same method, except that 0.84 x 10^6 cells were plated 6 hrs prior to transfection and 4ug of plasmid was added.

2.5 Chromatin Fractionation and co-immunoprecipitation

The final, optimized chromatin fractionation and co-IP protocol is described here; original conditions and changes made through the experiments are detailed in the main results section. Cells were washed twice with ice-cold PBS (Sigma Aldrich) and lysed in Buffer A (10 mM HEPES, 10mM KCl, 1.5 mM MgCl2, 0.34 M Sucrose, 10% Glycerol, 1mM DTT, 1mM PMSF/Pefabloc, protease inhibitor), supplemented with 0.1% T-X100, for 10 min on ice. Lysed cells were collected by scraping. Nuclei and cytoplasmic material were separated by centrifugation for 4 min at 1300 g at 4^{o}C. The supernatant was collected as the cytoplasmic fraction and cleared of any insoluble material with further centrifugation for 15 min at 20,000 g at 4oC. The nuclear pellet was washed once with buffer A before lysis in buffer B (3mM EDTA, 0.2mM EGTA, 1mM DTT, 1mM PMSF/Pefabloc, protease inhibitor) with rotation for 30 min at 4^{o}C. Insoluble nuclear material was spun down for 4 min at 1700 g at 4^{o}C and the supernatant taken as nuclear soluble fraction. The insoluble material was wash once with buffer B and then

resuspended in high-salt chromatin solubilization buffer (50mM Tris-HCl pH 7.5, 1.5 mM MgCl2, 500mM KCl, 1mM EDTA, 20% Glycerol, 0.1% NP-40, 1mM PMSF/Pefabloc, protease inhibitor). The lysate was vortexed for 2 min to aid solubilization. Nucleic acids were digested with 85U benzonase (Sigma-Aldrich) per 100 x 10^6 cells, with incubation for 10 min at 37°C and 20 min at 4°C. Chromatin was further solubilized with ultra-sonication for 3 x 10 sec at an amplitude of 30. The lysate was diluted to 200 mM KCl and insoluble material was removed by centrifugation at 15,000 RPM for 30 min at 4°C. Protein concentration was determined by Nanodrop.

For co-IP, antibodies were bound to Dynabead Protein A/G beads (ThermoFisher Scientific) for 10 min at room temperature and ~ 5 hr at 4°C. For mock IgG IPs, beads were incubated with serum from the same host type as the antibody of interest. 1mg of chromatin extract was incubated with the antibody-bead conjugate per IP for approximately 16 hr at 4°C. IPs were washed x5 with IP buffer (200mM chromatin solubilization buffer) and eluted by boiling in either 2x Laemmeli sample buffer (BioRad) or 4x NuPAGE LDS sample buffer (ThermoFisher Scientific). Proteins ≤ 250 kDa were separated by SDS-PAGE electrophoresis using 4–20% Mini-PROTEAN® TGX™ Precast Protein Gels (BioRad) and transferred to Immobilon-P PVDF Membrane (Merck Millipore) for detection. Proteins ≥ 250 kDa were separated by SDS-PAGE electrophoresis using Invitrogen NuPAGE 3-8% Tris-Acetate precast protein gels (see section 2.9 for details).

For analysis of co-purified nucleic acid molecules, elution was carried out in ChIP-seq Elution buffer (50mM Tris pH 8.0, 1mM EDTA, 1% SDS, 50nM NaHCO3) with incubation overnight at 65°C. TE buffer was added to reduce SDS present in the solution. RNase A or TURBO DNase treatments were carried out where described and according to the manufacturer's instructions. Then, Proteinase K was added and incubated at 45°C for 2hrs. Nucleic acids were then isolated by phenol-chloroform extraction and analysed on an Agilent High Sensitivity DNA or RNA 6000 Pico Chip, according to the manufacturer's instructions.

2.6 S9.6 IP and Dot Blot

Cells were fractioned and processed for S9.6 IP as described above in section 2.5, with the following modifications. To avoid digestion of RNA:DNA hybrids, samples were not treated with benzonase during chromatin solubilization and sonication was carried out for 10 min (Diagenode Biorupter) as in (Cristini *et al.*, 2018). Where indicated, chromatin samples were treated with Ribonuclease H enzyme (NEB) overnight at 37ºC to digest RNA:DNA hybrids in the extract. To avoid detection of single-stranded RNA by the S9.6 antibody, S9.6 IP samples were pre-treated with Purelink RNase A (Thermo Fisher Scientific) at 0.25ug/1mg chromatin extract for 1 hr 30 min at 4ºC. The reaction was stopped with addition of 143U Invitrogen SUPERase•In RNase Inhibitor (Thermo Fisher Scientific). RNA:DNA hybrid levels were assessed in chromatin samples by dot blot. Specifically, the chromatin lysate was directly wicked onto Amersham Protran nitrocellulose membrane (Merck) by pipetting small volumes above the membrane. Membranes were blocked in 5% (w/v) non-fat dry milk in PBS-0.1% Tween and incubated with S9.6 antibody overnight as for standard western blot. As above, detection was carried out using chemiluminescent fluorescence. RNase A-mediated digestion of RNA:DNA hybrids was performed using a non-ssRNA-specific enzyme (Thermo Scientific) at 1.5ug/25ug chromatin extract at 37ºC.

2.7 Protein isolation and immunoprecipitation with the GFP-Trap

Cells were washed x2 with ice-cold Gibco® Phosphate-Buffered Saline (PBS) (Life Technologies), scraped into PBS and spun down for 3 mins at 300g. Cell pellets were lysed in 20-pellet volumes of Buffer A (20mM HEPES, 10mM KCl, 1.5mM MgCl, 0.34M Sucrose, 10% Glycerol, 5mM beta-Mercaptoethanol, protease inhibitor) plus 0.02% Triton-X-100 for 10 mins on ice, with inversion every 2 mins. Nuclei were pelleted via centrifugation at 1300g for 5 mins at 4ºC before resuspension in 200ul Buffer B (50mM Tris pH 7.5, 500mM NaCl, 1mM EDTA, 0.1% NP-40, 20% Glycerol, 1mM DTT, protease inhibitor) with 5 – 10

strokes through a 19.5-gauge needle. Samples were incubated for 30 mins on ice in the presence of 30U of benzonase nuclease (Merck Millipore) with inversion every 5 mins. Following sonication for 5 minutes (30 seconds on, 30 seconds off) at high intensity using a Biorupter® sonicator, nuclear extracts were diluted to 150mM NaCl, using Buffer B minus NaCl. Lysates were cleared by centrifugation at 20000g for 10 mins at 4ºC and quantified by Qubit protein assay (ThermoFisher).

Unless otherwise stated, immunoprecipitation was carried out as follows. 500 – 570ug of protein was diluted to a final volume of 500 – 660ul in 150mM NaCl Buffer B. 5% of the diluted lysate was taken as a corresponding input sample. The remaining chromatin lysate was mixed with 15ul of equilibrated Chromotek® GFP-Trap beads (gtma-20; washed x3 with 150nM NaCl Buffer B) and incubated at 4ºC for 1-2hrs, with rotation. Immunocomplexes were harvested by magnetic separation and a sample of the supernatant taken as a corresponding 'non-bound' (NB) sample. Beads were washed once in Wash Buffer 1 (50mM Tris pH 7.5, 150mM NaCl, 1mM EDTA, 0.1% NP-40, 20% Glycerol, 1mM DTT, protease inhibitor) and once in Wash Buffer 2 (50mM Tris pH 7.5, 0-500mM NaCl, 1mM EDTA, 0.1% NP-40, 20% Glycerol, 1mM DTT, protease inhibitor). Beads were resuspended in 30ul of SDS-PAGE sample buffer and bound proteins eluted by boiling at 95ºC for 10mins.

2.8 DNA/Protein isolation by ChIP protocol

Samples prepared by 'ChIP protocol' were generated as follows, with reagent volumes altered according to the cell number collected for each sample (amounts listed correspond to 1 x 10^6 cells). Following electroporation, cells were washed with PBS and collected using trypin-EDTA for counting. Cells were pelleted at 4ºC and resuspended in 3ml of fresh media. 1/3 of each sample was set on ice as 'no fix – benzonase' samples. Formaldehyde was added to the remaining cells to a final concentration of 1%. Crosslinking in formaldehyde was carried out at room temperature for 10mins before quenching with glycine – added to a final concentration of 0.125M. Unfixed and fixed cells were washed x3 with ice-cold

PBS before resuspension in at least 10 volumes of Swelling buffer (2.5mM Hepes pH 7.8, 1.5mM $MgCl_2$, 10mM KCl, 0.1% NP-40, 1mM DTT, protease inhibitor). Samples were dounced ~20 times and centrifuged at 2000rpm for 5mins. The nuclear pellet was washed once with sonication buffer (without Triton X-100) and then resuspended in 100ul of sonication buffer (volume corresponding to 1 million nuclei per 100ul; 50mM Tris pH8.0, 140mM NaCl, 1mM EDTA, 1% Triton X-100, 0.1% Na-deoxycholate, 0.1% SDS, protease inhibitor). Fixed samples were split in half, to generate 'fix – benzonase' and 'fix – sonication' samples. 30U of benzonase was added to 'Fix – benzonase' and 'no fix – benzonase' samples, followed by incubation on ice for 30mins. Concurrently, 'fix – sonication' samples were sonicated in a Biorupter® Pico sonication device with 30sec on, 30 secs off, for 10 cycles. Triton X-100 was added to a final concentration of 1% and samples incubated on ice for 10mins. Insoluble material and debris was removed by centrifugation at 14000rpm for 15mins.

Protein concentration of the lysates was measured by Qubit assay, with 200ug of protein then diluted to 250ul for IP. 5% of each sample was taken as input. 15ul of GFP-Trap bead slurry was equilibrated by washing three times in ice-cold sonication buffer before addition of the diluted lysate and incubation for 1hr at 4ºC. Beads were collected by magnetic separation and 12.5ul of the supernatant taken as a NB sample. Immunocomplexes were washed x2 with 500ul Sonication buffer, with 5mins rotation at 4ºC. To remove non-specifically bound DNA, immunocomplexes were then washed x3 with 500ul of Wash Buffer A (50mM Tris pH 8.0, 500mM NaCl, 1mM EDTA, 1% Triton X-100, 0.1% Na-deoxycholate, 0.1% SDS, protease inhibitor) and x2 with Wash Buffer B (20mM Tris pH 8.0, 1mM EDTA, 250mM LiCl, 0-.5% NP-40, 0.5% Na-deoxycholate, protease inhibitor). Again, samples were rotated for 5mins at 4ºC for each wash. Finally, to remove detergents and salts from the previous washes, immunocomplexes were washed x2 with 500ul TE buffer + 50mM NaCl. Elution of bound material was carried out using Elution buffer (50mM Tris pH 8.0, 1mM EDTA, 1% SDS, 50nM NaHCO3) or SDS-sample buffer, as stated.

For isolation of DNA, samples were reverse cross-linked at 65ºC for at least 10hrs. 5ul of TE buffer was added to reduce SDS present in the solution. 30ug of RNAseA was added and incubated at 37ºC for 1hr before the reaction was

stopped with addition of 0.4ul of 0.5M EDTA. Then, 80ug of Proteinase K was added and incubated at 45ºC for 2hrs. Unless otherwise stated, the DNA was then cleaned using the Qiagen Qiaquick PCR purification kit, according to the manufacturer's instruction.

2.9 Sodium Dodecyl Sulphate-Polyacrylamide gel electrophoresis (SDS-PAGE) and western blotting

Protein samples were mixed 1:1 with 2x Lammeli-sample buffer (BioRad) and denatured at 95ºC for 10mins. Following boiling, proteins were separated on 12% hand-cast SDS-PAGE gels (made up as specified in Table 2) or pre-cast 4-20% Mini-PROTEAN® TGX™ gels (BioRad), as stated. Electrophoresis was carried out in 1x SDS-PAGE running buffer (193mM glycine, 25mM Tris, 0.1% (w/v) SDS, pH 8.6). Precision Plus Protein™ Dual Colour Standard (BioRad) was loaded to each gel to allow monitoring of gel separation and provision of molecular weight standards. Gels were run at 80 V until the proteins reached the end of stacking gel, and at 100-120V until the tracing dye reached the bottom of the gel.

Solution Component	Resolving Gel	Stacking Gel
	12%	4%
dH_2O	3.3 ml	1.7 ml
1.5M Tris (pH 8.8)	2.5 ml	-
0.5M Tris (pH 6.8)	-	420 µl
Acrylamide/bis-acrylamide, 30% solution	4 ml	330 µl
10% (w/v) SDS	100 µl	25 µl
10% (w/v) APS	100 µl	25 µl
N,N,N',N'-tetramethylethylenediamine (TEMED)	10 µl	2.5 µl

Table 2: Component make up of resolving and stacking gels used for SDS-PAGE. Each column denotes the volumes of 1 gel, with components added together in the order shown.

Following separation, gels were electrophoretically transferred onto Immobilon®-P 0.45µm polyvinylidene difluoride (PVDF) transfer membranes (Millipore) by wet transfer. Transfer was carried out in transfer buffer (192mM glycine, 25mM Tris, 20% (v/v) Methanol, pH 8.3) at 100 V for 1.1hrs. Western blotting of proteins ≥ 250 kDa was carried out as above, with the following changes. Samples were

boiled in NuPAGE™ LDS samples buffer + NuPAGE™ Sample Reducing Agent and separated on NuPAGE™ 3-8% Tris-Acetate Protein Gels (ThermoFisher). Gels were run in 1x NuPAGE™ Tris-Acetate SDS Running Buffer (ThermoFisher) plus NuPAGE™ Antioxidant at 150 V for ~1 hr 30 mins. Transfer was carried out overnight at 20V at 4^{o}C in 1x NuPAGE™ Transfer buffer (25mM Bicine, 25mM Bis-Tris (free base), 1mM EDTA, ph7.2) + 10% MeOH + NuPAGE™ Antioxidant.

To prevent non-specific antibody binding, membranes were incubated for a minimum of 1 hour at room temperature in 5% (w/v) non-fat dry milk in PBS-T (1x-PBS, 0.1% (v/v) Tween® 20). Incubation with primary antibodies (as listed in Table 3) diluted in 5% (w/v) non-fat dry milk in PBS-T was carried out overnight at 4^{o}C. Membranes were incubated for 1.5 hour in secondary antibodies made up to 1:10,000 in 5% (w/v) non-fat dry milk in TBS-T. Membranes were washed in PBS-T between each step. To visualize proteins of interest, membranes were incubated for ≤ 4 minutes in working solution of Thermo Scientific™ SuperSignal™ West Femto Maximum Sensitivity Substrate, made up by mixing equal parts of the kit's reagents. The HRP substrate was added in a ratio of 1 ml per full-sized membrane. Membranes were scanned using the ImageQuant™ LAS 4000 with precision exposure and high sensitivity/quality. For fluorescent secondary antibodies, incubation was carried out in the dark and membranes were scanned using the LI-COR® Odyssey® chemiluminescent western blot scanner using a resolution of 169um, medium quality, and a focus of 0.0. Densitometry was carried out using ImageStudio Lite software with statistical significance calculated by unpaired t test, unless otherwise specified. Fold enrichment quantifications were performed by first normalising the raw densitometry value to its corresponding Histone H3 quantification and the comparing between the samples indicated.

Protein	Company	Catalogue No.	Species	Dilution	
				IP	WB
SA1	Abcam	ab4455	mouse	10ug/IP	1:2000
SA2	Bethyl	A300-159	goat	10ug/IP	
SA2	Bethyl	A300-158	goat		1:2000
CTCF	Diagenode	C15410210	rabbit	7.5ug/IP	1:2500
Rad21	Abcam	ab992	rabbit		1:1000
mAID	MBL	M214-3	mouse		1:500
OsTIR	MBL	PD048	rabbit		1:500
SMC3	Abcam	ab9263	rabbit		1:2000
CHD6	Bethyl	A301-221A	rabbit		1:1000
MCM3	Bethyl	A300-124A	goat		1:2000
HNRNPUL2	Abcam	ab195338	rabbit		1:2000
EIF3I	Proteintech	11287-1-AP	rabbit		1:1000
YTHDC1	Abcam	ab122340	rabbit		1:1000
FTSJ3	Bethyl	A304-199A-M	rabbit		1:750
FANCI	Bethyl	A301-254A-M	rabbit		1:500
TAF15	Abcam	ab134916	rabbit		1:5000
HNRNPD	Abcam	Ab61193	rabbit		1:1000
DHX9	Abcam	ab26271	rabbit		1:2500
INO80	Proteintech	18810-1-AP	rabbit		1:750
ESYT2	Sigma-Aldrich	HPA002132	rabbit		1:750
s9.6	Kerafast	ENH001	mouse	5-10ug/IP	1:5000
s9.6	Millipore	MABE1095	mouse	5-10ug/IP	
RNASEH2A	Novus	NBP1-76981	rabbit		1:5000
AQR	Bethyl	A302-547A	rabbit		1:2000
Pol2	Covance	MMS-1289	mouse		1:2000
SETX	Bethyl	A301-105A	rabbit		1:1000
TOP1	Abcam	ab109374	rabbit		1:2000
MAU2	Abcam	ab183033	rabbit		1:1000
NIPBL	Abbiotec	250133	rat		1:1000
NIPBL	Bethyl	A301-779A	rabbit		1:1000
EGFP	Novus	NB600-308	rabbit		1:10000
FLAG	Sigma-Aldrich	F3165	mouse		1:5000
HA	Cell Signalling	3724	rabbit		1:5000
H3	Abcam	ab1791	rabbit		1:10000
Tubulin	Sigma-Aldrich	T9026	mouse		1:5000

Table 3: Details of antibodies used.

2.10 ChIP-seq sample analysis

ChIP lysates were prepared from HCT116 RmAC OsTIR1 cells treated with ethanol or IAA for 6hrs in two biological replicates. Samples were generated and processed by Dr. Stanimir Dulev.

Quality control of reads was preformed using FASTQC. Reads were aligned to the hg19 reference genome using Bowtie with 3 mismatches. PCR duplicates were detected and removed using SAMTOOLS. Bam files were imported into MISHA (v 3.5.6) and peaks were identified using a 0.995 percentile. Peaks that overlapped in both replicates were retained. Only replicate 1 of the SA1 library was used. Correlation plots of peaks across the genome from different ChIP libraries were compared with log-transformed percentiles plotted as a smoothed scatter plot. Comparison of peaks at regions of interest were carried out using deepTools (Version 3.1.0-2 (Ramírez *et al.*, 2016)). For input into deepTools, peak data was converted to bigwig format, with a bin size of 500, using the UCSC bedGraphtoBigWig package. The signal matrix was calculated for a window 2,000 bp up- and down-stream of the region of interest, missing data was treated as zero, and all other parameters were as default. Heatmaps were generated within deepTools, with parameters as default.

2.11 ChromHMM

ChIP-seq data for YY1, CBX3, SIN3A, POLR2A, POLR2AphosphoS5, H3K27ac, H3K4me3, H3K4me1, H3K27me3, EZH2, and H3K9me3 from HCT116 cells were obtained from ENCODE (Table 4). NIPBL data was obtained from Rao *et al.*, (2017). BAM files were binarized in ChromHMM using a bin size of 200 bp and a shift of 150 bp. Where input files were available, they were used in ChromHMM to determine the binarization threshold, else the ChromHMM default of a uniform background was assumed. The chromatin state model was generated for 15 states and compared to the hg19 genome assembly. All other parameters were as default.

Protein	Accession no.	Publication	Matched input
NIPBL (EtOH- and IAA-treated)	GSE104334	(Rao *et al.*, 2017a)	-
CBX1	GSM1010758	(Gertz *et al.*, 2013)	
EZH2	GSM3498250	(Dunham *et al.*, 2012)	GSM2308475; GSM2308476
POLR2A	GSM935426	(Dunham *et al.*, 2012)	GSM2308422
POLR2AphosphoS5	GSM803474	(Gertz *et al.*, 2013)	GSM803475
SIN3A	GSM1010905	(Gertz *et al.*, 2013)	
YY1	GSM803354	(Gertz *et al.*, 2013)	GSM803475
H3K4me1	GSM945858		GSM2308475; GSM2308476
H3K4me1	GSM2527549	(Dunham *et al.*, 2012)	GSM2308422
H3K4me3	GSM2533929	(Dunham *et al.*, 2012)	GSM2308475; GSM2308476
H3K4me3	GSM945304	(Thurman *et al.*, 2012)	GSM945287
H3K9me3	GSM2527565	(Dunham *et al.*, 2012)	
H3K9me3	GSM2308431	(Dunham *et al.*, 2012)	
H3K27ac	GSM2534277	(Dunham *et al.*, 2012)	GSM2308422
H3K27me3	GSM2308612	(Dunham *et al.*, 2012)	

Table 4: Chip-seq datasets used. GEO sample accession numbers for published ChIP-seq datasets analysed in this study. Acession numbers for samples with matched inputs processed for ChromHMM are indicated.

2.12 Hi-C data and contact hotspots analysis

Generating hotspots - Previously published Hi-C datasets derived from HCT116 RmAC OsTIR1 cells treated with ethanol or IAA by (Rao *et al.*, 2017a) were analyzed by Dr. Cristopher Barrington as previously described in (Barrington *et al.*, 2019). Custom R scripts were used to plot the density of the hotspot observed in control and auxin datasets. In addition, the density of the hotspots that overlapped with NIPBL (Rao *et al.* (2017)) and CTCF-SA co-bound sites were plotted.

2.13 Mass spectrometry (MS) sample preparation and analysis

2.13.1 Full lane SA1 IP-MS

SA1 immunoprecipitation samples were analysed by liquid chromatography–tandem mass spectrometry (LC-MS/MS). Five biological replicate experiments were carried out for MS and each included four samples, untreated (UT), treated with IAA for 4hrs, siCon, or siSA1, generated as described above. An extra sample treated with 10ul of NEB RNaseH was included for the first three replicates. Cells were fractionated to purify chromatin-bound proteins as in section 2.5 and immunoprecipitated with IgG- or SA1-bead conjugates. To maximise IP material for the MS, the antibody amount was increased to 15ug and the chromatin amount was increased to 2mg. The IP eluates were loaded into a pre-cast SDS-PAGE gel (4–20% Mini-PROTEAN® TGX™ Precast Protein Gel, 10-well, 50 µL) and proteins were run approximately 1 cm to prevent protein separation. Protein bands were excised and diced, and proteins were reduced with 5 mM TCEP in 50 mM triethylammonium bicarbonate (TEAB) at 37°C for 20 min, alkylated with 10 mM 2-chloroacetamide in 50 mM TEAB at ambient temperature for 20 min in the dark. Proteins were then digested with 150ng trypsin, at 37°C for 3 h followed by a second trypsin addition for 4 h, then overnight at room temperature. After digestion, peptides were extracted with acetonitrile and 50 mM TEAB washes. Samples were evaporated to dryness at 30°C and resolubilised in 0.1% formic acid.

LC-MS was performed by Amandeep Bhamra. Initial data analysis was also performed by Amandeep Bhamra, this paragraph describes a brief account of the analysis he conducted. Raw data was analysed with MaxQuant (Cox and Mann, 2008) version 1.5.5.1 where they were searched against the human UniProtKB database using default settings (http://www.uniprot.org/). To ensure high confidence identifications, PSMs, peptides, and proteins were filtered at a less than 1% false discovery rate (FDR). Statistical protein quantification analysis was done in MSstats (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 in the UTR-IAA comparison, 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 p-value change between UTR and IAA samples. Categories enriched in ii) are indicated on the network by dot lines.

Over-enrichment of the s9.6 interactome with SA1 was calculated separately using the hypergeometric distribution. S9.6 interactome data was obtained from Cristini *et al.*, (2018 and Wang *et al.*, (2018). 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.

2.13.2 Banded SA and CTCF IP-MS

Banded IP-MS experiments were carried out the same as the full lane samples, except for the difference detailed below. SA and CTCF IPs for the banded IP-MS experiments were run as in section 2.5. Proteins were separated on pre-cast 4-20% Mini-PROTEAN® TGX™ gels (BioRad) or NuPAGE™ 3-8% Tris-Acetate Protein Gels, as indicated. Gels were washed in MS grade water to remove SDS and stained in colloidal coomassie blue solution until visible bands were apparent. Background staining was removed by washing in acetic acid and methanol solution. Visible protein bands were excised and processed as above. Amandeep Bhamra performed the LC-MS and identification of peptides, using Proteome Discoverer. Statistical overrepresentation of protein classes within each IP was performed using Panther (version 15). Data was compared to the whole human genome using Fisher's exact test and Bonferroni correction for multiple testing. Non-redundant enriched protein classes were graphed as a pie chart using excel and all output was graph to a bar chart using ggplot in R.

2.14 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 Li *et al.*, (2020) to predict binding to SA proteins. The motif was input as [PFCAVIYL][FY][GDEN]F.{0,1}[DANE].{0,1}[DE]. Along with CTCF, five proteins found to contain the FGF-like motif, CHD6, MCM3, HNRNPUL2, EIF3I, and ESYT2 were validated for interaction with SA.

3

CTCF and SA can interact independently of the cohesin ring

3.1 Introduction

As discussed in detail in section 1.4.3, the SA proteins are thought to bridge interaction between the cohesin ring and CTCF, however, little is understood of the roles SA1 and SA2 play in cohesin activity. Furthermore, the current literature has predominantly investigated this interaction from tagged, truncated, or recombinantly expressed versions of the proteins – techniques that may not reveal the full story of the interaction and could influence how the proteins interact. To examine interaction of CTCF and the SA proteins in cells in unmodified conditions, I carried out co-immunoprecipitation (co-IP) using endogenous antibodies targeting CTCF, SA1, and SA2. While avoiding disruption due to epitope tagging, it is important to note that the endogenous antibodies used for IP may bind to functional sites of the proteins and also influence interactions. Reciprocal co-IP from CTCF and SA may help to identify such influence as only one of the interacting proteins will be targeted in each IP. Co-IP was tested in control and RAD21 knockdown conditions to assess the impact of the cohesin ring on interaction between CTCF and SA1/2. ChIP-seq and mass spectrometry methods were also used to build a more comprehensive understanding of the association between CTCF and SA1/2.

As discussed in detail in section 1.4.3, the SA proteins are thought to bridge interaction between the cohesin ring and CTCF, however, little is understood of the roles SA1 and SA2 play in cohesin activity. Furthermore, the current literature has predominantly investigated this interaction from tagged, truncated, or recombinantly expressed versions of the proteins – techniques that may not

reveal the full story of the interaction and could influence how the proteins interact. To examine interaction of CTCF and the SA proteins in cells in unmodified conditions, I carried out co-immunoprecipitation (co-IP) using endogenous antibodies targeting CTCF, SA1, and SA2. Co-IP was tested in control and RAD21 knockdown conditions to assess the impact of the cohesin ring on interaction between CTCF and SA1/2. ChIP-seq and mass spectrometry methods were also used to build a more comprehensive understanding of the association between CTCF and SA1/2.

3.2 Results

3.2.1 Characterisation of HCT116 RmAC OsTIR1 cells

HCT116 RmAC OsTIR1 cells were obtained from Natsume *et al.* (2016), in which, RAD21 is tagged with mini-Auxin-Inducible Degron (AID) sequences and mClover, and *Oryza sativa* TIR1 (OsTIR1) is expressed from the AAVS1 region. In the presence of auxin (also known as IAA), expressed OsTIR1 can form a functional SCF^{OsTIR1} E3 ligase complex with endogenously expressed proteins and polyubiquitinate AID sequences to target AID-tagged proteins for rapid, proteasomal-mediated degradation (Figure 5).

This cell line has now been used for multiple studies of cohesin activity. Most notably, Rao *et al.* (2017) determined that cohesin loss after 6 hrs of auxin treatment induced a loss of all chromatin loops, without affect to A/B compartment domains. They found gene expression to widely remain stable within this timeline, apart from new clustering of superenhancers and mis-expression of a minority of active genes. Oldach and Nieduszynski (2019), determined that acute depletion of RAD21 in these cells does not affect replication timing. The authors conclude that cohesin activity does not influence replication timing domains, however, their auxin-treatments seemed to be for 2-3.5 hrs, a shorter time frame compared to all other papers. Finally, Cremer *et al.*, (2020) used super-resolution imaging techniques to determine that with 21 hrs of auxin treatment endomitosis occurs as the chromosomes are duplicated but the cell nucleus does not divide. The

authors report that compartments can be successfully rebuilt despite the loss of chromatin loops, but increased heterogeneity and volume of replication domains occurs, implicating cohesin activity in the constraint of replication domains. Together these papers confirm that cohesin acts as an important regulator of chromatin loops and that its roles extend to chromatin organisation at the level of replication timing. They also indicate that, to avoid adverse effects, auxin treatment should be kept to short incubation times of ~6 hrs, which is sufficient to deplete cohesin and its mediated structures in the cell.

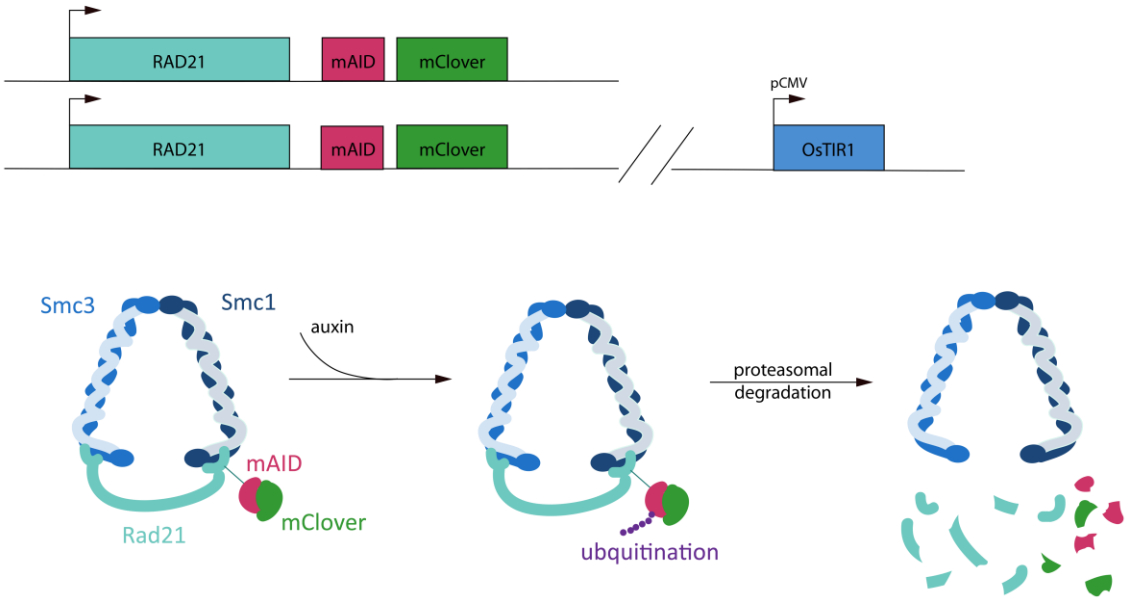


Figure 5: Schematic of RAD21 depletion in HCT116 RmAC OsTIR1 cells. Tagging of RAD21 with mAID and mClover and integration of OsTIR1 in the genome are shown (top). With addition of auxin to the cell media, a functional SCF^{OsTIR1} E3 ligase complex is generated and mAID is ubiquitinated. Thus, RAD21 is targeted for rapid proteasomal-mediated degradation. mAID = mini-Auxin-Inducible Degron.

To test auxin-mediated degradation of mAID-tagged RAD21 in the HCT116 RmAC OsTIR1 cell line, cells were grown to ~70% confluence and treated with either ethanol (as a solvent control) or auxin. Auxin treatment was tested at multiple timepoints to assess efficiency of knockdown over time. Cells were fractionated to obtain chromatin-bound proteins and effect on RAD21 and SMC3 was assessed by immunoblot. Using this system, Natsume *et al.* (2016) report that RAD21 can be rapidly removed from chromatin, with a half-life of 17 min. We found however, that RAD21 levels were never fully lost and further, began to recover over 24hrs in auxin-containing media (Figure 6). Correspondingly, SMC3 levels on chromatin were not significantly altered at any of the timepoints tested.

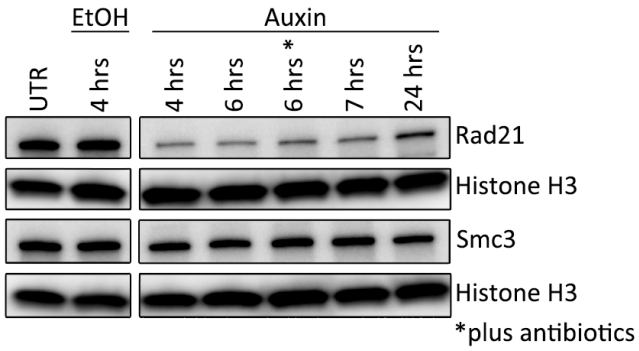


Figure 6: Depletion of RAD21 in HCT116 RmAC OsTIR1 cells – polyclonal cells. Timecourse of RAD21 depletion from chromatin following treatment with ethanol or auxin for the indicated times. The sample marked with an asterisk was generated from cells maintained in selection antibiotics for ~3 weeks. SMC3 levels were also assesed. UTR = Untreated. Histone H3 served as a loading control.

Several factors were considered to identify the cause of the lack of efficiency observed. Firstly, the use of CRISPR-Cas9 machinery to generate the RmAC OsTIR1 cell line may have introduced genetic instability into the cell line. Using long-read sequencing and fluorescence in situ hybridization (FISH)-based technologies multiple papers have now reported large scale deletions and rearrangements at Cas9 target loci (Kosicki, Tomberg and Bradley, 2018; Rayner *et al.*, 2019). These mutations are not always caught by polymerase chain reaction (PCR) methods which are commonly used to verify the genome edit. The authors further determined that these mutations are a consequence of Cas9 activity and not clonal or antibiotic selection (Rayner *et al.*, 2019).

The two RAD21 alleles in the RmAC OsTIR1 cell line are differentially tagged with hygromycin- and G418-resistance markers and OsTIR1 is tagged with a puromycin resistance marker. Thus, ideally, cells with homozygous-tagged RAD21 and integrated OsTIR1 can be selected for and maintained with triple antibiotic selection. Growth in the absence of antibiotics may have selected for cells with mutations or loss of the tags. To test this, the cells were grown for a prolonged period in antibiotics, however, this did not restore complete degradation of RAD21 (Figure 6, sample indicated by an asterisk). This suggested that major large-scale deletions had not occurred and the three antibiotic markers, at least, were retained.

Propensity to undergo Cas9-induced large-scale rearrangements can be influenced by the underlying chromosome stability of the cell line (Rayner *et al.*,

2019). HCT116 has been characterised as a near-diploid, chromosomally stable colon cancer cell line (Lengauer, Kinzler and Vogelstein, 1997), and, as such, is less likely to show very large-scale or karyotypic differences following CRISPR-Cas9 editing (Rayner *et al.*, 2019). Nevertheless, smaller-scale deletions and rearrangements that render the cells less responsive to auxin are still possible. HCT116 cells are established to be genetically instable as they are deficient in DNA mismatch repair (MMR) and show microsatellite instability (MSI or MIN) due to biallelic loss of MutL protein homolog 1 (MLH1). In fact, microsatellite mutation has been recorded at the markedly high rate of ~0.01 mutations per cell per generation (Bhattacharyya *et al.*, 1994). MISA, a web-based microsatellite prediction tool was used to predict microsatellites in the tagged RAD21 sequence (Beier *et al.*, 2017). RAD21-mAID-mClover was found to contain 17 microsatellites. Hence, the region could be subject to mutation over a relatively short period of growth. Such mutation may then have allowed the cell population to shift to become less responsive to auxin-mediated degradation, especially if these cells had a growth advantage over the parental clone.

To avoid residual RAD21 and questions of mutation complicating experiments, we obtained a new batch of HCT116 RmAC OsTIR1 cells from Natsume *et al.* (2016) and Dr. Yang Li, a postdoc in the Hadjur lab, employed fluorescence-activated cell sorting (FACs) to select for cells with a shift from strongest mClover fluorescence to no mClover with treatment of auxin. Four clones of sorted cells were obtained – termed H1, H2, H6, and H11 (Figure 7A). Based on efficacy of RAD21 knockdown and corresponding loss of SMC3 from chromatin, two clones – H2 and H11 – were expanded and used for future experiments. Experiments carried out in the original ‘polyclonal’ RmAC OsTIR1 cells, the H2 clone, or the H11 clone are indicated throughout. The H2 and H11 clones were broadly interchangeable, except for in the ‘reloading’ experiments in Chapter 5.

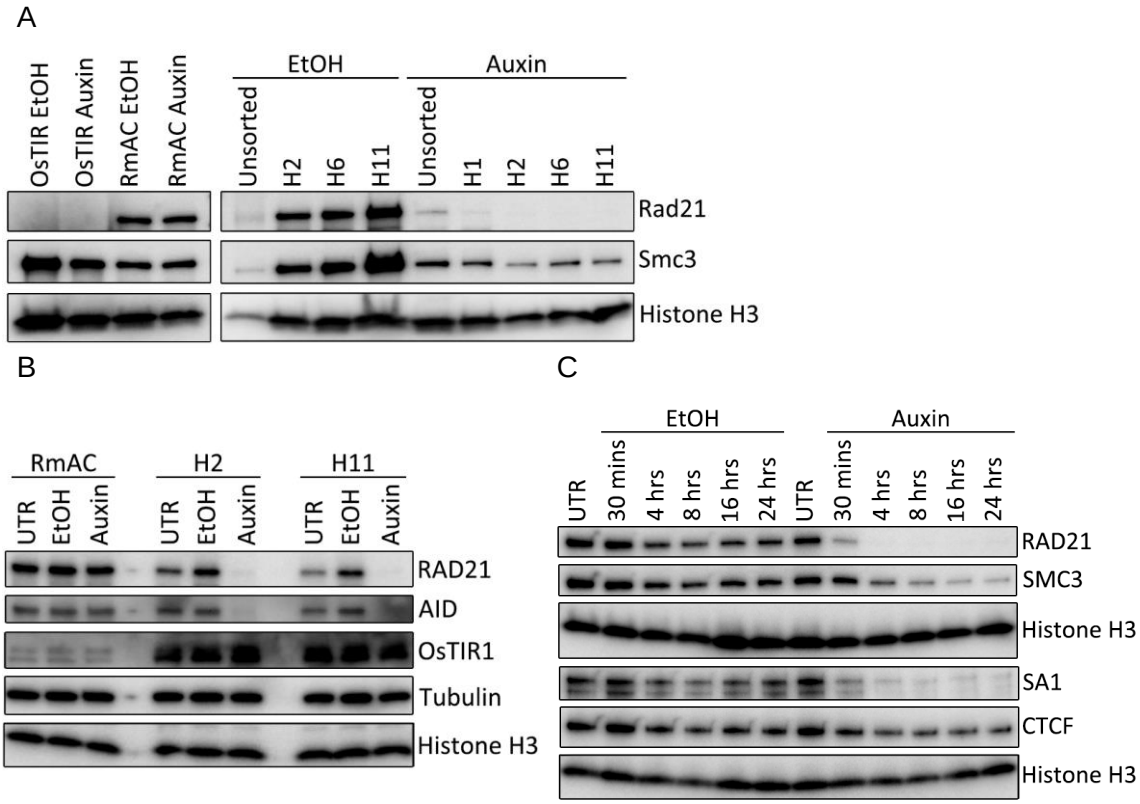


Figure 7: Depletion of RAD21 in HCT116 RmAC OsTIR1 cells – H2 and H11 clonal cells. (A) Comparison of RAD21 and SMC3 levels on chromatin in OsTIR1, RmAC, and RmAC OsTIR1 cell lines treated with ethanol or auxin. RAD21 signal is covered for the OsTIR1 samples due to the strength of the signal compared to that of the tagged-RAD21. H1, H2, H6, and H11 represent four RmAC OsTIR1 FACS-sorted clones. (B) WCE samples from RmAC and two of the RmAC OsTIR1 clones (H2 and H11) immunoblotted for RAD21, AID, and OsTIR1 proteins to confirm specificity of the SCF^{OsTIR1} E3 ligase in RmAC OsTIR1 cells. (C) Timecourse of effect of ethanol and auxin treatment on RAD21, SMC3, and CTCF levels on chromatin. UTR = Untreated. Tubulin and Histone H3 served as loading controls.

RAD21, AID, and OsTIR1 levels were tested in whole cell extracts (WCEs) from H2 and H11 clones to ensure correct expression and response to auxin treatment (Figure 7B). RmAC cells (also obtained from Natsume *et al.*, (2016)), which do not contain OsTIR1 in the AAVS1 locus, were also tested to ensure that, without formation of a functional ligase, auxin itself does not affect tagged RAD21 levels on chromatin. Similarly, ethanol controls were included to ensure that the auxin solvent did not affect the proteins tested. For both the H2 and H11 clones, RAD21 and AID were specifically degraded in the presence of OsTIR1 and auxin.

A timecourse of auxin treatment over 24hrs was carried out in H2 cells to test the longer-term affect of auxin treatment on RAD21, SMC3, SA1, and CTCF (Figure 7C). RAD21 was completely lost from chromatin by 4hrs, with no recovery observed across the 24hrs tested. Alongside RAD21 loss, SA1 and SMC3 levels on chromatin were reduced by 4hrs. Both proteins retained some ability to interact

with chromatin in the absence of RAD21 and consequent disruption of cohesins ring structure. CTCF levels have been reported to remain unchanged following cohesin loss from chromatin (Parelho *et al.*, 2008; Rao *et al.*, 2017), however, interestingly, CTCF levels on chromatin appeared slightly reduced in this case. CTCF reduction did not increase over the 24hr timecourse tested. From these experiments, 4hrs of auxin treatment was selected as an optimal timepoint for RAD21 degradation to limit effects on cell cycle progression while efficiently removing RAD21 from chromatin and reducing SMC3 and SA1 levels on chromatin.

3.2.2 Optimisation of co-immunoprecipitation protocol

Alongside determination of RAD21 depletion conditions, co-IP conditions were optimised. I optimised a co-IP protocol to detect interaction between members of the cohesin complex, and its regulators. Here, the starting protocol is described, followed by details of the modifications I tested and established for robust and reproducible co-IP in HCT116 RmAC OsTIR1 cells.

The co-IP protocol was as follows (Figure 8). Specific details can be found in the methods section 2.5. Dynabead™ Protein A/G were equilibrated in PBS-0.1% NP-40 buffer and bound to the antibody of choice. In the interim, chromatin was purified from the cells using an altered version of Mendez and Stillman (2000) subcellular fractionation protocol. Cells were lysed in a buffer containing sucrose, glycerol and a mild detergent, and K^+ and Mg^{+2} to preserve tertiary protein structure and ensure intact nuclei. Nuclei were then separated from the cytoplasmic solution by low speed centrifugation and burst in a buffer containing the chelating agents EDTA and EGTA. Insoluble nuclear material, including chromatin, was spun down. The chromatin fraction was isolated by solubilisation in a high salt buffer (containing 500mM KCl). Chromatin proteins were purified by nucleic acid digestion with benzonase (1U per 100 x 10^6 cells) and insoluble material was sheared by ultrasonication (3 x 10 secs). Remaining insoluble material was removed by centrifugation and the supernatant was retained as the purified chromatin protein fraction. The solubilisation buffer was diluted with a no-salt equivalent to achieve a mild salt buffer (300mM KCl) for IP. The bead-

antibody conjugate was then washed and incubated with 500ug of the purified chromatin protein overnight. Separation of the beads on a magnet isolated the IP material from the non-bound (NB)/flow through solution which was put aside for analysis. IP material was then washed on the beads before elution and western blot analysis or further downstream manipulation and analysis.

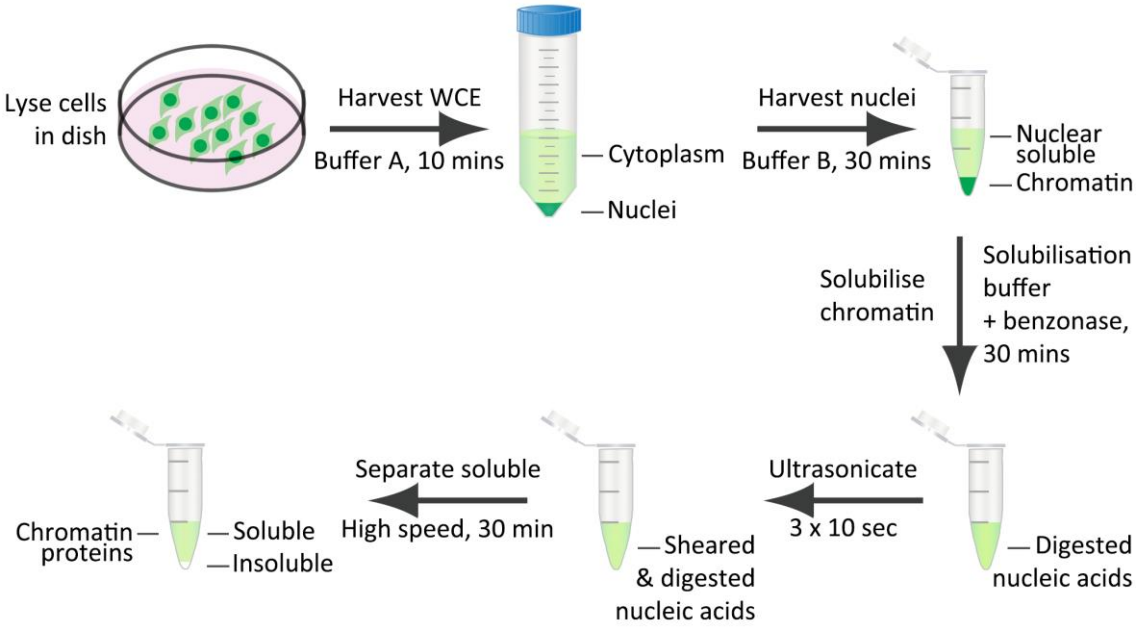


Figure 8: Schematic of chromatin fractionation protocol. A brief summary of the fractionation protocol used to obtain chromatin bound proteins is shown. WCE = Whole Cell Extract.

3.2.2.1 Co-IP of cohesin complex members

Using the co-IP protocol described above in HCT116 RmAC cells, IP of both CTCF and SA1 was achieved, with enrichment over input (Figure 9A). However, there was no co-IP of SA1 or SMC3 with CTCF. Similarly, only SMC3 was enriched with SA1, and at much lower levels than input. Hence, it was necessary to optimise the protocol to robustly co-IP cohesin components and additional interacting proteins such as CTCF.

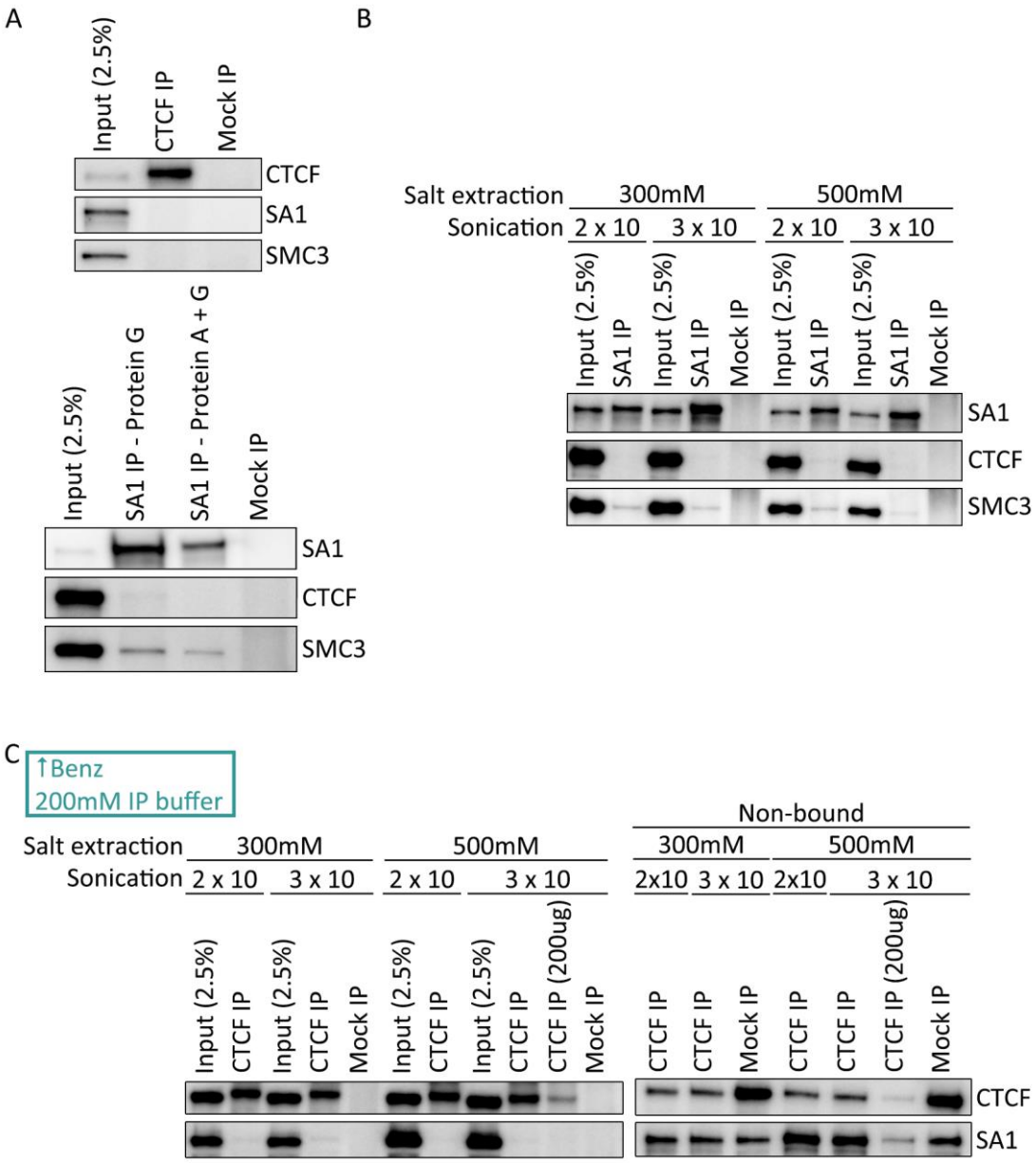


Figure 9: Optimisation of co-IP of cohesin complex members – Part 1. Changes to the fractionation and IP protocol between experiments are indicated in a green box. (A) IP of chromatin-bound proteins using endogenous CTCF (top) or SA1 (bottom) antibodies. For SA1 IP, two bead-antibody conjugate samples were used, one generated using just protein G beads and one generated using A and G protein beads. (B) Assessment of salt extraction and sonication conditions on SA1 IP and co-IP of CTCF and SMC3. IPs were incubated with chromatin-bound samples generated with the indicated differences in salt extraction of chromatin-bound proteins and shearing of nucleic acids. (C) Assessment of salt extraction and sonication conditions on CTCF IP and co-IP of SA1. IP (left) and non-bound (right, indicated) material are shown. For the CTCF IP (200ug) sample the concentration of protein incubated in the IP was reduced from 500ug to 200ug. Benz = Benzonase. Mock IP samples represent pull down with species matched IgG antibodies.

To determine if chromatin solubilisation conditions were disrupting protein interactions, chromatin solubilisation in lower salt conditions and reduced sonication was tested for effect on IP of SA1 and co-IP of CTCF and SMC3 (Figure 9B). SA1 IP was slightly increased in the reduced salt concentration and

the higher sonication condition, however, co-IP of SMC3 was equally poor in all samples and no co-IP of CTCF was observed. In addition, SA1 enrichment in this second experiment was reduced over input compared to the first experiment. It was unclear if this was caused by variation in chromatin solubilisation or use of a new tube of SA1 antibody. To avoid potential SA1 antibody variation, the next optimisation experiment was carried out using IP of CTCF. Although both the SA1 and CTCF antibodies used are polyclonal antibodies, which carry inherent variation, the CTCF antibody is classified as 'ChIP-seq Grade'. This means that the antibody has been highly validated for quality, including sensitivity and specificity. Thus, this antibody should have minimal variation across IPs.

The experiment was repeated with three changes to try to improve co-IP with CTCF (Figure 9C). Firstly, to potentially improve chromatin solubilisation, an increased benzonase ratio of 6U per 100 x 10^6 cells was used instead of the previously employed ratio of 1U per 100 x 10^6. Secondly, to preserve protein interactions during solubilisation and IP, the chromatin extract was diluted further to 200mM KCl instead of 300mM for insoluble material removal and IP. Thirdly, to avoid shocking proteins interactions during IP, the protein beads were washed and incubated in the 200mM chromatin solubilisation buffer rather than PBS-0.1% NP-40. As in the previous experiment, chromatin was extracted using either 300 or 500mM solubilisation buffer and sonicated for either 2 x 10 secs or 3 x 10 secs. Again, IP efficiency over input was reduced compared to the first experiment, however, CTCF IP was similar across the conditions tested. None of the conditions tested gave rise to co-IP of SA1. An extract of the non-bound solution was run on a western blot and showed that the CTCF antibody set-up was enriching the majority of CTCF from the chromatin material, however, SA1 levels were unchanged compared to a mock IP (Figure 9C). SMC3 was not assessed in this experiment. These results showed that the chromatin solubilisation conditions did not facilitate efficient co-IP.

A third optimisation experiment was set up to determine if the lack of co-IP observed thus far was due to loss of protein interactions during the fractionation process as a whole (Figure 10A). HCT116 RmAC cells were fractionated with the three alterations described above, however, 1/3rd of the collected cells were left in buffer A for ≥ 4hrs to burst all membranes and act as whole cell extract (WCE)

sample. In addition, the number of plates collected was increased from ~2/3 to 6 15cm dishes, to more closely match the amount used by others in the lab and to assess if the lack of co-IP observed thus far was due to insufficient material. No SA1 IP was detected from the WCE sample, whereas SA1 was IP'd from both 500ug and 200ug of chromatin material, albeit at differing amounts. In this instance, increased SMC3 co-IP was observed for the 500ug chromatin sample, suggesting that increasing the number of cells helped to preserve protein interactions, compared to previous experiments. Low level SMC3 co-IP was also present in the WCE sample, however higher SMC3 signal was also present in the WCE mock IP, thus this may represent non-specific pull down in the buffer A condition.

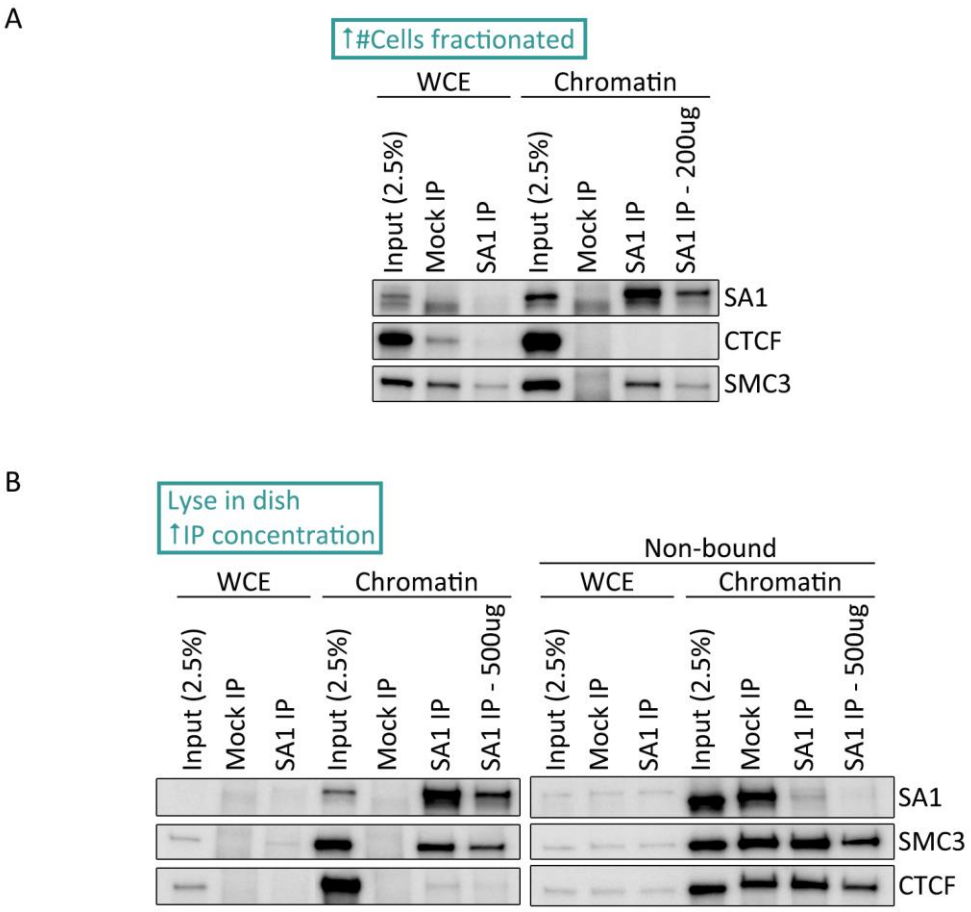


Figure 10: Optimisation of co-IP of cohesin complex members – Part 2. Changes to the fractionation and IP protocol between experiments are indicated in a green box. (A) SA1 IP from WCE and chromatin samples. The sample marked as 200ug had a reduced concentration of proteins incubated in its IP (from 500ug). (B) Successful co-IP of SMC3 with SA1. IP of WCE or chromatin proteins with endogenous SA1 antibody. Unless indicated, 1mg of proteins was incubated per IP. IP (left) and Non-bound (right, indicated) material are shown. Benz = Benzonase. Mock IP samples represent pull down with species matched IgG antibodies.

The final optimisation experiment was set-up to further preserve protein interactions and make the fractionation process less stressful. A second high cell number experiment was performed and here the cells were lysed directly in the dish rather than collection by trypsin and lysis in a falcon tube (Figure 10B). Cell number could not be calculated in this case so, except for buffer A, the volumes of all buffers were kept the same as for the previous experiment with an estimate of collecting ~120 x 10^6 cells in total from the 6 dishes. To minimally cover the surface of the 15cm dish the ratio of buffer A to cell number had to be doubled compared to lysing in a tube. To try to maximise the IP, the amount of antibody used for IP and the concentration of purified protein added to the IP were doubled, with one SA1 IP included with the original 500ug of chromatin protein added to the increased antibody amount. To again try to test if the fractionation process was preventing co-IP, a WCE sample was set up as above, with two modifications. The WCE sample was spun down at 12,000g for 10mins to remove any large insoluble aggregates that may interfere with the IP and the WCE sample was diluted in the chromatin 200mM KCl IP buffer to try to prevent interference from buffer A itself. Finally, non-bound material from the IP samples were also run on a western blot to determine if the co-IP proteins were not pulled down with the beads or were lost during washing stages.

No SA1 was left in the non-bound fraction for both the 500ug or 1mg samples, and increased SA1 IP was seen for the 1mg sample, indicating that this was an optimised IP set-up (Figure 10B). Similar co-IP enrichment of SMC3 over input was obtained for the 500ug sample compared to the previous experiment, indicating that 500ug is saturated with 5ul of SA1 antibody. SMC3 co-IP scaled approximately with the SA1 IP, suggesting that increasing chromatin amount for the IP increases the total signal but does not affect the efficacy of co-IP. Overall, this experiment demonstrated that increasing cell number and lysing directly in the dish increased co-IP of SMC3 with SA1, potentially by allowing the proteins to remain in complex during the fractionation process.

No efficient SA1 IP or SMC3/CTCF co-IP was observed for the WCE sample, however, non-bound protein levels for the WCE samples were also very low, so this sample was likely not generated correctly as a WCE and cannot be used to comment on the effect of the fractionation process on co-IP.

RAD21 levels on chromatin are decreased in HCT116 RmAC and RmAC OsTIR1 cells compared to HCT116 OsTIR1 cells (obtained from Natsume *et al.*, (2016)), in which, RAD21 is not modified (Figure 11A). Reduced RAD21 on chromatin may be due to decreased expression or increased degradation compared to endogenous levels. To check if reduced RAD21 levels were affecting co-IP of SMC3, at least, with SA1, an identical co-IP experiment was set up using OsTIR1 cells (Figure 11B). A very large nuclear pellet was obtained, requiring double the volume of chromatin solubilisation buffer for resuspension and an extra round of sonication for solubilisation. For this experiment, an SA2 IP was also included to assess co-IP of SMC3 and CTCF in comparison to SA1. For the SA1 IP, no increase in SMC3 co-IP was observed with the change in cell type, and CTCF co-IP was still not achieved, indicating that RAD21 levels were not affecting co-IP results. SA2 pull down approximately doubled SMC3 co-IP, demonstrating the efficacy of the optimised protocol for co-IP of cohesin complex members. Yet, co-IP of CTCF was absent from both SA1 and SA2 IPs, illustrating the need for further optimisation.

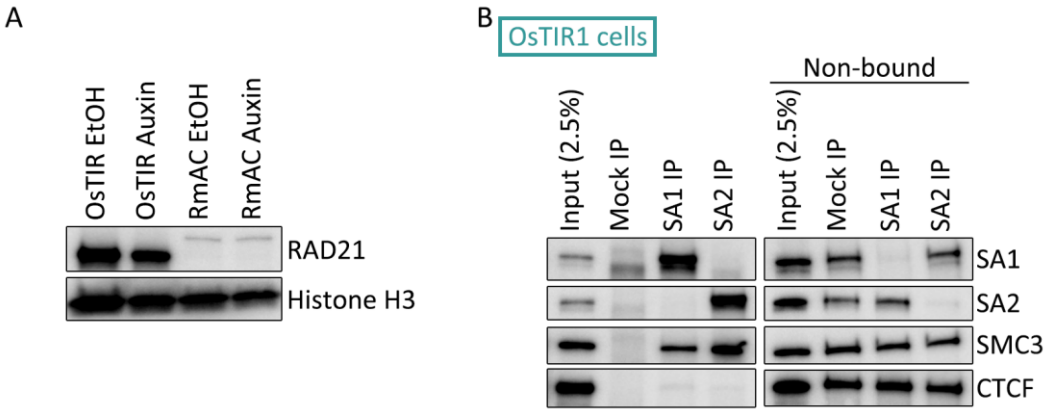


Figure 11: Increased RAD21 levels do not improve co-IP of with SA proteins. (A) Chromatin samples from Figure 1 B with OsTIR1 samples uncovered to show the difference in RAD21 levels on chromatin between untagged (OsTIR1) and tagged (RmAC) cell lines. Histone H3 is blotted as a loading control. (B) IP of chromatin-bound proteins using endogenous SA1 or SA2 antibodies from OsTIR1 cells. IP (left) and Non-bound (right, indicated) material are shown. Changes to the fractionation and IP protocol between experiments are indicated in a green box. Mock IP samples represent pull down with species matched IgG antibodies.

3.2.2.2 Co-IP of cohesin and interacting proteins (CTCF)

Despite having now optimised co-IP of a cohesin ring member with SA1, co-IP of CTCF remained very weak (Figure 10B and Figure 11B). In addition, equal levels of SMC3 and CTCF were observed in mock and IP non-bound samples, meaning

that the protocol was still not completely optimised for co-IP and interaction between the proteins was still being lost during processing or IP. Therefore, further optimisation was still required to ascertain a protocol for co-IP of SA1 and CTCF.

Thus, the optimised ‘lysis in plate’ conditions were combined with a titration of benzonase concentration in the HCT OsTIR1 cells to determine if revised conditions of chromatin digestion would now increase SA1 and CTCF co-IP (Figure 12A). Cells were processed in batches of two 15cm dishes to allow multiple benzonase conditions to be tested without requiring overwhelming amounts of material. In addition, the volume of chromatin solubilisation buffer was reduced back to the original amount to prevent negative impact on the co-IP. 0U of benzonase greatly reduced SA1 input and IP levels, indicating the critical importance of nucleic acid digestion to the experiment. 2 and 6U of benzonase maintained the successful SMC3 co-IP observed in the previous few experiments and here resulted in successful co-IP of CTCF. While CTCF levels in the 2 and 6U benzonase non-bound mock and SA1 IP samples were still even, the protocol could now be used to co-IP CTCF and SA1. This successful 6U benzonase co-IP protocol was then used to test for reciprocal pull down of CTCF and SA2 (Figure 12B). Minimal co-IP of CTCF and SA2 was observed. Perhaps due to weaker interaction of CTCF and SA2 or perhaps due to a requirement for different conditions to capture CTCF and SA2 in complex.

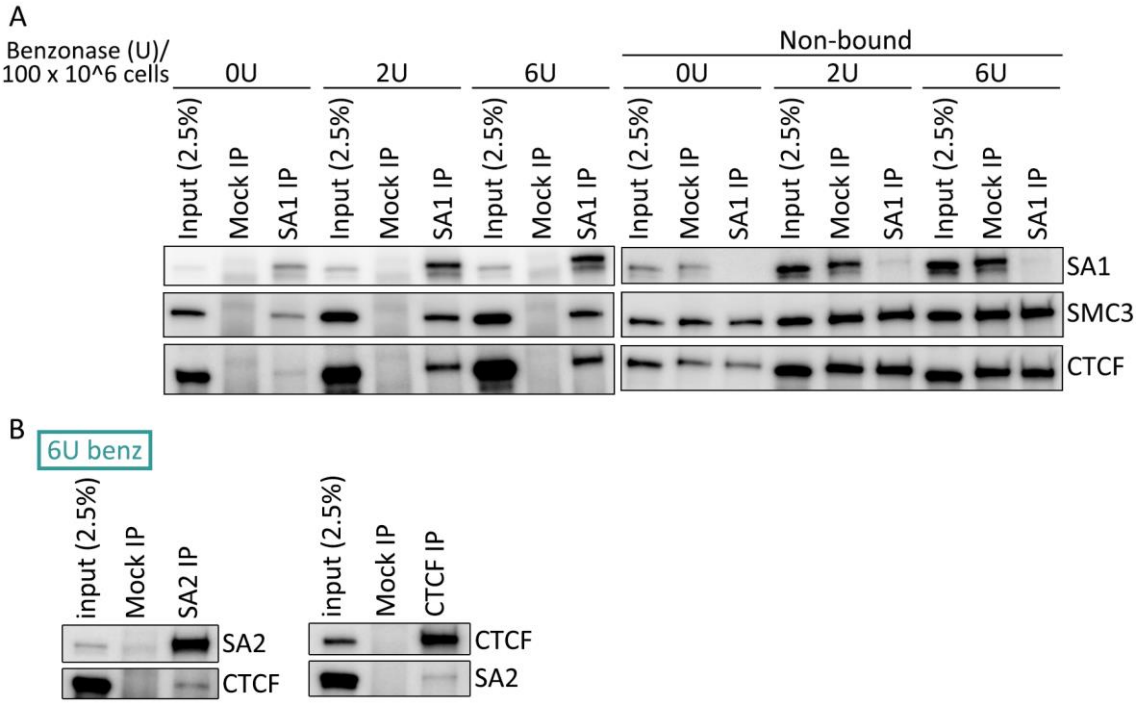


Figure 12: Optimisation of CTCF co-IP with SA proteins. (A) Optimisation of benzonase digestion during sample fractionation for SA1 IP and CTCF/SMC3 co-IP. Samples were generated with treatment of 0, 2, or 6U of benzonase per 100x10^6 cells during fractionation. IP (left) and Non-bound (right, indicated) material are shown. (B) Assessment of the 6U of benzonase per 100x10^6 cells condition on SA2 and CTCF co-IP. IP samples are generated from chromatin protein samples. Changes to the fractionation and IP protocol between experiments are indicated in a green box. Mock IP samples represent pull down with species matched IgG antibodies.

3.2.3 CTCF and SA1 interact in the absence of RAD21

To investigate CTCF–SA interaction in the presence or absence of RAD21 and the cohesin ring, HCT116 RmAC OsTIR1 cells were treated with ethanol or auxin for 4-6hrs before fractionation to obtain purified chromatin-bound proteins. The first tests were performed in the polyclonal RmAC OsTIR1 cells because the clonal lines were not produced at the time. Chromatin proteins were IP'd with an antibody for CTCF, SA1, or SA2 and then immunoblotted for co-IP (Figure 13 A - C). Different amounts of chromatin were used for IP in each of the three experiments due to the availability of material, with 750ug used in Figure 13 A and C, and 500ug used in Figure 13B. Note, in Figure 13C, * indicates 290ug of chromatin was used for the SA2 auxin IP. Auxin treatment was altered slightly between the experiments due to time constraints of additional experiments. Cells were treated for 4, 6, and 5 hrs for the three experiments, respectively. The co-IP method used was the optimised 6U benzonase version from the previous section.

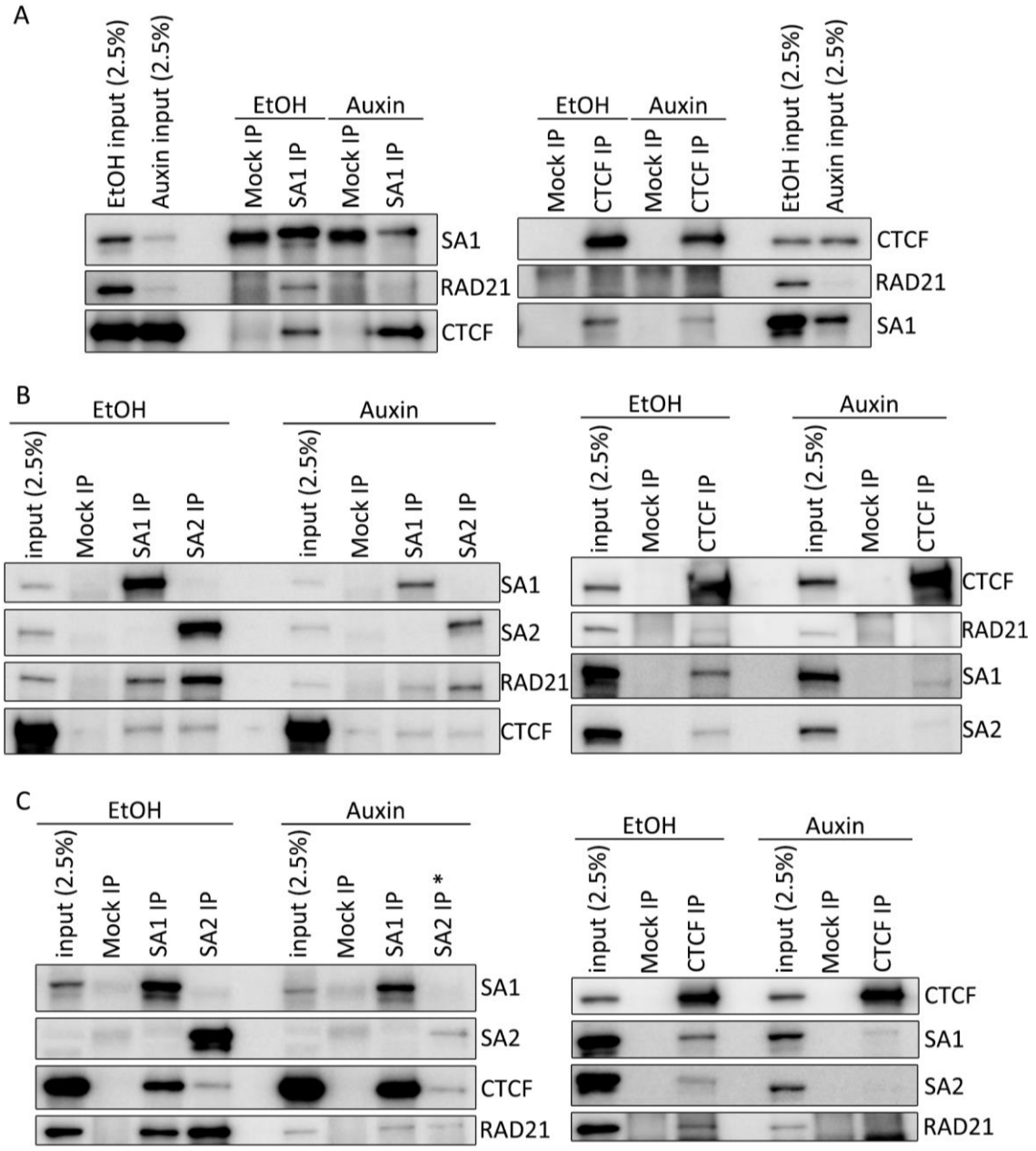


Figure 13: Co-IP of CTCF and SA in polyclonal HCT116 RmAC OsTIR1 cells treated with ethanol or auxin. (A) IP from 750ug of chromatin material using endogenous SA1 (left) or CTCF (right) antibody or species matched IgG (mock). Cells were treated with ethanol or auxin for 4 hrs prior to collection and corresponding IP samples are indicated. (B) Repeat of (A) with 500ug of chomatin material incubated per IP and inclusion of an endogenous SA2 IP. (C) Repeat of (B) with 750ug of chromatin material per IP. *SA2 IP marked by an asterisks was incubated with a reduced concentration of chromatin material (290ug) due to a lack of available material. EtOH = Ethanol.

Across all three experiments input samples showed that SA1 levels on chromatin were reduced with auxin treatment. Despite this decrease, SA1 could still be IP'd, indicating that residual SA1 remained on chromatin, albeit at reduced levels compared to ethanol-treated controls. SA1 IP in the first experiment was relatively weak, as evidenced by the strength of the non-specific band in the mock IP samples that is localised just below the SA1 band (Figure 13A). When the SA1

IP has worked well (and enrichment is high) this band is not visible, as in Figure 13 B and C. In contrast, CTCF levels on chromatin remained constant with equal levels observed in input and IP samples regardless of treatment.

RAD21 was detected in all of the ethanol-treated IP samples, although its signal was more difficult to detect in CTCF IPs due to higher levels of background signal from cross-reaction of the secondary antibody. Input and co-IP levels of RAD21 were reduced with auxin-treatment, indicating the rapid degradation of RAD21 from the chromatin. However, as discussed in section 3.2.1, residual RAD21 was observed in auxin-treated polyclonal HCT116 RmAC OsTIR1 cells and was evident in the SA IPs shown here. This is especially true of the SA2 IP in Figure 13 B and the SA1 IP in Figure 13 C.

Robust co-IP of CTCF was observed in ethanol-treated samples from SA1 IPs where 750ug of chromatin material was used (Figure 13 A & C). Interestingly, co-IP of CTCF and SA1 was observed in both ethanol and auxin conditions from all three experiments. In fact, in Figure 13 A and C, in which robust ethanol co-IP signal was obtained, the auxin CTCF co-IP was even enhanced over ethanol. In contrast, despite the fact that SA2 was reproducibly IP'd in these conditions, CTCF and SA2 showed weaker co-IP, even in ethanol conditions. Co-IP of RAD21 was stronger for SA2 IP samples than SA1 IP samples, indicating that the weakness of CTCF and SA2 co-IP was not just a consequence of poor co-IP conditions for SA2.

The residual RAD21 observed in the auxin SA IPs above raised the question of whether the retained SA1-CTCF co-IP was simply a function of left-over RAD21 on chromatin. When the new clones were generated, H2 and H11 were used to repeat the co-IP experiment, now with complete loss of RAD21 at the 4hr auxin timepoint. A first test was done with CTCF IP and immunoblot for co-IP of SA1 and SA2 (Figure 14A). Unfortunately, CTCF signal in the ethanol H2 and H11 samples burnt on the membrane, making it impossible to examine their relative IP efficiency in this case. Although, the fact that they burnt before the auxin sample likely suggests a higher level of CTCF in these two samples. Co-IP of RAD21, SA1, and SA2 was weak across all samples. Nuclei were accidentally burst in half the volume of buffer B than previously, perhaps negatively impacting

on isolation of chromatin bound protein complexes. Despite the technical difficulties, co-IP of SA1 was stronger than SA2, again suggesting that SA1 interacts with CTCF more strongly that SA2. With the complete loss of RAD21 in auxin-treated samples, SA2 was now strongly depleted from chromatin. Alongside the relative strength of SA2-RAD21 co-IP shown above, this suggests that SA2 interacts more strongly with RAD21.

Before a full co-IP experiment was run, an SA1 optimisation check was carried out in the H2 clones to determine if the previously optimised co-IP conditions remained optimal in these new clones (Figure 14B). As previously, cells were collected in batches of two 15cm dishes and used to test three chromatin solubilisation conditions; i) as previously, solubilisation in 500mM KCl solubilisation buffer, vortex for 2mins, and 3 x 10 sec bursts of sonication, ii) a vortex test with the same buffer and sonication conditions but no vortexing, and iii) a low salt, low sonication test with solubilisation in 150mM KCl solubilisation buffer, vortex for 2 mins, and 2 x 10 sec bursts of sonication. SA1 IP was similar across the conditions, although slightly reduced in the lower salt, lower sonication sample. CTCF co-IP remained optimal in the previously optimised chromatin collection conditions.

A full test of CTCF and SA1 interaction was hence run using the H2 and H11 clones and the optimised co-IP protocol (Figure 14C). In these cells, RAD21 signal was abolished in the input of auxin-treated samples, confirming complete loss from chromatin. SA1 and CTCF were both enriched over input in their respective IPs, indicating good IP. As in the experiments above, input samples indicated that SA1 levels on chromatin were reduced with auxin-treatment while CTCF remained constant. Despite the decrease in SA1 levels, it could still be IP'd, indicated retained stability on chromatin. IP of SA1 was similar in the H2 and H11 clones. IP of CTCF was also similar in the H2 and H11 clones, although there did appear to be an issue with the H11 CTCF ethanol IP. A replicate H11 CTCF IP did not show the same lack of signal, so this was perhaps personal error in the IP set up or a western blot loading or detection issue (Supplemental Figure 1). Further results are not shown from the replicate CTCF IP due to downstream technical issues.

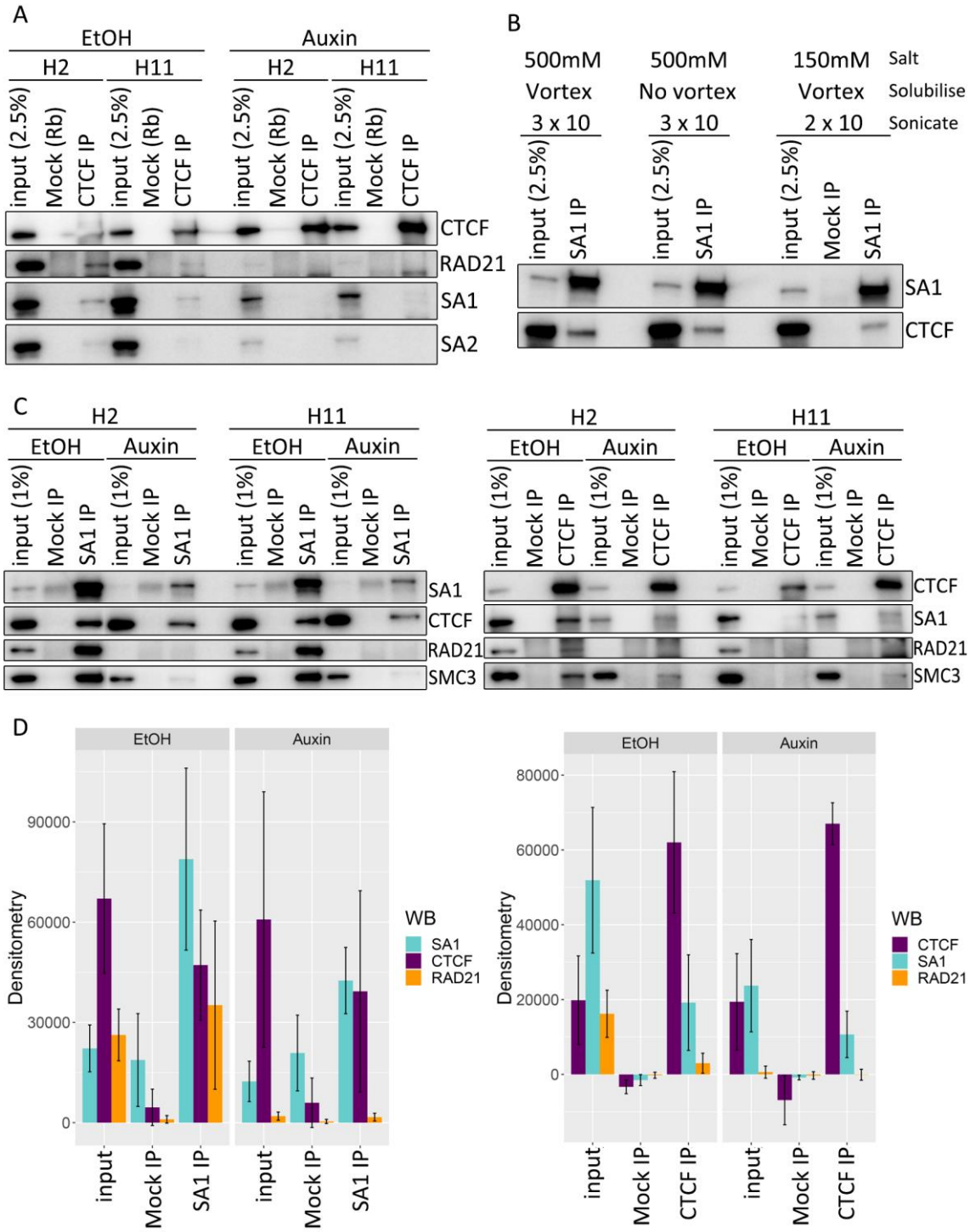


Figure 14: Co-IP of CTCF and SA in FACs-sorted clones of HCT116 RmAC OsTIR1 cells treated with ethanol or auxin. (A) IP of chromatin-bound proteins using endogenous CTCF antibody from either H2 or H11 clone cell lines, following treatment with ethanol or auxin for 4 hrs. (B) Optimisation of chromatin solubilisation conditions in the new sorted H2 cell line. Altered conditions are indicated and affect on CTCF co-IP determined by western blot. (C) Full assessment of SA1 and CTCF reciprocal co-IP in H2 and H11 cell lines treated with ethanol or auxin for 4hrs. (D) Average of raw densitometry values (arbitrary units) for SA1 and CTCF IPs from Figure 13 A & C and Figure 14C (n=4). Graphs were generated using the ggplot function in R.

RAD21 was enriched over input in SA1 ethanol-treated IPs and completely absent from SA1 auxin-treated IPs. Similarly, RAD21 was detected in CTCF IP from ethanol-treated samples and not detected from auxin-treated samples. Issues with the H11 CTCF IP make it difficult to assess RAD21 co-IP in this sample. As previously, cross-reaction of the RAD21 secondary antibody with the CTCF IP signal make the co-IP signal more difficult to detect over background.

Co-IP of CTCF and SA1 was retained despite the increased loss of RAD21, indicating the stability of the CTCF–SA1 interaction. Co-IP was similar between the H2 and H11 clones. As this could not be completely confirmed for the CTCF IPs, for all future co-IPs, the H2 clone was used to ensure best results.

As shown in Figure 7C, some residual SMC3 remained on chromatin at the 4hr auxin timepoint used here. Now that RAD21 was complete depleted from the SA1-CTCF interaction, SMC3 was also immunoblotted to assess any role in the co-IP observed. Similar to Figure 7C, input samples indicate retained SMC3 on chromatin. A small fraction of this SMC3 also remained in complex with CTCF in auxin-treated samples. However, SMC3 was completely absent from the SA1 auxin-treated IPs, indicating the specificity of the SA1-CTCF interaction in the absence of cohesin ring proteins.

Co-IP results from four CTCF and SA1 replicate co-IPs, from polyclonal and sorted cell populations (Figure 13 A & C and Figure 14C), were quantified to consider reproducibility of the interaction (Figure 14D). Overall, these experiments and densitometry results demonstrate interaction of SA1 and CTCF in the absence of RAD21.

3.2.4 CTCF and SA1/SA2 colocalise on chromatin in the absence of cohesin

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) was used to further analyse CTCF and SA interaction and determine where in the genome the proteins may be interacting. Two biological replicate chromatin fractions were prepared from cells treated with ethanol or auxin for 6hrs and chromatin IP (ChIP)

was performed using endogenous antibodies for CTCF, SA1, SA2, RAD21, SMC3, and IgG by Dr. Stanimir Dulev (a post-doc in the Hadjur Lab). The ChIP samples were processed into Next Generation sequencing libraries and sequenced on an Illumina platform. Sequenced reads were aligned to the human genome, quality controlled, and identification of binding sites or 'peaks' was carried out using MISHA, an R package for the analysis of genomic data. Using this package, I compared the genome-wide profile of binding sites for each ChIP-seq library as a comparison between a) biological replicates, b) ethanol and auxin treatment, and c) between proteins. The correlation plots in Figure 15, Figure 16, Figure 17, and Figure 18 represent the percentile normalised ChIP reads per chromosome across the entire genome and thresholded for the top 99.5% to identify the peaks.

Correlation of replicate peak calls was observed for all samples except for SA1 (Figure 15). As expected from the co-IP experiments, only background signal was identified in RAD21 and SMC3 auxin-treated samples while peaks were detected in CTCF, SA1, and SA2 auxin-treated samples. Lack of correlation in the SA1 replicates did not match background type signal (like seen for RAD21 and SMC3) indicating that peaks were still identified in both SA1 replicates. A high number of PCR duplicates were identified in SA1 replicate 2. Although removed prior to peak detection, these duplicates indicated poor IP of DNA and resulted in poor coverage across the genome, perhaps accounting for the lack of correlation with replicate 1. To avoid any erroneous input from this sample, only SA1 biological replicate 1 was considered for all further analyses.

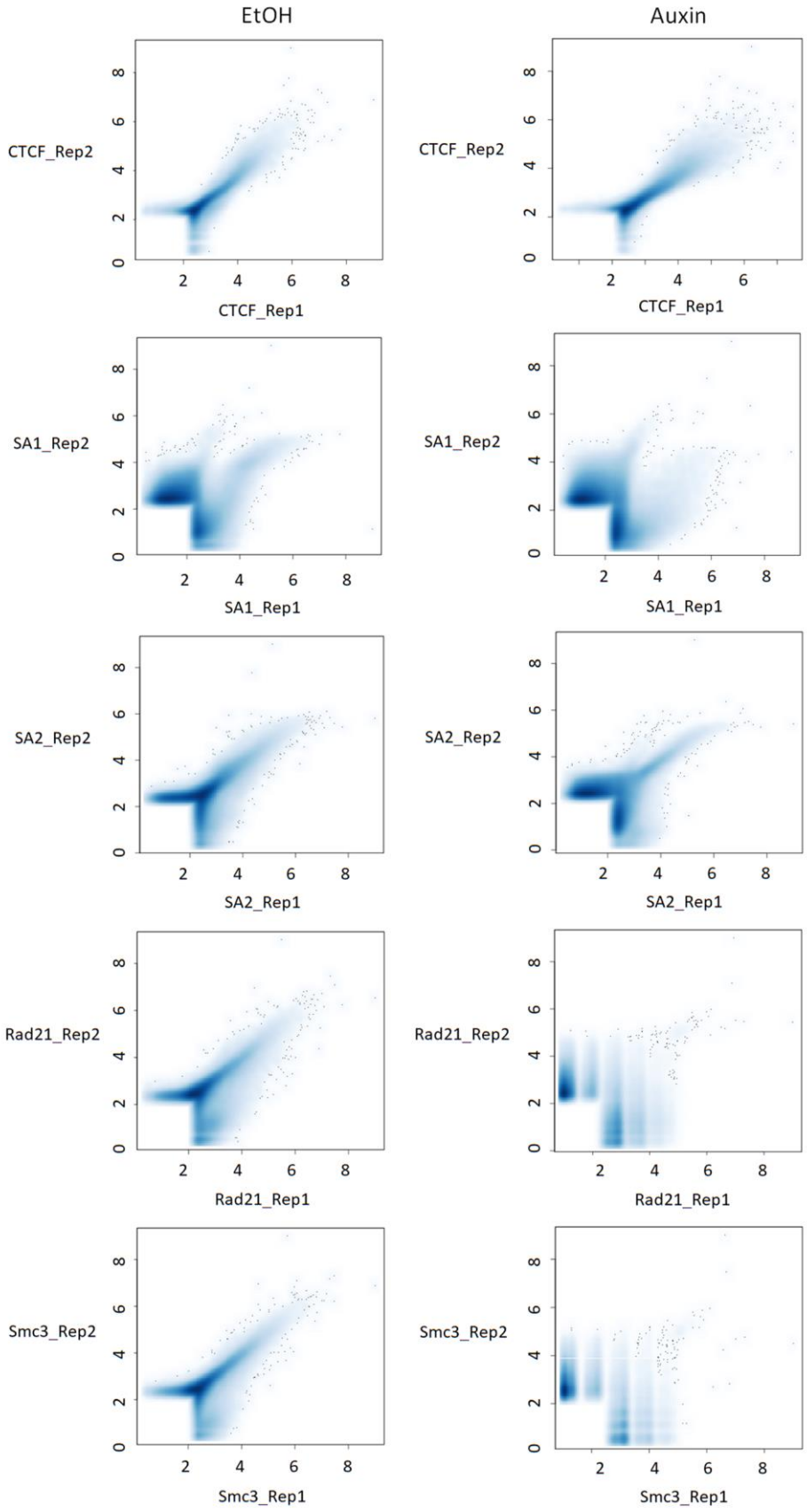


Figure 15: Correlation analysis of cohesin and CTCF ChIP-seq replicates in ethanol- and auxin-treated samples. Correlation plots comparing ChIP peaks between biological replicates of each protein. Peaks were identified using MISHA and a threshold of 0.995. Plots were generated with smoothScatter colour density representation. Comparison between ethanol-treated samples are shown on the left (EtOH) and comparison between auxin-treated samples are shown on the right (Auxin). Rep1 = biological replicate 1. Rep2 = biological replicate 2.

To determine the effect of ethanol and auxin treatment on each of the proteins, replicate peaks were merged and correlation of ethanol and auxin peaks was assessed for all of the proteins (Figure 16). CTCF peaks were highly correlated, with a Pearson's correlation co-efficient of 0.94, indicating that auxin treatment had little effect on CTCF peaks. SA2 also showed some correlation of peaks, with a Pearson's correlation co-efficient of 0.51, indicating that a proportion of SA2 peaks were unchanged with auxin treatment and a proportion of SA2 peaks were changed with auxin treatment, compared to control. SA1 showed similar correlation of a proportion of peaks. In contrast, RAD21 and SMC3 peaks showed no correlation between ethanol and auxin samples. Here auxin treatment induced a shift of peaks to low-level signal away from the diagonal. All of these findings were expected given the co-IP results, although SMC3 signal was now reduced more similarly to RAD21.

The degree of correlation between proteins was also assessed. As expected, in ethanol conditions, all of protein were correlated with one another, with a Pearson correlation coefficient of ~0.7 (Figure 17). SMC3 and RAD21 showed the highest correlation (coefficient of 0.96), followed by SA2 and SMC3/RAD21 (both with coefficient of 0.85). SA1 and SA2 were also correlated (coefficient of 0.75). While it is still not known if SA1 and SA2 can interact with the cohesin ring together, ChIP-seq from various cell types have been published which also report overlap of SA1 and SA2 peaks, albeit to varying extents (Cuadrado *et al.*, 2019; Holzmann *et al.*, 2019; Casa *et al.*, 2020). Localisation of cohesin-SA1 and cohesin-SA2 on the same fragment of DNA analysed could also account for such overlap.

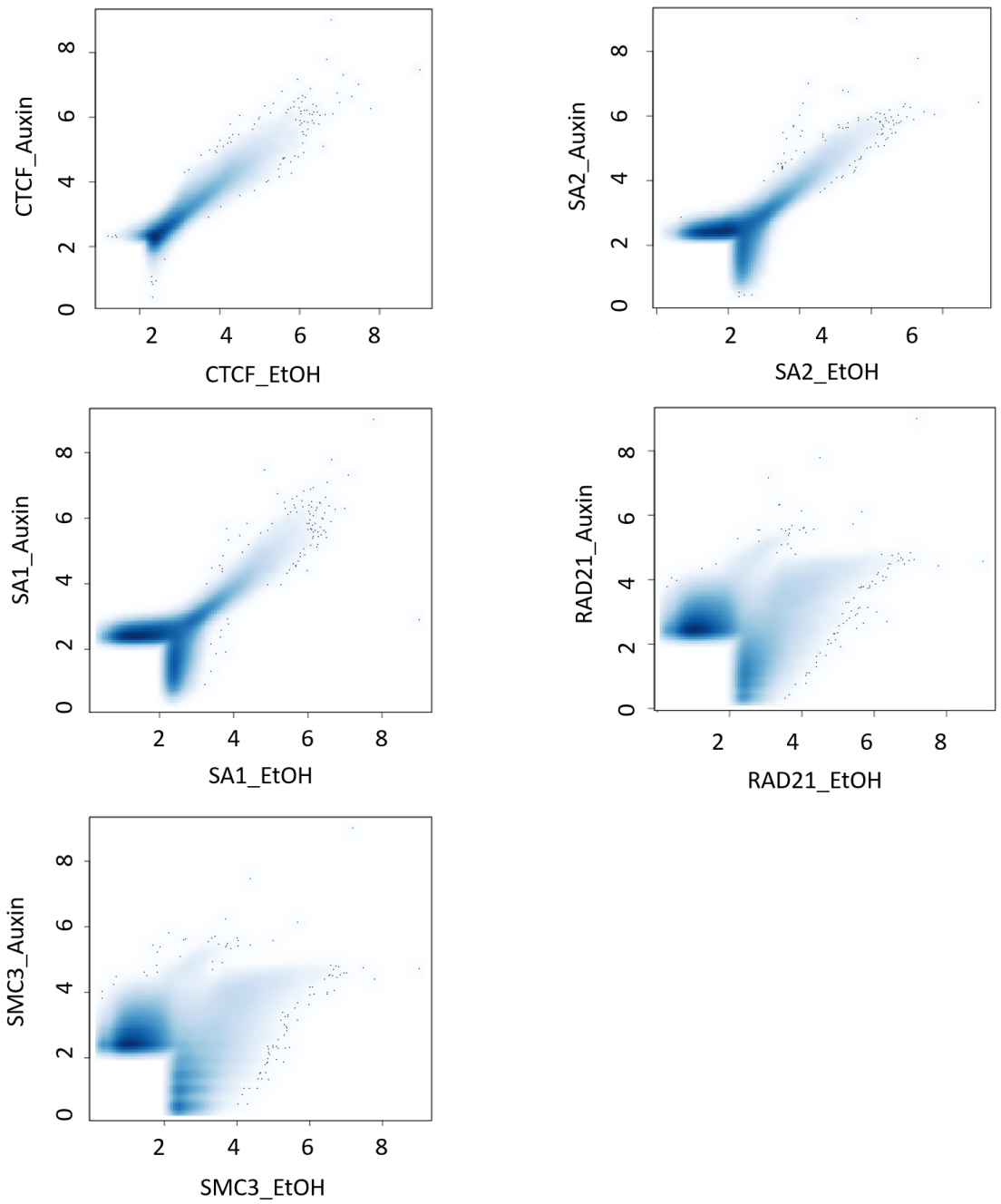


Figure 16: Correlation analysis of the effect of auxin treatment on cohesin and CTCF ChIP-Seq samples. Correlation plots comparing ChIP peaks between ethanol and auxin treatment of each protein. Each sample represents the merge track of the two biological replicates in Figure 15, except for SA1, for which only replicate 1 was used. Peaks were identified using MISHA and a threshold of 0.995. Plots were generated with smoothScatter colour density representation.

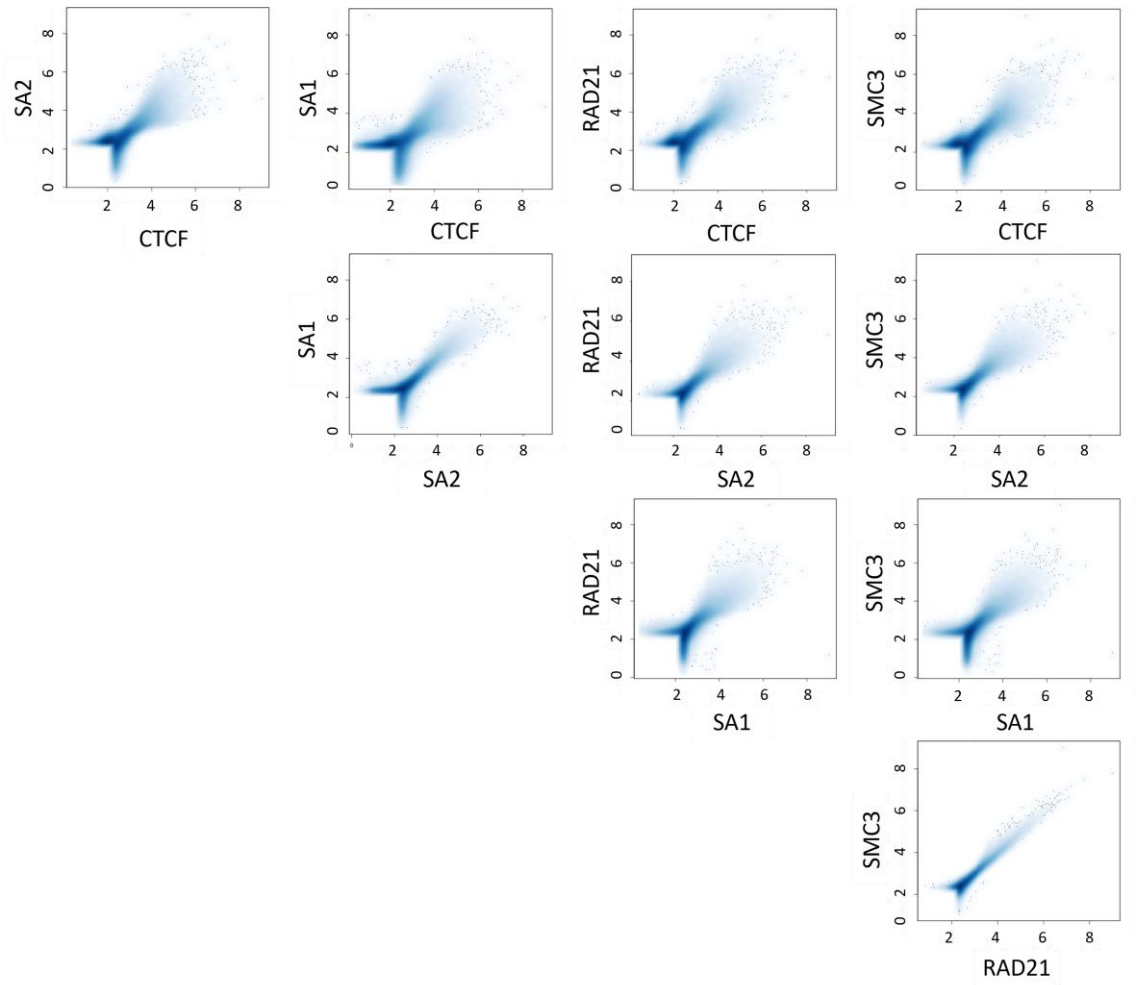


Figure 17: Cohesin and CTCF peaks are correlated in ethanol conditions. Correlation plots comparing ChIP peaks between the different ethanol-treated protein samples. Each sample represents the merge track of the two biological replicates in Figure 15, except for SA1, for which only replicate 1 was used. Plots were generated with smoothScatter colour density representation.

In auxin conditions, correlation of RAD21 and SMC3 with each other and the other proteins was lost (Figure 18). This was expected given the loss of signal for these proteins with auxin treatment (Figure 16). In agreement with co-IP results, CTCF and SA1 peaks showed some correlation. Whereas, in contrast to the above co-IP results, SA2 peaks showed correlation with CTCF similar to that of SA1. Correlation of SA1 and SA2 was also observed, indicating that some sites can still be occupied by both SA proteins, even in the absence of RAD21.

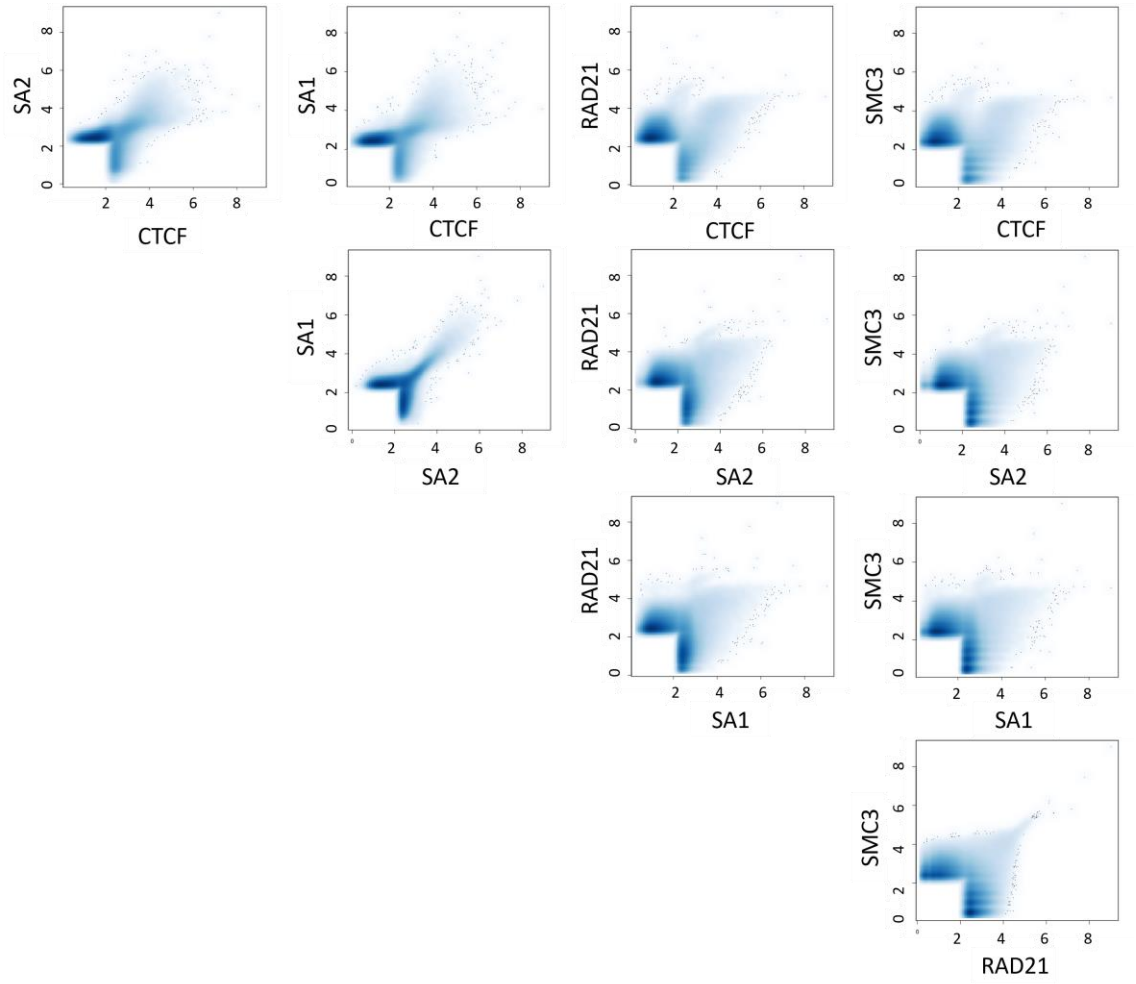


Figure 18: CTCF and SA are correlated in auxin conditions. Correlation plots comparing ChIP peaks between the different ethanol-treated protein samples. Each sample represents the merge track of the two biological replicates in Figure 15, except for SA1, for which only replicate 1 was used. Plots were generated with smoothScatter colour density representation.

The correlation plots in Figure 18 also illustrated that not all CTCF–SA colocalisation sites were retained upon auxin treatment. To view retained and lost colocalisation sites, peaks from two regions on Chromosome 6 were mapped using SHAMAN (Mendelson Cohen *et al.*, 2017) (Figure 19). The peaks were identified using MISHA as above. The height of each ChIP track was normalised to CTCF which showed the highest signal peaks in both regions. In both cases, CTCF, SA1, SA2, RAD21, and SMC3 peaks were aligned in ethanol-treated (control) tracks. As expected from the correlation plots above, SA1 signal was reduced compared to the other proteins. In the region shown on the left, CTCF, SA1, and SA2 peaks were also aligned at the same sites in auxin conditions, whereas RAD21 and SMC3 peaks were completely lost. In the region shown on the right, CTCF peaks show similar signal to the ethanol control sample, but only

low SA1 and SA2 peaks were observed, indicating a loss of these proteins at this site in the absence of RAD21 and SMC3.

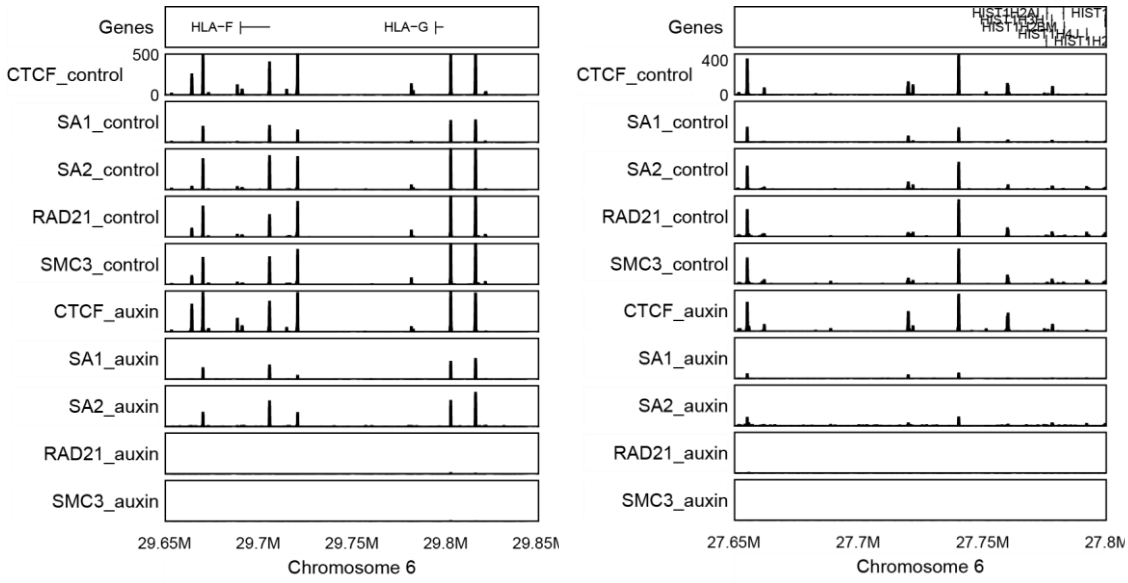


Figure 19: Colocalisation of CTCF and SA in ethanol and auxin conditions. Example regions of retained (left) and lost (right) SA peaks at sites of CTCF binding. Peaks were identified using MISHA and mapped in SHAMAN. Track height for each protein is normalised to CTCF.

In conjunction with the genome-wide correlation plots, DeepTools suite was used to focus on ChIP signal at a given set of regions (Figure 20). Regions defined by CTCF binding in ethanol conditions were colocalised by SA1, SA2, RAD21, and SMC3 in control conditions, as expected (Figure 20A). Surprisingly, SA2 was detected at a higher proportion of the CTCF sites than SA1. In contrast to the ethanol conditions, in auxin conditions these same sites lost colocalisation with RAD21 and SMC3 and retained binding for SA1 and SA2. Similarly, regions defined by SA binding in control conditions showed only CTCF and SA signal in auxin conditions (Figure 21). Regions defined by binding of both CTCF and SA1 in auxin conditions, showed RAD21 and SMC3 signal in control samples, demonstrating that the CTCF–SA observed in auxin-treated cells is localised to sites that cohesin occupies in physiological conditions, and is not localised to a completely different set of binding sites (Figure 20B).

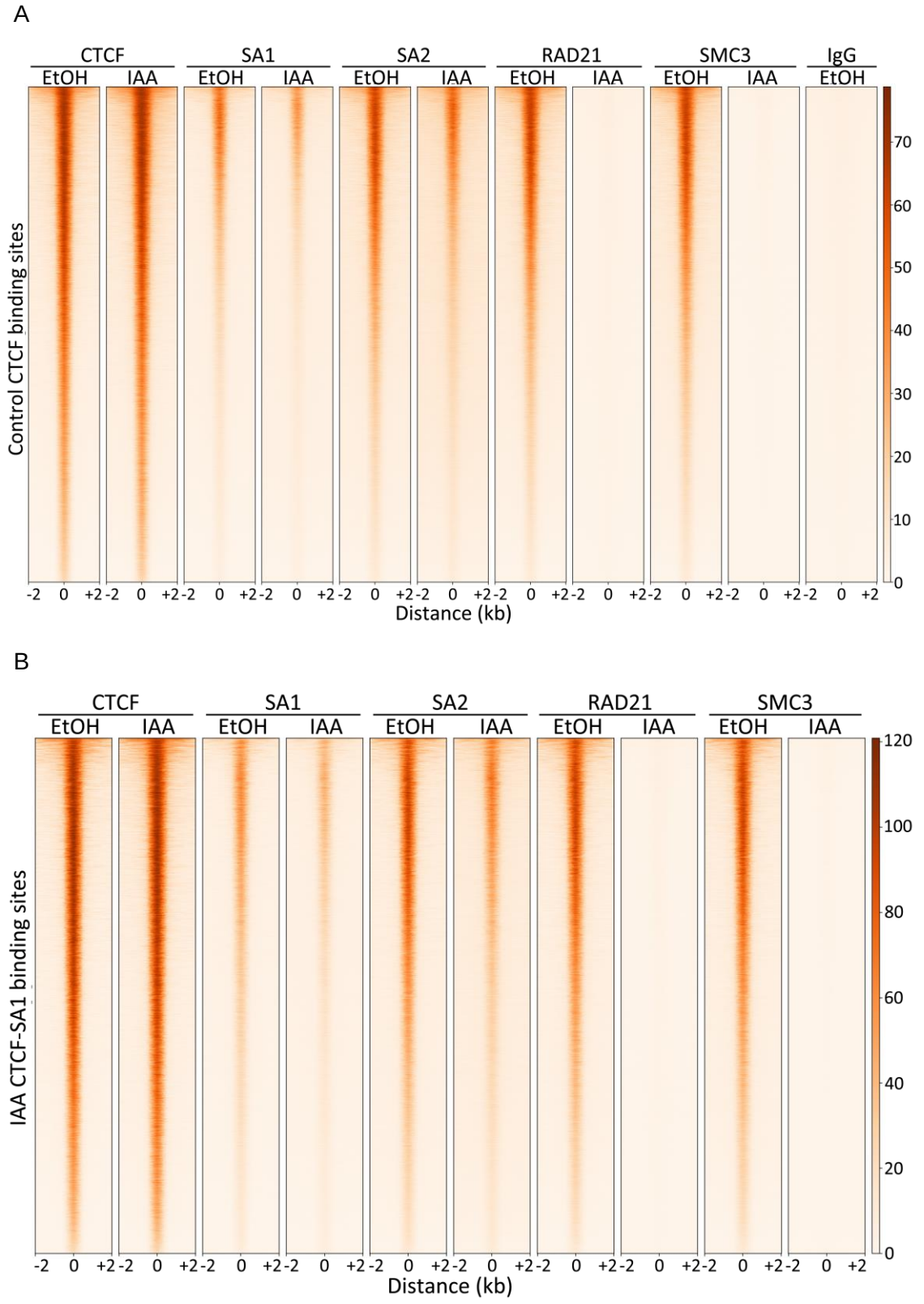


Figure 20: ChIP-Seq analysis of cohesin and CTCF in ethanol and auxin conditions (6 hrs). (A) Heatmap of CTCF, SA1, SA2, RAD21, SMC3, and IgG signal distribution in ethanol- and auxin-treated samples at sites bound by CTCF in control conditions. A window of 2 kb is shown either side of the CTCF peak location. (E) Heatmap of CTCF, SA1, SA2, RAD21, SMC3, and IgG signal distribution in ethanol- and auxin-treated samples at sites defined by binding of CTCF and SA1 in auxin conditions. A window of 2 kb is shown either side of the CTCF-SA1 peak location.

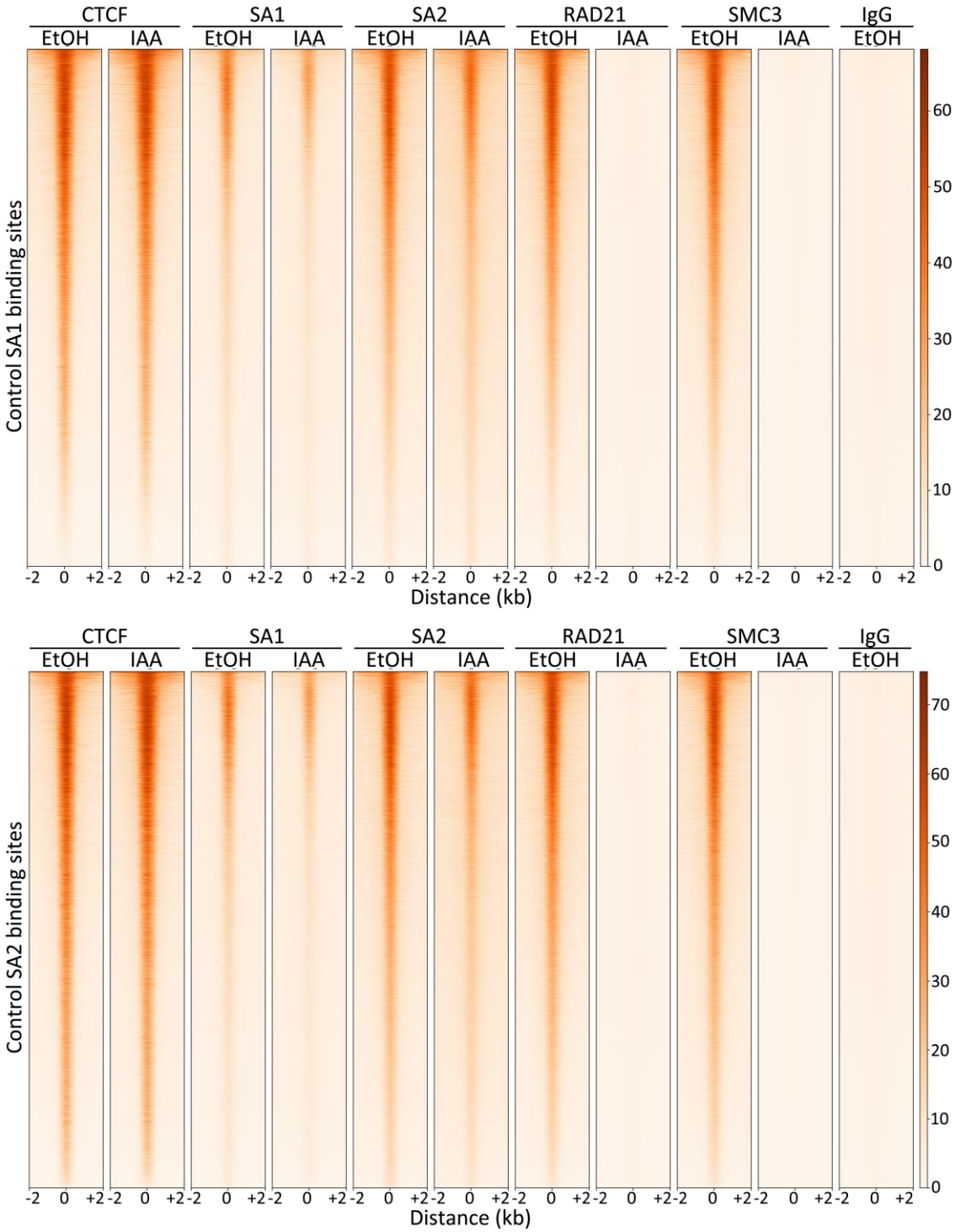


Figure 21: ChIP-Seq analysis of cohesin and CTCF in ethanol and auxin conditions (6 hrs) – SA heatmaps. Heatmap of CTCF, SA1, SA2, RAD21, SMC3, and IgG signal distribution in ethanol- and auxin-treated samples at sites defined by binding of SA1 (top) and SA2 (bottom) in ethanol conditions. A window of 2 kb is shown either side of the SA peak location. EtOH = Ethanol; IAA = Auxin; kb = kilobase.

Thus, when cohesin is depleted from HCT116 cells, CTCF and SA1 can be observed in complex by both co-IP and ChIP-seq, two different but complimentary techniques. CTCF and SA2 were also detected together, but only by ChIP-seq; raising the question of whether CTCF and SA2 directly interact *in vivo* and, if not,

what differences in SA1 and SA2 account for this discrepancy. This will be discussed further in section 3.2.6.

3.2.5 BiFC-ChIP, a new method to identify protein-protein interactions on chromatin

The population level nature of 'bulk' ChIP-seq makes it imperfect for identification of the genomic colocalisation of interacting proteins on chromatin. Colocalisation of proteins is inferred from comparison of global binding patterns across the population, however the location of specific interactions in a given cell cannot be determined. Re-ChIP methods were developed that allow detection of multiple proteins bound to a single DNA sequence by sequential IP of the proteins of interest prior to library preparation and sequencing (Geisberg and Struhl, 2005; Truax and Greer, 2012). However, loss of material is seen with each sequential IP and subsequently, libraries may have high PCR duplication levels and low complexity. Importantly, Re-ChIP only determines that proteins localise to the same region of DNA, no direct interaction can be established. More recently developed single-cell ChIP-seq technologies can address the question of co-localisation in a given cell, however, specific interaction can still not be confirmed and sensitivity is reduced compared to bulk sequencing (Rotem *et al.*, 2015; Grosselin *et al.*, 2019). On the other hand, immunofluorescence techniques such as Stochastic Optical Reconstruction Microscopy (STORM) allow high-resolution visualisation of interacting molecules in single cells across an imaged population. However, the number of proteins that can be investigated at any one time is severely limited. In the early stages of my PhD research I developed a method to robustly identify protein-protein interaction on chromatin by combining bimolecular fluorescent complementation (BiFC) with ChIP. This work was carried out with the aim to investigate colocalization of cohesin and its regulator proteins, including investigation of CTCF-SA1 vs CTCF-SA2 colocalisation.

BiFC is an assay that allows visualisation of protein-protein interactions across different cell types and organisms (Ghosh, Hamilton and Regan, 2000; Hu, Chinenov and Kerppola, 2002). BiFC analysis depends on the interaction of two non-fluorescent fragments of a fluorescent protein, driven by their fusion to a pair

of interacting proteins. If the two proteins of interest interact, the fluorescent protein will be reconstituted and a fluorescent signal seen (Figure 22). Historically BiFC has been used to visually confirm protein interactions by immunofluorescence. I developed this method further by showing that the reconstituted fluorescent protein can be selectively recognised, without recognising the N- or C-terminal fragments alone, using a nanobody IP system – the Chromotek® GFP-Trap. This specific recognition allowed selective identification and enrichment of the interacting protein pair. Moreover, this work determined that the selective IP is retained in ChIP-seq conditions, meaning that identification of directly interacting proteins and the chromatin environment in which they are interacting can be achieved.

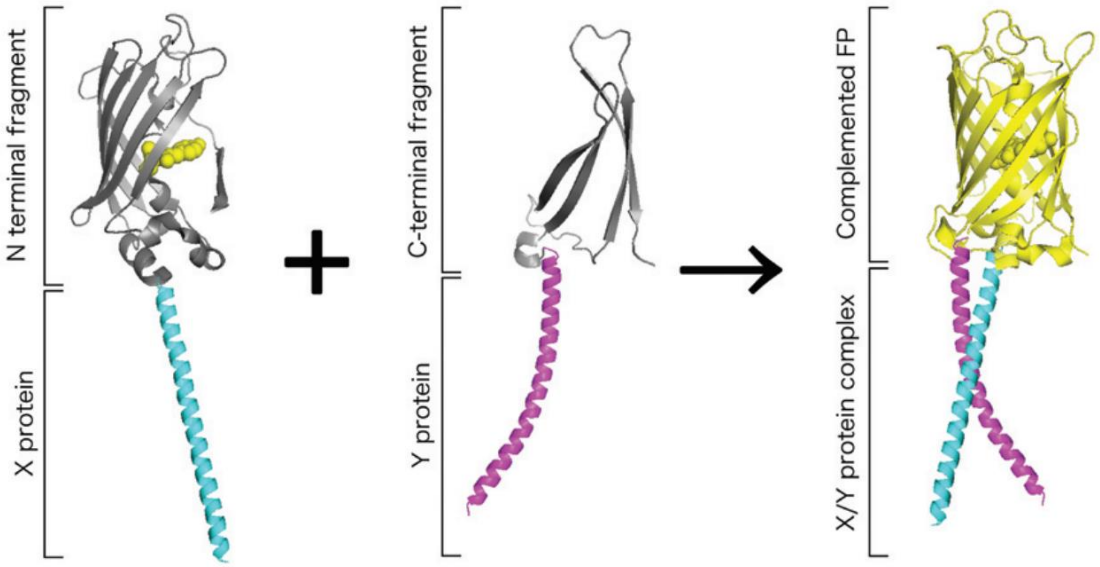


Figure 22: Schematic of BiFC assay obtained from Kodama and Hu (2012). BiFC occurs when two non-fluorescent fragments of a fluorescent protein are brought together by fusion to a pair of interacting proteins (X and Y) – the fluorescent protein will be reconstituted and a positive signal observed. FP = Fluorescent Protein.

The BiFC assay has been developed using a range of fluorescent proteins and has been widely validated via fusion of split fragments to the bZIP and Rel family transcription factors, Fos and Jun (Hu, Chinenov and Kerppola, 2002). Therefore, I validated the BiFC method for use with ChIP using these two proteins. The BiFC plasmids pBiFC-bJunVN173 and pBiFC-bFosVC155 were obtained from Addgene, in which, N- and C-terminal fragments of the Venus fluorescent protein have been fused to FLAG-tagged bJun and HA-tagged bFos, respectively (Shyu *et al.*, 2006). These constructs were chosen for use in this project as we

hypothesized that the reconstituted Venus protein may be recognised by the ChromoTek GFP-Trap® IP system.

All plasmids were prepped using Addgene's standard instructions. To test for expression of the plasmids and BiFC using the split Venus protein, U2OS cells were transiently transfected with plasmids expressing bFosVC155 and bJunVN173, either individually or in tandem (Figure 23A). bJun fused to an alternative N-terminal Venus fragment (bJunVN155(I152L)), containing an isoleucine to leucine mutation, was also tested in combination with bFosVC155. A fluorescent signal was obtained only in cells expressing both N- and C-terminal fragments of the Venus protein. Similar to published BiFC results for Fos and Jun, punctate staining on chromatin was observed. Transfection with CTCF-EGFP was carried out as a positive control for fluorescence. In addition, double transfection of bJunVN173 with bFosDeltaZIPVC155, a mutant version of Fos that is unable to interact with Jun due to deletion of a portion of its ZIP domain, was also tested, to control for specificity of the complementation reaction. In this case, no fluorescence was observed, confirming that interaction between Fos and Jun was required for a signal to be achieved.

Average fluorescent intensity measurements were calculated for each of the samples using CellProfiler™, with 3 separate images used for the bFosVC155 – bJunVN173 sample and 2 images used for each of the other samples (Figure 23B). This quantification confirmed that fluorescence was only obtained in the CTCF-EGFP and double transfected samples. Increased fluorescence was obtained for complementation using the VN155(I152L) fragment compared to the VN173 fragment. This finding is in line with initial characterization of the I152L mutation for its reduced signal-to-noise ratio in BiFC (Shyu *et al.*, 2006).

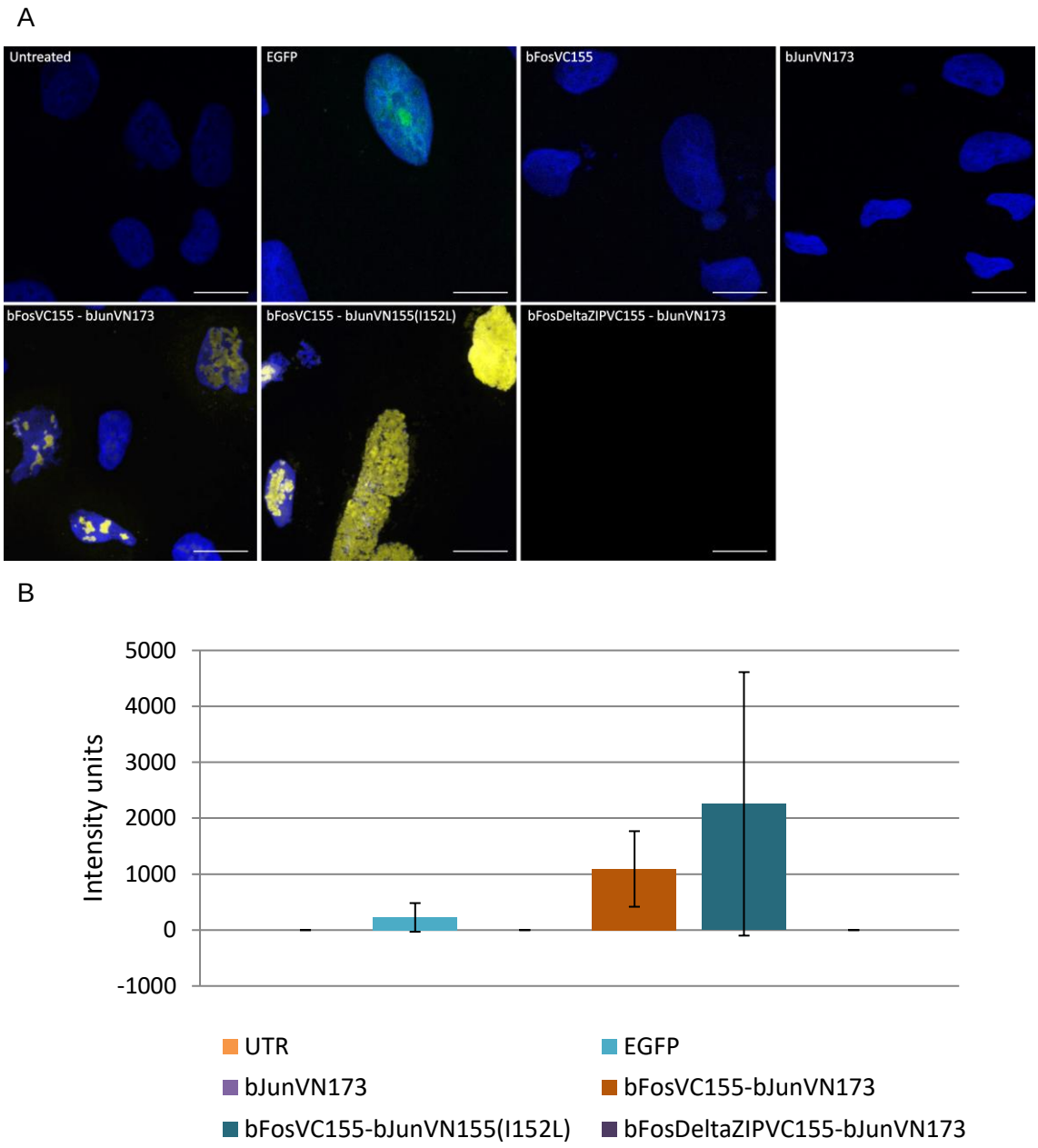


Figure 23: Validation of BiFC method. (A) Fluorescent images and (B) average fluorescent intensity results (N=2 or 3) are shown from transiently transfected U2OS cells, following removal of soluble proteins with 0.25% triton X-100 and imaging with a Zeiss LSM880 confocal microscope. DAPI staining is shown in blue. Scale bar = 10μm. UTR = untreated; EGFP = enhanced green fluorescent protein.

To determine if the complemented Venus protein could be selectively immunoprecipitated using the GFP-Trap system, U2OS cells were electroporated with bFosVC155, bJunVN173, or bFosVC155 and bJunVN173, chromatin samples were prepared (see methods section 2.7), and ChromoTek GFP-Trap® was used to IP Venus. U2OS cells stably expressing GFP-tagged CTCF were also processed as a positive IP control.

Immunoprecipitation was achieved for cells expressing CTCF-EGFP and both bFosVC155 and bJunVN173 (Figure 24). In contrast, neither bFosVC155 nor bJunVN173 were pulled down from single transfected cells, suggesting that the reconstituted Venus protein can be selectively IP'd without recognition of either of its half-fragments. Non-bound (NB) samples showed only low levels of bFosVC155 in the single transfected sample, leaving it unclear if some pull-down of the C-terminal end of the Venus protein might occur at higher levels (Supplemental Figure 2A).

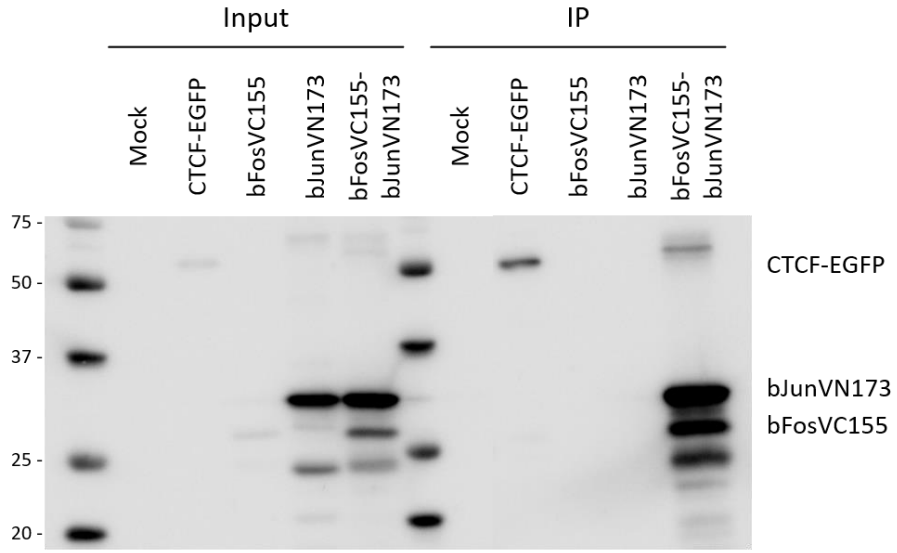


Figure 24: Complemented Venus protein can be selectively immunoprecipitated. Input (2.5%), and IP samples were run on 12% SDS-PAGE gels and blotted for GFP. CTCF-EGFP was included as a positive IP control for the nanobody used. The mock sample is processed from cells electroporated in the presence of no DNA. UTR = untreated; EGFP = enhanced green fluorescent protein.

Multiple bands were present in the IP and NB samples in the experiment discussed above. To better assess what each of these bands represented, the experiment was repeated, and membranes blotted for the HA and FLAG tags present on bFosVC155 and bJunVN173, respectively. Fluorescent secondary antibodies were used to allow visualisation of HA and FLAG simultaneously. In addition, multiple wash conditions were evaluated to determine how stably these additional products bind to the GFP-Trap - samples were washed with 150mM, 300mM, or 500mM NaCL wash buffer. Finally, electroporation with bFosDeltaZIPVC155 and bJunVN173 was also included in this experiment as a negative control to assess the specificity of the Venus-GFP-Trap interaction.

IP of material from double transfected cells was once again successful, even following washing with 500mM NaCl wash buffer (Figure 25A). Increasing the concentration of the salt wash did not reduce the pull down of additional, smaller molecular weight proteins. As corresponding bands were not visible in input samples and these bands were detected by anti-Flag and anti-HA antibodies (not just anti-EGFP), these proteins may be degradation products formed during the IP process. An additional band was also observed at ~65kDa, which appears to be detected by both the FLAG and HA antibodies. It is possible that this band represents the entire bFosVC155-bJunVN173 complex held together due to the strength of interaction between the VC155 and VN173 fragments. Thus, the multiple bands may represent degradation and complexed versions of the tagged proteins.

In this experiment, faint IP was also observed from cells transfected with either bFosVC155 or bJunVN173. This observation suggests that, under certain conditions, the GFP-Trap can recognise the N- and C-terminal fragments of the Venus protein. However, much lower levels of protein were present in these IP samples compared to double transfected samples, suggesting that optimisation of the IP protocol would prevent pull down of these proteins. In addition, NB bFosVC155 and bJunVN173 samples contained much higher levels of the protein than their corresponding IP samples, indicating that the majority of these proteins are not bound by the GFP-Trap (Supplemental Figure 2B). Of note, bFosDeltaZIPVC155 and bJunVN173 were pulled down by the GFP-Trap in the bFosDeltaZIPVC155-bJunVN173 IP sample (Figure 25A). In this case, Fos and Jun do not interact and thus the Venus protein should not be formed in these cells. However, IP was observed at a higher efficiency than in single transfected cells, suggesting some reconstitution of the Venus protein had occurred. Observation of these cells under a fluorescent microscope confirmed reconstitution of the Venus protein to its fluorescent form (Figure 25B). This likely represents background complementation driven by random interaction of the VC155 and VN173 fragments.

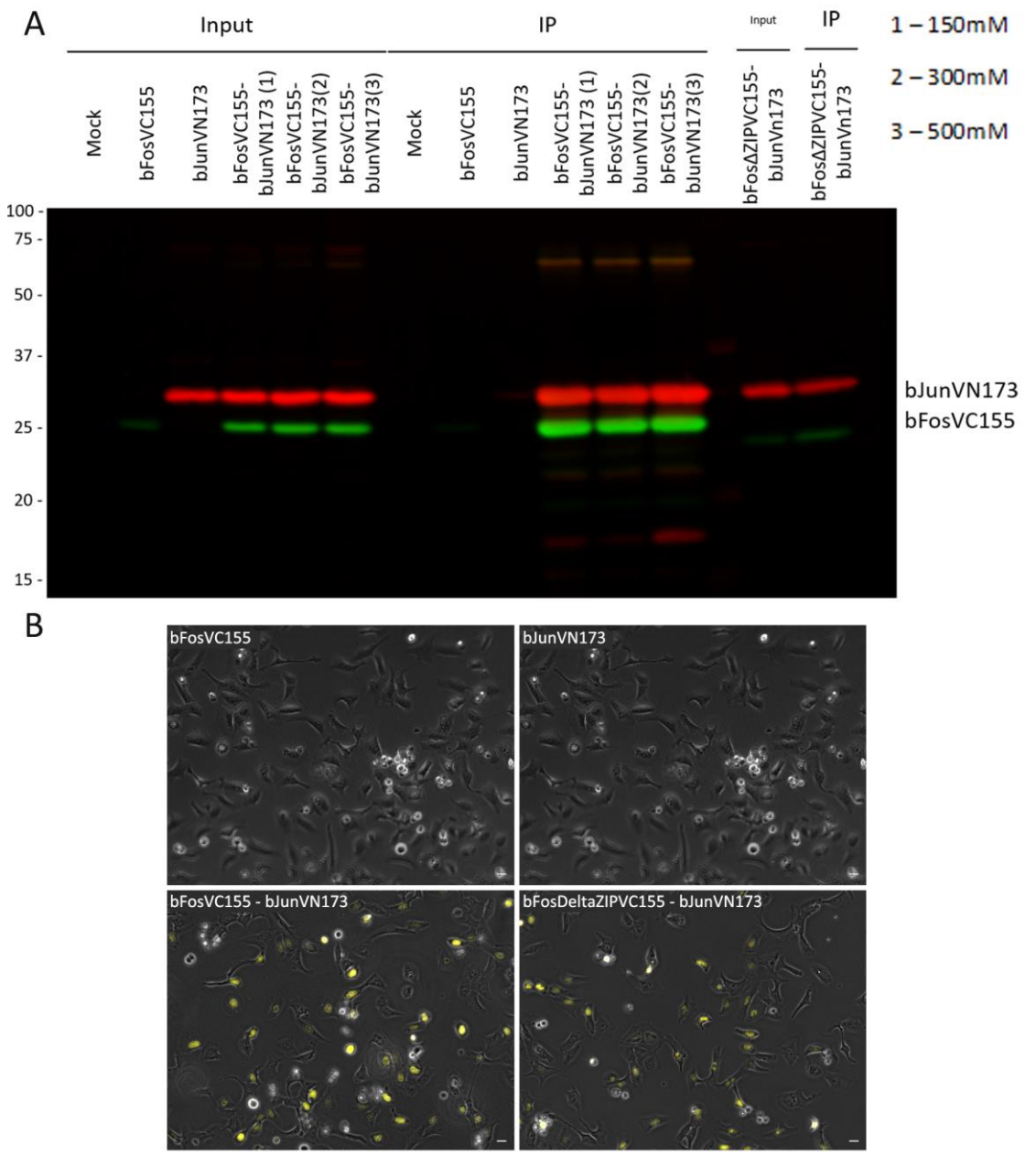


Figure 25: Optimisation of BiFC-IP wash conditions – Part 1. (A) Input (2.5%) and IP samples were run on 12% SDS-PAGE gels and blotted for HA (green) and FLAG (red). Mock samples are processed from cells electroporated in the presence of no DNA. (B) Fluorescent images of a subset of the samples are shown. Images were captured using an Axio Observer Z1 microscope prior to collection. bFosΔZIPVC155-bJunVN173 = bFosDeltaZIPVC155-bJunVN173.

In order to optimise the IP protocol and identify a wash strategy that prevented pull down of the N- and C-terminal fragments alone, single and double transfected samples were processed and washed with either 500mM NaCl wash buffer or 500mM NaCl wash buffer + 0.1% SDS. To reduce background Venus reconstitution and degradation of protein during IP, the amount of DNA electroporated into the cells was reduced and incubation time with the GFP-Trap was reduced from 2hrs to 1hr, respectively. With increased brightness, western

blot results showed that bFosVC155 was still IP'd by the GFP-Trap following washing with 500mM NaCl wash buffer (Figure 26). Inclusion of 0.1% SDS in the wash buffer reduced this pull down. Thus, washing with 500mM salt wash buffer containing 0.1% SDS was identified as the most efficient wash for reducing IP of the half-fragments of Venus. Decreasing the incubation time of the IP to 1 hr reduced the presence of additional, smaller molecular weight proteins in double transfected samples, potentially by limiting the generation of degradation products. Reducing the amount of DNA electroporated into the cells did not prevent IP of bFosDeltaZIPVC155-bJunVN173, meaning that background Venus reconstitution could occur under these conditions and further optimisation was required. Fluorescent Venus was visible in these cells under the microscope, again suggesting that this recognition by the GFP-Trap is a consequence of background Venus reconstitution.

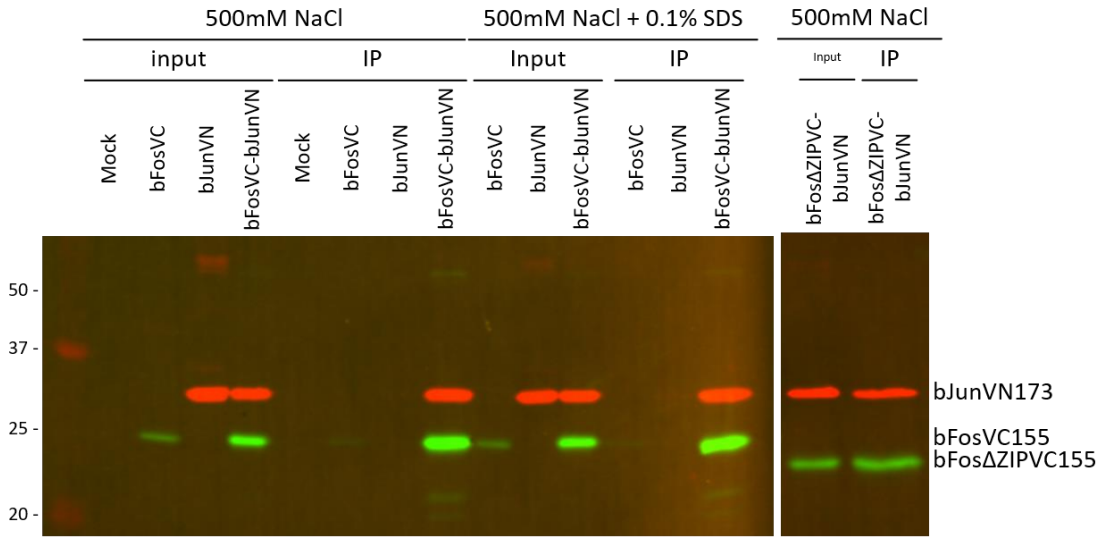


Figure 26: Optimisation of BiFC-IP wash conditions – Part 2. Input (2.5%) and IP samples were run on 12% SDS-PAGE gels and blotted for HA (green) and FLAG (red). Salt and SDS concentration of the wash buffer used for each set of samples are shown above the corresponding lanes. bFosVC = bFosVC155; bJunVN = bJunVN173; bFosΔZIPVC155-bJunVN173 = bFosDeltaZIPVC155-bJunVN173.

Given the successful IP of complemented Venus in a standard IP condition, IP in the ChIP conditions was next assessed. Incubation of the GFP-Trap with bFosVC155-bJunVN173 in ChIP conditions ('ChIP buffer'; 250nM NaCl, 0.1% SDS, and 1% Triton X-100) did not affect recognition of the Venus protein (Figure 27 A-C). Selective IP of the Venus protein was observed following incubation in buffer containing 150mM NaCl – same condition as all previous IPs -, 250mM

NaCl, and ChIP buffer. In previous experiments, the amount of overall protein incubated with the GFP-Trap was ~500ug and reducing this amount to 200ug in this experiment may account for the elimination of IP of bFosVC155 and bJunVN173 in singly transfected samples in all three conditions. Western blotting of corresponding NB samples confirmed the presence of both proteins in the IP solution, however, low levels of bFosVC155 contributing to its lack of IP could not be ruled out here.

To determine if selective recognition of Venus can be achieved following a full ChIP protocol (see methods section 2.7), bFosVC155, bJunVN173, and bFosVC155-bJunVN173 samples were processed under three conditions: 1) ChIP protocol, no fixation, no sonication, + benzonase treatment; 2) ChIP protocol, no sonication, + fixation, + benzonase treatment; 3) ChIP protocol + fixation, + sonication. Selective recognition of the reconstituted Venus protein was observed under all three conditions (Figure 28 A & B). Of note, under condition 1 ('No fix – benz'), high levels of bFosVC155 were present in the lysate and did not result in IP of the single-transfected bFosVC155 (Figure 28A). Thus, low levels of bFosVC155 in previous experiments are not likely to account for the lack of pull down in corresponding IP samples. Importantly, in fixed cells isolated by the ChIP protocol ('Fix – sonication'), IP was observed only in double transfected cells (Figure 28B), with the faint signal seen in the bJunVN173 IP lane caused by spill over of the bFosVC155-bJunVN173 sample.

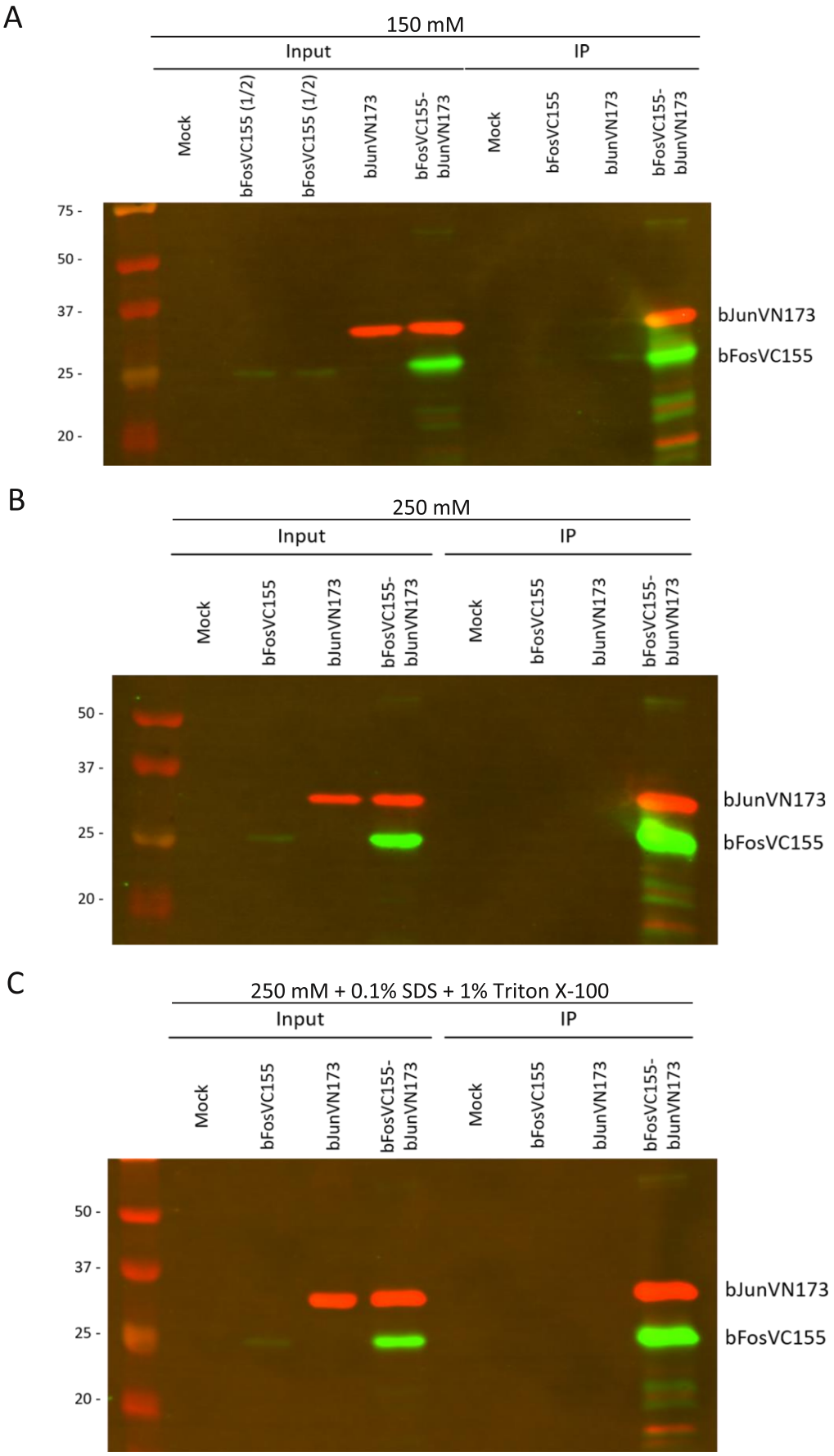


Figure 27: BiFC-IP in ChIP buffer conditions. (A) – (C) Input and IP samples were run on 12% SDS-PAGE gels and blotted for HA (green) and FLAG (red). Samples were incubated with the GFP-Trap in IP buffer containing (A) 150mM NaCl, (B) 250mM salt, or (C) 250mM NaCl + 0.1% SDS + 1% Triton X-100 (ChIP buffer).

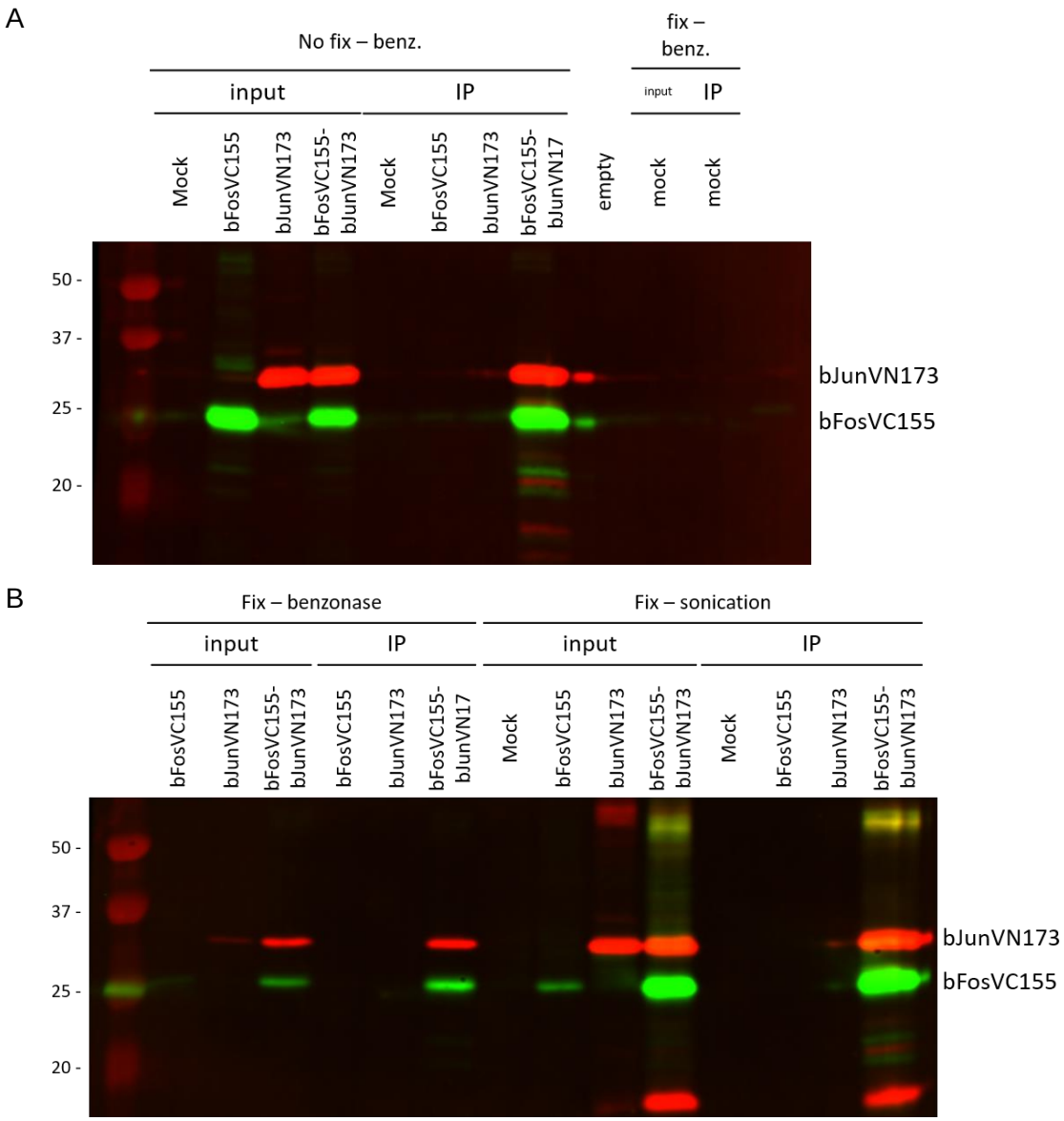


Figure 28: BiFC-IP in ChIP conditions. (A) and (B) Input and IP samples were run on 12% SDS-PAGE gels and blotted for HA (green) and FLAG (red). Processing methods for each set of samples are shown above the corresponding lanes. Benz = benzonase.

As background Venus reconstitution was apparent in previous IP conditions (Figure 23C & Figure 25C), electroporation with 2, 1, 0.5, and 0.25ug of bFosDeltaZIPVC155 and bJunVN173 was tested to identify whether transfection with a lower volume of DNA would alleviate background Venus formation. Fluorescent signal was greatly reduced following electroporation with 0.5 or 0.25ug of DNA. To confirm loss of background Venus reconstitution with the reduced electroporation volume, U2OS cells were electroporated with 0.5ug of bFosVC155, bJunVN173, bFosVC155-bJunVN173, or bFosDeltaZIPVC155-bJunVN173. In addition, an extra bFosDeltaZIPVC155-bJunVN173 sample was set up with electroporation of 0.25ug of plasmid DNA, in case 0.5ug still resulted

in IP of background Venus. 300ug of chromatin lysate from each sample was incubated with the GFP-Trap for IP. Following elution in 100ul of Elution buffer, 33.33ul of each sample was mixed with SDS-sample buffer for western blotting, with the remaining 66.67ul was de-crosslinked and purified for DNA.

Only a faint western blot signal was detected for the bFosVC155-bJunVN173 double transfected sample (Figure 29). No signal was detected for any of the other samples, including the bFosDeltaZIPVC155-bJunVN173 double transfected samples. However, without a robust bFosVC155-bJunVN173 signal, this cannot be accounted for purely by a lack of IP. Elution in Elution buffer instead of SDS-sample buffer, elution for only 20 minutes, and the reduced level of plasmid DNA transfected into the cells may all have contributed to the faintness of the western blot signals observed. DNA purified from the remaining 66.67ul of eluted solution was quantified by qubit assay and nanodrop. DNA levels were all too low for further analysis (such as qPCR or ChIP-seq) but confirmed that DNA could be purified from the IP.

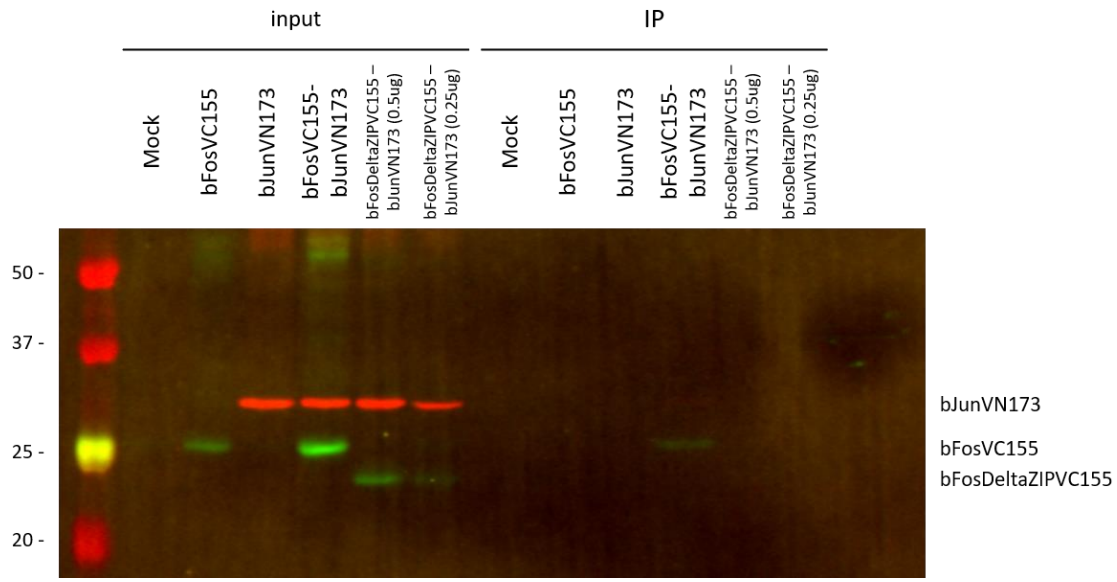


Figure 29: Optimising transfection concentration for BiFC-ChIP – Part 1. Input (10%) and IP samples were run on 12% SDS-PAGE gels and blotted for HA (green) and FLAG (red). All samples were electroporated with 0.5ug of plasmid DNA unless indicated in the lane annotation.

For a robust ChIP-seq experiment a higher yield was required, but needed to be achieved without re-introducing background Venus IP. From the same starting material as the above experiment, a second GFP-Trap was set-up, this time with 1.82 X the amount of chromatin (545ug) incubated per IP. In addition, incubation

time was extended to 2hrs and samples were eluted in SDS-sample buffer. Increased IP of bFosVC155- bJunVN173 was observed, with only slight pull-down of bFosDeltaZIPVC155- bJunVN173 (0.5ug) (Figure 30A). Western blotting of 4% of the NB material confirmed the presence of each construct in IP samples (Figure 30B), confirming that in these conditions the reconstituted Venus protein was selectively IP'd with only a small amount of background.

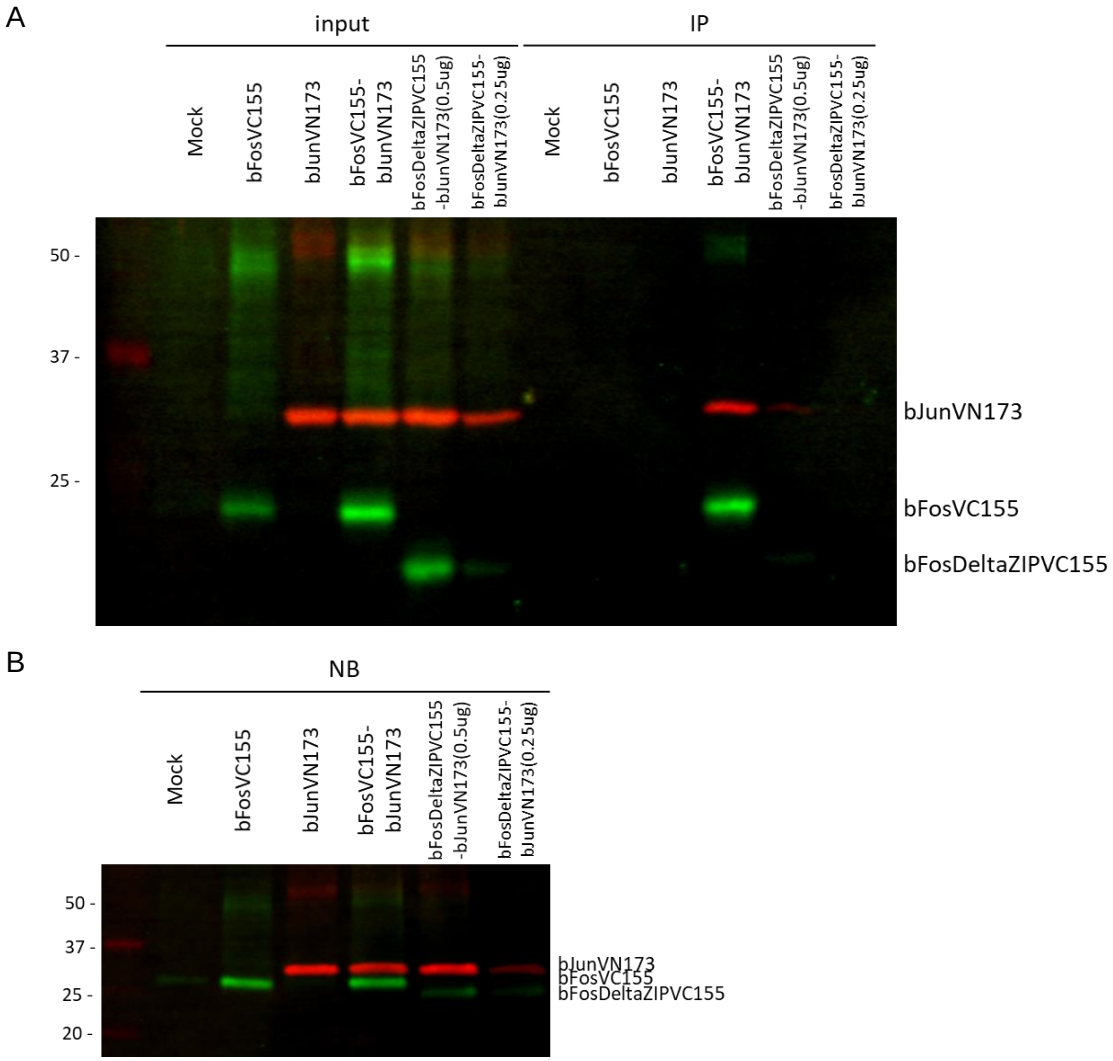


Figure 30: Optimising transfection concentration for BiFC-ChIP – Part 2. (A) Input (4%) and IP samples were run on 12% SDS-PAGE gels and blotted for HA (green) and FLAG (red). All samples were electroporated with 0.5ug of plasmid DNA unless indicated in the lane annotation. (B) 4% of NB samples were run on a 4–20% Mini-PROTEAN® TGX™ precast protein gel and blotted for HA (green) and FLAG (red).

The same conditions were hence used for a scaled-up experiment to collect sufficient DNA for qPCR. 800ug of chromatin lysate was obtained for incubation with the GFP-Trap. To maximise DNA yields, conjugated complexes were eluted in 200ul of Elution buffer overnight at 65^{o}C, before decrosslinking of the DNA and

purification by phenolchloroformisoamyl alcohol purification and isopropanol precipitation. Once again, DNA yields were too low to measure accurately and needed to be concentrated using an Eppendorf® 5301 concentrator. For 10ul of concentrated solution, DNA yields of 5.9 and 1.9ng of DNA were measured for bFosVC155-bJunVN173 and bFosDeltaZIPVC155-bJunVN173, respectively. Hence, electroporation with 0.5ug of DNA generated enough reconstituted Venus protein for downstream analysis. Increased electroporation amount, an alternate transfection amount, or integration into the genome may also improve this yield and allow assessment of whether tagging with the split Venus constructs alters the localisation signal of the tagged proteins, either by qPCR or ChIP-seq.

Overall, this worked validated a method to enrich specifically interacting proteins for downstream use in ChIP-seq analysis.

3.2.6 SA2 interacts with CTCF but not as robustly as SA1

Colocalisation of CTCF and SA2 ChIP-Seq peaks suggested that CTCF and SA2 interact and the lack of co-IP observed was due to the conditions used. Technical differences between the ChIP-seq and co-IP protocols include the harshness of the buffers used and the extent of sonication, both of which could differentially affect the stability of interacting proteins. An additional technical difference is the fixation step of ChIP-seq. Prior to collection proteins are fixed to the DNA during ChIP. If CTCF and SA2 interaction is dynamic or weak in nature this fixation step could account for increased detection together.

As initial co-IP optimisation experiments focused on CTCF-SA1 interaction and in order to test the technical differences mentioned above, different sonication and salt conditions for chromatin solubilisation were tested to determine if more mild conditions could reveal CTCF and SA2 co-IP (Figure 31). Some CTCF co-IP was observed with 150mM salt and 1 x 10 secs sonication, suggesting that CTCF and SA2 interact in accessible chromatin (harsh conditions were not required to solubilise these proteins) and/or that CTCF-SA2 interaction is disturbed in the harsher conditions. However, the CTCF co-IP signal was still very weak compared to input and significantly different to that observed with SA1.

Using a different SA2 antibody for IP (SA2 IP (Rb)) also slightly increased co-IP of CTCF, but again, detection was still much below input. Therefore, neither reducing buffer salt concentration nor chromatin sonication time resulted in robust CTCF–SA2 co-IP.

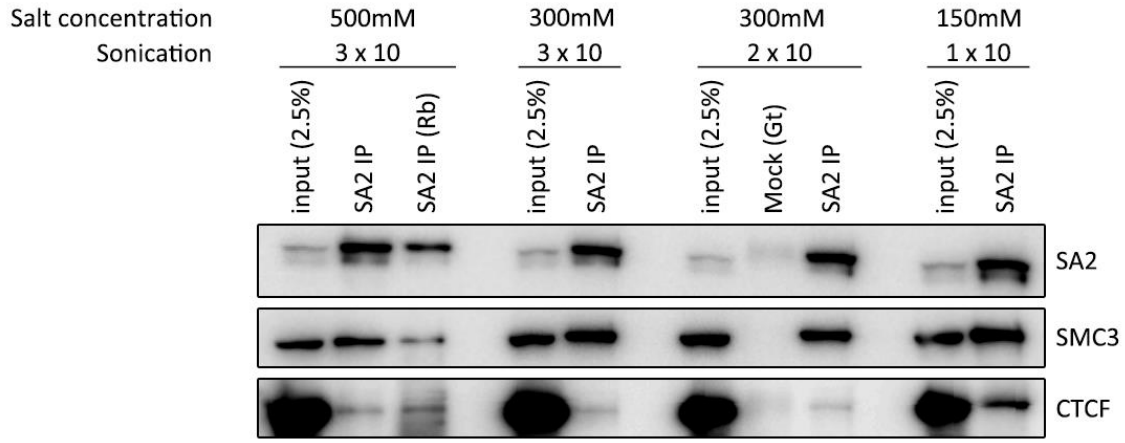


Figure 31: Effect of salt concentration and sonication conditions on co-IP of CTCF with SA2. Optimisation of SA2 IP conditions to preserve interaction with CTCF. Cells were fractionated for chromatin under four conditions, with changes to the salt concentration of the chromatin solubilisation buffer and sonication conditions indicated. An alternative SA2 antibody raised in rabbit was also tested (SA2 IP (Rb)) for efficacy of pull down and enrichment of CTCF. SMC3 was blotted as a positive control for interaction with SA2.

As the co-IP differences between SA1 and SA2 with CTCF did not seem to be obviously technical, biological differences between the SA proteins were considered. The idea of a biological contribution was support by the CTCF and SA1 co-IP results in Section 3.2.2 and 3.2.3, where, as well as the canonical CTCF and SA1 bands, additional bands were consistently present in the input and these bands co-IP'd with varying efficiency (Figure 32A). Notably, some of these bands became enriched upon auxin treatment. For example, the SA1 band termed 's3' co-IP'd with CTCF more efficiently than the SA1 band termed 's2', and the CTCF band termed 'c3' co-IP'd with SA1 more efficiently than the CTCF band termed 'c2'.

It is possible that these additional bands represent modified or alternatively spliced versions of CTCF and SA1. Protein degradation, insufficient membrane blocking, or excess secondary antibody may also contribute to the presence of multiple bands on a western blot. These contributions should be minimal in western blots through this study however as, samples were generated on ice with

protease inhibitors added to each buffer, increasing the concentration of the milk block from 5% to 10% did not alter the banding pattern, and secondary antibody was used at a dilution of 1:10,000 and all membranes are washed for 5 x 5 mins in PBS – 1% Tween prior to imaging. Cross-reactivity of the primary antibody with off-target proteins may also contribute to additional band detection, however, for co-IP'd bands, the off-target protein would need to interact with CTCF/SA1 and the SA1 s2 and s3 bands, at least, were specifically reduced with siRNA-mediated knockdown of SA1 (samples in Figure 67). For these reasons, we hypothesised that at least some of these bands may represent variants of SA1 and CTCF and that the variants impact interaction.

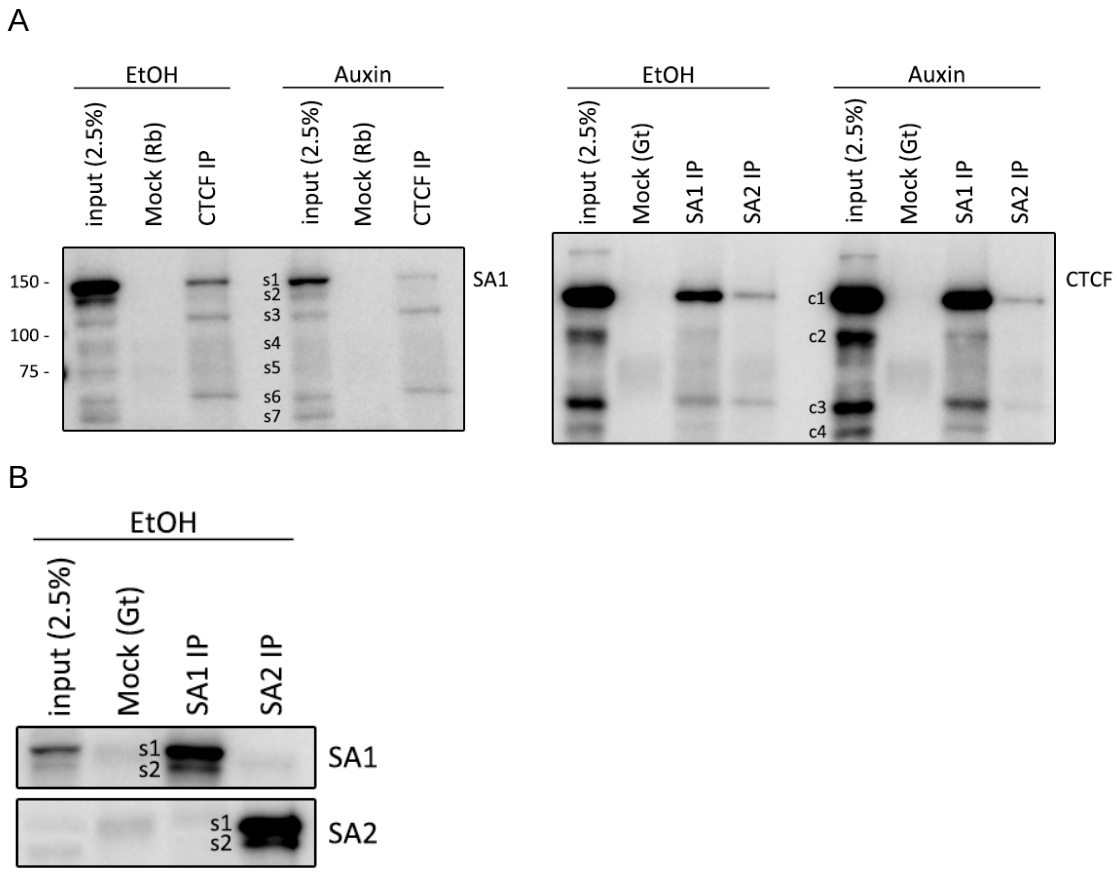


Figure 32: Analysis of naturally occurring variants of SA1 and SA2 and the contribution of a conserved C-terminal domain to interaction with CTCF. (A) Full SA1 (left) and CTCF (right) western blots from the CTCF and SA IPs shown in Figure 13C, respectively. SA1 putative variant bands are indicated as s1-s7 and CTCF putative variant bands are indicated as c1-c4. (B) Comparison of putative s1 and s2 variants of SA1 and SA2. Again, the blots refer to the experiment shown in Figure 13C.

Mass spectrometry is an analytical technique used to identify the mass-to-charge ratio of molecules in a sample, allowing identification of the precise molecular weight and chemical structure of the molecules (Mellon, 2003). In proteomics,

target proteins can be digested enzymatically to expose charged amino acids and produce a characteristic set of peptide fragments, known as a 'peptide map' (Kellner and Houthaeve, 1999). Further to mass determination, tandem mass spectrometry allows sequencing of detected peptides and accurate protein identification by comparison to existing proteome databases (Domon and Aebersold, 2006). Hence, mass spectrometry could be used to distinguish variant splicing isoforms of SA1 and SA2 if the splicing isoforms digest to produce a peptide map distinct to the canonical protein. This would then allow determination of the nature of interaction between specific variants of each protein and the contribution of specific regions of each protein to their interactions together.

Notably, two isoforms of SA1 and two isoforms of SA2 have been described on the UCSC Genome Browser. The canonical SA1 isoform has a mass of 144 kDa and its variant isoform, which is missing amino acids 1150-1186 (exon 31), has a mass of 140 kDa. In contrast, the canonical SA2 isoform has a mass of 141 kDa and its variant isoform, which contains an additional exon (exon 32) at amino acid 1156, has a mass of 145 kDa. HCT116 ENCODE RNAseq data was analysed by Dr. Wazeer Varsally, a postdoc in the Hadjur lab. He determined that in HCT cells around 96% of SA1 transcripts contain exon 31 while 75% of SA2 transcripts lack exon 32. Altogether this raised the question of whether, in western blots, SA1 band 's1' represents canonical SA1 and SA1 band 's2' represents the exon31-delta version of SA1, and if SA2 band 's1' represents the variant exon 32+ version of SA2 and SA2 band 's2' the canonical exon32-delta version of SA2 (Figure 32B). Upon IP, the s1 band of SA1 had a stronger signal than band s2, whereas SA2 band s2 was more equal to band s1.

When SA1 and SA2 are aligned, either globally or at the local exon level, exon 31 of SA1 aligns with exon 32 of SA2 (Supplemental Figure 3A). SA1 exon 31 shares 33% identity and 66% amino acid similarity with SA2 exon 32 (Table 5). In contrast, SA1 exon 31 only shares 25% similarity with SA2 exon 31 and 26% similarity with SA2 exon 33 (Supplemental Figure 3 B & C). Dr. Dubravka Pezic, a postdoc in the Hadjur lab, computed the theoretical isoelectric point (pI) of both exons using ExPASy – SA1 exon 31 has a theoretical pI of 10.42 and SA2 exon 32 has a theoretical pI of 9.97. Therefore, while not identical, the amino acid properties of the two exons are similar and both exons are basic in nature (pI >7).

Given that SA1 band s2 does not co-IP well with CTCF, we hypothesized that exon 31/32 may play an important role in the difference between SA1 and SA2 interaction with CTCF (Figure 32A).

NW Score	Identities	Positives	Gaps
50	13/39(33%)	26/39(66%)	4/39(10%)

```
SA1 exon31  1   PQMQISW-LGQPKLEDLNRK-DRTGMNYMKVRTGVRHAV  37
                 Q  ++W L Q + E+  ++ +R  M+Y+K+RT ++HA+
SA2 exon32  1   TQ--VTWMLAQRQQEEARQQQERAAMSYVKLRTNLQHAI  37
```

Table 5: **Alignment of SA1 exon31 and SA2 exon32.** Blast® Needleman-Wunsch alignment results for compaison of the protein sequence of SA1 exon31 and SA2 exon32. Summary statistics for the alignment are shown at the top of the table. Sequences for each of the exons were obtained from the UCSC genome browser Human GRCh38 track. Alignment of the two sequnces is shown in the middle line of sequence; a letter indicates an identical match, + indicates a positive match of similarity, and white space indicates mismatch.

In silico analysis of the two SA1 and SA2 isoforms, performed by Amandeep Bhamra from the Cancer Institute proteomics department, determined that trypsin digestion would produce distinct peptide maps for each isoform. Specifically, SA1 digestion could produce four potential peptides from the exon 31 region for the canonical isoform – ENSRPMGDQIQEPESEHGSEPDFLHNPQMQISWLGQPK, LEDLNRK, TGMNYMK, or DRTGMNYMK – and one potential peptide for the exon31-delta isoform – ENSRPmGDQIQEPESEHGSEPDFLHNR. Congruently, SA2 digestion could produce one potential peptide from the exon 31 – exon 32 junction for the canonical isoform – LRPEDSFMSVYPMQTEHHQTPLDYNR – and one potential peptide for the exon 32 region for the exon32+ isoform – TNLQHAIR. This suggested that mass spectrometry could be used to determine if a particular band from the SA1 IP was the canonical or the exon31-delta version of SA1 and if a particular band from the SA2 IP was the canonical or exon32+ version of SA2. The variant specific peptides and their presence in downstream experiments are summarised in Table 6 and Table 7.

A pilot SA1 mass spec experiment (MS1) was run in the Cancer Institute Proteomics facility using purified chromatin protein lysate from the HCT116 RmAC OsTIR1 H2 clone with the SA1 IP set up as in section 3.2.3. The IP was eluted in 2x Lammeli sample buffer and run on a pre-cast Bio-Rad Mini-PROTEAN® TGX™ gradient gel. Coomassie staining was carried out for around 3 hrs before destaining as in methods section 2.13.2. Stained bands were a weak

blue and could not be visualized well using a camera, however, they were visible by eye. Four bands were excised from the gel at; a) 150 kDa, a wide band to capture the potential canonical and delta-exon31 bands, b) 100 kDa, c) 75 kDa, and d) just below 75kDa. Peptides were isolated from the bands by in-gel digestion with trypsin and analyzed by tandem mass spectrometry.

Peptide mapping was carried out by Amandeep Bhamra using Proteome Discoverer software. Results are summarized in the MS1 column of Table 6. Peptides mapping to SA1 were identified in all four bands analysed. Bands A and D contained the peptide ENSRPMGDQIQEPESEHGSEPDFLHNPQMQISWLGQPK, band A contained the peptide LEDLNRK, and bands A, B, and D contained the peptide TGMNYMK. Thus, the SA1 variants in bands A, B, and D contained exon 31. Band A also contained the peptide ENSRPmGDQIQEPESEHGSEPDFLHNR, meaning an SA1 variant without exon 31 was also present in the band. Hence, this pilot experiment demonstrated that exon31+/- SA1 variants could be distinguished by mass spectrometry. It also highlighted that an optimized Coomassie staining protocol would be required to visualize the s1 and s2 bands.

SA1 variant	Peptide sequence	MS1	MS2	
		Band A	Band A	Band A.2
+ exon 31	ENSRPMGDQIQEPESEHGSEPDFLHNPQ MQISWLGQPK	Y	-	-
	LEDLNRK	-	Y	Y
	TGMNYMK	Y	Y	Y
	DRTGMNYMK	-	-	Y
- exon 31	ENSRPmGDQIQEPESEHGSEPDFLHNR	Y	-	Y

Table 6: Detection of naturally occurring variants of SA1. Summary of peptides that distinguish SA1 + exon 31 and SA1 - exon 31 and their presence (Y) or absence (-) in two SA1 IP plus mass spectrometry experiments, termed MS1 and MS2.

A second mass spec experiment (MS2) was run to optimize staining and cut apart the SA1 and SA2 doublets for analysis. SA2 IP was carried out as in section 3.2.3. To try to ensure visualization and isolation of the two SA1 bands of interest, IP was tested at two increased concentrations: 1) chromatin and antibody concentration were doubled and bead volume increased from 30ul to 35ul; and 2) three IPs were carried out using the same concentrations as earlier IPs, but

were then sequentially eluted into the same sample buffer. In addition, IP samples were eluted in 4x NuPAGE LDS buffer and run on a pre-cast NuPAGE™ 3-8% Tris-Acetate gel, which should separate high molecular weight proteins with high resolution and sensitivity. Coomassie staining after 2 hrs was slightly stronger than the previous experiment but was still weak, so, the gel was left in Coomassie stain overnight at room temperature. Destaining was carried out for ~2hrs before imaging and cutting of the bands indicated in Figure 33. Due to the extended staining time, the gel pieces were further destained in 50:50 TEAB and Acetonitrile overnight to prevent interference of the Coomassie dye in the mass spec analysis. To maximise protein detection, bands from both SA1 lanes were joined, however, low levels of trypsin peptides were then detected in the digested samples, indicating that digestion may not have been fully efficient. Gel pieces were subjected to a second round of digestion to better breakdown proteins within the gel pieces and the resulting peptide solution added to the first round of samples.

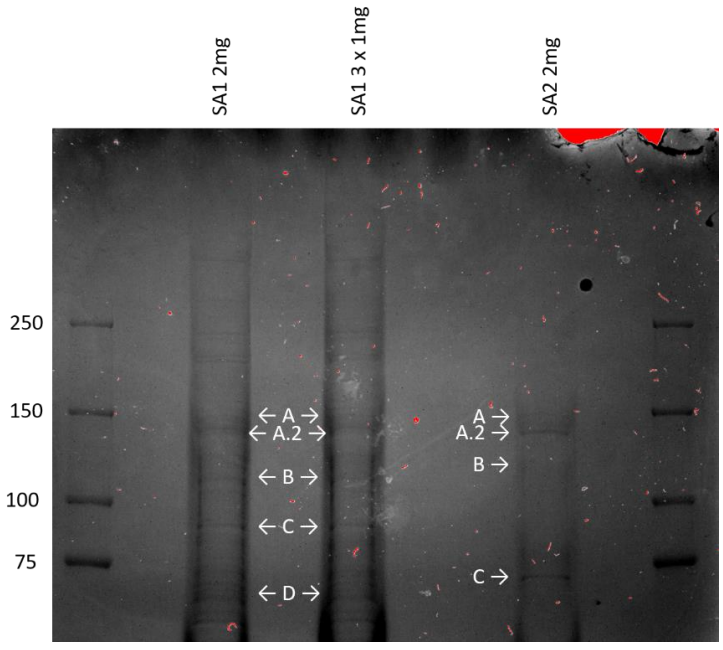


Figure 33: Analysis of naturally occurring variants of SA1 and SA2 – mass spectrometry of IP bands. Coomassie stained SDS-PAGE gel of SA1 and SA2 IPs. Indicated bands were cut and processed for LC-MS.

Results of the SA1 IP are summarized in the MS2 columns of Table 6 above and results of the SA2 IP are summarized in Table 7 below. SA1 peptides were detected in all five SA1 IP bands. The peptides LEDLNRK and TGMNYMK were

detected in bands A and A2, and the peptide DRTGMNYMK was detected in band A2. ENSRPmGDQIQEPESEHGSEPDFLHNR was also found in band A2. Thus, band A contained an SA1 variant with exon 31, while band A2 contained exon31+/- SA1 variants. SA2 peptides were found from all four SA2 IP bands. LRPEDSFMSVYPMQTEHHQTPLDYNR was detected in band A. Whereas, the peptide TNLQHAIR was identified in bands A and A2. The peptide QTEHHQTPLDYNR was also observed in band A2; although this peptide was not predicted by the *in-silico* digestion, it also spans the exon 31–exon 33 junction, thus, bands A and A2 contained exon32+/- SA2 variants. Therefore, the naturally occurring exon31/32 +/- variants of SA1 and SA2 are expressed in the HCT116 RmAC OsTIR1 cells. Quantification of these variants was not possible in these experiments as a distinct form of mass spectrometry is required to quantify proteins, in which, known quantities of labelled peptides are included in the samples analysed.

SA2 variant	Peptide	MS2	
		Band A	Band A.2
- exon 32	LRPEDSFMSVYPMQTEHHQTPLDYNR	Y	-
	QTEHHQTPLDYNR	-	Y
+ exon 32	TNLQHAIR	Y	Y

Table 7: Detection of naturally occurring variants of SA2. Summary of peptides that distinguish SA2 - exon 32 and SA2 + exon 32 and their presence (Y) or absence (-) in the SA2 IP plus mass spectrometry experiment, termed MS2.

The presence of multiple isoform specific peptides from a single band may suggest that the distinct bands observed on the western blots are cause by a post-translation modification that can be deposited on both isoforms. However, streaking of proteins in SDS-PAGE gels can cause smearing of peptides down through the gel, and, given the sensitivity of mass spectrometry and the way that it aligns peptides to proteins, it is not possible to distinguish such contamination. For example, in the case where a peptide unique to SA1 exon31+ and a peptide unique to SA1 exon31- where both present in the same band, every other peptide that mapped to other regions of SA1 would be mapped to both the SA1 exon31+ and the SA1 exon31- variants, meaning that contamination with one single peptide of SA1 exon31+ into a SA1 exon31- band could not be identified.

Finally, to assess whether the alternative exons played a role in CTCF/SA interactions, mass spec of a CTCF IP was carried out, to see which SA1 and SA2 peptides would be present in its pull down, and if these contained peptides for the alternatively spliced exons (Figure 34). Coomassie staining of the CTCF IP is shown. The bands marked as 3 and 4 correspond to the size of CTCF. The bands marked as 2 and 3 correspond to the size of the SA1 and SA2 exon31/32 splicing variants. Additional bands indicated on the gel were also cut to assess protein content.

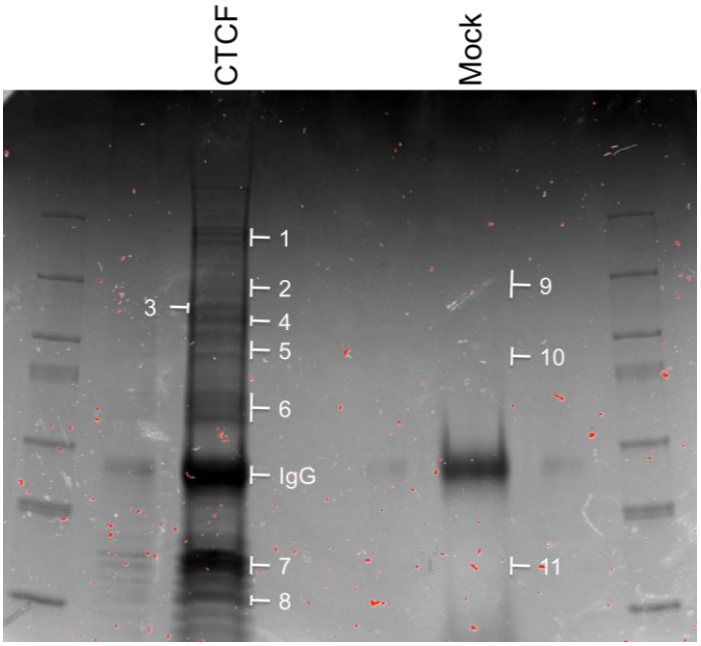


Figure 34: Analysis of naturally occurring variants of SA1 and SA2 – mass spectrometry of CTCF IP bands. Coomassie stained SDS-PAGE gel of CTCF and IgG (Mock) IPs. Indicated bands were cut and processed for LC-MS.

Peptides from SA1 were identified in bands 2 and 3 as expected, however, no peptides were present to distinguish the presence or absence of exon 31 in this SA1 (Table 8). Peptides from SA2 were identified in bands 2, 3 and 4. As for SA1, no peptides were present to distinguish the presence or absence of exon 32 in this SA2 (Table 8). Regardless, the identified peptides still indicate regions of the proteins that are present in the SA molecules that interact with CTCF, with the caveat that different regions of proteins are relatively 'mass spec-able', depending on factors such as amino acid sequence and accessibility to trypsin. Peptides were relatively evenly across SA1 (Table 8). In contrast, no peptides were identified in the last 319 amino acids of SA2 (Table 8).

Canonical SA1 Isoform sequence (includes exon 31 – highlighted in grey)

MITSELPVLQDSTNETTAHSDAGSELEETEVKGKRKRGRPGRPPSTNKKPRKSPGEKSRIEA
GIRGAGRGRANGHPQQNGEGEPVTLFEVVKLGKSAMQSVVDDWIESYKQDRDIALLDLINF
FIQCSGCRGTVRIEMFRNMQNAEIIRKMTEEFDEDSGDYPLTMPGPQWKKFRSNFCEFIGVL
IRQCQYSIIYDEYMMDTVISLLTGLSDSQVRAFRHTSTLAAMKLMTALVNVALNLSIHQDNTQ
RQYEAERNKMIGKRANERLELLLQKRKELQENQDEIENMMNSIFKGIFVHRYRDAIAEIRAICI
EEIGVWMKMYSDAFLNDSYLKYVGWTLHDRQGEVRLKCLKALQSLYTNRELFPKLELFTNR
FKDRIVSMTLDKEYDVAVEAIRLVTLILHGSEEALSNEDCENVYHLVYSAHRPVAVAAGEFLH
KKLFSRHDPQAEEALAKRRGRNSPNGNLIRMLVLFFLESELHEHAAYLVDSLWESSQELLKD
WECMTELLLEEPVQGEEAMSDRQESALIELMVCTIRQAAEAHPPVGRGTGKRVLTAKERKT
QIDDRNKLTEHFIITLPMLLSKYSADAEKVANLLQIPQYFDLEIYSTGRMEKHLDALLKQIKFVV
EKHVESDVLEACSKTYSILCSEEYTIQNRVDIARSQLIDEFVDRFNHSVEDLLQEGEEADDDD
IYNVLSTLKRLTSFHNAHDLTKWDLFGNCYRLLKTGIEHGAMPEQIVVQALQCSHYSILWQLV
KITDGSPSKEDLLVLRKTVKSFLAVCQQCLSNVNTPVKEQAFMLLCDLLMIFSHQLMTGGRE
GLQPLVFNPDTGLQSELLSFVMDHVFIDQDEENQSMEGDEEDEANKIEALHKRRNLLAAFSK
LIIYDIVDMHAAADIFKHYMKYYNDYGDIIKETLSKTRQIDKIQCAKTLILSLQQLFNELVQEQGP
NLDRTSAHVSGIKELARRFALTFGLDQIKTREAVATLHKDGIEFAFKYQNQKGQEYPPPNLAF
LEVLSEFSSKLLRQDKKTVHSYLEKFLTEQMMERREDVWLPLISYRNSLVTGGEDDRMSVN
SGSSSSKTSSVRNKKGRPPLHKKRVEDESLDNTWLNRTDTMIQTPGPLPAPQLTSTVLREN
SRPMGDQIQEPESEHGSEPDFLHNPQMQISWLGQPKLEDLNRKDRTGMNYMKVRTGVRH
AVRGLMEEDAEPIFEDVMMSSRSQLEDMNEEFEDTMVIDLPPSRNRRERAELRPDFFDSAA
IIEDDSGFGMPMF

Variant SA2 isoform sequence (includes exon 32 – highlighted in grey)

MIAAPEIPTDFNLLQESETHFSSDTDFEDIEGKNQKQGKGKTCKKGKKGPAEKGKGGNGGG
KPPSGPNRMNGHHQQNGVENMMLFEVVKMGKSAMQSVVDDWIESYKHDRDIALLDLINFFI
QCSGCKGVVTAEMFRHMQNSEIIRKMTEEFDEDSGDYPLTMAGPQWKKFKSSFCEFIGVLV
RQCQYSIIYDEYMMDTVISLLTGLSDSQVRAFRHTSTLAAMKLMTALVNVALNLSINMDNTQR
QYEAERNKMIGKRANERLELLLQKRKELQENQDEIENMMNAIFKGVFVHRYRDAIAEIRAICIE
EIGIWMKMYSDAFLNDSYLKYVGWTMHDKQGEVRLKCLTALQGLYYNKELNSKLELFTSRF
KDRIVSMTLDKEYDVAVQAIKLLTLVLQSSEEVLTAEDCENVYHLVYSAHRPVAVAAGEFLYK
KLFSRRDPEEDGMMKRRGRQGPNANLVKTLVFFFLESELHEHAAYLVDSMWDCATELLKD
WECMNSLLLEEPLSGEEALTDRQESALIEIMLCTIRQAAECHPPVGRGTGKRVLTAKEKKTQ
LDDRTKITELFAVALPQLLAKYSVDAEKVTNLLQLPQYFDLEIYTTGRLEKHLDALLRQIRNIVE
KHTDTDVLEACSKTYHALCNEEFTIFNRVDISRSQLIDELADKFNRLLEDFLQEGEEPDEDDA
YQVLSTLKRITAFHNAHDLSKWDLFACNYKLLKTGIENGDMPEQIVIHALQCTHYVILWQLAKI
TESSSTKEDLLRLKKQMRVFCQICQHYLTNVNTTVKEQAFTILCDILMIFSHQIMSGGRDMLE
PLVYTPDSSLQSELLSFILDHVFIEQDDDNNSADGQQEDEASKIEALHKRRNLLAAFCKLIVYT
VVEMNTAADIFKQYMKYYNDYGDIIKETMSKTRQIDKIQCAKTLILSLQQLFNEMIQENGYNFD
RSSSTFSGIKELARRFALTFGLDQLKTREAIAMLHKDGIEFAFKEPNPQGESHPPLNLAFLDIL
SEFSSKLLRQDKRTVYVYLEKFMTFQMSLRREDVWLPLMSYRNSLLAGGDDDTMSVISGIS
SRGSTVRSKKSKPSTGKRKVVEGMQLSLTEESSSSDSMWLSREQTLHTPVMMQTPQLTSTI
MREPKRLRPEDSFMSVYPMQTEHHQTPLDYNTQVTWMLAQRQQEEARQQQERAAMSYV
KLRTNLQHAIRRGTSLMEDDEEPIVEDVMMSSEGRIEDLNEGMDFDTMDIDLPPSKNRRERT
ELKPDFFDPASIMDESVLGVSMF

Table 8: Analysis of naturally occurring variants of SA1 and SA2 – SA peptides identified in mass spectrometry of CTCF. Sequence of SA1 (top) and SA2 (bottom) with variant exons highlighted in grey and peptides identified in the CTCF IP indicated in orange text.

Multiple peptides were identified in this C-terminal region of SA2 in the SA2 MS2 IP, indicating that lack of detection in the CTCF IP is not a consequence of a lack of trypsin-digestible amino acids in this region of SA2 (Table 9). The lack of C-terminal SA2 peptides detected could indicate that this region is variable spliced

in SA2 molecules interacting with CTCF, resulting in no robust identification of peptides due to dilution of any peptides below detectable levels. Given the differences between SA1 and SA2 detection with CTCF in co-IP and ChIP-seq experiments, this difference may suggest an important role for the C-terminus of SA1 in stabilisation of interaction with CTCF.

Peptides from this region of SA2 may also be absent from the CTCF IP due to post-translational modifications, poor recovery in this sample, or co-elution with more abundant peptides that mask them from sequencing. Given the vast numbers of peptides analysed in each MS-IP sample, it is also possible that the algorithm may simply not have matched peptides from this region by chance. These caveats are especially important as this result is from a single replicate.

Variant SA2 isoform sequence (includes exon 32 – highlighted in grey)

MIAAPEIPTDFNLLQESETHFSSDTDFEDIEGKNQKQGKGKTCKKGKKGPAEKGKGGNGGG
KPPSGPNRMNGHHQQNGVENMMLFEVVKMGKSAMQSVVDDWIESYKHDRDIALLDLINFFI
QCSGCKGVVTAEMFRHMQNSEIIRKMTEEFDEDSGDYPLTMAGPQWKKFKSSFCEFIGVLV
RQCQYSIIYDEYMMDTVISLLTGLSDSQVRAFRHTSTLAAMKLMTALVNVALNLSINMDNTQR
QYEAERNKMIGKRANERLELLLQKRKELQENQDEIENMMNAIFKGVFVHRYRDAIAEIRAICIE
EIGIWMKMYSDAFLNDSYLKYVGWTMHDKQGEVRLKCLTALQGLYYNKELNSKLELFTSRF
KDRIVSMTLDKEYDVAVQAIKLLTLVLQSSEEVLTAEDCENVYHLVYSAHRPVAVAAGEFLYK
KLFSRRDPEEDGMMKRRGRQGPNANLVKTLVFFFLESELHEHAAYLVDSMWDCATELLKD
WECMNSLLLEEPLSGEEALTDRQESALIEIMLCTIRQAAECHPPVGRGTGKRVLTAKEKKTQ
LDDRTKITELFAVALPQLLAKYSVDAEKVTNLLQLPQYFDLEIYTTGRLEKHLDALLRQIRNIVE
KHTDTDVLEACSKTYHALCNEEFTIFNRVDISRSQLIDELADKFNRLLEDFLQEGEEPDEDDA
YQVLSTLKRITAFHNAHDLSKWDLFACNYKLLKTGIENGDMPEQIVIHALQCTHYVILWQLAKI
TESSSTKEDLLRLKKQMRVFCQICQHYLTNVNTTVKEQAFTILCDILMIFSHQIMSGGRDMLE
PLVYTPDSSLQSELLSFILDHVFIEQDDDNNSADGQQEDEASKIEALHKRRNLLAAFCKLIVYT
VVEMNTAADIFKQYMKYYNDYGDIIKETMSKTRQIDKIQCAKTLILSLQQLFNEMIQENGYNFD
RSSSTFSGIKELARRFALTFGLDQLKTREAIAMLHKDGIEFAFKEPNPQGESHPPLNLAFLDIL
SEFSSKLLRQDKRTVYVYLEKFMTFQMSLRREDVWLPLMSYRNSLLAGGDDDTMSVISGIS
SRGSTVRSKKSKPSTGKRKVVEGMQLSLTEESSSSDSMWLSREQTLHTPVMMQTPQLTSTI
MREPKRLRPEDSFMSVYPMQTEHHQTPLDYNTQVTWMLAQRQQEEARQQQERAAMSYV
KLRTNLQHAIRRGTSLMEDDEEPIVEDVMMSSEGRIEDLNEGMDFDTMDIDLPPSKNRRERT
ELKPDFFDPASIMDESVLGVSMF

Table 9: Analysis of naturally occurring variants of SA1 and SA2 – SA2 peptides identified in IP-MS of SA2. Sequence of SA2 with variant exon highlighted in grey and peptides identified in the SA2 IP band A indicated in orange text. The peptide that distinguished splicing out of exon 32 is indicated in green and the peptide that distinguished splicing in of exon 32 indicated in red. Amino acids that overlap between these two peptides are indicated in blue. Some of the sequences in orange also represent overlapping peptides, this is not idicated for simplicity.

3.2.7 Optimised nucleic acid digestion conditions for identification of CTCF and SA in complex

As expected, core cohesin complex members and regulators of cohesin, such as PDS5B, were detected in the CTCF and SA mass spec experiments. These preliminary experiments also suggested that many additional proteins were also co-purified with SA proteins and were predominantly RNA and DNA-binding proteins. These results are discussed in section 4.2.1. Given the abundance of nucleic acid binding proteins identified in the preliminary mass spectrometry experiments, co-IP of nucleic acids was analysed to determine if benzonase was efficiently digesting DNA and RNA or if molecules of CTCF/SA1 could be pulling down large amounts of interacting nucleic acids. Briefly, two set of cells were fractionated, one treated with benzonase and one without. Chromatin proteins were pulled down using IgG or an SA1 antibody, and the IP material purified for DNA. DNA samples were run on an Agilent High Sensitivity DNA Chip (Figure 35).

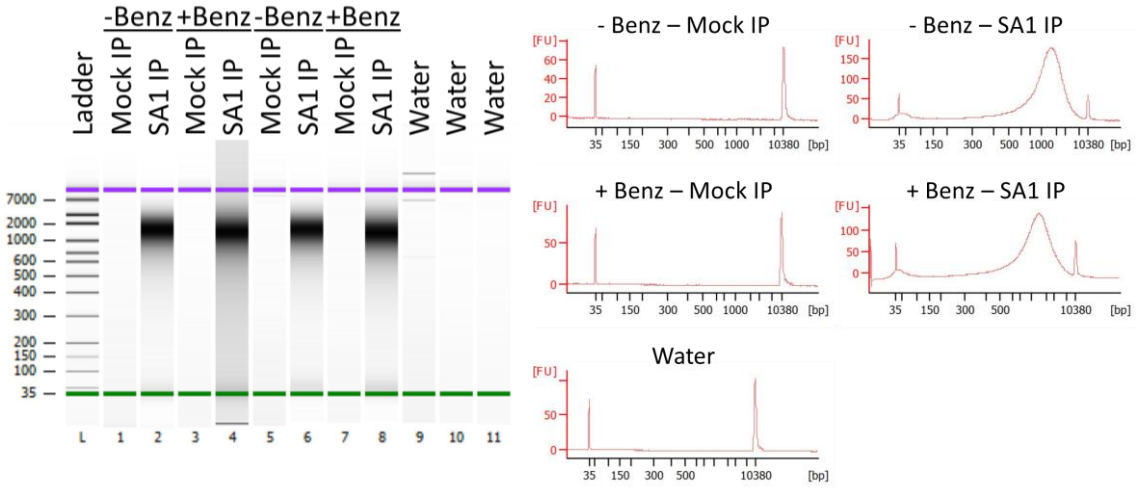


Figure 35: DNA content of the SA1 IP. Excerpt of Bioanalyzer result from Agilent High Sensitivity DNA Chip. Two sets of cells were fractionated, one with no benzonase digestion (-Benz) and one with benzonase digestion (+ Benz) during chromatin solubilisation. The resulting chromatin samples were enriched in an SA1 or IgG (mock) IP. IP material was purified for DNA and analysed on a DNA chip. An image of the gel is shown on the left and DNA distribution graphs are shown on the right.

Treatment with benzonase reduced the overall concentration of DNA in the SA1 IP 3.5-fold compared to the undigested control (Figure 35). While reduced, DNA was still present at a concentration of 1.32 ng/ul in the benzonase-treated sample, a higher level than expected. Moreover, the average DNA fragment size in both

conditions was ~1,600bp, and the distribution of DNA fragment size was similar regardless of benzonase treatment. Therefore, more large fragments of DNA were present in the IP than expected.

Colleagues in the institute suggested a higher concentration of benzonase that might digest nucleic acids more efficiently. To test the effect of the increased benzonase concentration on co-IP of CTCF with SA1 and on the presence of DNA and RNA in the IP, four SA1 IP conditions were assessed: 1) a control for expected co-IP with the previously optimised condition of 6U benzonase per 100 x 10^6 cells, 2) the new increased benzonase condition of 85U per 100 x 10^6 cells, 3) a positive control for complete digestion of RNA with 6U benzonase per 100 x 10^6 cells plus RNase A that would also assess the specific contribution of the remaining DNA to the co-IP, and 4) a positive control for the complete digestion of DNA with 6U benzonase per 100 x 10^6 cells plus TURBO™ DNase that would also assess the specific contribution of the remaining RNA to the co-IP. While benzonase can effectively digest nucleic acids at 4^{o}C, RNase A and TURBO™ DNase require higher temperatures to work well. Hence, to try and optimise digestion while minimising protein degradation, all digests were carried out at 37^{o}C for 10 mins then transferred to 4^{o}C for 20 minutes. Immediately thereafter, the samples were supplemented with EDTA to inhibit RNase A and TURBO™ DNase activity and try to prevent off-target digestion of DNA and RNA during the overnight IP. The IP eluates were each split in two, with one half processed for DNA and RNA and the other half run on a western blot to assess SA1 IP and CTCF co-IP in the different conditions.

Purified DNA was run on an Agilent High Sensitivity DNA Chip (Figure 36). By comparison of the 6 and 85U benzonase samples, increasing the benzonase concentration during fractionation increased the concentration of DNA co-IP'd with SA1 and decreased the average fragment size from 1,338 bp to 482 bp. The distribution of DNA fragment size was also altered with increased benzonase concentration. Firstly, a higher, broader peak was observed ~35–100 bp, indicating an increase in short DNA fragments. Secondly, ~100–6,000 bp, DNA fragment size shifted from a right-skewed, unimodal distribution to a wider and flatter normal distribution, indicating a shift from predominantly larger DNA fragment to a range of DNA fragment sizes. Addition of RNase A to sample 3 did

not affect the concentration and fragment size distribution of DNA compared to 6U benzonase alone, indicating the specificity of the enzyme for RNA. Addition of TURBO DNase to sample 4 decreased DNA concentration ~3-fold compared to the 6U benzonase alone. The remaining DNA was still present as two peaks: a normal peak at 35–100 bp and a right skewed peak at 600–6,000 bp. However, the TURBO DNase had a greater effect on the longer DNA fragments – the concentration of DNA in the 35–100 bp peak was half, whereas the concentration of DNA at ~1,400 was reduced X20.

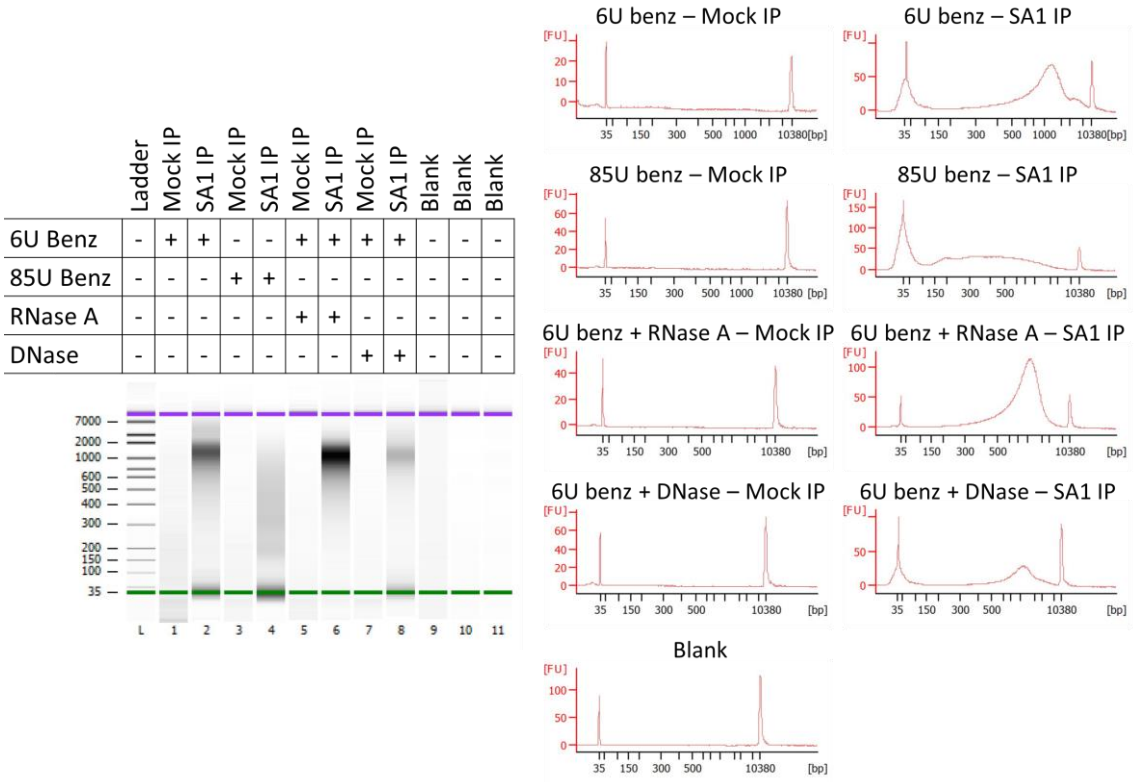


Figure 36: Dependence of SA1-CTCF interaction on specific nucleic acid digestion – DNA analysis. Excerpt of Bioanalyzer result from Agilent High Sensitivity DNA Chip. Four sets of cells were fractionated and treated with either 6U benzonase per $100x10^6$ (6U benz), 85U benzonase per $100x10^6$ (85U benz), 6U benzonase per $100x10^6$ and RNase A (6U benz + RNase A), or 6U benzonase per $100x10^6$ and TURBO DNase (6U benz + DNase). The resulting chromatin samples were enriched in an SA1 or IgG (mock) IP. DNA purified from the IP was analysed on an Agilent High Sensitivity DNA Chip. An image of the gel is shown on the left and DNA distribution graphs are shown on the right.

Purified RNA was run on an Agilent RNA 6000 Pico Chip. By comparison of the 6 and 85U benzonase samples, increased benzonase during fractionation increased the overall concentration of RNA in the IP and reduced the average size of RNA fragments co-IP'd with SA1 (Figure 37). The distribution of RNA length was considerably shifted towards smaller RNAs. In contrast, addition of

RNase A to sample 3 reduced the concentration of small RNA fragments compared to the 6U benzonase treatment alone. This decrease was predominantly from shorter RNAs of <1,000 bp with RNA between 1,000 and 2,000 bp retained at a level similar to the 6U benzonase sample. The RNase A used was procured from a previous member of the lab and may have unwittingly been specific to single-stranded RNA – RNAs longer than 200bp may be more likely to self-hybridise and form dsRNAs, hence, protecting them from digestion in this condition and explaining the distribution of digestion (Li, Zhu and Luo, 2016). Addition of TURBO DNase to sample 4 did reduce the overall concentration of RNA compared to the 6U benzonase sample, however, the size distribution of the remaining RNA fragments was not affected.

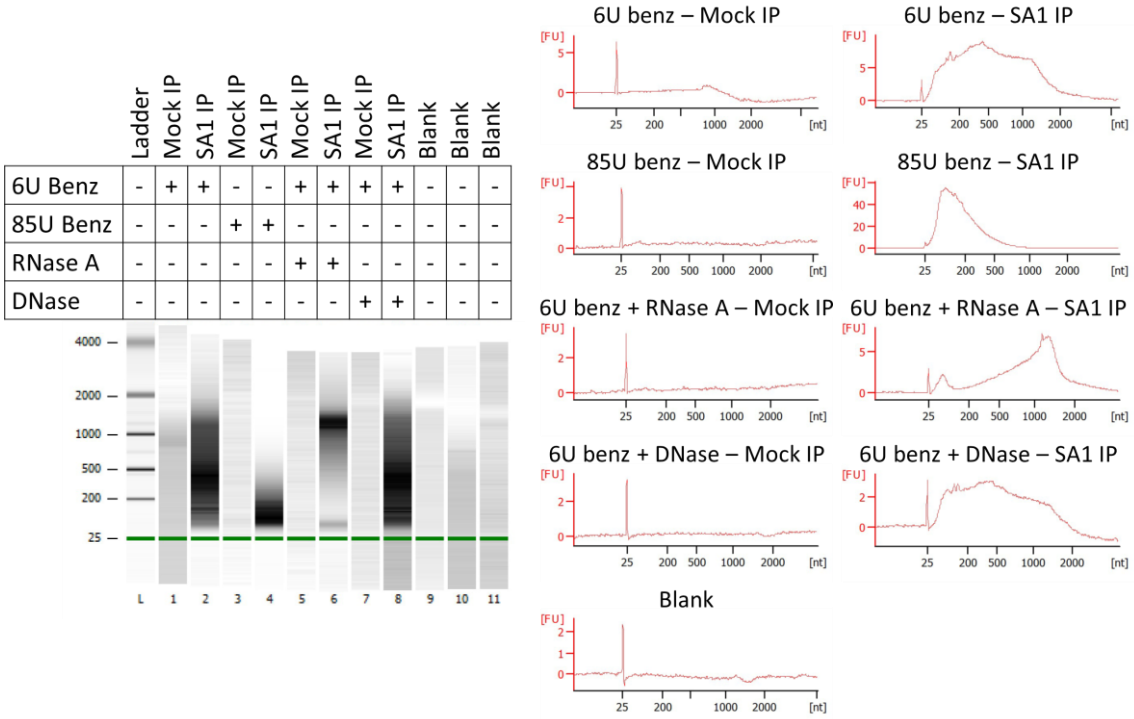


Figure 37: Dependence of SA1-CTCF interaction on specific nucleic acid digestion – RNA analysis. Excerpt of Bioanalyzer result from Agilent RNA 6000 Pico Chip. Four sets of cells were fractionated and treated with either 6U benzonase per $100x10^6$ (6U benz), 85U benzonase per $100x10^6$ (85U benz), 6U benzonase per $100x10^6$ and RNase A (6U benz + RNase A), or 6U benzonase per $100x10^6$ and TURBO DNase (6U benz + DNase). The resulting chromatin samples were enriched in an SA1 or IgG (mock) IP. RNA purified from the IP was analysed on an Agilent RNA 6000 Pico Chip. An image of the gel is shown on the left and RNA distribution graphs are shown on the right.

Interestingly, western blot of the different IP samples revealed that while none of the treatments effected the amount of SA1 IP'd, co-IP of CTCF was greatly improved with the 85U benzonase treatment (Figure 38). This was likely

facilitated by the increase of either the range of shorter DNA fragments or the short RNA fragments, or a combination of the two. Loss of short RNA fragments with RNase A addition in sample 3 resulted in a similar co-IP efficiency as the 6U benzonase alone, indicating that the short RNA fragments are not required for co-IP of CTCF with SA1. Whereas, addition of TURBO DNase to sample 4 decreased the efficiency of CTCF co-IP compared to the 6U benzonase sample. This addition of TURBO DNase removed the majority of larger DNA fragments and reduced the levels of smaller DNA fragments, indicating that at least one of these components is required for co-IP of CTCF with SA1. The concentration of RNA fragment of all sizes was also reduced in this sample and could have played an additional role in the decrease in co-IP observed. All together these results suggest that the SA1–CTCF interaction is best captured with digestion to a range of short and long DNA fragments or short RNA fragments, or both.

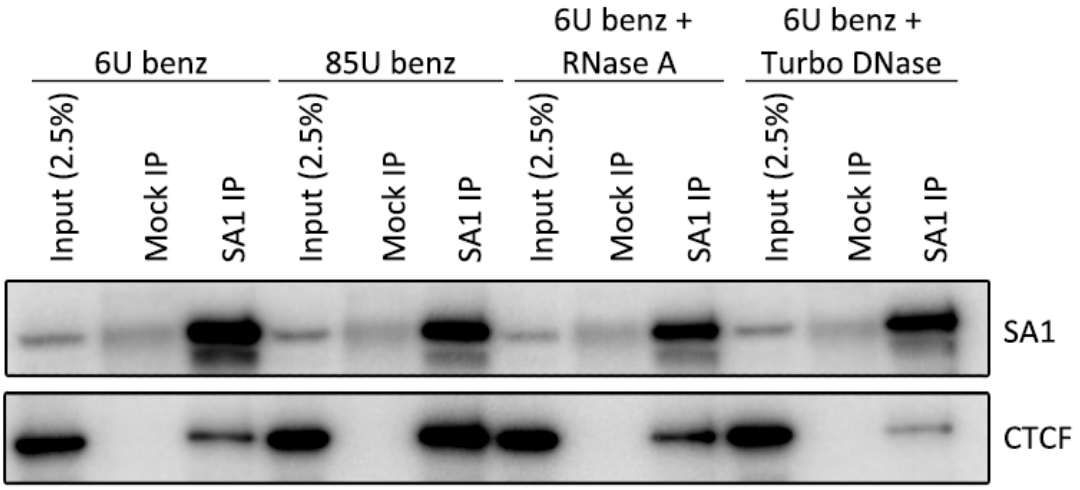


Figure 38: Dependence of SA1-CTCF interaction on specific nucleic acid digestion – WB analysis. Western blot of the IP samples from Figure 36 and Figure 37. The membrane was blotted for SA1 and CTCF to assess IP and co-IP in the different solubilisation conditions, respectively.

A final DNA and RNA co-IP assay was set up test the effect of the 85U benzonase condition on CTCF–SA2 interaction and to determine if an even higher benzonase amount would further increase co-IP. The 85U benzonase condition resulted in successful co-IP of SA2 and increased co-IP of SA1 with CTCF compared to earlier experiments (Figure 39A). As co-IP and ChIP-seq results thus far suggested the stability of CTCF-SA2 interaction affects its capture, this suggested that more stable CTCF-SA was captured with nucleic acid fragmentation with the 85U benzonase condition.

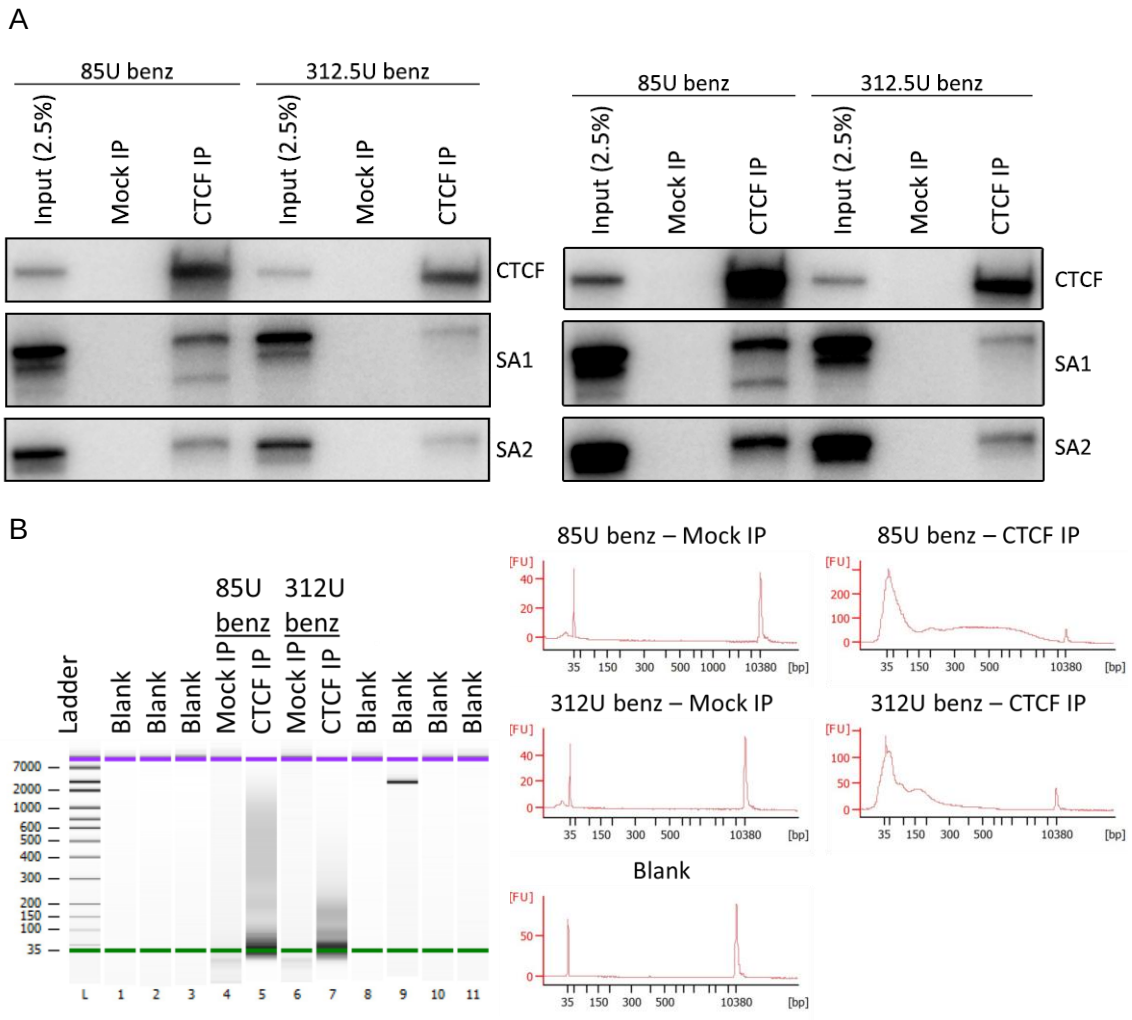


Figure 39: Dependence of CTCF-SA2 interaction on specific nucleic acid digestion. Two sets of cells were fractionated, one with 85U benzonase per 100x10^6 (85U benz) and one with 312U benzonase per 100x10^6 (312U benz) during chromatin solubilisation. The resulting chromatin samples were enriched in a CTCF or IgG (mock) IP. (A) Western blot of the IP samples. The membrane was blotted for CTCF to assess IP and for SA1 and SA2 to assess co-IP. Low (left) and high (right) exposures of the same blots are shown to allow comparison with other experiments and visualisation of changes to SA co-IP, respectively. (B) DNA chip analysis of the IP samples. Excerpt of Bioanalyzer result from Agilent High Sensitivity DNA Chip. An image of the gel is shown on the left and DNA distribution graphs are shown on the right.

Increasing the benzonase concentration even higher was detrimental and reduced co-IP of both the SA proteins. Treatment with 312U benzonase per 100 x 10^6 cells decreased DNA fragments to < 167 bp (Figure 39B). This suggested that the increased co-IP at 85U benzonase was not simply a consequence of solubilising more proteins with increased digestion of nucleic acids. Instead a specific digestion appeared to be required. This experiment also suggests that the DNA fragments of 187-~1000 bp facilitate CTCF–SA interaction capture in the 85U benzonase condition, as these are the fragments that are lost with 312U. However, the concentration of DNAs < 167 bp was reduced compared to the 85U

benzonase condition, indicating that these short DNA fragments could also play a role in CTCF–SA co-IP. Unfortunately, RNA digestion could not be assessed due to technical difficulties. A more controlled experiment with DNA and/or RNA fragments of specific lengths would be required to tease out the exact nucleic acids that are causing the greatest effects in these benzonase experiments. Overall, it is clear that nucleic acids are involved in the interaction and that specific degradation conditions are required to allow optimal detection of the interaction.

Using the new 85U benzonase digestion condition optimised above, co-IP of CTCF and SA1/SA2 was repeated in the presence or absence of RAD21 and the cohesin ring (Figure 40A). Under ethanol conditions, CTCF and RAD21 were both enriched with SA1 and SA2 IP. Differential enrichment of the two proteins was evident with the two SA proteins, indicating a higher affinity of CTCF for SA1 and a higher affinity of RAD21 for SA2, as was observed in earlier experiments (Figure 13C). With auxin treatment RAD21 was degraded from the cells and, now, CTCF enrichment was increased compared to endogenous conditions, despite considerably lower IP levels of both SA proteins. This suggests a shift in SA localisation to CTCF. The reciprocal IP, with pull-down of CTCF, confirmed interaction in both RAD21-positive and RAD21-negative conditions.

To assess reproducibility of the above results and to help visualise co-IP of SA1 and SA2 with CTCF more robustly, a biological replicate of the IPs was run and half the input material loaded in the gel (Figure 40B). Unfortunately, SA1 and SA2 IP signal was patchy, potentially due to damage to the membrane from being stripped multiple times, making it difficult to assess and compare the efficiencies of their IP. However, CTCF and RAD21 signal matched that of the SA IPs above. This suggested that the IPs had worked similarly and the co-IP results were robust.

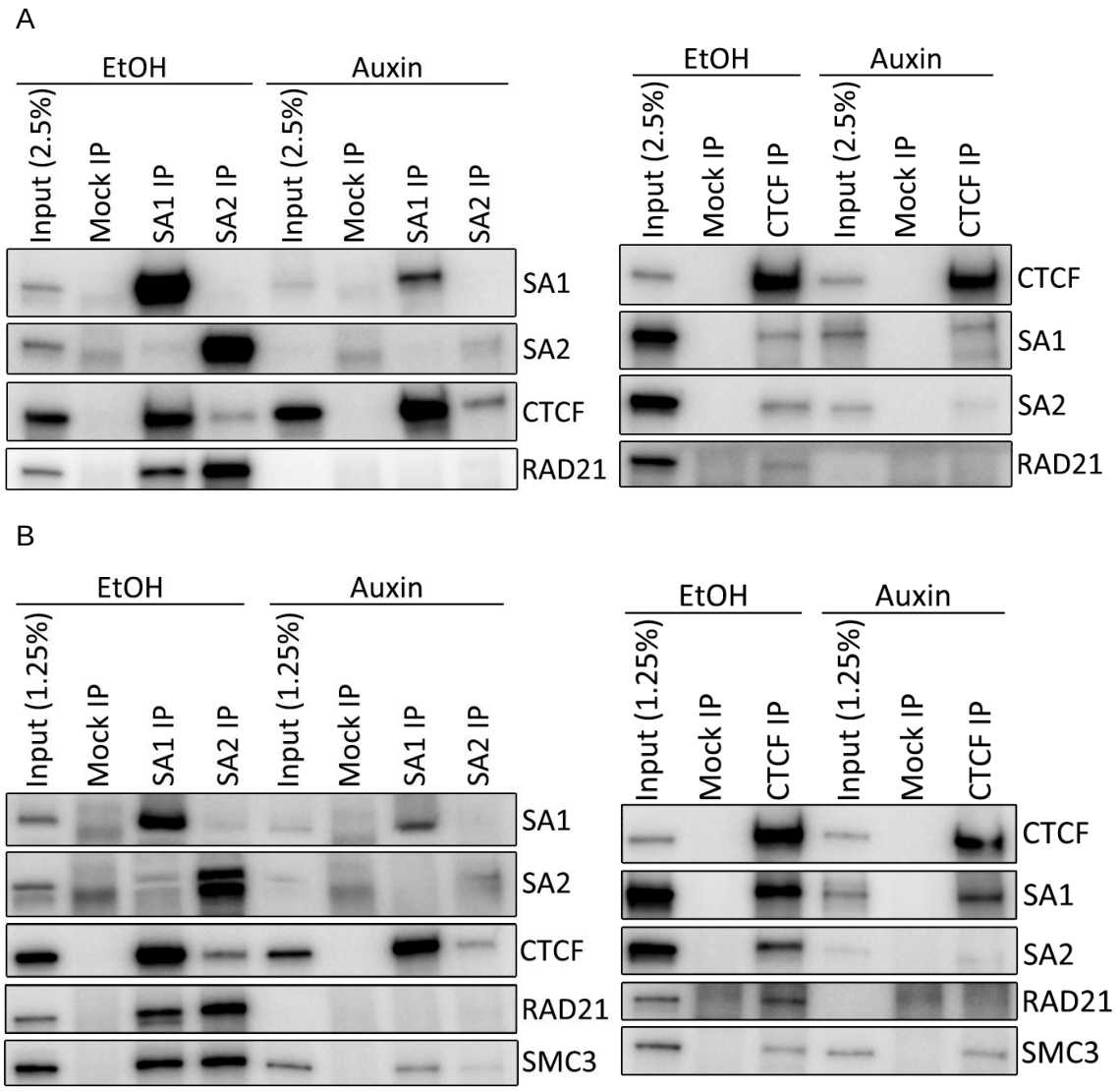


Figure 40: Co-IP of CTCF and SA in ethanol and auxin conditions (4 hrs) with further optimised nucleic acid digestion. (A) Cells were fractionated for chromatin using the newly optimised 85U benzonase per 100x10^6 condition. Chromatin proteins were pulled down using endogenous antibodies for SA1 and SA2 (left) and CTCF (right). RAD21 was blotted to assess efficiency of the auxin treatment. (B) Biological repeat of (A) with SMC3 immunoblotted for co-IP and just 1.25% input loaded in the gel to allow better visualisation of the SA co-IP signal with CTCF.

SA co-IP with CTCF was more visible with decreased input. As in previous experiments, SA1 co-IP with CTCF was retained with degradation of RAD21, and for both replicates, enrichment over input was increased in the auxin condition. Despite increased SA2 co-IP in ethanol conditions, co-IP of SA2 was abolished with loss of RAD21 in this experiment. In light of the differential enrichment of CTCF and RAD21 in SA1 and SA2 ethanol IPs, this difference makes sense and reveals the specificity of the strength of the CTCF–SA1 RAD21-independent interaction.

SMC3 did not contribute to co-IP of CTCF with SA1 at the lower benzonase condition, as shown in Figure 7C. To assess if this remained the case for the increased enrichment of CTCF observed in the 85U benzonase condition, membranes from this replicate were immunoblotted for SMC3 also. As observed in Figure 7C and Figure 14C, a fraction of SMC3 remained on chromatin in the absence of RAD21, although highly reduced compared to control. With the increase in benzonase, SMC3 could be detected in the auxin SA IP, however, it was not enriched over input nor was it enriched compared to co-IP in ethanol conditions. Therefore, this SMC3 was not likely to significantly contribute to the SA1-CTCF interaction observed. As in Figure 14C, SMC3 retained some interaction with CTCF in auxin conditions, however it was not enriched over input comparably with SA1. Hence, specificity of the SA1-CTCF cohesin-independent interaction was still observed with increased solubilisation of chromatin.

3.3 Discussion

Interaction of the SA proteins with CTCF has now been reported by multiple labs in multiple cell systems (Xiao, Wallace and Felsenfeld, 2011; Saldaña-Meyer *et al.*, 2014; Li *et al.*, 2020; Wutz *et al.*, 2020). However, these studies have primary used recombinant versions of the proteins and all but Wutz *et al.*, (2020) tagged either CTCF or the SA proteins. In this thesis, a protocol to detect interaction of cohesin subunits and their regulators under endogenous conditions was developed. It was determined that 40 x 10^6 cells (2x15cm dishes), 200mM salt IP buffer, and specific digestion of DNA and RNA was required to robustly IP the proteins in complex. Co-IP of cohesin subunits was more easily achieved than co-IP with CTCF, indicating that altered conditions can be used to tune the fidelity of the interactions pulled down.

Using HCT116 cells, interaction between endogenous, unlabelled CTCF and SA was confirmed. This is an important finding as it validates interaction data from recombinant studies in native conditions. CTCF-SA1 interaction was more easily detected than CTCF-SA2, which was barely detectable in initial experiments. A recently published paper, Wutz *et al.,* (2020), reported the same findings from

Hela cells. Wutz *et al.*, (2020) carried out chromatin IP in a similar IP buffer, using endogenous antibodies targeting SA1 and SA2 and reported much higher levels of CTCF interacting with SA1 than with SA2. They further quantified these interactions using label-free quantitative mass spectrometry and confirmed CTCF was significantly overrepresented in SA1 IP compared to SA2 IP. As IP is carried out overnight, these experiments suggest that CTCF-SA1 interaction is more stable than CTCF-SA2 interaction, at least under the conditions used. Investigation of the crystal structure of SA2 and RAD21 in complex with a fragment of the N-terminus of CTCF shows interaction via the 'Conserved Essential Surface' (CES) of SA2 (Li *et al.*, 2020). As this portion of SA2 is characterised for high conservation with SA1, the different strengths of CTCF co-IP was surprising and indicates that the interaction is mediated by more than just this region of SA.

RAD21 is tagged with AID and mClover in the HCT116 cell line used (Natsume *et al.*, 2016). The AID system allows rapid, proteasomal-mediated degradation of the tagged proteins simply by the addition of a plant hormone, auxin, to the cell media. Hence, contribution of the cohesin ring to CTCF-SA interaction could be investigated with acute depletion of RAD21 and without the need to effect cell health or cell cycle dynamics with transfection of siRNAs. Surprisingly, rather than abolishing interaction, co-IP of CTCF and SA was sustained, and in fact co-IP of CTCF with SA1 was enriched compared to the interaction seen in the presence of RAD21. This enrichment suggests that upon loss of interaction with the cohesin ring, the fraction of SA1 in complex with CTCF in the cells is increased. These experiments further show that the amino acids of RAD21 included in the crystallisation paper discussed above are not required for the interaction of CTCF and SA in endogenous conditions.

Following auxin-treatment SA and SMC3 remained bound to chromatin. It is not clear if the signal observed represents stably bound fractions of these proteins that can maintain interaction with chromatin following the loss of RAD21 or newly associated molecules that interact with chromatin in the absence of RAD21. Multiple studies indicate that in mammalian cells cohesin dissociates from chromatin with a half-life of ~13-25 mins (Gerlich *et al.*, 2006; Hansen *et al.*, 2017; Holzmann *et al.*, 2019). Hence, the majority of cohesin molecules in the cell

population should be lost from chromatin within the 4 hr auxin treatment timeframe. A more stable residence time of ~6-8 hrs has also been recorded for cohesin in cells in G2 phase (Gerlich *et al.*, 2006; Holzmann *et al.*, 2019). Thus, G2 cells in the population may account for the SA and SMC3 signal observed following RAD21 loss. Cell synchronisation and auxin treatment specifically in G1 and G2 cell populations would help to distinguish if this were the case.

Stabilised binding of cohesin during G2 is disrupted by proteolytic cleavage of RAD21 by seperase, therefore, it is unlikely that this stabilisation could withstand complete degradation of RAD21 (Uhlmann, Lottspelch and Nasmyth, 1999; Hauf, Waizenegger and Peters, 2001). While stabilised binding of cohesin is likely disrupted, different effects of RAD21 cleaveage on cohesin subunits have been reported in the literature. In human cells, RAD21 and SA2 dissociate from chromatin with similar kinetics, as shown by IF analysis of RAD21 and SA2 distribution during the cell cycle (Prieto *et al.*, 2002). In yeast cells, Scc1 is naturally absent during early G1 due to proteolytic cleave by sepearase (Ciosk *et al.*, 2000). Scc3 levels on chromatin are significantly reduced at this timepoint, suggesting that in yeast cells Scc3 requires Scc1 to interact with chromatin. This dependency has been observed in additional studies in yeast (Tóth *et al.*, 1999). Hence, RAD21 degredation is thought to trigger release of SA proteins from chromatin. This highlights the novelty of the SA behaviour uncovered in this thesis and suggests that i) SA proteins have evolved this ability in human cells and ii) SA may be reassociating with chromatin and CTCF after RAD21 depletion.

In contrast to the Scc1-Scc3 interaction described above, Smc1 can still be observed on chromatin in early G1 in the absence of Scc1 in yeast cells (Ciosk *et al.*, 2000). This suggest that interaction with chromatin in the absence of Rad21 is a conserved ability of the Smc proteins. In yeast cells, the Smc1 observed on chromatin in the absence of Scc1 was reduced in an *scc2-4* mutant, suggesting that the loader complex may influence this chromatin association. In addition, select reports of retained DNA interaction for the Smc3 protein until late anaphase have been made in yeast and the parasite *Trypanosoma brucei* (Tanaka *et al.*, 1999; Bessat and Ersfeld, 2009). In *T. brucei*, this Smc3 signal was assessed by IF and found to be detergent-sensitive, indicating that Smc3 may only interact with DNA weakly (Bessat and Ersfeld, 2009). This again validates the idea that

the SMC and SA proteins obseved on chromatin in the absence of RAD21 represent newly associating proteins. The chromatin-bound proteins assessed in this thesis were extracted in high salt conditions (500 mM) indicating that they can also interact with DNA more strongly in cells.

Bessat and Ersfeld (2009) suggest that the Smc3 signal they observed represents a soluble pool of Smc3 that is ready to reassociate with chromatin upon interaction with newly translated Scc1. The interaction of SA and CTCF observed here in the absence of RAD21 may thus help to rapidly target reassembled cohesin to CTCF-bound sites in the genome. Such a mechanism could help to ensure fast contacts between CTCF-bound sites, however, this does not correspond with the loop extrusion model of loop formation which postulates that cohesin loads at distinct sites from CTCF and extrudes a loop until it encounters convergent CTCF barrier elements (Alipour and Marko, 2012; Sanborn *et al.*, 2015; Fudenberg *et al.*, 2016b). A newly emerging idea suggests that cohesin-SA1 may be involved in the formation of chromatin condensates via intrinsically disordered regions in SA1 and multivalent interaction in cohesin (Weitzer, Lehane and Uhlmann, 2003; Davidson *et al.*, 2019; Pežić *et al.*, 2021; Ryu *et al.*, 2021). This raises the intriguing question of whether SA1 can form the seed sites for these condensates in the absence of cohesin (discussed further in Chapters 4 and 5).

To determine if SMC3 contributed to the CTCF-SA interaction observed in the absence of RAD21, a number of the CTCF and SA IPs were also immunoblotted for SMC3. Like SA, SMC3 could be detected in CTCF IPs, however, it was not enriched over input as robustly as SA1. SMC3 detection in SA IPs was variable and when observed it was at much lower levels than in the presence of RAD21, indicating that although SA and SMC3 proteins could still bind to DNA, it was not likely they were all doing so in complex. Given the contrast to CTCF enrichment in SA1 IPs and the fact that CTCF co-IP was observed even in the absence of SMC3 co-IP, SMC3 could not account for CTCF-SA interaction observed. What role SMC3 plays on chromatin in the absence of RAD21 and SA interactions remains to be established.

CTCF-SA1 interaction in the absence of the cohesin ring is an important, novel finding as it reveals new insight into the function of SA1. As discussed in section 1.4.4, such cohesin-independence has been previously described during S-phase at SA1-mediated cohesion of telomeres. Here interaction of SA1 with DNA, via its AT-hook, and the protein TRF1 are required for faithful cohesion, while the cohesin ring proteins are not required (Bisht, Daniloski and Smith, 2013; Lin *et al.*, 2016). Discovery of further cohesin-independent activity with CTCF indicates that this function is not confined to S-phase and likely represents a fundamental mechanism of SA1 activity.

Detection of protein interactions by co-IP allows detection of interactions in their native confirmation in unmodified conditions however, it may fail to detect low-affinity or short-lived interactions. An alternative technique to investigate overlap of proteins on chromatin is chromatin immunoprecipitation followed by sequencing (ChIP-seq). In ChIP-seq experiments, proteins are crosslinked to the DNA, thereby stabilising and holding in place transient interactions at a snapshot in time. As a second, distinct investigation of CTCF and SA interaction, ChIP-seq of CTCF, SA, and cohesin in ethanol- and auxin-treated cells was carried out. This experiment allowed assessment of the distribution of CTCF and SA (+/- cohesin) across the genome and determination of the levels of colocalisation of the proteins, at a population level.

Although co-IP of CTCF with SA2 was much weaker than with SA1, ChIP-seq analysis detected both SA1 and SA2 at sites bound by CTCF. Proteins are fixed onto the DNA during the ChIP-seq protocol, potentially allowing detection of SA2 and CTCF together. Alternatively, SA2 may localise to the same position as CTCF, but not interact directly enough to be detected by co-IP. Moreover, the population level nature of ChIP-seq experiments means any colocalised signal may not come from exactly the same cells, but just shows that across the population a given position is bound by both proteins. SA2 was enriched to more CTCF sites than SA1. This was unexpected given the relative strength of CTCF co-IP with SA1 compared to SA2. Issues with sequencing depth in the SA1 ChIP-seq samples and different antibody efficiencies for ChIP-seq may account for this discrepancy. Importantly, with auxin treatment both SA1 and SA2 remained at CTCF-bound sites, while SMC3 and Rad1 were ablated to control IgG levels.

Hence, the ChIP-seq analysis suggested that SA2 may also interact with CTCF, but in a manner that could not be detected by the co-IP conditions used.

Alteration of salt and sonication conditions did not result in efficient co-IP of CTCF-SA2. In conjunction with this work, Yang Li, a postdoctoral researcher in the lab tested for co-IP of CTCF with YFP-tagged SA2 over a short, 2 hr immunoprecipitation using the efficient nanobody-based GFP-Trap system. The efficiency of CTCF co-IP over input was not increased with the shorted incubation time. While the different IP systems used makes comparison imperfect, together these experiments suggested that the reduced co-IP was representative of a biological variable rather than a technical variable of the IP. In fact, further adaptation of nucleic acid digestion by increasing benzonase added prior to IP increased co-IP between CTCF and SA1 and SA2. Isolation of chromatin under conditions producing DNA fragments ~35-1000 bp and RNA ~25-200bp was most conductive to efficiently capture both interactions. The fact that further digestion with even higher amounts of benzonase did not increase co-IP even more suggests that this is not simply a matter of solubilising more proteins but instead that interaction between CTCF and SA is dependent on a specific nucleic acid landscape. As CTCF, SA1, and SA2 have all been described to interact with DNA under specific conditions, such a dependency may make sense. One alternative explanation is that different amounts of benzonase penetrate chromatin structures differently, as tightly packed chromatin may exclude the enzyme. In such a case, the 85U benzonase condition may optimally penetrate to the level of chromatin structure mediated by CTCF-SA. Structural experiments under the different benzonase conditions would be required to investigate such a mechanism.

Dependence of the interaction with CTCF on a conserved, basic domain in the C-terminus of SA1 and SA2 that is more frequently spliced out of SA2 than SA1 was also explored. Immunoprecipitation followed by mass spectrometry was used to test the banding pattern of SA1 and SA2 on an SDS-PAGE gel for the presence of the alternatively spliced variants. Initial experiments from two SA1 IPs indicated that alternative splice isoforms for this exon were present in HCT116 cells and were represented by distinct bands on a western blot. The canonical SA1 band contained exon 31, whereas a secondary SA1 band

contained exon 31+ and exon 31- SA1 peptides. Smearing of peptides in the gel is possible, especially from higher to lower molecular weights. Given the abundance difference between the canonical and secondary SA1 bands and the sensitivity of mass spectrometry, it is possible that the exon 31+ peptides are contaminants from the canonical band. As the secondary band was not co-IP'd with CTCF, this suggests that the presence of this exon in SA1 plays a contributing role in its efficient interaction with CTCF. Further investigation with tagged version of the exon 31+/- variants of SA1 would be required to confirm this definitively.

While the alternative splice isoforms of SA2 could also be detected in HCT116 cells, separation of exon32+/- variants into distinct bands on the western blot was not observed. This suggests that the two SA2 bands observed by western blot represent a different variation in SA2, perhaps post-translational modification. Lack of separation of the SA2 exon 32 variants means that characterisation of the effect of the exon on interaction with CTCF is not possible by this method. In addition, no peptides distinguishing either SA1 or SA2 variants could be detected in mass spec analysis of a CTCF IP. If the peptides detected represent the population of either SA1 or SA2 variants interacting with CTCF, such a result would suggest that there is diversity in this region, as no consensus exon splicing was present at levels high enough to detect. Dr. Yang Li cloned and overexpressed the different SA2 variants and observed no effect on CTCF co-IP between the two, suggesting that this exon is not sufficient for regulation of the interaction. As discussed in the introduction section 1.4.3, conflicting reports have been published regarding domains in CTCF that are required for interaction with the SA proteins. Hence, it is likely that multiple interaction domains within CTCF and the SA proteins mediate their interaction and diversification of these domains in SA1 and SA2 may allow varied dynamics/stability of interaction. Given the dependence of CTCF-SA interaction on nucleic acids, varied nucleic acid binding domains could allow further regulation of the interaction. Importantly, co-IP in the absence of RAD21 determined that contribution of RAD21 amino acids is not required to mediate interaction of whole, endogenous versions of the proteins.

The BiFC-ChIP method validated in this chapter represents a good candidate for identification of transient CTCF–SA2 interactions. In a positive BiFC assay,

formation of the complex begins with interaction between the two proteins of interest, however, following fluorophore reconstitution, a stable fluorescent protein is formed that will then hold the complex together irreversibly. This process is advantageous for the detection of transient or weak interactions and hence may reveal new insight into CTCF-SA1 and CTCF-SA2 complexes. Further, the ChromoTek GFPTrap® is a nano-trap that utilises an Alpaca GFP-binding single variable domain antibody (also termed a nanobody). Nanobodies show high specificity for their epitope and can be produced in batch, alleviating the specificity and batch variability complications of polyclonal antibodies (Duc *et al.*, 2012). Thus, use of the GFP-Trap should mitigate any discrepancies in results that arise from antibody-based variable ChIP efficiencies.

The background Venus reconstitution observed in the BiFC-ChIP experiments are likely a consequence of the transient transfection method used and overexpression of the tagged proteins. For a proper BiFC-ChIP-seq investigation, a stable cell line expressing inducible, siRNA-resistant versions of the proteins of interest should help to alleviate background Venus reconstitution and detrimental effects on cell viability from permanent formation of the Venus protein. Unfortunately, only an average of the complexes formed in the time between expression and complex formation will be captured, meaning that real time detection of rapid interactions and the dynamics between competing interactions may be lost (Hu, Chinenov and Kerppola, 2002). Overall, BiFC-ChIP represents a promising method to investigate CTCF-SA interaction further, and while it could be used to determine the average complex differences in the presence or absence of the cohesin ring, it could not be used for investigation of the real-time effect on the dynamics of SA interactions.

It has previously been shown that the yeast orthologs of SA and RAD21 (Scc3 and Scc1, respectively) together show sequence independent interaction with dsDNA fragments of 32 bp and disruption of this interaction reduced cohesin levels on chromatin by 40% despite proper assembly of the cohesin complex (Li *et al.*, 2018). As loss of the basic patch in the C-terminus of SA1 appears to reduce interaction with CTCF and basic regions of proteins are important for interaction with nucleic acids, and the SA proteins have structural similarity with NIPBL, we hypothesis that SA1 may first interact with CTCF before recruitment

of cohesin and loading onto the DNA (Chapter 5). First, interaction of SA1 with other chromatin-bound proteins in the absence of the cohesin ring was explored to determine if this represents a fundamental mechanism of its behaviour during interphase that extended beyond CTCF (Chapter 4).

4

SA1 interacts with a wide variety of proteins and with nucleic acids independently of RAD21

4.1 Introduction

Most biological functions are carried out by the combined actions of multiple proteins; single proteins and multiprotein complexes work together to form molecular machines or pathways that execute complex processes. As discussed in the introduction (Chapter 1), a number of cohesin regulators have been identified that determine the association and stability of cohesin on chromatin. However, non-canonical cohesin regulators have also been described that are better known for their primary biological functions. For example, the nucleosome remodellor protein SNF2H/SMARCA5 interacts with cohesin and induces cohesin occupancy at Alu-rich sites (Hakimi *et al.*, 2002). Muñoz *et al.*, (2019) further determined that in yeast cohesin and its loader complex interact with the chromatin remodelling complex known as RSC, and this interaction is required for cohesin loading. Similarly, the MCM2-7 replisome complex members guide cohesin loading during S-phase (discussed in more detail in section 1.5.5; Zheng *et al.*, 2018). Finally, in fission yeast Swi6 has been shown to direct cohesin enrichment at heterochromatic regions via interaction with the SA ortholog Psc3 (Nonaka *et al.*, 2002). Hence, cohesin association with chromatin can be coordinated by more than just its canonical regulators and there is evidence that NIPBL and SA proteins mediate interaction with the non-canonical regulators.

As discussed in section 3.1, tandem mass spectrometry allows identification of proteins based on mass determination and sequencing. Immunoprecipitation

followed by mass spectrometry (IP-MS) can be used to identify the interactome of a protein in an unbiased manner. By a similar technique, the RAD21 interactome has been reported from Hela cells (Panigrahi *et al.*, 2012). Three groups of proteins were enriched, namely, cohesin complex and regulator proteins, ubiquitin–proteasome pathway proteins, and replication proteins. Various proteins spanning a range of biological processes were also identified, including transcription regulation, RNA processing, and DNA-damage response. A systematic investigation of cohesin protein interactomes has also been reported from HCT116 cells (Kim *et al.*, 2019). Here the authors tagged SMC1A, SMC3, RAD21, SA1, SA2, WAPL, PDS5A, PDS5B, sororin, NIPBL, and MAU2 and subjected each protein to IP-MS. The authors report proteins identified in ≥ 4 of the samples as the cohesin interactome, which was highly enriched for splicing factors and RNA-binding proteins. In this chapter, the interactome of SA1 was assessed using mass spectrometry to identify potential non-canonical regulators of cohesin and to determine the full range of proteins that SA1 can interact with in the absence of RAD21. Specificity of these interactions was confirmed with SA1 knockdown and dependence on RAD21 was assess by treatment with auxin.

As discussed in the introduction section 1.4.3, putative cohesin interactors can be identified by an FGF-type motif that supports binding to the CES of SA in complex with RAD21 (Li *et al.*, 2020). Proteins identified by this study in yeast overlapped with the SA1 interactome and were validated for interaction in the presence or absence of RAD21. Finally, given the importance of nucleic acid digestion for efficient co-IP of CTCF with SA1 (section 3.2.7) and the abundance of nucleic acid binding proteins identified in the SA1 interactome, interaction of SA1 with canonical and non-canonical RNA structures was probed.

4.2 Results

4.2.1 Banded mass spectrometry reveals a snapshot of SA1, SA2, and CTCF interactomes

Despite the fact that the mass spec experiments in section 3.2.6 were performed to specifically assess isoform peptides and not to identify interacting proteins, aside from SA1, 1252 proteins were detected in SA1 MS1 and 1935 proteins were detected in SA1 MS2 (referring to experiments from Table 6 and Figure 33). With the caveat in mind that only proteins the same size as the cut bands would be detected, these proteins represent a snapshot of the potential SA1 interactome. These experiments were not originally run to investigate SA interactors and so no controls to differentiate true interactors from sticky proteins were included. Panther protein class overrepresentation analysis was used to characterise this preliminary interactome.

Despite altered cutting of bands between the two SA1 IP-MS experiments, similar protein classes were enriched in the replicates MS1 and MS2 (Figure 41). As expected, an abundance of chromatin-binding and -regulatory proteins were detected, however, many RNA binding proteins were also co-purified. Gene-specific transcriptional factors represented a large proportion of the overrepresented protein classes, followed by translation, cytoskeletal and DNA binding protein classes.

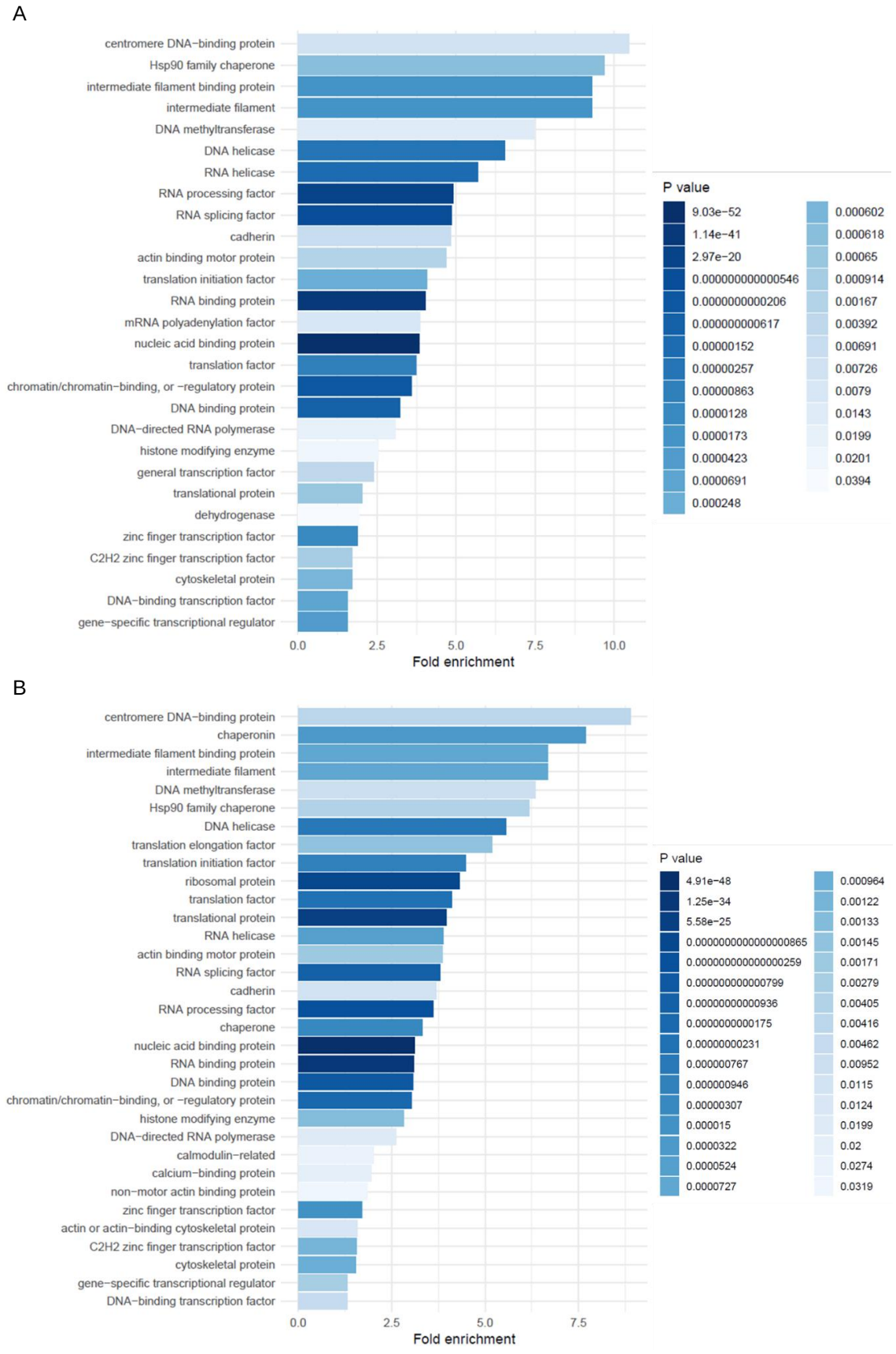
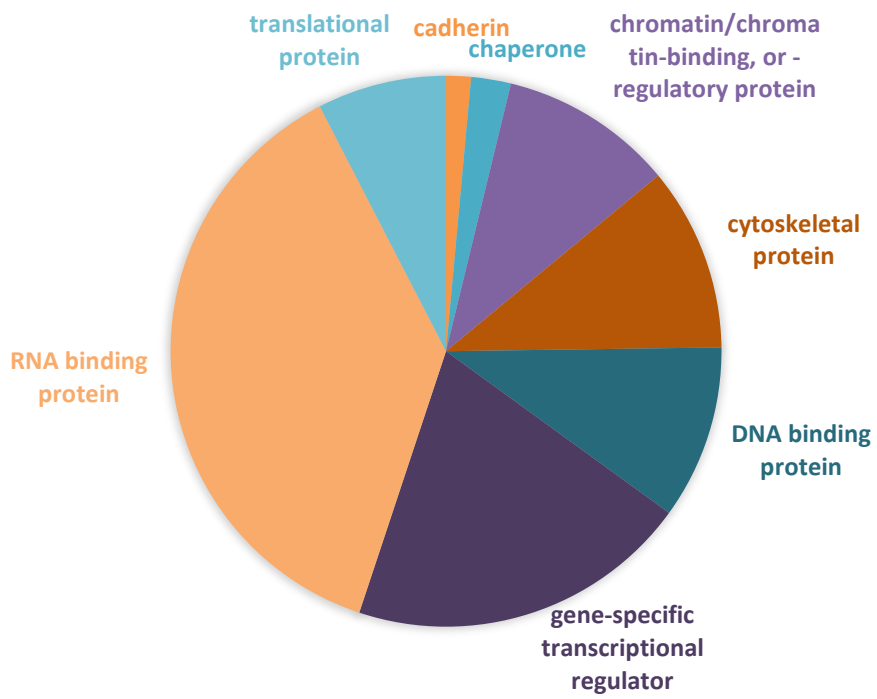


Figure 41: Panther protein classes enriched in SA1 banded MS1 and MS2. Bar graph of enrichment of Panther protein classes overrepresented in the SA1 IP MS1 (A) and MS2 (B) IPs. Overrepresentation was calculated using Fisher's Exact test with Bonferroni correction for multiple testing. P-values for each protein class are coloured based on the blue heatmap shown in the legend.

To analyse the reproducibly enriched proteins, the same analysis was carried out on proteins that were detected in both SA1 MS1 and MS2. Panther overrepresentation analysis of protein classes in the joint list determined that RNA binding proteins accounted for 37% of the overrepresented proteins (Figure 42A). Within the RNA binding protein class, RNA helicases, RNA processing factors, RNA splicing factors, mRNA polyadenylation factors, DNA-directed RNA polymerase factors, and general transcription factors were also significantly overrepresented. 20% of the overrepresented proteins were classed as gene-specific transcriptional regulators, within which, DNA-binding, zinc finger, and C2H2 zinc finger transcriptional regulators were also statistically enriched. Chromatin-binding or -regulatory proteins, cytoskeletal proteins, DNA binding proteins, and translational proteins all accounted for ~10% of the overrepresented proteins. DNA methyltransferases and DNA helicases were overrepresented within the DNA binding group and translational elongation and initiation factors were overrepresented within the translational protein group. Cytoskeletal proteins showed the highest levels of enrichment, but the RNA binding protein groups showed the highest statistical significance (Figure 42B).

Similarly, aside from SA2, 937 proteins were detected in the SA2 IP bands and were analysed as a snapshot of the SA2 interactome (referring to experiment from Table 7 and Figure 33). 20% of the overrepresented proteins from the SA2 interactome were classed as metabolite interconversion enzymes, with oxidoreductase, peroxidase, ligase, acetyltransferase and dehydrogenase protein classes also overrepresented within this grouping (Figure 43A). Translational proteins were also strongly represented as 19% of the overrepresented proteins, including, overrepresentation of ribosomal proteins, translation factors, and translation initiation and elongation factors. Similar to SA1, RNA binding proteins were a major overrepresented protein class, including RNA processing factors, RNA splicing factors, and RNA helicases. In contrast to SA1, transcription factors were not overrepresented in the SA2 interactome, and in fact, were statistically underrepresented.

A



B

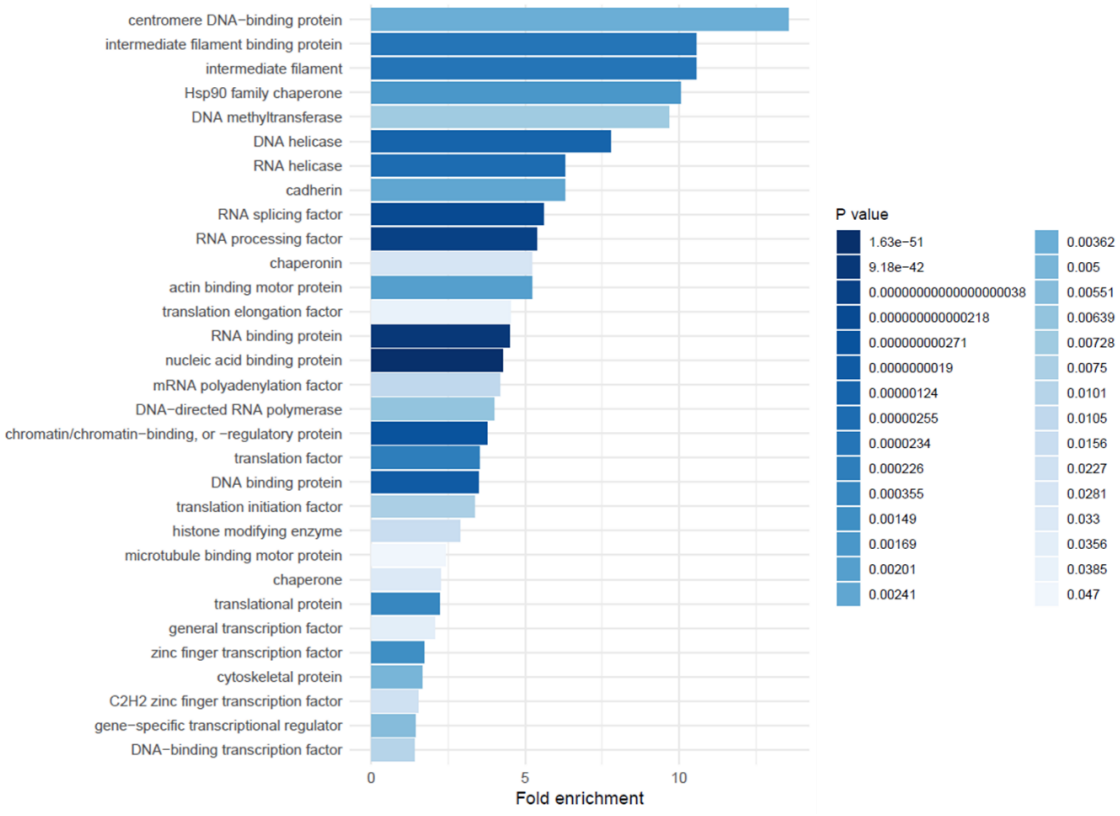
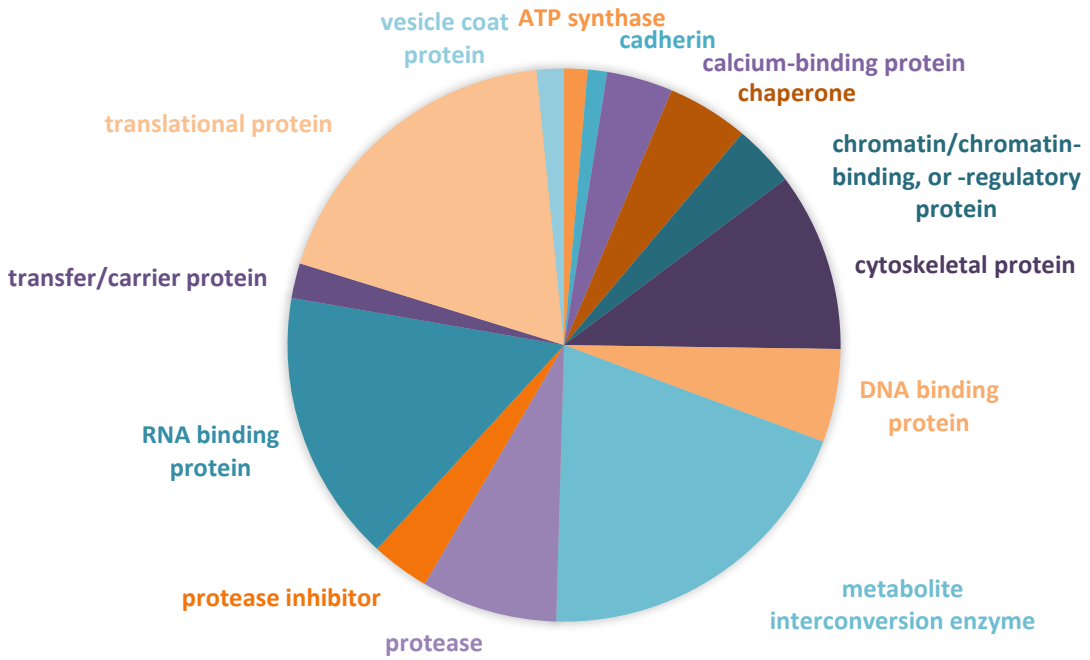


Figure 42: Panther protein classes enriched in overlap of SA1 banded MS1 & MS2. (A) Pie chart of proportion of proteins in non-redundant Panther protein classes statistically enriched in the overlap of SA1 banded MS1 & MS2. Overrepresentation of protein classes was calculated using Fisher's Exact test with Bonferroni correction for multiple testing. (B) Bar graph of enrichment of each protein class with p-value coloured based on the blue heatmap shown in the legend.

A



B

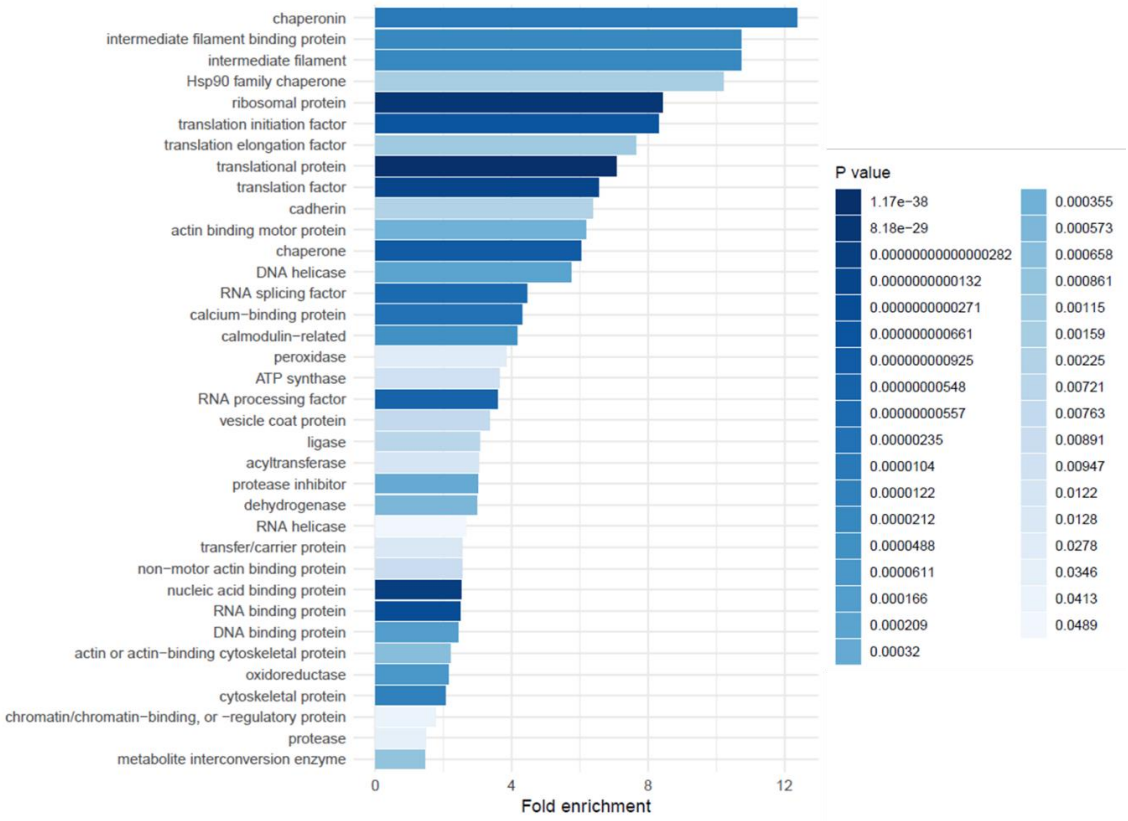
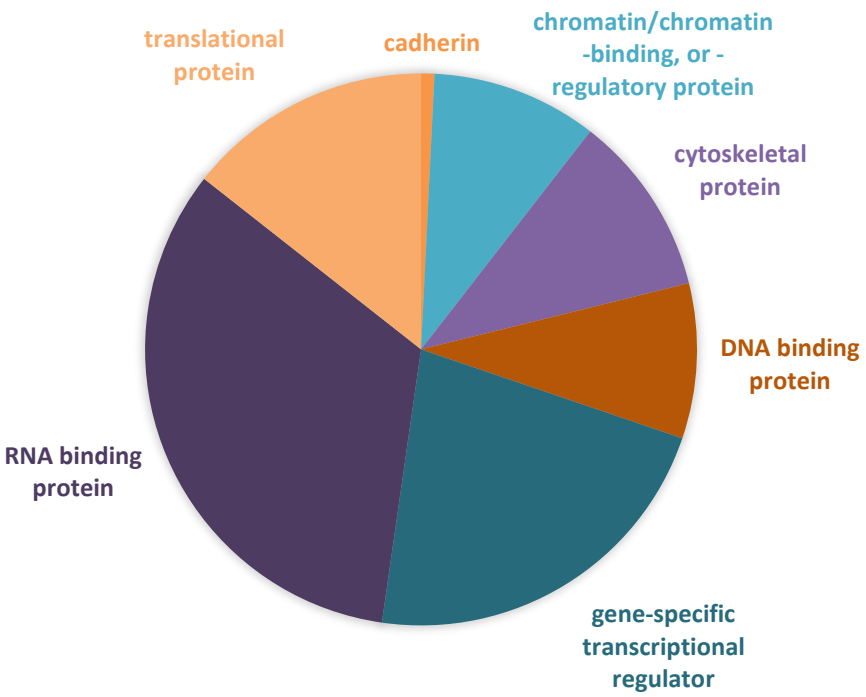


Figure 43: Panther protein classes enriched in SA2 banded MS. (A) Pie chart of proportion of proteins in non-redundant Panther protein classes statistically enriched in the SA2 banded MS. Overrepresentation of protein classes was calculated using Fisher's Exact test with Bonferroni correction for multiple testing. (B) Bar graph of enrichment of each protein class with p-value coloured based on the blue heatmap shown in the legend.

Fold enrichment over expected and Bonferroni corrected p value are shown for each of the protein classes overrepresented in the SA2 IP in Figure 43B. Ribosomal and translation protein groups were the most significantly enriched. This snapshot revealed potential similarities and differences between the SA1 and SA2 interactomes – chromatin-binding/regulatory, RNA binding, DNA binding, translation, cadherin, cytoskeletal and chaperone proteins were overrepresented in both SA1 and SA2 IPs, whereas transcriptional regulators were only identified in the SA1 IPs and a range of enzymes were only identified in the SA2 IP.

Finally, aside from CTCF, 2545 proteins were detected in the CTCF IP and were used to analyse the CTCF interactome (referring to experiment from Figure 34). From the CTCF IP, 33% of overrepresented proteins were RNA binding proteins, including RNA processing and splicing factors, RNA helicases, exoribonucleases, DNA-directed RNA polymerase, mRNA polyadenylation factor and general transcription factors (Figure 44A). Gene-specific transcriptional factors also accounted for a large percentage of the overrepresented proteins (22%), and included transcriptional cofactors, DNA-binding, zinc finger, and C2H2 zinc finger transcription factors. Translational, cytoskeletal, chromatin-binding and -regulatory, and DNA binding proteins were all ~10% of the overrepresented proteins. RNA binding proteins were most significantly enriched of all the classes (Figure 44B). Thus, all of the protein classes overrepresented in this potential CTCF interactome were also identified in the SA1 interactomes above. Whereas the SA2 interactome contained distinct classes not enriched with SA1 or CTCF, including metabolite interconversion enzymes, transfer, and protease proteins. These experiments revealed a view of the SA interactomes and suggest that SA1 and SA2 interact with many more proteins than currently thought. Whole lane mass spectrometry is thus required to determine the full interactomes of these proteins, whether the differences revealed were caused by the specific bands cut for each protein or by each protein binding to a distinct set of protein partners, and to control for co-IP of non-specific interactors

A



B

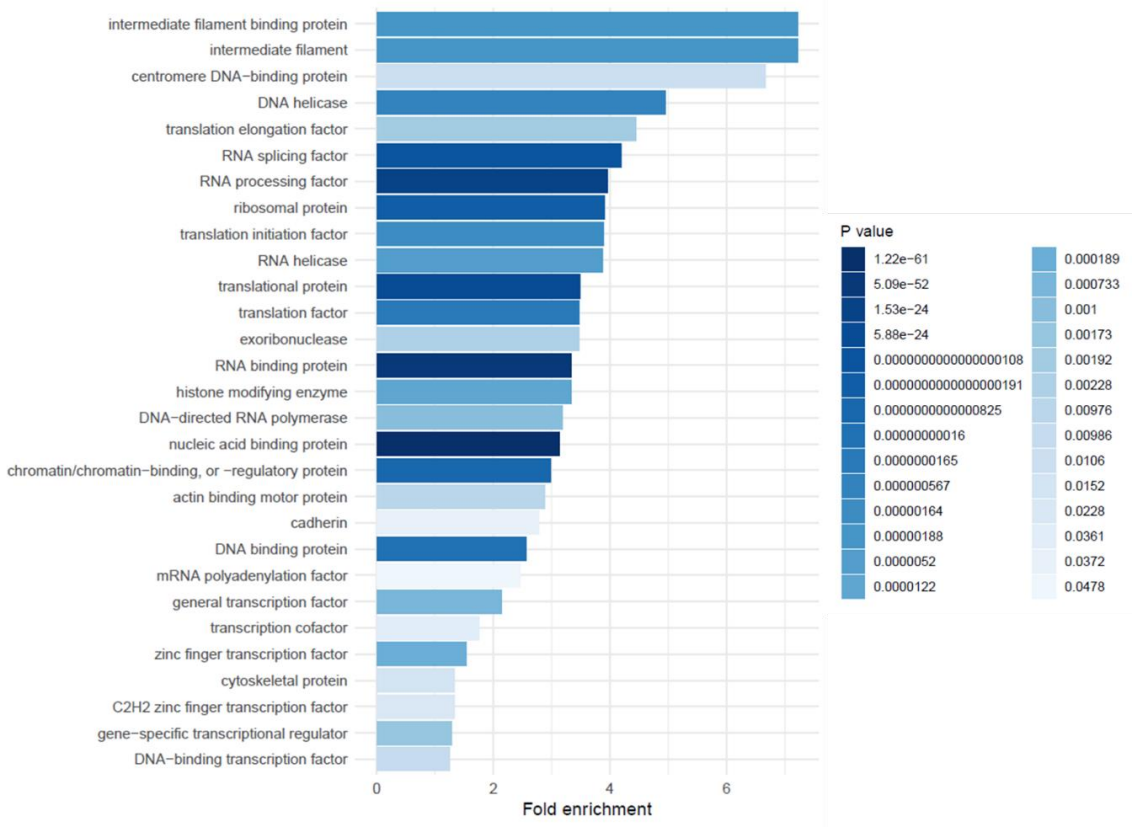


Figure 44: Panther protein classes enriched in CTCF banded MS. (A) Pie chart of proportion of proteins in non-redundant Panther protein classes statistically enriched in the CTCF banded MS. Overrepresentation of protein classes was calculated using Fisher's Exact test with Bonferroni correction for multiple testing. (B) Bar graph of enrichment of each protein class with p-value coloured based on the blue heatmap shown in the legend.

4.2.2 Full lane SA1 mass spectrometry

Following on from the results of the banded mass spectrometry experiments, full lane mass spectrometry experiments were carried out to determine the SA1 interactome. Similar to the banded IP-MS, chromatin material was fractionated from H2 cells, now using the increased benzonase condition. IPs were run using 15ug of SA1 antibody and 2mg of chromatin material and were eluted in 2x Lammeli sample buffer. The eluted material was run on Bio-Rad pre-cast gels, however, in this case, samples were run only until the loading dye had travelled 1cm through the gel to concentrate the proteins in the stacking portion of the gel. For each sample, the 1cm of gel was extracted and cut up into 1mm square portions. Once diced, peptides were isolated from the gel pieces by in-gel digestion with trypsin and analyzed by tandem mass spectrometry as previously. Peptide mapping was carried out by Amandeep Bhamra using MaxQuant software and MStats was used for statistical analysis of the final runs.

Five replicates were generated in total with each replicate containing 5 samples; a mock IP, an SA1 IP from untreated cells (UTR), an SA1 IP from cells treated with auxin for 4 hrs (IAA), an SA1 IP from cells treated with scrambled siRNA (siCon), and an SA1 IP from cells treated with siRNA targeting SA1 (siSA1). For the first three replicates an additional sample treated with RNase H was included to digest specific RNA structure, however, the amount of RNase H was too low and investigation of the influence of RNA structure on SA1 activity was left for separate experiments (see section 4.2.4 for this work). As the RNase H treatment was too low to effect RNA structures, these samples instead represented technical replicates of the UTR sample or were left out of the analysis. Mock IP samples were included to allow detection of sticky proteins that were pulled down despite not interacting with SA1. Auxin-treated samples were included to determine if any other proteins acted like CTCF, i.e. could interact with SA1 in the absence of the cohesin ring. siSA1 samples were included to allow detection of SA1 interactome specificity – as the level of any protein that truly interacts with SA1 should be altered in this condition. However, transfecting cells with siRNA can change their behaviour and could alter SA1 levels and activity itself, so siCon samples were included to control for the siRNA transfection.

Initially just three replicates were generated, however, upon processing 10% of each sample as an initial test of the material, it was clear that many less proteins were detected in replicates 2 and 3 than in replicate 1 (Table 10 and Figure 45A). A profile plot of the log-transformed intensities of detected proteins for each sample was used to assess variation in the data (Figure 45A). Each grey dot represents a specific protein and if the protein is present across samples a connecting line is drawn. Cohesin members are highlighted in colour. From this profile plot it was apparent that many more proteins (grey dots) were detected in replicate 1 samples and proteins that were detected across samples were detected at lower levels in replicates 2 and 3. There was concern about cell growth and the appearance of unusual cell morphology in the cell culture at the time of processing of replicates 2 and 3, so it was decided to disregard these samples from further analysis and to generate two more biological replicates to give an overall n = 3. These two new replicates were termed replicates 4 and 5.

The presence and abundance of proteins in replicate 4 was similar to replicate 1, however, replicate 5 was not as efficient for protein detection (Table 10 and Figure 45B). To determine if new columns in the mass spectrometry machine or simply the injection of 2x the amount of replicate 5 material would alleviate the discrepancy in detection, another set of test injections were run after a check-up of the machine (Figure 45C). Replacement of columns in the mass spectrometer did not increase detection in replicate 5, however, loading 2x the amount of peptide extract did increase protein detection and levels. Hence, full injections were run with ~equal amounts of replicate 1 and 4, and 2x replicate 5. The RNase H-treated sample from replicate 1 was included as a technical replicate of the SA1 UTR replicate 1 sample, as the amount of RNase H added had negligible effect on RNA structures and SA1 interactions (henceforth referred to as UTR_Rep2).

Condition	Rep1	Rep2	Rep3	Rep4	Rep5
Mock	43	56	67	77	93
SA1 UTR	1370	630	606	1399	722
siCon	1398	889	868	1451	1188
siSa1	1473	1220	1055	1632	1110
IAA	1468	867	753	1502	1213
RNaseH	1561	661	567	-	-

Table 10: Count of proteins detected in the SA1 full lane mass spectrometry replicates. Total number of proteins detected per sample condition for each replicate experiment.

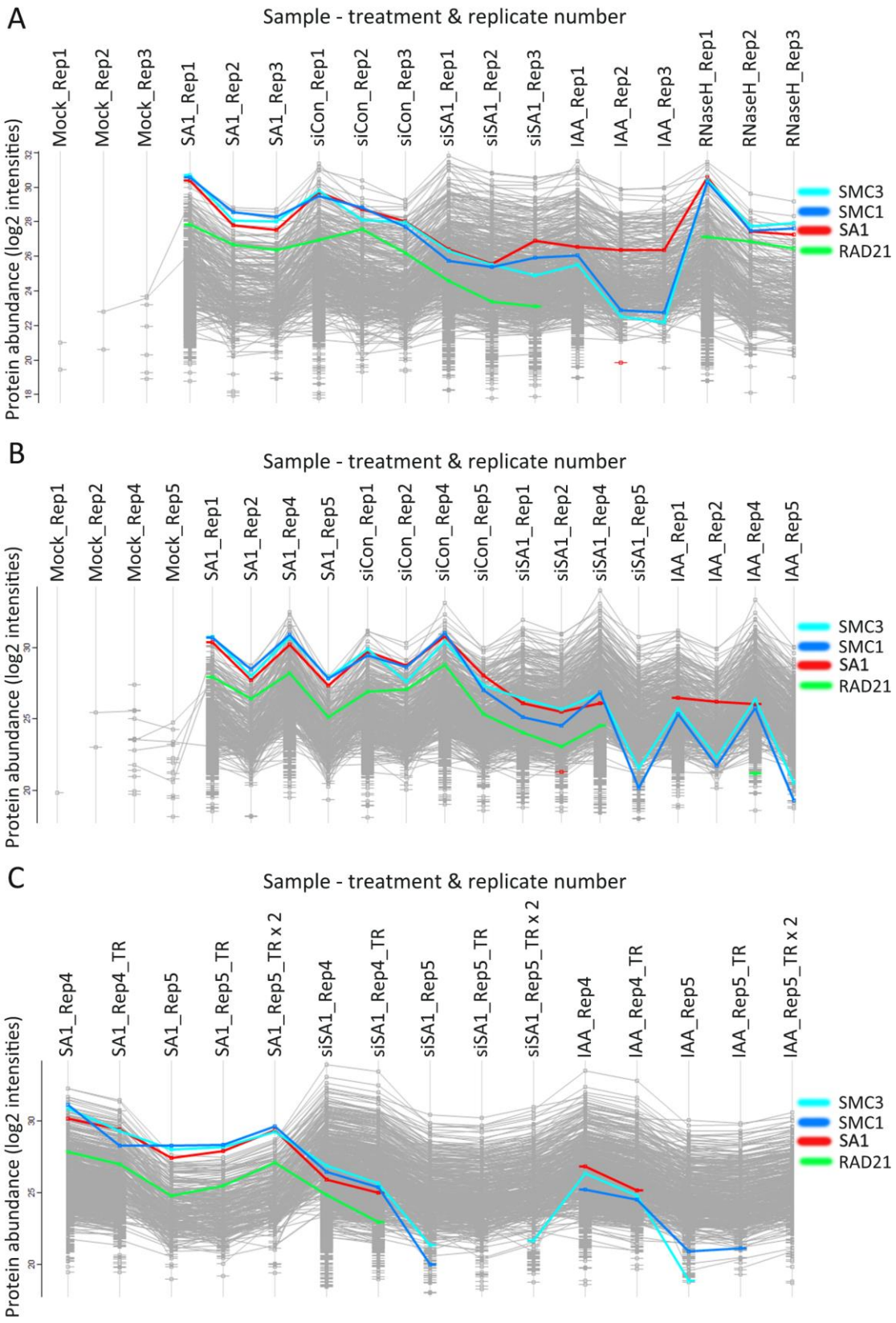


Figure 45: Quality control analysis of SA1 full lane mass spectrometry replicates. (A) – (C) Protein profile plots were generated in MaxQuant software by Amandeep Bhamra and show the abundance of proteins (represented by grey dots) across the samples and replicates (joined by grey line if present across samples). SMC3, SMC1, SA1, and RAD21 are highlighted in light blue, blue, red, and green, respectively. (A) Protein profile plot of Mock, UTR (SA1), siCon, siSA1, IAA, and RNase H samples for replicates 1, 2, and 3. (B) Protein profile plot of Mock, UTR (SA1), siCon, siSA1, and IAA samples for replicates 1, 2, 4, and 5. (C) Protein profile plot of UTR (SA1), siSA1, and IAA samples for replicate 4, a re-run of replicate 4, replicate 5, a re-run of replicate 5, and a second re-run of replicate 5 where double the concentration of total protein was run.

Peptide intensities were measured by the mass spectrometer and MStats software used linear modelling to log transform the data and calculate the fold change between treatments. Using the log transformed intensities as a measure of protein abundance, the coefficient of correlation was calculated between sample replicates and plotted as a heatmap using the corrplot function in R (Figure 46A). Correlation of replicates 1 and 4 was ≥ 0.61 for all conditions, however, comparison with replicate 5 showed reduced correlation. For IAA, siSA1, and UTR samples, correlation of replicates 1 and 4 with replicate 5 was between 0.43 – 0.49. Hierarchical clustering was used to group the samples according to their correlations and to reveal the relationships between samples (Figure 46B). The lower half of the correlation heatmap is shown, with clusters of similar samples seen as triangles of darker blue off the diagonal. The upper half of the heatmap shows the corresponding correlation plots from which the coefficient is calculated.

Four clusters of correlation were apparent in the heatmap; i) the three mock samples, which show very high correlation, ii) the replicate 5 samples, except for siCon which instead groups with replicates 1 and 4 samples, iii) the replicate 1 samples, including the RNase H-treated sample termed UTR_Rep2, and iv) the replicate 4 samples. As well as showing the most divergence between replicates, the lowest levels of correlation within the replicate clusters was measured for replicate 5, corroborating that these samples contained the most variation. In contrast, replicate 1 and replicate 4 samples show quite high correlation within themselves, despite samples being treated differently. This implied that across the samples, biological variance or variance in the SA1 IP was higher than variance caused by the different treatments. Such variance could cloud calculation of differences between treatments. Therefore, it was decided to calculate statistics using a paired analysis, in which, relationship between and not just across replicate conditions is assumed. Taking account of this relationship like so minimises the impact of the variance.

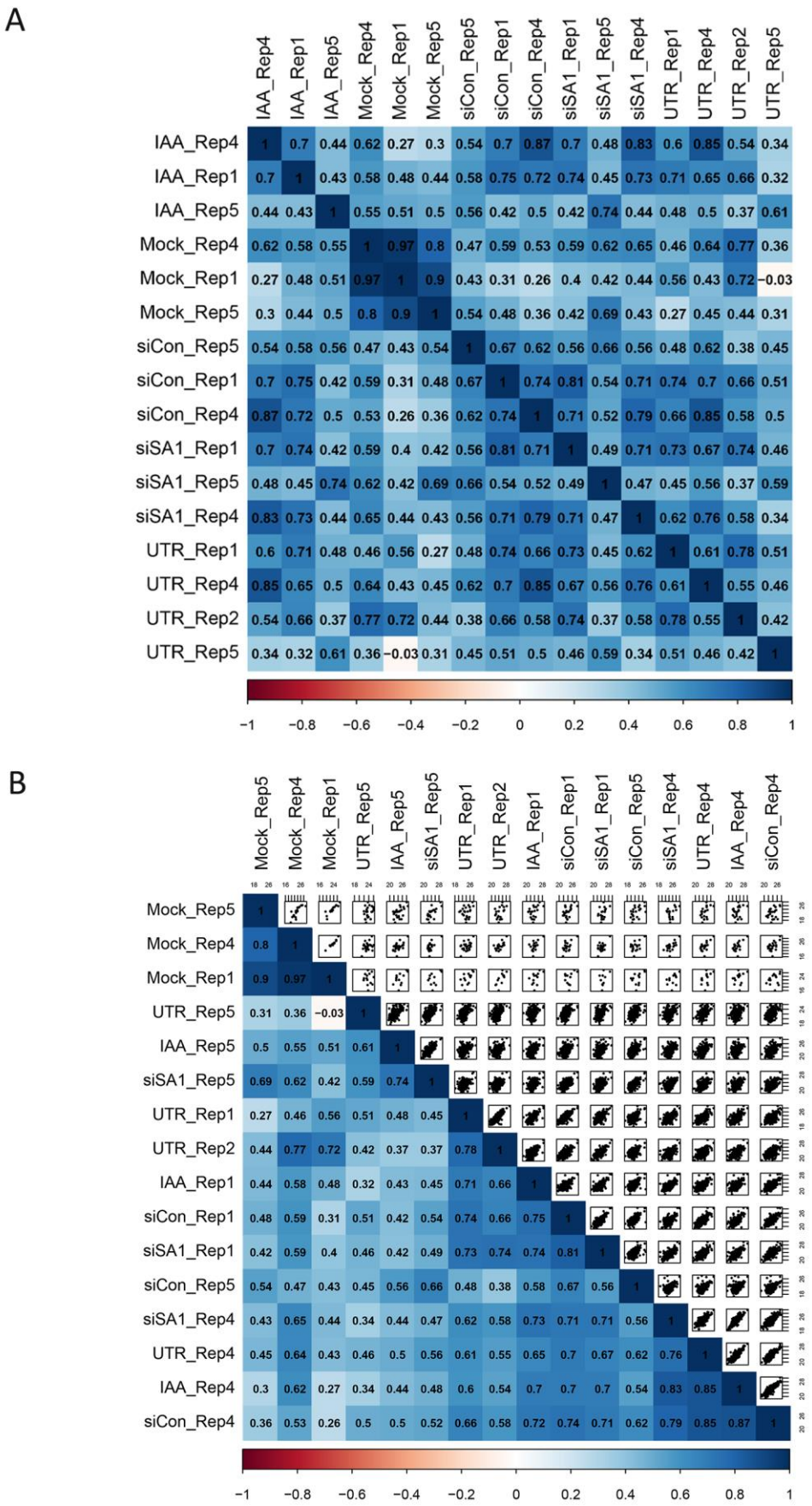


Figure 46: Correlation of SA1 IP-MS samples. (A) Heatmap of coefficient of correlation between samples replicates plotted using the corrplot function in R. (B) Hierarchical clustered version of (A) with the upper portion of the heatmap showing the underlying correlation plots from which the coefficient is calculated.

1327 proteins were identified across all the samples with a false discovery rate (FDR) of 1%. 44 of these proteins were identified in at least one of the mock IP samples and so were disregarded from downstream analysis. The remaining 1283 proteins represent the SA1 interactome. These 1283 proteins were predominantly identified across all of the different conditions (i.e. UTR, IAA, siCon, siSA1) but at differing quantities depending on the treatment. 1246 of the proteins had corresponding gene names and 1236 of these were recognised by Search tool for the retrieval of interacting genes/proteins (STRING). STRING was used to annotate protein interactions and functional enrichments in the SA1 interactome.

STRING computes protein interactions based on seven 'channels' of information; co-expression, experiments, databases, and textmining aggregate protein interaction data from a wide range of datasets and databases, while neighbourhood, fusion, and gene co-occurrence channels input predicted interactions (Szklarczyk *et al.*, 2019). The predicted interactions are calculated by comparing evolution events between genomes and searching for non-random association of data for genes across the events, indicating functional association. STRING can also be used to detect enriched functional categories within the protein network, with results from multiple classifications, including GO, KEGG, and InterPro. To adjust for multiple testing, Benjamini and Hochberg correction is applied within each classification group. To evaluate the proteins enriched in the SA1 interactome, a network of the 1236 proteins was generated and GO term enrichment calculated at a high confidence threshold (0.7). Cytoscape was used to curate a visual representation of the STRING network and enrichments for GO biological processes and GO molecular functions. To allow visualisation of the large network, proteins with the highest average quantification values were selected for each of the major enriched GO biological processes (Figure 47A). Log transformed adjusted p-values (FDR value) for each of the major processes are also shown (Figure 47B).

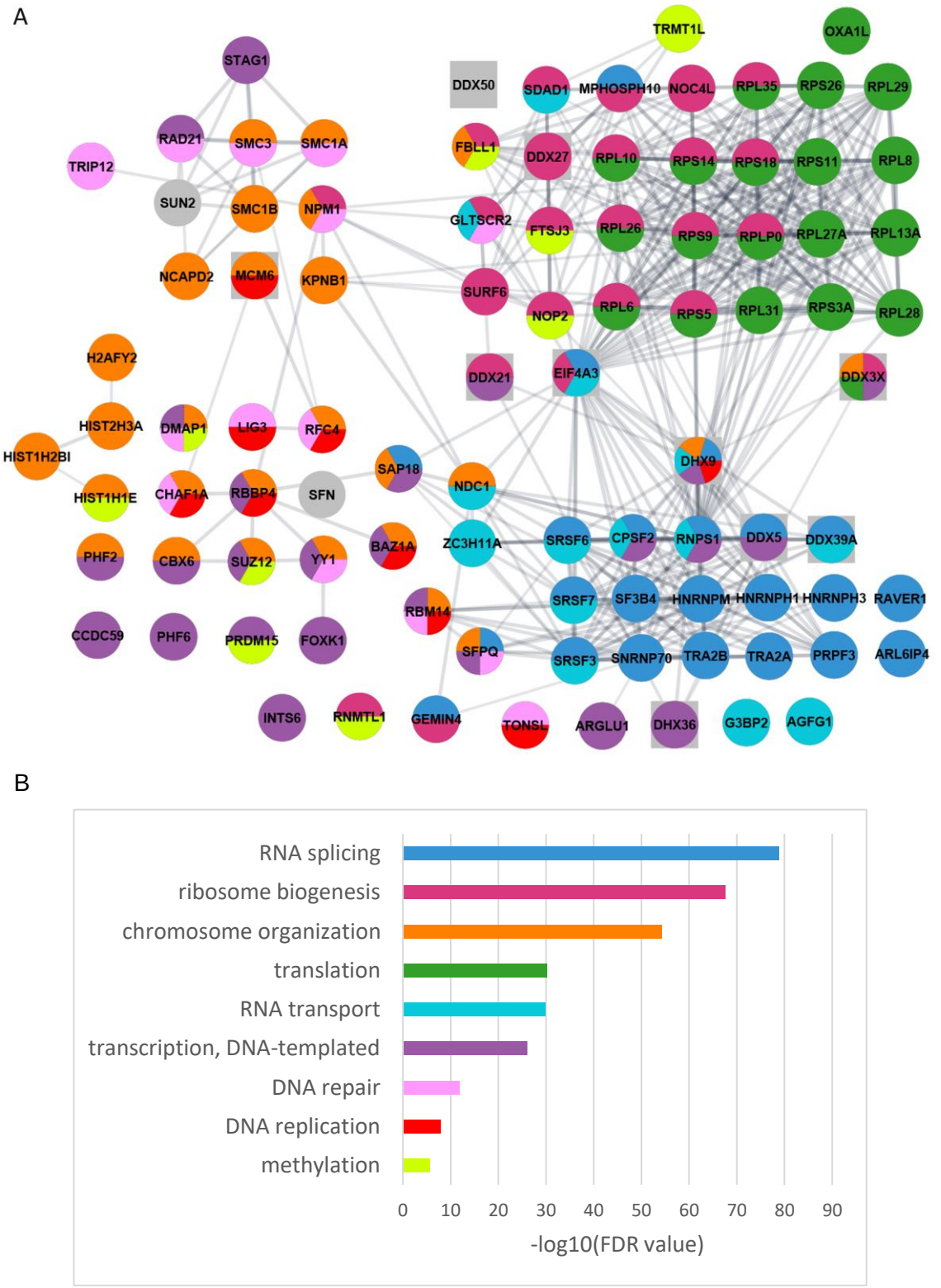


Figure 47: Subset view of the HCT116 SA1 interactome. Network of proteins identified in SA1 IP samples. Protein interactions and GO term enrichment analysis was generated using STRING. Node colours denote the major enriched categories (see (B)), with squares signifying helicase proteins. Proteins within each enriched category were subset from the full interactome network based on their average abundance, with the top 10-20 most abundant protein selected per category. The full network cannot be shown as >1000 proteins creates an unintelligible mass of proteins. (B) Bar chart of log10-transformed FDR values for each enriched GO biological process coloured in the network. Colours in the chart match those in the network. FDR values are calculated by STRING and represent p-values adjusted for multiple testing.

The most significant enrichments were gene expression and RNA processing (FDR= 7.28E-139 and 5.44E-136, respectively). In addition, ribosome biogenesis (2.96E-68), translation (7.22E-31), chromosome organization (FDR= 4.57E-30), transcription (7.48E-27), DNA repair (1.22E-12), and DNA replication (1.26E-08) processes were enriched. Within these enrichment groups, high abundance of condensin (NCAPD2), MCM2-7 complex (MCM6), and Polycomb group (CBX3, SUZ12, YY1) members were detected. Helicase, HNRN, RPL, and RPS proteins were also significantly enriched. Hence, SA1 was identified to interact with diverse proteins across the nucleoplasm and nucleolus.

MStats software was run by Amandeep Bhamra to carry out a paired analysis, for which linear modelling was used to compute the fold change and p-value between conditions for a merge of the replicate abundances. The abundance of 40 proteins was significantly changed between UTR and IAA samples (log2FC ≥ 1 and ≤ -1 and p-value of < 0.05). 6 out of the 40 proteins were reduced with IAA treatment (SMC3, STAG1, SMC1A, PDS5B, RPS9, and FMR1) representing cohesin-dependent SA1 interactors. The remaining 34 proteins show increased association with SA1 in the absence of the cohesin ring and were termed cohesin-independent interactors. With the same criteria of significant change, the abundance of 273 proteins was significantly changed between siCon and siSA1 samples. As the levels of these proteins changed with SA1 depletion they can be considered true interactors of SA1. 12 proteins overlapped between both comparisons and represent true SA1 interactors – two of which are cohesin-dependent, SMC3 and FMR1, and 9 of which are cohesin-independent (Table 11). Annotation using Panther and String determined that half of the 12 proteins were chromatin binding, three of which have roles in cell division and four of which are involved in response to DNA damage stimulus. 8 of the 12 proteins were classed as regulators of gene expression, with roles in the regulation of splicing and RNA metabolic processes also. Therefore, stringent filtering identified a small SA1 interactome with links to chromatin regulation, gene expression, DNA damage response, and RNA processing, and a large proportion of which interacts with SA1 independently of the cohesin ring.

Protein	Panther Protein Class	String Gene Ontology
SSRP1		Chromatin binding, cellular response to DNA damage stimulus, regulation of RNA metabolic process, nucleic acid metabolic process, regulation of gene expression
hCG_31253		
HM13	aspartic protease	
UBAP2		Regulation of gene expression
CBX2		Chromatin binding, regulation of RNA metabolic process, nucleic acid metabolic process, regulation of gene expression
ZNF326	scaffold/adaptor protein	Regulation of RNA splicing, regulation of RNA metabolic process, nucleic acid metabolic process, regulation of gene expression
ZMYM4	zinc finger transcription factor	Regulation of RNA metabolic process, regulation of gene expression
SMC3	chromatin/chromatin-binding, or -regulatory protein	Chromatin binding, cell division, cellular response to DNA damage stimulus, nucleic acid metabolic process
FTSJ3	RNA methyltransferase	Nucleic acid metabolic process
STAG1	chromatin/chromatin-binding, or -regulatory protein	Chromatin binding, cell division, regulation of RNA metabolic process, nucleic acid metabolic process, regulation of gene expression
AATF	chromatin/chromatin-binding, or -regulatory protein	Cell division, cellular response to DNA damage stimulus, regulation of RNA metabolic process, regulation of gene expression
FMR1	Translation factor	Chromatin binding, regulation of RNA splicing, cellular response to DNA damage stimulus, regulation of RNA metabolic process, nucleic acid metabolic process, regulation of gene expression

Table 11: Ontology of high confidence SA1 cohesin-independent interactors. Significantly depleted (grey) and enriched (white) proteins co-purified with SA1 (green) in IAA conditions compared to UTR conditions that also had a significantly altered abundunce in siSA1 conditions compared to siCon. Significance was considered as log2FC ≥ 1 and ≤ -1 and p-value of < 0.05. Panther protein class and STRING gene ontology annotations are shown for each of the proteins.

Using more lenient cut-off values (log2FC ≥ 0.581 and ≤ -0.58 and a p-value of 0.05), a wider view of the putative SA1 interactome can be viewed. With the wider parameters, 137 proteins were significantly changed between UTR and IAA samples. 134 of these proteins were identifiable by STRING and were used to

generate a network of the SA1 interactome in the absence of cohesin ($SA1^{\Delta Coh}$ interactome). As expected, members of the core cohesin complex, SMC3 and SMC1A, were depleted and RAD21 was degraded such that no peptides were detectable. In addition, known regulators of cohesin were depleted, including PDS5B, NuMA protein NUMA1, and FACT complex subunit SSRP1 (Kong *et al.*, 2009; Garcia-Luis *et al.*, 2019). In line with the enrichment observed for CTCF in IAA-conditions, the vast majority of the $SA1^{\Delta Coh}$ interactors were enriched for binding with SA1 in IAA conditions (117 of 136).

Gene Ontology (GO) enrichment analysis was used to categorise the functions of the proteins and colour nodes in the network. For visual clarity, half of the proteins within each major enriched GO biological process were subset from the network, based on p-value change between UTR and IAA samples (Figure 48A; full network shown in Supplemental Figure 4). Proteins involved in chromosome organisation, transcription, RNA processing, ribosome biogenesis, and translation retained interaction with SA1 in the absence of RAD21. Comparison with the whole genome determined that these biological processes were enriched in the $SA1^{\Delta Coh}$ interactome compared to the background of the cells (Figure 48B). Similar to the CTCF ChIP results, this suggests that SA1 maintains interaction with proteins that it localizes with in the presence of cohesin, albeit at different abundances. Comparison to the UTR SA1 interactome discussed above determined that there was a significant increase in proteins involved in RNA processing (FDR=0.0298), ribosome biogenesis (FDR=0.0197), ribonucleoprotein complex biogenesis (FDR=0.0298) and rRNA processing (FDR=0.0409) with IAA treatment (Figure 48, A - dotted lines & C). This indicates that a larger fraction of the SA1 present binds to these proteins in the absence of cohesin.

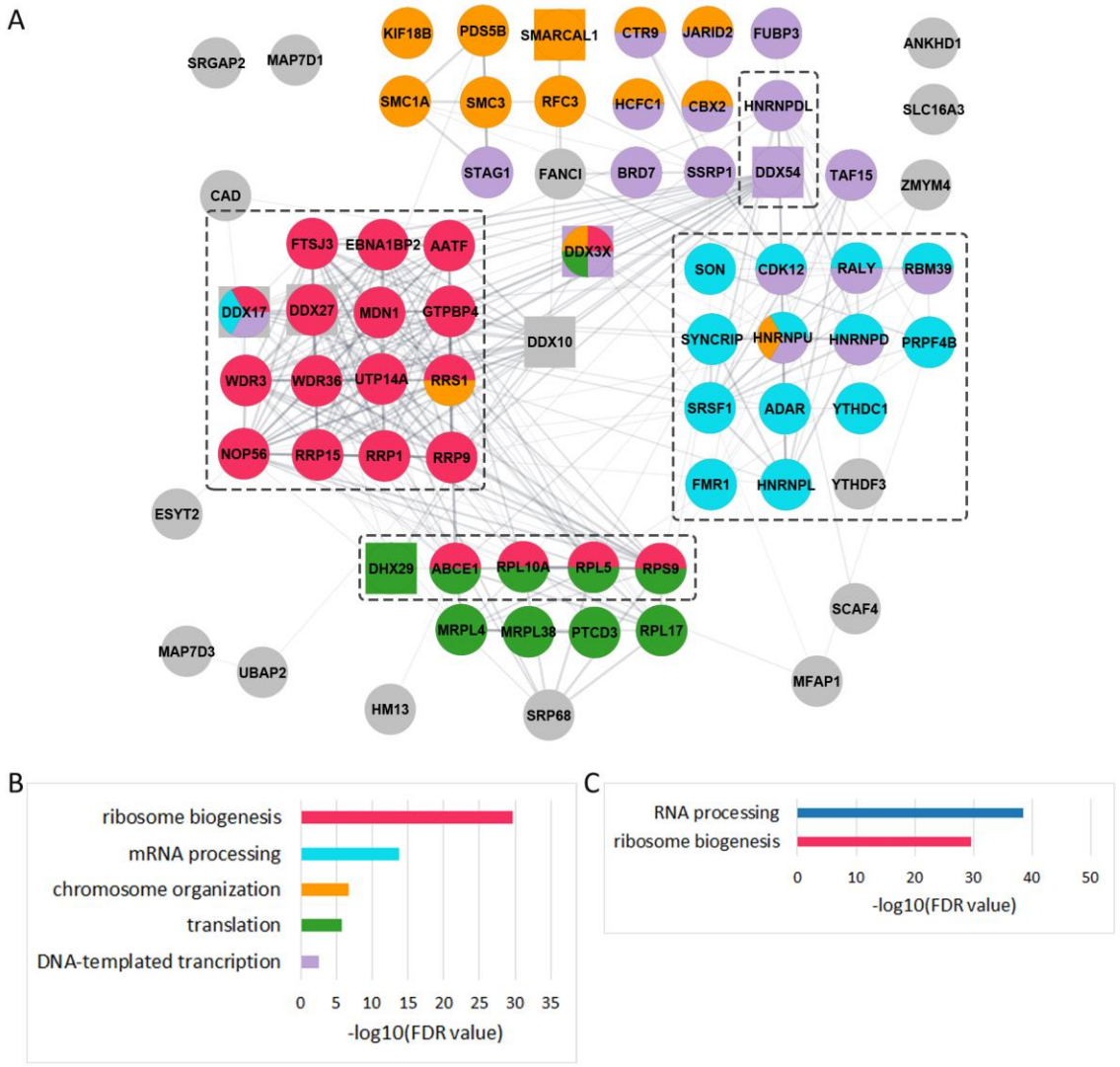


Figure 48: Subset view of SA1$^{\Delta Coh}$ interactome. Network of proteins co-purified with SA1 and with altered abundance in IAA conditions compared to UTR conditions. Proteins were considered to have altered abundance for log2FC ≥ 0.581 and ≤ -0.58 and a pvalue of < 0.05. Protein interactions and GO term enrichment analysis was generated using STRING. Node colours denote the major enriched categories compare to whole genome background (see (B)), with square nodes signifying helicase proteins. Dotted lines encompass the processes within the network that are enriched within the SA1 interactome itself with IAA treatment, compared to the UTR SA1 interactome. Half the proteins within each enrichment category were subset in the network based on p-value change between IAA and UTR samples. See Supplemental Figure 4 for full network. (B) Bar chart of log10-transformed FDR values for the enriched GO biological processes coloured in the network. Colours in the chart match those in the network. FDR values are calculated by STRING and represent p-values adjusted for multiple testing. (C) Bar chart of log10-transformed FDR values for GO biological processes enriched within the SA1 interactome following treatment with IAA.

As in the UTR SA1 interactome, RNA processing was one of the most enriched biological processes in the SA1 IAA interactome compared to the background of the cells (FDR=3.62E-39), and included proteins involved in RNA modification (YTHDC1, ADAR1, FTSJ3), mRNA stabilization and export (SYNCRIP, FMR1), and several RNA splicing regulators (SRSF1, SON). Accordingly, there was a

significant enrichment for DNA and RNA helicases (FDR=3.54E-08; MCM3, DHX9, etc) as well as RNA binding proteins (FDR=9.11E-11) within which were many hnRNP family members. Proteins associated with ribosome biogenesis and translation were a large component of the network (FDR=2.20E-30 and 1.64E-06, respectively), including both large and small subunit components (RPL5, 17, 29, RPS9), rRNA processing factors (BOP1, NOP56), and components of the snoRNA pathway (FDR=4.39x10^{-05}, WDR3, NOP56).

Overall, these results showed that the SA1$^{\Delta Coh}$ interactome was enriched not only for transcriptional and epigenetic regulators, but also predominantly for RNA processing and modification, ribogenesis and translation pathways. In addition, SA1 is further enriched to proteins involved in RNA processing and ribosome biogenesis in the absence of cohesin. Accordingly, this suggests that SA1 may facilitate an aspect of cohesin regulation at a variety of functionally distinct cellular locations through its association with these diverse proteins.

To assess the impact of IAA treatment on individual proteins, volcano plots were generated comparing abundance with UTR (Figure 49). Log2FC is plotted on the x-axis and p-value is plotted on the y-axis. Almost all downregulated proteins are members of or known regulators of the cohesin complex. Whereas a wide range of proteins were upregulated.

A number of the highest upregulated proteins were validated for interaction with SA1 by IP of SA1 in ethanol- and auxin-treated cells (Figure 50A). These proteins were also chosen for their diverse functions – enriched GO biological processes associated with each protein are listed to their right and coloured according to the networks in Figure 47 and Figure 48. Densitometry of the blots confirmed an increase in signal intensity in the SA1 IAA YTHDC1, TAF15, INO80, ESYT2, FTSJ3, and FANCI bands, compared to their corresponding ethanol controls. Apart from FANCI, the increase was detected for both the input and IP samples. Conversely, RAD21 and SA1 levels were strongly decreased with auxin treatment. To calculate the enrichment of each protein, IP intensity values were divided by the corresponding input intensity and then normalised to SA1 enrichment values to account for the differential IP amount in ethanol- and auxin-treated samples (Figure 50B). All of the proteins tested had increased enrichment

in auxin-treated conditions compared to ethanol-treated conditions. Due to the level of signal in the FANCI and HNRNPD mock lanes, the enrichment of these two proteins should be considered cautiously and would require further investigation with pre-clearing of the lysate on beads with no bound antibody to determine if this signal remains when signal in the mock is removed. YTHDC1, TAF15, INO80, ESYT2, and FTSJ3 validate the interaction of SA1 with numerous proteins spanning a range of biological processes, in the presence and absence of the cohesin ring.

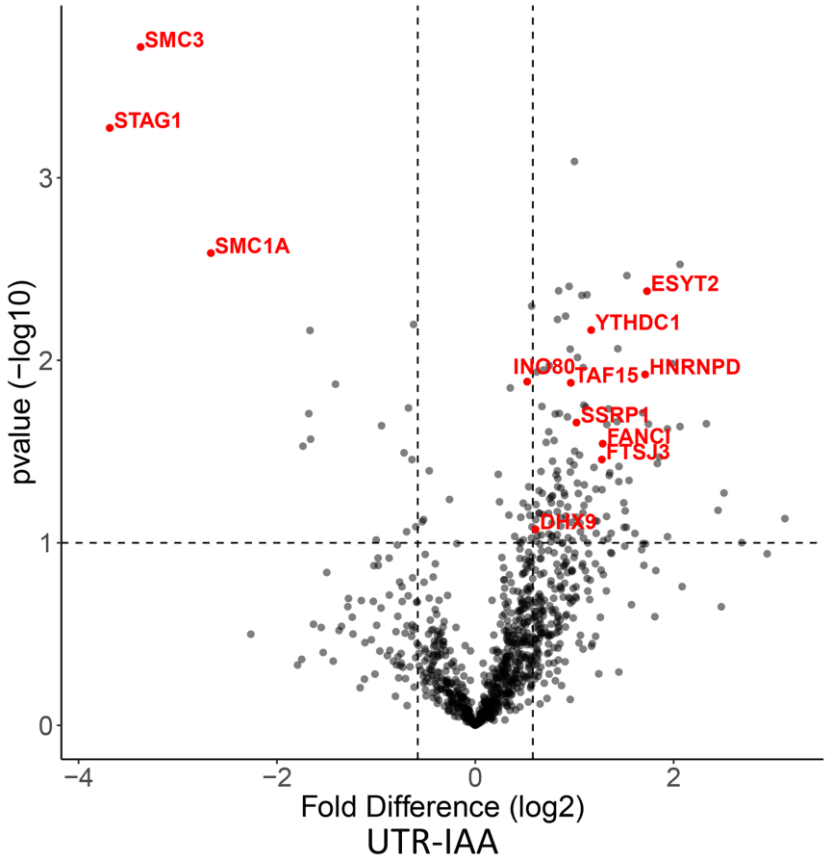


Figure 49: Effect of IAA treatment on SA1 interactors. Volcano plot of SA1 interactors -log10-transformed p-value and log2-transformed fold change comparing UTR vs. IAA conditions. Vertical dashed lines represent changes of 1.5-fold. Horizontal dashed line represents a p-value of 0.1. Cohesin complex members and validated high-confidence proteins have been highlighted.

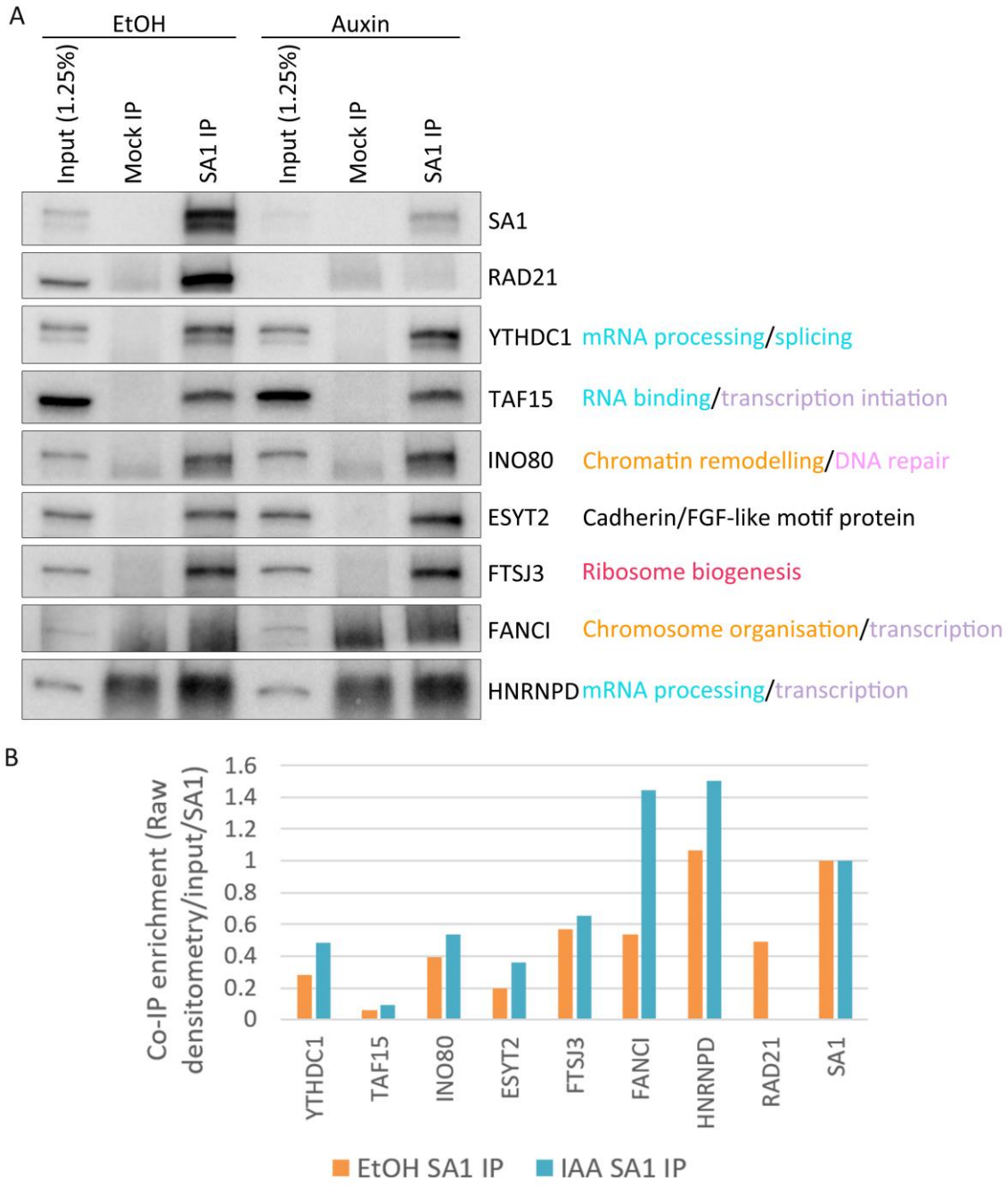


Figure 50: Validation of SA1 interactors in the presence and absence of RAD21. (A) Validation of proteins that are significantly enriched with chromatin SA1 IP in the absence of RAD21, according to Figure 49 and represent each of the enriched biological processes in Figure 23A. Biological processes attributed to each of the proteins tested are listed with colours according to the network in Figure 23A. (B) Bar plot of co-IP enrichment values for each of the validated proteins in (A). Raw values were calculated in ImageStudio Lite. Values were calculated as raw densitometry value in the SA1 IP normalised to corresponding input value and then all normalised to the quantification of SA1 itself.

4.2.3 SA1 interacts with CES-binding-motif-containing proteins in the presence and absence of RAD21

As discussed in the introduction section 1.4.3, Li *et al.* (2020) identified a regular expression motif from binding partners of cohesin that is predicted to identify proteins that can bind to the 'conserved essential surface' (CES) of SA1 and SA2. The authors confirmed interaction of proteins containing the motif by peptide array. A number of these proteins were noticed to overlap with the mass spectrometry experiments discussed above, such as HNRNPUL2, MCM3, CHD6, and ESYT2. SLiMSearch was used to identify human proteins containing the FGF-like, CES-binding motif and 35 proteins were identified (Supplemental Figure 5). These included the known regulators of cohesin CTCF, WAPL, and shugoshin 1 (SGO1). A further subset of proteins were significantly enriched in the SA1 interactome and were also identified in the banded mass spectrometry experiments from section 4.2.1 (Table 12).

		Banded mass spectrometry experiments			
CES-binding protein	**SA1$^{\Delta Coh}$**	**SA1 MS1**	**SA1 MS2**	**SA2 MS**	**CTCF MS**
HNRNPUL2/SAF-A2	Y	Y	Y	Y	Y
MCM3	Y	Y	Y	Y	Y
CHD6	Y	Y	-	-	Y
EIF3I	-	-	-	-	Y
CTCF	Y	Y	Y	-	Y
ESYT2	Y	-	Y	-	Y
MDC1	Y	Y	Y	-	Y
POGZ	Y	Y	Y	-	Y
SGO1	-	Y	Y	-	Y
WAPL	-	Y	Y	Y	Y
DPP3	-	Y	Y	Y	-
ZGPAT	Y	-	-	-	Y

Table 12: FGF-like motif proteins identified in SA and CTCF MS experiments. 12 FGF-like motif proteins that were identified in at least one of the MS experiments are listed with their presence or absence in each experiments indicated. Y = Yes/Present, - = No/Not present.

These FGF-like-motif-containing proteins had a range of different biological functions, including, chromatin remodelling, RNA binding, transcription regulation, and replication. Interestingly, apart from EIF3I, ESYT2, and DPP3 all of these proteins have a known role in DNA damage response (Coster and Goldberg, 2010; Polo *et al.*, 2012; Baude *et al.*, 2015; Han *et al.*, 2015; Murakami-Tonami *et al.*, 2016; Hilmi *et al.*, 2017; Moore *et al.*, 2019; Benedict *et al.*, 2020;

Kargapolova *et al.*, 2020). Furthermore, CTCF, MCM3, HNRNPUL2, CHD6, MDC1, and SGO1 have known RNA-binding activity (Nioi *et al.*, 2005; Lutz, Stöger and Nieto, 2006; Liu *et al.*, 2015; Conrad *et al.*, 2016; Jiang *et al.*, 2017), with CTCF and MCM3 further linked to R-loops (see below for further investigation; Sanz *et al.*, 2016; Hamperl *et al.*, 2017). Therefore, through their interaction with the SA proteins, the FGF-like-motif proteins may act to recruit cohesin to a variety of locations across the genome.

Antibodies for proteins related to HNRNPUL2 and CHD6, namely, SAF-A (also known as HNRNPU), CHD4, and CHD1, were readily available from Professor Richard Jenners laboratory at the institute and were used to test for co-IP with SA1 (Figure 51). Using the old 6U benzonase co-IP set up, all three proteins were pulled down with SA1 in cells treated with ethanol or auxin, with CTCF also blotted for as a positive control. This suggested that SA1 could interact with proteins from the HNRN and CHD family groups in the presence or absence of the cohesin ring.

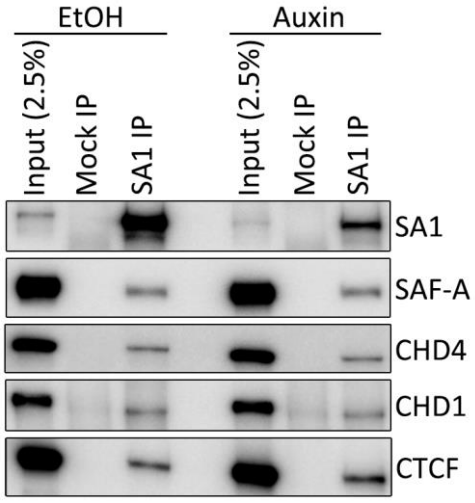


Figure 51: FGF-like motif protein family members interact with SA1 in the presence or absence of cohesin. Antibodies for SAF-A, CHD4, and CHD1 were obtain from Prof. Richard Jenners lab. SAF-A (also known as HNRNPU) is from the same protein family as the FGF-like motif protein HNRNPUL2 (also known as SAF-A2). CHD4 and CHD1 are from the same protein family as the FGF-like motif protein CHD6. Chromatin proteins was solubilised using 6U benzonase per $100x10^6$ cells and pulled down with SA1 or IgG (mock) antibodies. CTCF was included as a positive control for interaction with SA1.

Further to CTCF and ESYT2 tested in previous experiments, HNRNPUL2, CHD6, MCM3, and EIF3I were validated for interaction with SA1 and SA2 by co-IP in the optimised 85U benzonase conditions in two biological replicate experiments. The

first replicate is the same samples and membranes shown in Figure 40. Cells were treated with ethanol or auxin for 4 hrs to assess reliance of interaction with the SA proteins on cohesin. EIF3I was included as a technical test as it was not identified in any of the SA mass spectrometry experiments and is most commonly thought to function in the cytoplasm. CTCF, HNRNPUL2, and MCM3 were pulled down by SA1 in both experiments, in both ethanol and auxin conditions, demonstrating their ability to interact with SA1 in the presence or absence of cohesin (Figure 52 A & B). RAD21 was degraded in both experiments and reproducibly showed higher levels of co-IP with SA2 than SA1.

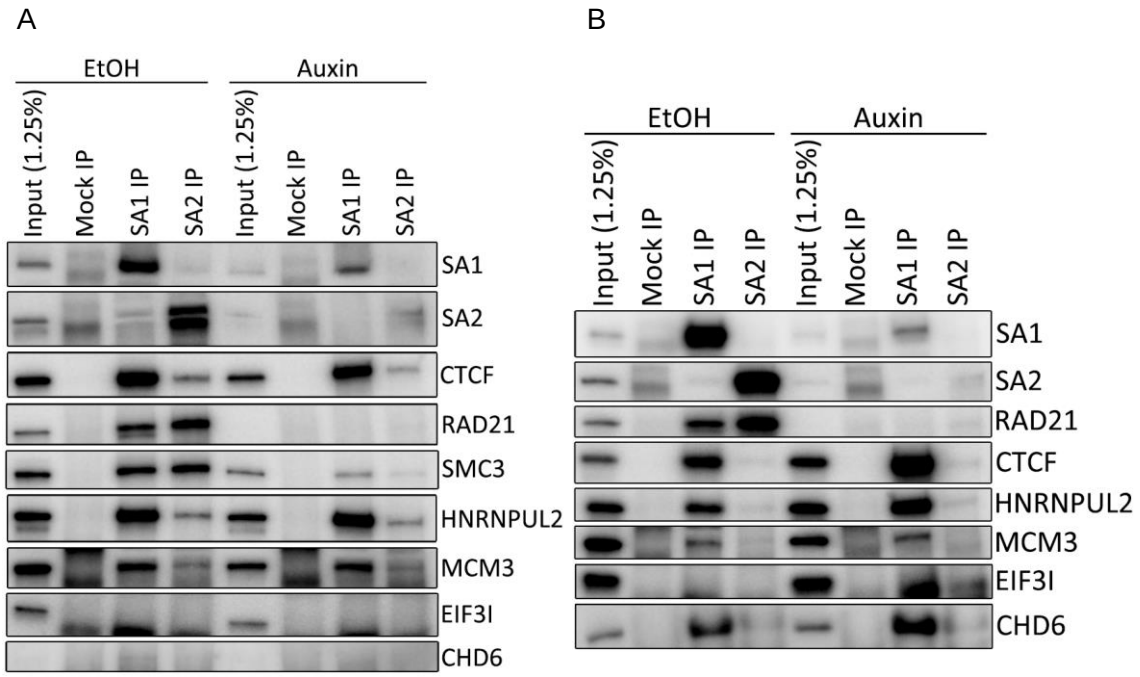


Figure 52: FGF-like motif proteins interact with SA1 in the presence or absence of cohesin. Proteins containing an FGF-like motif are predicted to interact with the SA proteins via their CES domain (Li et al., 2020). (A) Validation of interaction of the FGF-like motif proteins HNRNPUL2, MCM3, EIF3I, and CHD6 in IgG (mock), SA1, and SA2 chromatin IP from cells treated with ethanol or auxin (4 hrs). Chromatin proteins was solubilised using 85U benzonase per $100x10^6$ cells. CTCF was blotted for as a positive control for interaction of an FGF-like motif protein with SA. RAD21 was blotted for to determine efficacy of the auxin treatment. (B) Biological replicate of (A).

Despite the strong co-IP of RAD21 with SA2 in both experiments, co-IP of CTCF, HNRNPUL2, and MCM3 was observed only in experiment 1 (although at significantly lower levels than SA1), perhaps due to a difference in cell cycle distribution, cell state, or the extent of chromatin solubilisation. Although co-IP levels were reduced compared to the corresponding SA1 IP, the three proteins

remained in complex with SA2 in the auxin conditions, indicating that SA2 can also interact with these proteins independently of RAD21.

CHD6 could not be detected in experiment 1 due to its high molecular weight. In experiment 2 half of the eluate was run on a tris-acetate gel under conditions optimised for detection of high molecular weight proteins (discussed in section 5.2.1), here CHD6 could be observed and, like CTCF, showed increased interaction with SA1 in the RAD21-depleted IP. EIF3I was not definitely enriched in either experiment. While its input signal was clear, only a smear of signal at a slightly lower molecular weight was observed in IP lanes. It was not clear if this smear of signal was EIF3I signal, cross-reactivity of the secondary antibody with the light chain of IgG, or, possibly, a combination of the two. Lack of EIF3I co-IP suggests that the presence of an FGF-like motif in a protein is not sufficient to induce interaction with SA and additional factors likely contribute.

As a group, the 35 putative CES binding proteins were statistically over-enriched for proteins involved in cohesion and endocytosis – however, apart from these two enrichments the majority of the proteins are involved in distinct cellular processes. Thus, interaction of SA1 with members of this group further illustrates its ability to interact with a diverse range of proteins, independently of cohesin. The western blot from Figure 38 was stripped and re-probed for HNRNPUL2 to determine if it's co-IP with SA1 was optimal under the same DNA and RNA digestion conditions as CTCF. Optimal co-IP of HNRNPUL2 was observed under the same 85U benzonase condition as CTCF, suggesting that the interaction has the same reliance on DNA/RNA (Figure 53).

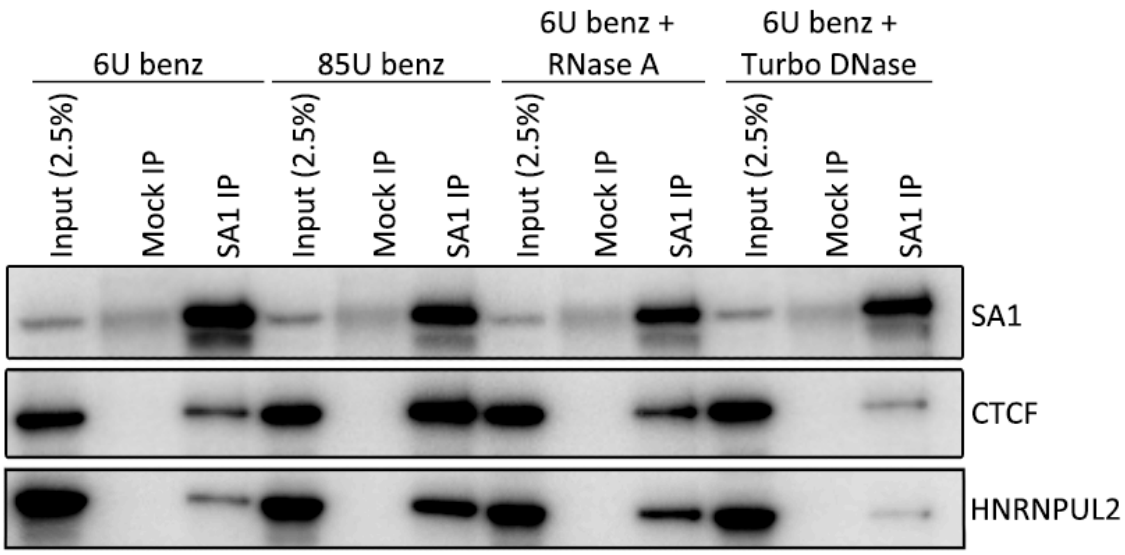


Figure 53: Solubilisation conditions for co-IP of FGF-like motif proteins with SA1. Figure 38 with the membrane now stripped and re-probed for HNRNPUL2.

4.2.4 SA1 interacts with R-loop associated proteins

In all of the above SA1 IP-MS experiments, from banded and full lane set-ups, RNA binding/processing proteins were very significantly enriched. Regulation of gene expression and transcription processes were also significantly overrepresented processes. As discussing in the introduction section 1.5.5, cohesin loading has been shown to occur at sites of specific nucleic acid structures. For example, in yeast cohesion, cohesin captures the second strand of DNA via a single-strand intermediate at the replication fork (Murayama *et al.*, 2018). Given the abundance of RNA binding proteins and transcription factors detected in the SA1 mass spectrometry experiments, we hypothesised that cohesin might also recognise RNA at transcriptionally active genes during loading onto DNA and that interaction of SA1 with the aforementioned proteins may facilitate targeting of cohesin to such sites. As discussed in the introduction section 1.5.7, during transcription, newly synthesised mRNA can thread back to bind to the exposed template strand of DNA, forming an intermediate RNA:DNA conformation, known as an R-loop. In mammalian cells, R-loops are predominately detected at the promoter and termination sites of active genes (Ginno *et al.*, 2012; Sanz *et al.*, 2016). As such, R-loops represent a very specific nucleic acid structure that is comparable to the replication fork by the presence of an intersection of a single strand of DNA opposite a double-stranded nucleic acid molecule (Figure 54). Formation of G-quadruplex (G4) structures on the

unannealed DNA strand may even play a role in the recognition process in the place of the Okazaki fragments on the lagging strand of replicating DNA, although G4 structures may be too distinct in nature for this to be the case. Nucleic acids within the transcription bubble also form a similar type of structure, however, this structure is thought to be enclosed within the polymerase machinery. Hence, we tested whether SA1 may be interacting with such nucleic acid structures by investigating R-loops specifically.

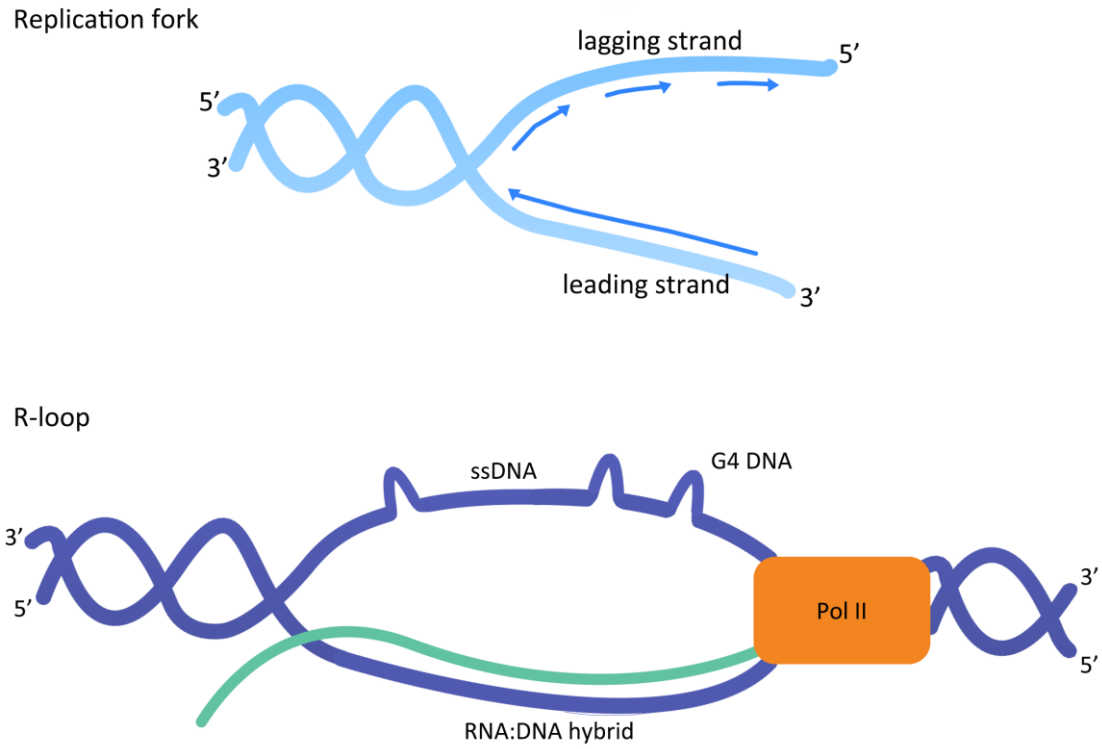


Figure 54: Schematic of the structural similarity between the replication fork and an R-loop. A schematic of DNA unwinding at the replication fork is shown in blue (top). Leading and lagging strand synthesis is indicated with arrows. A schematic of DNA unwinding at an R-loop is shown in purple (bottom). The transcribed RNA that hybridises to the DNA is shown in green. G4 DNA structure that may form in the free ssDNA are shown as peaks. Similar to the Okazaki fragments formed on the lagging strand of the replication fork these G4 structures signify that the displaced single-strand of an R-loop may also form distinct nucleic acid structures. G4 = G-quadruplex; PolII = RNA Polymerase II; ssDNA = single-stranded DNA.

RNA:DNA hybrids can be specifically recognised by the s9.6 antibody, and as such, interacting proteins can be detected by co-IP from an s9.6 pull-down. Two mass spectrometry analyses of R-loop-associated proteins have previously been reported. Cristini *et al.* (2018) identified the RNA:DNA hybrid interactome in Hela cells by IP from nuclear extracts with the s9.6 antibody followed by mass spectrometry. In contrast, Wang *et al.* (2018) generated two specific R-loops previously identified by RNA:DNA IP and sequencing (DRIP-seq) - namely, the

5′ end of the BAMBI gene and the 3′ end of the DPP9 gene. The authors added biotinylated version of the two RNA:DNA hybrids to B cells extracts before IP and mass spectrometry. Despite the different experiment methods, 197 proteins overlapped from both studies and represent strong candidates of R-loop associated proteins. As an initial test for SA1 binding at R-loops, the 197 R-loop associated proteins were checked for overlap with the banded SA1 mass spectrometry experiments. Of these 197 proteins; 136 overlapped with SA1 MS1, 155 overlapped with SA1 MS2, and 130 overlapped with SA1 MS1 and MS2. This suggested that SA1 also interacts with R-loop binding proteins. Of the 130 proteins that overlapped between the R-loop and SA1 mass spectrometry datasets, helicases, eukaryotic initiation factors (eIFs), heterogeneous nuclear ribonucleoproteins (HNRNPs), and RNA splicing factors are some of the groups of proteins that were co-purified. R-loop associated proteins from all the different functional groups were found in the banded SA1 mass spectrometry experiments, suggesting that SA1 can associate with R-loops in a range of different biological scenarios.

Proteins detected in the full lane SA1 mass spectrometry experiments were also compared with the 197 s9.6 interacting proteins. Comparison of the SA1 interactome and the $SA1^{\Delta Coh}$ interactome with the Cristini, Wang, and Cristini and Wang overlap s9.6 interactomes revealed significant overenrichment of R-loop binding proteins in the SA1 interactome in the presence and absence of cohesin (Figure 55). This suggests that SA1 localises to RNA:DNA hybrids in the genome, even when the cohesin trimer is not present. Furthermore, many of the proteins with the largest fold-change value between the UTR and IAA samples have known roles in R-loop regulation, including TAF15, FANCI, and HNRNPD, SSRP1, and INO80 (Britton *et al.*, 2014; Herrera-Moyano *et al.*, 2014; Alfano *et al.*, 2019; Liang *et al.*, 2019; Prendergast *et al.*, 2020). As for the banded mass spectrometry experiments, the overlapping proteins encompassed a range of functions, including helicases, hnRNPs, RNA processing and splicing factors, ribosome biogenesis proteins, and transcriptional regulators. This again implicates SA1 binding to R-loops via a diverse range of proteins.

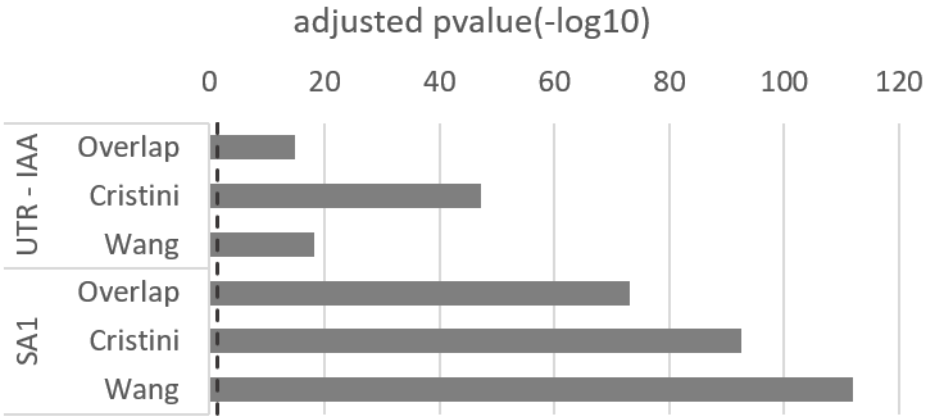


Figure 55: Enrichment of R-loop proteins in the SA1 interactome. Two s9.6 interactome datasets were obtained from the literature; namely, Cristini et al. (2018) and Wang et al. (2018). A high confidence R-loop interactome list was generated from the intersection of the two datasets (termed 'Overlap'). Enrichment of the overlap list and the individual s9.6 interactomes within the UTR SA1 interactome and the SA1$^{\Delta Coh}$ interactome was calculated using the hypergeometric distribution. -log10-transformed adjusted p-value (FDR value) of enrichment are plotted.

4.2.5 SA1 interacts with R-loops

Given the significant enrichment of R-loop-binding proteins, direct interaction of SA1 with R-loop structures was tested. The s9.6 antibody was used to enrich R-loops from chromatin extracts, with known R-loop interacting protein immunoblotted as positive controls, including, POL2, SETX, AQR, MCM3, HNRNPUL2, TOP1, DHX9, and Histone H3 (Skourti-Stathaki, Proudfoot and Gromak, 2011; Sollier *et al.*, 2014; Hamperl *et al.*, 2017; Cristini *et al.*, 2018; Wang *et al.*, 2018).

4.2.5.1 Optimisation of s9.6 IP

In order to assess whether SA proteins were interacting with R-loops, the correct conditions for s9.6 IP in HCT116 RmAC OsTIR1 H2 cells was first optimised. This included generation of a specificity control by digestions of R-loops with an RNase H enzyme, which specifically digests the RNA portion of RNA:DNA hybrid structures (Stein and Hausen, 1969; Wahba *et al.*, 2011). Digestion was tested either with overexpression of a plasmid expressing EGFP-tagged RNASEH1 (obtained from Addgene) or addition of recombinant RNase H enzyme to the chromatin extract (obtained from NEB). Technical differences between all the s9.6 IPs performed are summarised in Table 13.

Figure		Chromatin solubilisation		R-loop digestion		RNase A pre-treatment		IP buffer salt (mM)
		Benz (U/100 x10^6 cells)	Sonication	EGFP-RNASEH1 (ug/1x10^6 cells)	NEB RNase H (ul/IP)	Enzyme	Condition	
56		85	Probe	2	10 or 20	-	-	200
57	A	85 or 6	Probe	1 or 2	-	-	-	200
	B	6 or 0	Probe or Biorupter	-	-	-	-	200
58		0	Biorupter	2, 4, or 6	-	Purelink	0.08ng to IP O/N	200
59	A	0	Biorupter	-	-	EN0531	7ug, 1hr 30 mins at 4°C + IP O/N	300
							7ug, 1hr at RT + IP O/N	
							1ug, 30mins at 37°C + IP O/N	
							7ug, 30mins at 37°C + IP O/N	
	B	0	Biorupter	-	-	EN0531	0.25, 2, or 7ug 1hr 30mins at 4°C + IP O/N	300
60	A & B	0	Biorupter	2 or 4	-	EN0531	2ug, 1hr 30mins at 4°C + IP O/N	300
61	B	0	Biorupter	-	5	Purelink	0.25ug, 1hr 30mins at 4°C	200
62	A	0	Biorupter	-	10, O/N	Purelink	0.25ug, 1hr 30mins at 4°C	200
	C	0	Biorupter	-	20, O/N	Purelink	0.25ug, 1hr 30mins at 4°C	200

Table 13: Summary of technical differences between s9.6 IP experiments. Technical differences between Figures 48 – 54. Sets of similar experiments are highlight in the same colours. NEB RNase H digestion was carried out at 37°C. Unless otherwise indicated digestion was carried out for 30mins. Purelink RNase A is a ssRNA-specific RNase A and EN0531 is an RNase A that digests ssRNA at salt concentrations ≥ 300mM. U = units; Benz = Benzonase; O/N = Overnight.

HCT116 cells were fractionated to obtain chromatin-bound proteins and s9.6 IP undertaken to test for co-IP of SA1. Two commercially available s9.6 antibodies were trialled, one from Kerafast and one from Millipore. Two s9.6 IP samples were prepared for comparison; i) IP from cells transfected with the lipofectamine transfection reagent and no plasmid (termed Vehicle), in which R-loops levels should be as endogenous, and ii) IP from negative control cells transfected with EGFP-RNASEH1 to deplete R-loops. To help confirm observation of R-loop loss, two additional negative controls were included from untransfected cells that were treated with the recombinant RNase H enzyme during chromatin solubilisation

(termed NEB RNase H). POL2, SETX, and HNRNPUL2 were immunoblotted for co-IP as positive controls for interaction with R-loops.

Efficiency of co-IP was equal from both of the s9.6 antibodies, thus, the Kerafast antibody was used for future IPs due to a preference for this antibody in the existing R-loop literature (Figure 56). POL2 was not detected in any of the IP samples. SETX and HNRNPUL2 were observed in all IPs. Similarity of the co-IP signal in Vehicle and EGFP-RNASEH1 transfected cells suggested that the RNASEH1 overexpression was not sufficient to digest R-loops. SA1 and SA2 co-IP was also observed in all samples, suggesting that both SA proteins can interact with R-loops. Co-IP levels of SETX, HNRNPUL2, and SA were all reduced with treatment of the chromatin with the NEB RNaseH enzyme, suggesting specific interaction with R-loops. Interestingly, co-IP of SA1 band s3 was completely lost, suggesting an increased interaction of this putative SA1 variant with R-loops. However, the lack of POL2 co-IP and the residual co-IP signal in negative control samples indicated that the experiment set-up required further optimisation.

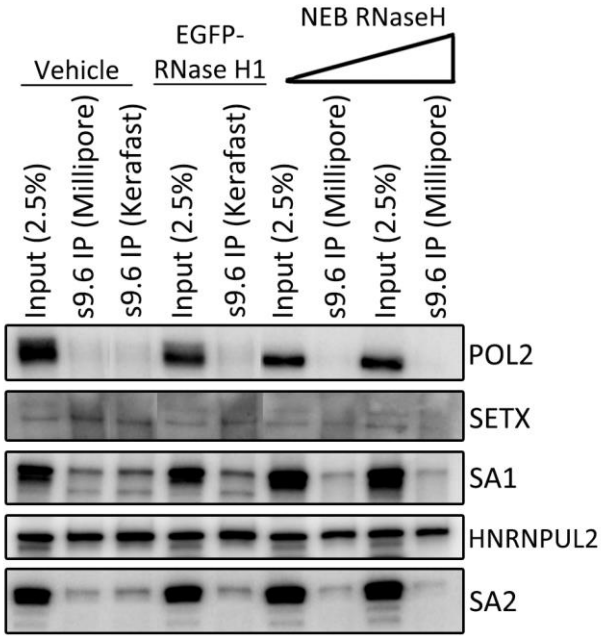


Figure 56: Overexpression of RNase H1 does not reduce enrichment of R-loop binding proteins with s9.6 – Intial test. IP of chromatin bound proteins with s9.6 antibodies obtained from Millipore or Kerafast. POL2 and SETX were blotted for as positive controls for interaction with RNA:DNA hybrids. HNRNPUL2 is an RNA scaffold protein that is predicted to also interact with R-loops and so was included as a positive control. Co-purification of SA1 and SA2 with s9.6 was assessed. Control cells were transfected with no plasmid (vehicle) while a negative control was attempted by transfection of EGFP-tagged RNase H1. Two additional negative control IPs were generate from untransfected cells that were treated with 10 or 20 ul of recombinant RNase H during chromatin solubilisation (NEB RNaseH). EGFP = enhanced green fluorescent protein.

As discussed in section 3.2.7, increased benzonase concentration during chromatin solubilisation facilitates co-IP of SA1 and CTCF, however, benzonase indiscriminately digests all DNA and RNA forms and so may digest the R-loops trying to be isolated. Hence, the effect of benzonase on R-loops was assessed by solubilisation of chromatin for IP with either 85U or 6U benzonase per 100 x 10^6 cells. Transfection with EGFP-RNASEH1 was again included to try to generate a negative control for the presence of R-loops. Decreasing the concentration of benzonase used to solubilise the chromatin facilitated co-IP of POL2 with the s9.6 antibody and left levels of SETX co-IP unchanged (Figure 57A). This suggested that R-loops were more efficiently IP'd in these conditions as POL2 should be found at all R-loops (correspondence with Dr. Konstantina Skourti-Stathaki). POL2, SETX and HNRNPUL2 co-IP was now reduced for RNASEH1-transfected samples, indicating a reduction of R-loops in the cell population. However, SA enrichment levels were reduced in all samples, including the previously used 85U benzonase condition, suggesting a change in SA activity across all the samples, perhaps due to different cell cycle distribution in the seeding population, and meaning that specificity of SA interaction with s9.6 could not be confirmed.

For co-IP and mass spectrometry analyses of R-loops, Cristini *et al.*, (2018) solubilised their chromatin with 10 mins of sonication in a Diagenode Biorupter and no benzonase. Thus, a test of these IP conditions compared to the probe sonication and 6U benzonase condition used above was carried out to assess if enrichment of R-loop associated proteins could be improved further. MCM3 was included as an additional positive control for interaction with R-loops (Hamperl *et al.*, 2017; Cristini *et al.*, 2018; Wang *et al.*, 2018). Input levels of POL2 and SETX varied with the different chromatin solubilisation methods, however, the ratio of their enrichment over input remained similar across all three conditions tested (Figure 57B). Withdrawing benzonase treatment completely did not affect input or co-IP levels of MCM3, SA1, or HNRNPUL2, whereas, switching to Biorupter sonication did reduce input levels of SA1 and HNRNPUL2. Despite reduced input, enrichment of SA1 and HNRNPUL2 was increased in the Biorupter sonication condition. Similarly, co-IP of MCM3 was notably increased in the Biorupter sonication condition. Altogether, these changes designated that no benzonase

treatment and 10mins of Biorupter sonication were enhanced conditions for IP of R-loops and their interacting proteins.

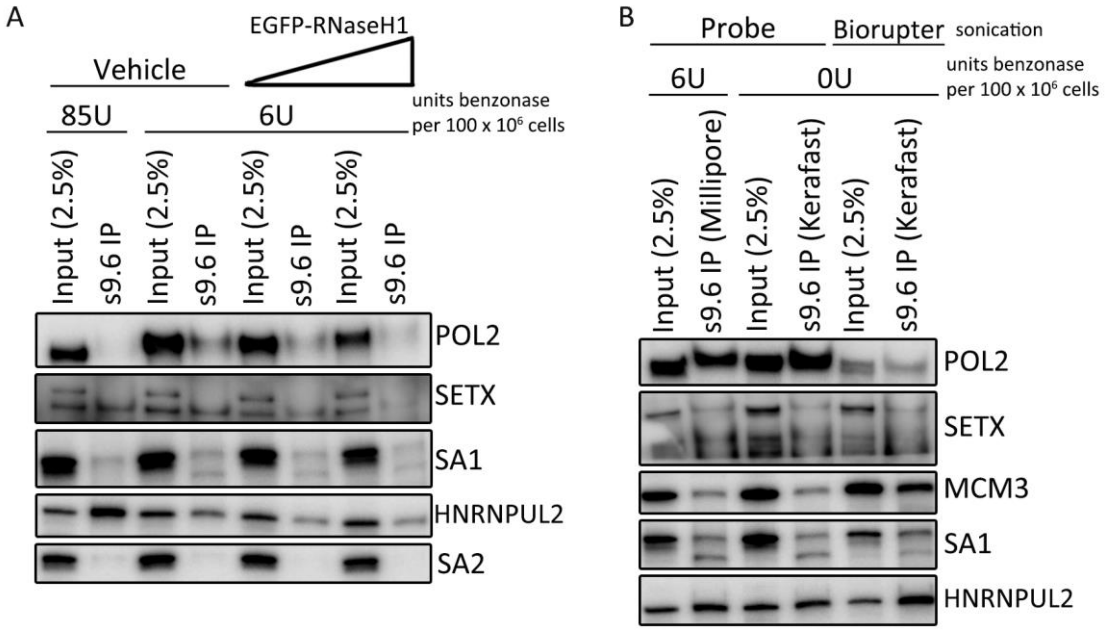


Figure 57: Overexpression of RNase H1 does not reduce enrichment of R-loop binding proteins with s9.6 – Optimisation of chromatin solubilisation conditions. (A) Effect of benzonase treatment on co-IP with s9.6. Samples were treated with 85 or 6U of benzonase per 100x10^6 cells to determine the effect of nucleic acid digestion on co-IP of the positive control proteins with s9.6. Again overexpression of EGFP-RNase H1 was attempted as a negative control for the presence of RNA:DNA hybrids in the cells. (B) Effect of sonication on co-IP with s9.6 and further optimisation of benzonase treatment. Probe and Biorupter sonication methods and treatment 6 or 0U of benzonase per 100x10^6 cells during chromatin solublisation were tested for affect on co-IP of the positive control proteins with s9.6. MCM3 was included as an additional positive control for interaction with RNA:DNA hybrids. EGFP = enhanced green fluorescent protein; U = units.

A titration of EGFP-RNASEH1 transfection was tested under the new enhanced conditions to assess specificity of the co-IP signal observed. To evaluate non-specific co-IP, a mock IP was also run for the untreated chromatin sample in this experiment. H3, TOP1, and DHX9 were also included as additional positive controls for co-IP with R-loops (Cristini *et al.*, 2018). Co-IP levels did not change with lower RNASEH1 transfection amounts and variable enrichment and depletion of the R-loop interacting proteins was observed with higher transfection amounts (Figure 58). Hence, specific interaction could still not be confirmed under the new conditions. Interestingly, RNASEH1 had the opposite effect on the SA1 s3 band compared to treatment with NEB RNaseH during chromatin solubilisation in Figure 56 – here the s3 band was further enriched compared to the canonical

s1 band and compared to its own co-IP in the untreated control. This again suggested a relationship between this potential SA1 variant and R-loops.

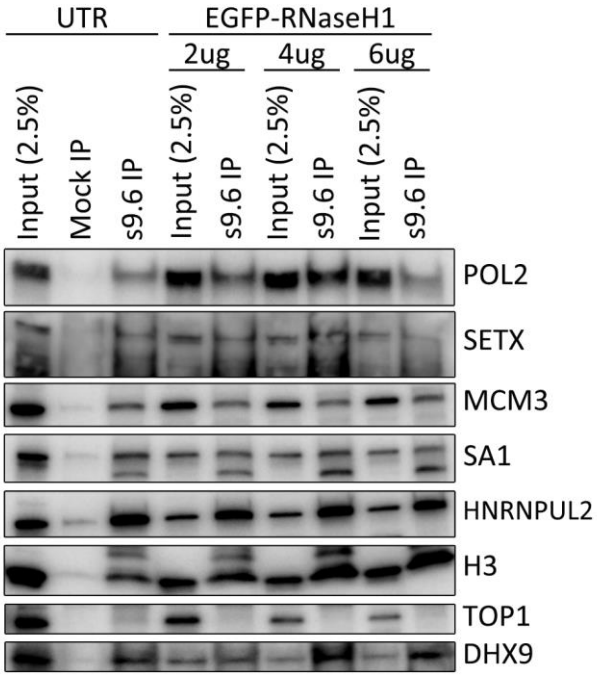


Figure 58: Overexpression of RNase H1 does not reduce enrichment of R-loop binding proteins with s9.6. Using the optimised conditions from Figure 57, co-IP of chromatin proteins with s9.6 was tested in UTR cells or cells transfected with a titration of EGFP-RNase H1. Histone H3 (H3), TOP1, and DHX9 were included as additional positive controls for interaction with RNA:DNA hybrids. An IgG matched antibody raised in the same species as the s9.6 was used for the Mock IP (anti-B2E2). UTR = Untreated; EGFP = enhanced green fluorescent protein; U = units.

Most of the proteins being used to validate SA1 binding to R-loops have roles outside of R-loop regulation and interaction with contaminating ssRNA may account for the observed enrichments. The Kerafast s9.6 antibody is validated for no cross-reaction with double-stranded DNA (dsDNA) or ssDNA and only minor cross-reaction with AU-rich double-stranded RNA (dsRNA). Cross-reactivity with ssRNA is not described by the manufacturer but has been recorded in the literature – Zhang *et al.*, (2015) tested for IP of inducible R-loops at the immunoglobulin heavy chain locus (IgH) class switch regions in a range of RNase-treated conditions. They determined that RNase A pretreatment is required to eliminate background binding to the s9.6 antibody. The authors speculate that this background may be caused by free ssRNA annealing to DNA to form non-physiological RNA:DNA hybrids or folding of the RNA upon itself to form RNA:RNA duplexes, which have previously been shown to be recognised

by s9.6 (Phillips *et al.*, 2013). Accordingly, a literature search was conducted and four potential RNase A pretreatment protocols tested for effect on s9.6 IP. For this optimisation experiment, co-IP was lost almost across the board, including in the untreated control. perhaps because the RNase A enzyme used required the salt concentration of IP buffer be increased to 300mM to digest ssRNAs (Figure 59A). Despite the global loss of signal, there was rescue of DHX9 co-IP in the 4ºC RNase A pretreated sample, suggesting further optimisation could allow enrichment of R-loops.

Subsequently, three concentrations of RNase A treatment at 4ºC were tested for rescue of co-IP. Specifically, pretreatment with 0, 0.25, 2, and 7ug of RNase A were tested. Co-IP of POL2, MCM3, DHX9, and H3 were restored, especially with 0.25 and 2ug of RNase A (Figure 59B). SA1 signal could not be reliably assessed due to damage to the membrane from strong DHX9 co-IP signal.

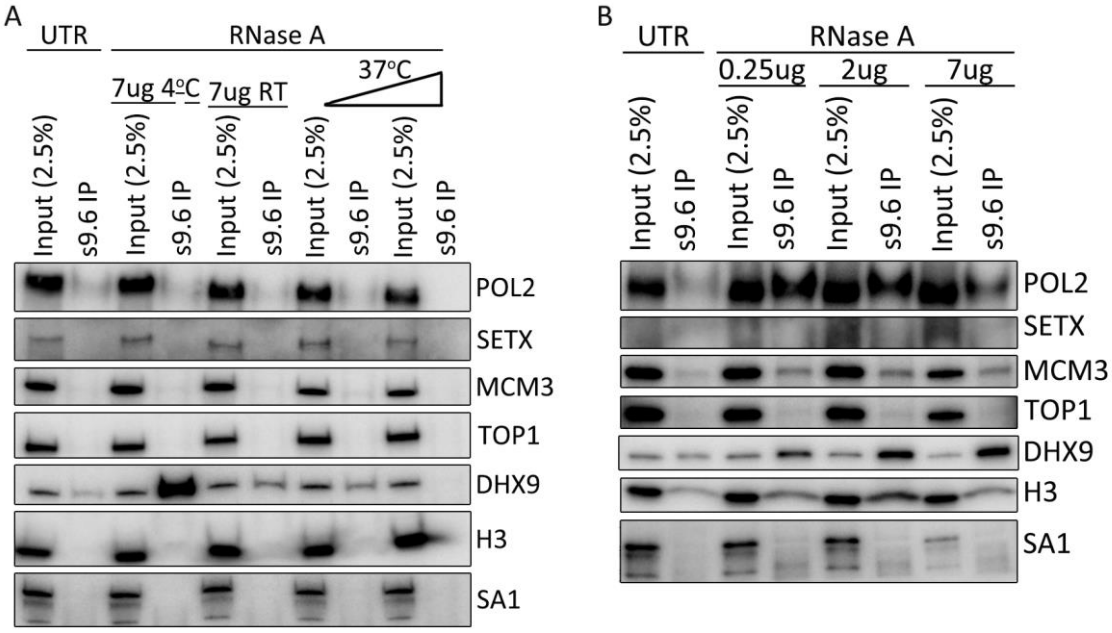


Figure 59: Optimisation of RNase A pre-treatment for s9.6 IP. To ensure ssRNA species were not contaminating the s9.6 IPs a range of RNase A pre-treatments were tested for effect on co-IP of the positive control proteins. (A) s9.6 IP from UTR or RNase A pre-treated chromatin extracts. Chromatin extracts were treated with 7ug of RNase A for 1hr 30 mins at 4ºC, 7ug of RNase A for 1hr at RT, 1ug of RNase A for 30mins at 37ºC, or 7ug of RNase A for 30mins at 37ºC. Thermofisher EN0531 RNase A was used and the salt concentration of the solubilisation/IP buffer was diluted to 300mM to ensure digestion of ssRNAs. (B) s9.6 IP from UTR and RNase A pre-treated chromatin extracts. RNase A treatment was carried out at 4ºC for 1hr 30mins with 0.25, 2, or 7ug Thermofisher EN0531 RNase A. As above, the salt concentration of the buffer was adjusted to 300mM. UTR = Untreated; RT = Room Temperature.

To test the specificity of the rescued signal, the 2ug RNase A pretreatment condition was used for s9.6 IP from vehicle- and EGFP-RNASEH1-transfected cells. For this experiment, HCT116 OsTIR1 cells were used to allow visual confirmation of EGFP-RNASEH1 expression and two time points of EGFP-RNASEH1 expression were tested for efficiency. Variable growth and cell health was observed for the OsTIR1 cells, however, expression of EGFP was observed in all EGFP-RNASEH1-transfected cells, indicating that lack of expression was probably not an issue in previous experiments (Figure 60A). Despite confirmed expression, reduction of co-IP in the EGFP-RNASEH1-transfected samples compared to the Vehicle samples was not observed (Figure 60B). In fact, co-IP of MCM3, DHX9, and H3 was increased rather than decreased, at both the 40 and 65hr timepoints. POL2 enrichment was decreased with RNASEH1 overexpression at 40hr but increased in the 65hr sample. Therefore, overall a clear impact on interacting proteins was not observed, perhaps due to variability in the s9.6 IP or variability in R-loop levels in the cells and maybe even upregulation with the over-expression of the RNA:DNA hybrid nuclease. No strong SA1 enrichment was observed in this experiment, but due to the variable levels of all the other proteins no inference can be made about its interaction with R-loops. At this point it was decided to leave RNASEH1 overexpression as the cells did not tolerate it well at higher levels or longer time-points and no consistent effect could be observed at the conditions tolerated.

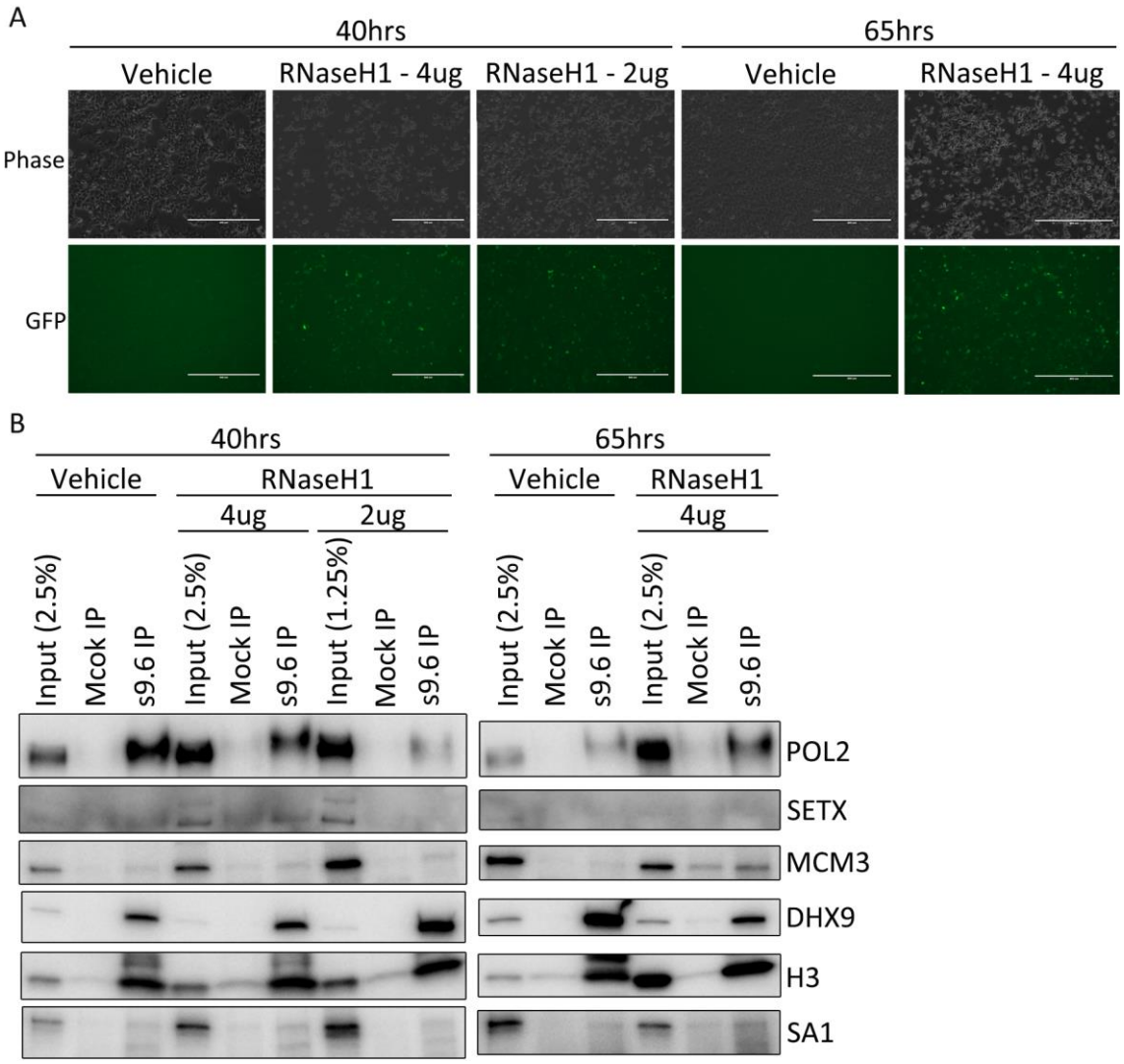


Figure 60: Overexpression of RNase H1 does not reduce enrichment of R-loop binding proteins with s9.6 following RNase A pre-treatment. (A) Zeiss brightfield inverted bench-top microscope images of OsTIR1 cells transfected with no plasmid (vehicle), 4, or 2ug of EGFP-RNase H1 for 40 or 65hrs. Phase and GFP channels are shown. Scale bar = 400 um. (B) s9.6 or IgG-matched (mock) IP of chromatin extracts obtained from the samples in (A) and pre-treated with the 2ug RNase A condition from (Figure 59B).

4.2.5.2 Testing specificity of SA1 enrichment in s9.6 IP with the NEB RNaseH enzyme as a negative control

Given the difficulties in generating a consistent negative control by overexpression of EGFP-RNASEH1, two new avenues of investigation were embarked upon. Firstly, the NEB RNase H enzyme was used to digest RNA:DNA hybrids within the purified chromatin extract to produce a negative control sample for R-loop IP. Secondly, a protocol to dot blot the chromatin lysate was developed to allow measurement of R-loop levels directly in the samples, rather than by proxy with enrichment of interacting proteins with s9.6 IP.

Dot blots are an assay that can be used to detect proteins or nucleic acids on membranes without electrophoresis to separate the molecules in the sample. The lysate is simply blotted on to the membrane in a small dot under conditions that allow the protein or nucleic acid to bind to the membrane. Due to its relative simplicity, the protein dot blot method was used to detect R-loops associated with membrane-bound proteins. Specialised dot blot apparatus was not used and the lysates were transferred to the membrane directly by pipetting. Using input material from the previous experiment, a dot blot for R-loops and a dot blot for SA1 were trialled. Signal could be detected for both antibodies (Figure 61A). For two of the samples, indicated on the membrane by an asterisk, no signal was observed as the lysate did not transfer to the membrane at all due to technical issues. Apart from this, these dot blots worked as a proof of concept to show that proteins, such as SA1, and the R-loops interacting with them can be detected by dot blot even without the expensive apparatus. This dot blot also corroborated what was observed by IP, expressly, that R-loops were not being digested with the overexpression of EGFP-RNASEH1. Hence, the dot blot could be used to assess efficiency of the NEB RNase H digestion in future experiments.

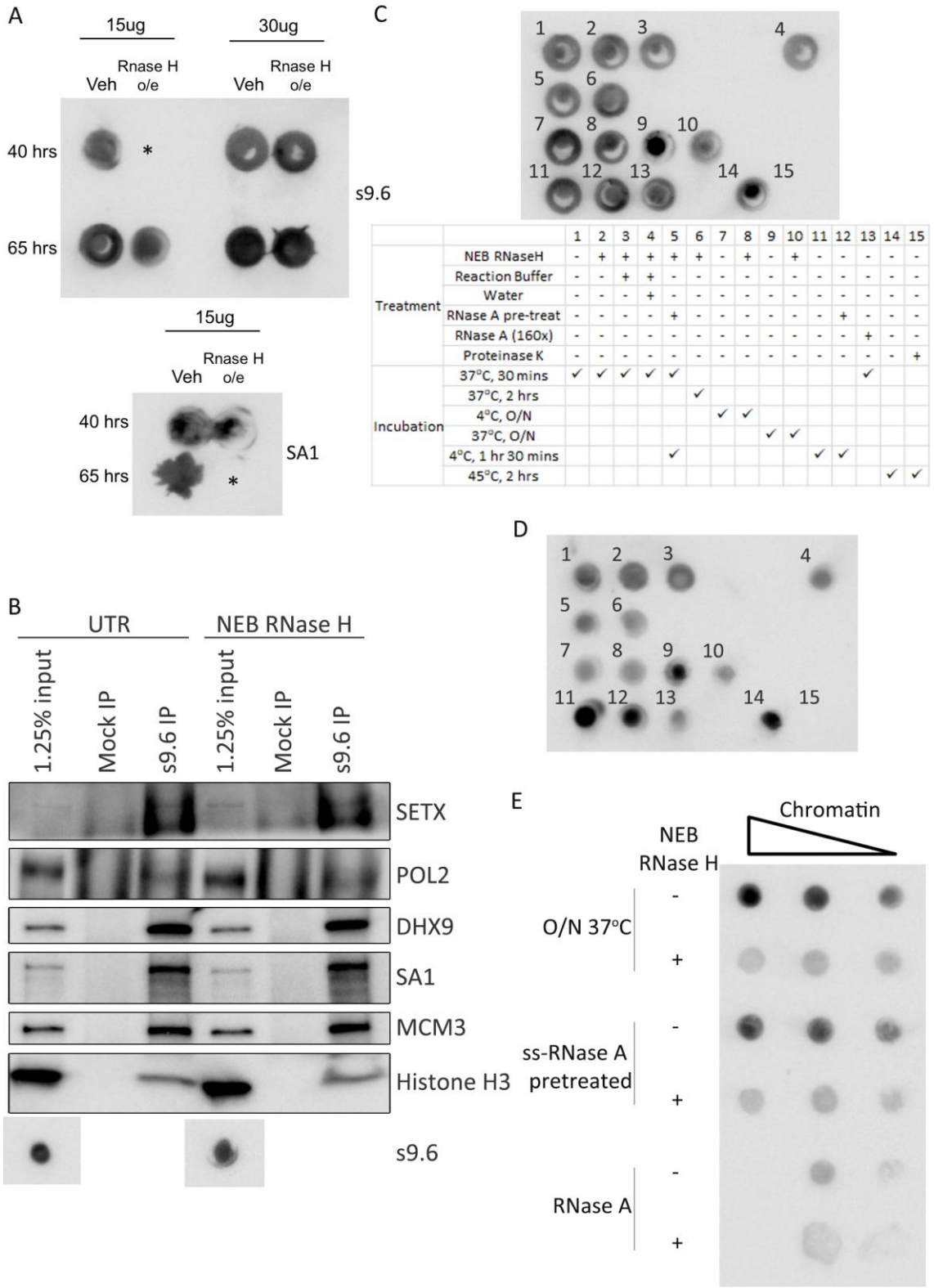


Figure 61: Optimisation of NEB RNase H digestion of R-loops. (A) Proof-of-concept for a protein dot blot protocol. Dot blot of chromatin extracts from Figure 60. Blots were probed for s9.6 (top) or SA1 (bottom). N.B. * indicates samples that did not transfer to membrane. (B) Chromatin IP with s9.6 or IgG (Mock) antibodies. Chromatin samples were pre-treated with a ssRNA-specific RNase A. Chromatin was also UTR or treated with 10ul NEB RNaseH just prior to IP. Dot blot of the input is included. (C) Optimisation of NEB RNaseH digestion conditions. Chromatin extracts were treated as signified in the table, dotted onto the membrane, and probed for s9.6. (D) Technical dot blot repeat of (C). (E) Dot blot of 12.5, 6.25, and 3.125ug of chromatin, evaluating the effect of NEB RNase H and ssRNase A pre-treatment on R-loop levels. A positive control for the digestion of R-loops was included with a high concentration and temperature global RNase A digestion sample. UTR = Untreated; o/e = overexpression; O/N = Overnight.

Using the conditions optimised in section 4.2.5.1, s9.6 IP was carried out +/- digestion with the NEB RNase H. RNase A pretreatment was kept to remove contaminating free RNA, however, a ssRNA-specific RNase A was used as it did not require a higher salt concentration that might have affected s9.6 IP efficiency in the previous experiments. An RNase inhibitor was also added to each IP to prevent continued digestion overnight. Enrichment levels of all of the proteins tested was unchanged with NEB RNaseH digestion (Figure 61B). Dot blot for s9.6 confirmed that R-loops levels were retained in the + NEB RNase H sample, with only a slight decrease in signal compared to the untreated lysate. Hence, optimisation of the NEB enzyme digestion was required.

Now that s9.6 dot blot was being trialled, multiple NEB enzyme conditions could be tested with a small amount of chromatin lysate and enzyme. Modification of the IP buffer, digestion time, and temperature were tested, as indicated (Figure 61C). Sample 1 was included as a negative control for digestion of R-loops. Samples 14 and 15 were used to test how degradation of proteins from the lysate would affect s9.6 signal on the blot. The majority of the samples appeared to have similar amounts of R-loops (Figure 61C). Complete loss of signal in sample 15 demonstrated that the R-loops detected were in complex with proteins that are bound to the membrane. Sample 10, which was incubated O/N at 37^{o}C showed the most reduction compared to its no enzyme control (sample 9). This suggested that in these cells and buffer conditions, an increased digestion time was required. This type of elongated digestion has been described in the literature for DRIP-seq experiments (Chen *et al.*, 2017; Halász *et al.*, 2017; Abakir *et al.*, 2020). To confirm efficacy of the O/N at 37^{o}C digestion and avoid any 'coffee ring' signal, a technical replicate of the dot blot was generated (Figure 61D). Signal on the blot varied slightly to the first replicate, perhaps from freezing and thawing of the samples. Regardless of the discrepancies in signal, the highest change in +/- NEB RNaseH was again between samples 9 and 10, confirming this as the optimal digestion strategy.

Overnight digestion was confirmed in a biological replicate (Figure 61E). Digested material was treated with the ssRNA-specific RNase A to assess effect of the pretreatment procedure on this material. Here the RNase A pretreatment reduced

s9.6 signal in the -NEB RNaseH sample compared to the original -NEB RNaseH samples. This indicates that the s9.6 antibody can recognise some material that is digested by RNase A. To generate a positive control for R-loop digestion, an aliquot of the sample was also digested with RNase A under conditions to the digestion all RNA species, including those in R-loop structures. s9.6 signal in this sample matched that digested by the NEB RNaseH enzyme, confirming loss of R-loops in this sample.

S9.6 IP was repeated with two chromatin lysates incubated O/N at 37^{o}C, one including the NEB RNaseH enzyme and one without (UTR). Dot blot of the input material confirmed that R-loops were reduced with the new digestion protocol (Figure 62A). Accordingly, enrichment of the RNA helicase AQR, SETX, POL2, MCM3, Histone H3, and SA1 was reduced in the RNase H-treated IP compared to the UTR IP. Hence, specific interaction of SA1 with R-loops was validated alongside known R-loop interactors. DHX9 did not show reduced enriched with RNase H treatment. DHX9 translocates on RNA with a 3’ single-stranded tail – a structure that may be increased in the lysate by the action of the RNaseH enzyme, possibly explaining the increase of DHX9 present in the negative control IP (Lee and Hurwitz, 1992).

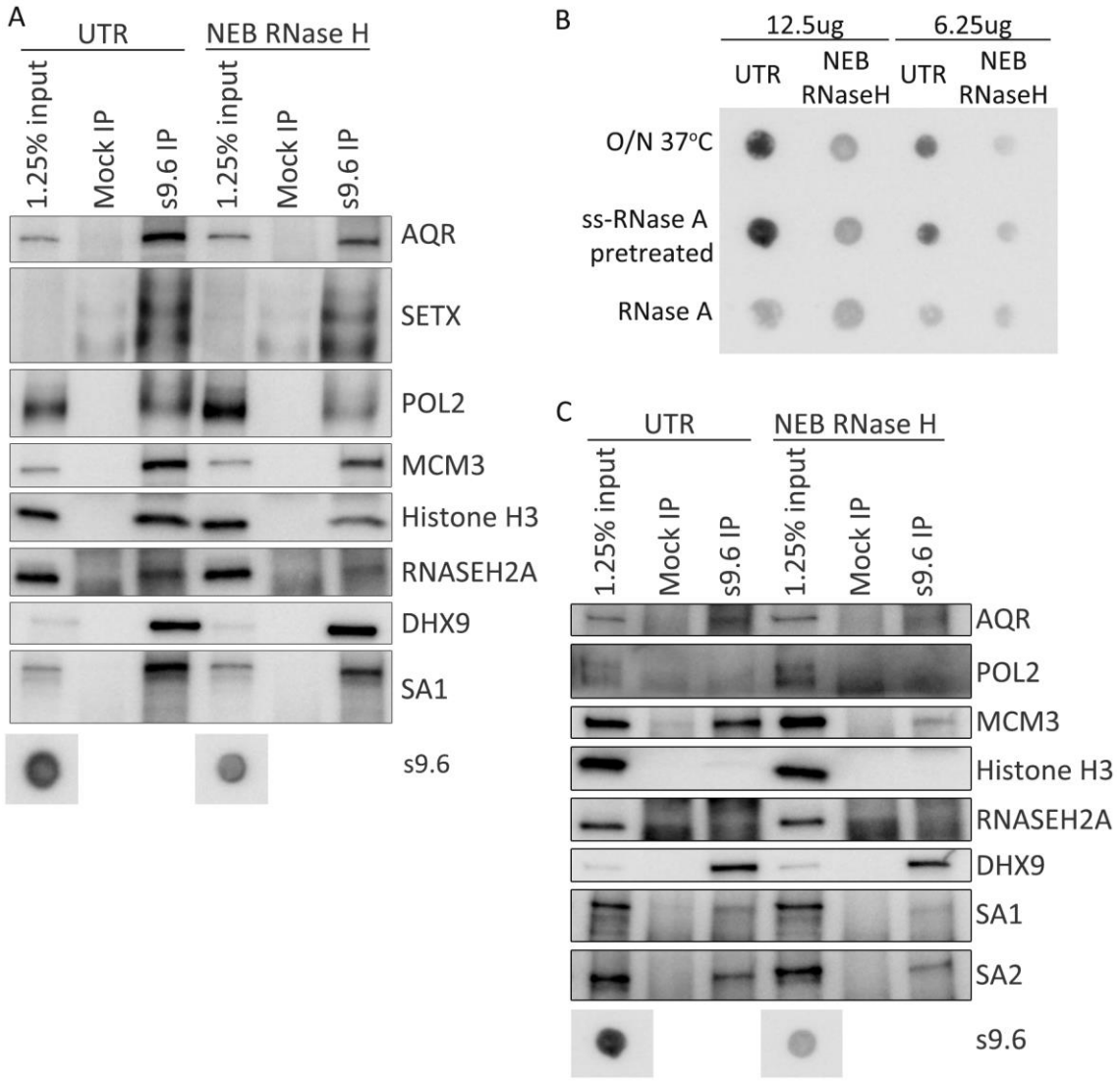


Figure 62: SA1 and SA2 co-IP with s9.6. (A) Chromatin IP with s9.6 or IgG (Mock) antibodies. AQR, SETX, POL2, MCM3, Histone H3, RNASEH2A, and DHX9 are blotted for as positive controls for interaction with RNA:DNA hybrids. Chromatin samples were UTR or treated with 22ul of NEB RNase H O/N at 37^{0}C and were pre-treated with a ssRNA-specific RNase A. Dot blot of input material is included. (B) Dot blot of 12.5 and 6.25ug of input chromatin material for (C evaluating the effect of NEB RNase H digestion and ssRNase A pre-treatment on R-loop levels. A positive control for the digestion of R-loops was included with a high concentration and temperature global RNase A digestion sample. (E) Biological repeat of (A) with 40ul of NEB RNase H. UTR = Untreated; O/N = Overnight.

Now that reduction of R-loop signal had been achieved, the experiment was repeated to determine the reproducibility of SA1 co-IP with R-loops. Due to time constraints imposed by the novel coronavirus pandemic the experiment was carried out on previously frozen down chromatin material. s9.6 dot blot was used to assess efficacy of the NEB RNase H digest. Pre- and post-RNase A pretreated samples were blotted on the membrane, as well as corresponding samples that underwent a ‘full’ RNase A digest. All NEB RNase H-digested samples matched that of the full RNase A digest, indicating that R-loops were efficiently degraded

from the lysate (Figure 62B). In this experiment, the ssRNA-specific RNase A pretreatment to remove ssRNA did not alter s9.6 signal compared to the corresponding UTR sample, perhaps because the freeze-thaw cycle this chromatin sample underwent already degraded ssRNA sufficiently to prevent cross-reaction. Co-IP signal for AQR, POL2, Histone H3, RNASEH2A, and SA1 was reduced in the UTR IP compare to Figure 62A, this may have been due to loss of RNA:DNA hybrids with freezing of the sample (Figure 62C). Notwithstanding the reduced signal in the UTR sample, co-IP of AQR, MCM3, DHX9, SA1, and SA2 was reduced with RNaseH treatment. Therefore, this experiment validated the enrichment of SA1 and SA2 with s9.6 IP and the sensitivity of this signal to RNaseH-mediated digestion of R-loops. Hence, the SA proteins were confirmed as interactors of R-loops.

4.2.6 SA1 interacts with RNA

Given the interaction of SA1 with so many RNA binding proteins and the dependence of co-IP on specific nucleic acid digestion conditions, we hypothesized that SA1 may in fact directly interact with RNA. In collaboration with Professor Richard Jenner and Dr. Manuel Beltran-Nebot, UV cross-linking and immunoprecipitation (CLIP) was used to investigate interaction of SA1 and SA2 with RNA. HCT116 RmAC OsTIR1 H2 cells were treated with scramble siRNA (siCtl), siSA1, or siSA2 for 72hrs. The samples were then processed for CLIP by Dr. Manuel Beltran-Nebot. SA1 and SA2 were effectively knocked down in their respective siRNA samples, as observed in both WCE input samples and IP samples (Figure 63A).

For both SA1 and SA2, a distinct band of radioactive RNA was detected in complex with the protein, meaning that both SA proteins can interact with RNA. The specificity of the RNA band was confirmed by loss of the signal in the siRNA knockdown samples. Dependence of SA interaction with RNA on the cohesin ring was then tested by CLIP of ethanol- and auxin-treated cells (Figure 63B). Again, the cells were grown and treated by myself and processed for CLIP by Dr. Manuel Beltran-Nebot. WCE input material is shown on a separate blot and confirmed complete degradation of RAD21 from the samples and reduction of SA1 and SA2. IP samples were loaded on the gel to equalise SA1 and SA2 enrichment in

ethanol and auxin samples – this allowed best visualisation of residual RNA signal in the auxin-treated samples. With equalised SA IP levels, RNA was proportionally associated in the auxin conditions despite loss of the cohesin ring. This experiment determined that SA1 and SA2 can specifically interact with RNA independently of RAD21.

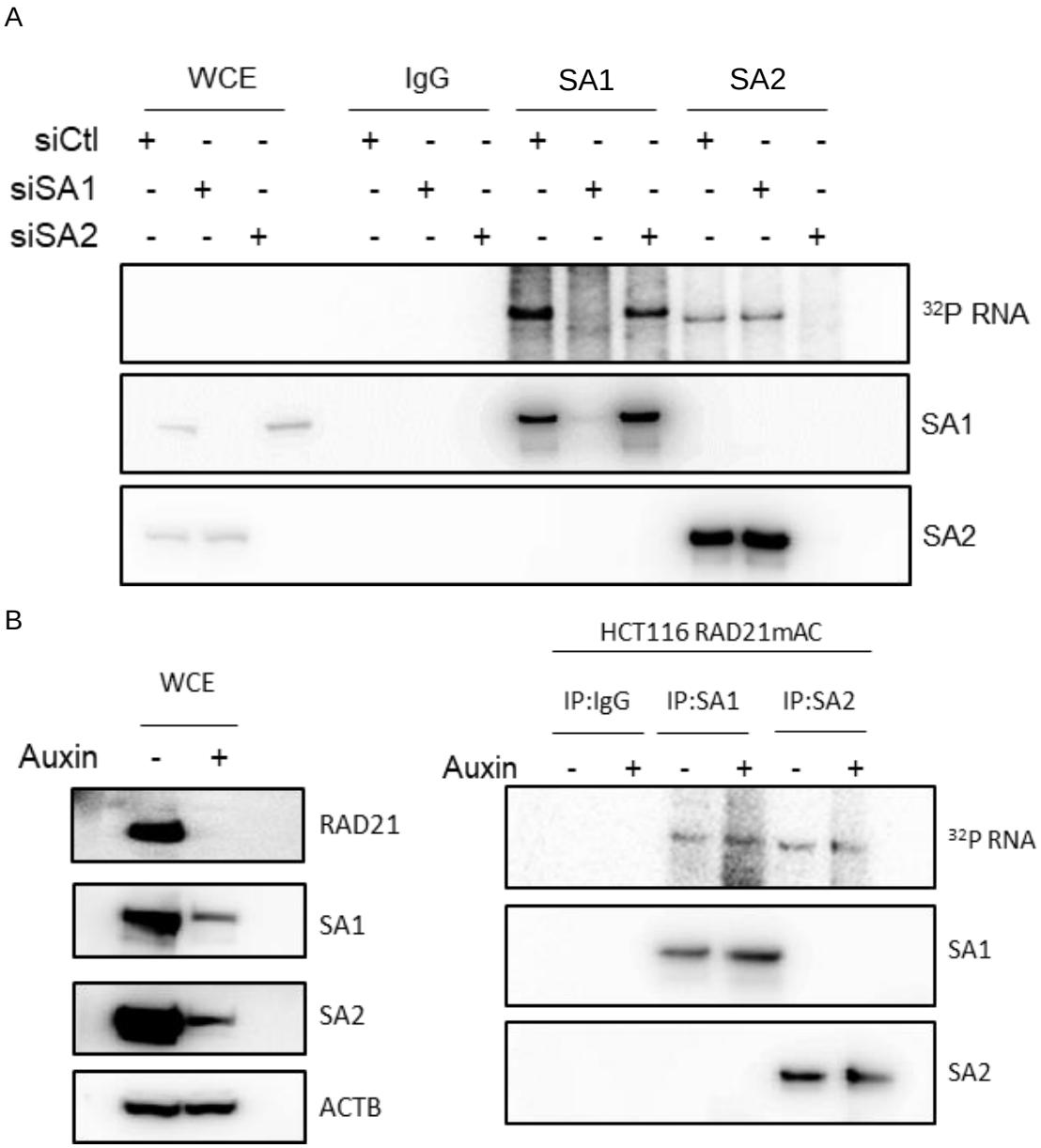


Figure 63: SA1 and SA2 interact with RNA in the presence and absence of RAD21. (A) CLIP of WCE material with IgG, SA1, or SA2 antibodies. Cells were treated with control scramble siRNA (siCtl), siRNA targeting SA1 (siSA1), or siRNA targeting SA2 (siSA2). Input material is labelled as WCE. (B) CLIP of WCE material with IgG, SA1, or SA2 antibodies. Input samples are shown on the left (WCE) and IP samples are shown on the right. Samples were treated with ethanol (-) or auxin (+) for 4 hrs prior to collection.

4.3 Discussion

IP-MS of Coomassie-stained bands from SA1, SA2, and CTCF IPs gave insight into the potential interactomes of these proteins. Protein groups enriched in the SA1 and CTCF IPs were strikingly similar, both in the proteins classes and the proportion of proteins in each class. In contrast, SA2 showed a more diverse set of enriched proteins classes. Given the relative strength of CTCF co-IP with SA1 compared to SA2 observed in Chapter 3, this suggests that multiprotein interactions of CTCF and SA1 may help to stabilise their interaction. RNA binding proteins and gene-specific transcription regulators were strongly represented in the CTCF and SA1 IPs suggesting such regulation might occur at sites of active transcription.

While a large proportion of the SA2 enriched proteins were also RNA binding proteins, the SA2 IP was most strongly enriched for different classes of enzymes, predominantly metabolite interconversion enzymes. The role of metabolism in chromatin biology has been widely studied in the context of histone modification (reviewed in: Schvartzman, Thompson and Finley, 2018). For example, the methyl groups deposited on histones are derived from two metabolic pathways, serine-glycine-one carbon metabolism and the methionine cycle. Similarly, metabolite products are used to fuel demethylation enzymes. Proteins from both serine-glycine-one carbon metabolism and the methionine cycle were detected in the SA2 IP, for example, SHMT2, PSPH, and PHGDH from serine-glycine-one carbon metabolism and MAT2A and AHCY from the methionine cycle. The snapshot of the SA2 interactome viewed in this banded IP-MS experiment suggests that SA2 interacts with enzymes involved in metabolism, perhaps directing cohesin to sites with specific histone modifications in response to environmental cues, or directing the metabolite conversion enzymes to sites of gene loops, where it is thought to be localised (Kojic *et al.*, 2018).

6 of the 8 major proteins classes overrepresented in the banded SA1 potential interactome were also enriched in the full lane SA1 interactome, namely, chromatin -binding and -regulatory, RNA binding, gene-specific transcriptional regulatory, DNA binding, translation, and cytoskeletal protein classes. Thus,

despite only analysing a subset of the IP material, the banded IP-MS experiments likely provide biologically relevant insight into the interactomes of these proteins.

As expected from the literature, chromatin-binding and -regulation proteins, transcriptional regulators, and DNA repair proteins were enriched in the SA1 interactome. Examples include the suz12 component of PRC2, which has been shown to interact with cohesin. Loss of this interaction has important effects on 3D organisation of chromatin and gene regulation and has been suggested to play a role in the tumourigenesis of cohesin loss in AML (Fisher *et al.*, 2017; Cuadrado *et al.*, 2019). Similarly, the transcription factor YY1 has been shown to interact with cohesin and with CTCF and Pax5 at long-range chromatin contacts (Medvedovic *et al.*, 2013; Pan *et al.*, 2013). The MCM6 DNA helicase was also identified among the highest abundance proteins in the SA1 interactome, and multiple additional members of the MCM2-7 complex, which is known to localise cohesin loading to replicated sites during S-phase, were also identified (Zheng *et al.*, 2018). Interestingly, the NCAPD2 protein, a member of the condensin complex was also a highly abundant protein. This was similar observed in a study of the RAD21 interactome in Hela cells (Panigrahi *et al.*, 2012), although cohesin and condensin are not commonly assumed to interact together. Despite being one of cohesins most well-established interactors, CTCF was only identified at a low abundance across all of the SA1 IPs analysed, suggesting that additional protein interactors of cohesin may be under-appreciated in the literature. Alternatively, the sequence and structure of CTCF may make it difficult to detect by mass spectrometry, however, robust detection in the banded CTCF IP-MS suggest that at high abundance CTCF can be analysed by mass spectrometry.

Following gene expression, RNA processing was the most significantly enriched biological process in the SA1 interactome network, with many sub-groups of RNA processing also highly enriched, including but not limited to RNA splicing, RNA transport, and RNA stabilisation. Many SR splicing factors, HNRN proteins, and DEAD-box helicases were highly enriched in the SA1 interactome. While cohesin is thought to load onto chromatin at sites of active transaction, perhaps by co-opting the open nature of chromatin at such sites, these enrichments suggest that SA1-cohesin localises to the transcription bubble at sites of RNA processing. Previous interrogation of the cohesin interactome has shown a similar enrichment

of RNA processing and metabolism, validating the importance of these processes in cohesin function (Panigrahi *et al.*, 2012; Kim *et al.*, 2019).

Unexpectedly, ribosome biogenesis and translation were also highly enriched processes within the SA1 interactome network. The enrichment of ribosome proteins and rRNA processing factors suggests a role for cohesin in the regulation of chromatin structure within the nucleolus. While overlooked as a consequence of stickiness in previous cohesin interactome studies, functional importance in the nucleolus has been shown previously in yeast cells (Bose *et al.*, 2012; Harris *et al.*, 2014). Mutations to cohesin that prevent stabilisation on chromatin were shown to disrupt nucleic acid structure in the nucleolus, decrease rRNA processing, and reduce ribosomal biogenesis, suggesting a functional role for cohesin in ribogenesis. Perhaps most unexpected, translation factors were enriched. This is unusual as the central dogma is that translation occurs in the cytoplasm. However, this is certainly not the first description of activity of translation factors in the nucleus, and it has been suggested that translation within the nucleus may occur as a means of quality control of mRNAs (reviewed in: Reid and Nicchitta, 2012). Interaction of SA1 with an abundance of RPL and RPS proteins suggest that cohesin is also active at such sites. Ribosomal proteins are highly abundant proteins, hence, even though their quantity in the SA1 IP changes with siRNA-mediated knockdown of SA1, more formal verification of interaction, such as by BiFC, would help to clarify if any were identified as false positives.

Investigation of the SA1 interactome in cells depleted of RAD21 determined that the cohesin-independent SA1 interactome is also enriched for chromosome organisation, transcription, RNA processing, ribosome biogenesis, and translation processes. Similar to the CTCF ChIP results, this suggests that SA1 is not switching localisation in the absence of cohesin, but instead suggests that SA1 may interact with these proteins upstream of cohesin. Comparison of the SA1 +/-cohesin interactomes revealed a further enrichment of proteins involved in RNA processing and ribosome biogenesis in auxin conditions. While the majority of enriched processes were maintained in the absence of the cohesin ring, the abundance of proteins in each group was altered. Whether this is a

function of the dynamic nature of IP experiments or a biological shift in binding in the presence of cohesin is not clear.

ESYT2, YTHDC1, TAF15, HNRNPD, SSRP1, FANCI, FTSJ3, and INO80 represent some of the proteins that were mostly strongly enriched in auxin compared to UTR conditions. These proteins span the wide variety of biological processes enriched in the SA1 interactome, indicating that SA1 contains an inherent ability to interact with proteins across these processes, regardless of cohesin binding. As examples, the potential implications of SA1 interaction with three of these highly enriched proteins is discussed below.

YTHDC1 is a 'reader' protein that recognises N6-methyladenosine (m6A) modifications on mRNA (Xu *et al.*, 2014). m6A is a prevalent RNA modification and has been implicated in the processing, export, translation, and stability of mRNAs (reviewed in: Zhao, Roundtree and He, 2016). In the nucleus, YTHDC1 binding to m6A regulates splicing of the mRNA by recruitment of splicing factors that promote exon inclusion (SRSF3) and expulsion of splicing factors that promote exon skipping (SRSF10). Alongside the significant enrichment of splicing factors in the SA1 interactome, this suggests that SA1 may function at splice sites of mRNAs. YTHDC1 and SA1 have both been found to regulate heterochromatin in mESCs via effect on H3K9me3 deposition and compaction, respectively, indicating additionally potential importance of this interaction for chromatin organisation and cell fate (Liu *et al.*, 2021; Pežić *et al.*, 2021).

FTSJ3 is an RNA 2′-O-methyltransferase that is required for processing of pre-rRNA to mature 18S rRNA (Morello *et al.*, 2011). Thus, SA1 is also recruited to proteins involved in RNA processing in the nucleolus. FTSJ3 has recently been identified as a potential driver of breast cancer, where its upregulation may induce cancer cell survival and progression. Hence, interaction with FTSJ3 likely target SA1 to rRNA sites important for these processes.

INO80 is the ATPase subunit of the chromatin remodeller INO80 complex (Shen *et al.*, 2000). The INO80 complex binds to nucleosomes with adjacent free DNA and shifts the nucleosomes to bind evenly along the DNA (Udugama, Sabri and Bartholomew, 2011). Thus, SA1 interaction with INO80 may facilitate localisation

of cohesin at free DNA for loading in the absence of dense nucleosomes. The histone variant H2A.Z accumulates at sites of active transcription, following disruption of the replicative variant H2A, wherein it can maintain open chromatin and gene activation via mutual exclusivity with DNA methylation (Zilberman *et al.*, 2008; Weber, Henikoff and Henikoff, 2010). The INO80 complex can regulate turnover of the histone variant H2A.Z at sites of transcription, replication, DNA repair, and possibly at telomeres (Yu *et al.*, 2007; Papamichos-Chronakis and Peterson, 2008; Papamichos-Chronakis *et al.*, 2011; Cao *et al.*, 2015). Hence, INO80 and SA1 are both active at the same genomic locations. As discussed in section 1.3.1.1.1, nucleosome occupancy affects CTCF-DNA interaction and nucleosomes flanking CTCF are enriched for H2A.Z (Fu *et al.*, 2008). Identification of SA1-INO80 interaction adds to our understanding of this relationship. It is possible that INO80 recuits SA1 to sites of H2A.Z and nucleosome turnover to stabilise CTCF binding and mediate chromatin looping. Alternatively, SA1 may recruit INO80 to CTCF loop anchors to remodel nucleosomes. Nucelosome remodelling has been linked to sister chromatid cohesion and cohesin loading, indicating the downstream importance of this relationship (Hakimi *et al.*, 2002; Muñoz *et al.*, 2019).

FGF-like-motif proteins have been shown to interact with SA2 via its CES (Hara *et al.*, 2014; Li *et al.*, 2020). Given the conserved nature of the CES these proteins are also expected to SA1. In this thesis, interaction of FGF-like motif proteins with SA1 and SA2 was confirmed in human cells. These IPs furthered identified that enrichment of the FGF-like motif proteins was increased with SA1 compared to SA2 in human cells. Given the differential efficiency of FGF-like motif proteins co-IP with SA1 and SA2, it would be interesting to determine if the same range of biological processes enriched with SA1 are enriched with SA2 in auxin conditions. This would help to determine if both SA paralogs have such an inherent ability to interact with proteins involved in these processes or it is specific to SA1. It would also be interesting to determine if any of the FGF-like motif enrich more efficiently with SA2.

As discussed in the introduction section 1.2.2, sororin-mediated stabilization of cohesin during sister chromatid cohesion is thought to occur via competitive binding of FGF-motifs in sororin and WAPL with PDS5 proteins (Nishiyama *et al.*,

2010). However, displacement of WAPL in chromatin extracts was not observed leaving the molecular mechanism of stabilization unclear. As FGF-motif porteins have since been shown to interact with SA, this raises the question of whether the SA proteins may be involved in sororin-mediated stabilization of cohesin and WAPL antagonism. Indeed sororin depletion induces cohesion defects similar to depletion of SA1 and SA2, perhaps suggesting similar function (Nishiyama *et al.*, 2010) and sororin has been shown to co-IP SA2 *in vitro* (Zhang and Pati, 2015). Enrichment of the FGF-like motif proteins was increased with SA1 compared to SA2, suggesting that sororin may interact more strongly with SA1 in human cells. However, sororin was not detected in any of the full lane SA1 IP-MS experiments. The predicted molecular weight of sororin is 28 kDa, which is much smaller that the lowest band cut in the SA2 banded IP-MS (~65kDa). Thus, it would be interesting to probe an SA2 IP for sororin and to assess any role of the potential interaction in cohesion and WAPL antagonism.

Interaction of SA1 with SAF-A and the FGF-like protein HNRNPUL2 (SAF-A2) points to further links between cohesin and phase separation. SAF-A oligomerizes with chromatin associated RNAs to form condensates that are required for open chromatin conformation at active genes (Nozawa *et al.*, 2017; Fan *et al.*, 2018). SAF-A contains intrinsically disordered regions that may initially seed these condensates by locally concentrating multiple molecules of SAF-A (Nozawa *et al.*, 2017; Michieletto and Gilbert, 2019). A functional role for cohesin in this process was suggested as SAF-A, RAD21, and CTCF were identified in reciprocal co-IPs and depletion of SAF-A reduced RAD21 binding at 3816 sites and increased RAD21 binding at 535 sites by ChIP-seq (Fan *et al.*, 2018). This thesis determines that SA1 interacts with SAF-A in the presence or absence of RAD21, perhaps suggesting that SA1 interacts with SAF-A to subsequently localise RAD21 and the cohesin ring. As SA1 also contains intrinsically disordered regions, it may also contribute to the formation of SAF-A-mediated condensates. Bridging-induced condensate formation by yeast cohesin *in vitro* suggests that multivalency within the cohesin complex may also contribute to this phase seperation. Multiple HNRN proteins have been shown to form phase separated condensates similar to SAF-A (Molliex *et al.*, 2015; Ryan *et al.*, 2018; Batlle *et al.*, 2020). In this thesis SA1 is shown to interact with SAF-A,

HNRNPUL2, and perhaps HNRNPD, suggesting a functional link between SA1 and the HNRN proteins, perhaps to organise chromatin into condensates.

The biological processes enriched in the SA1 +/-cohesin interactomes overlapped with published R-loop interactomes, suggesting functional interaction of SA1 and R-loops. In addition, the proteins TAF15, FANCI, HNRNPD, SSRP1, and INO80 that were highly enriched in the SA1 interactome in auxin conditions have known roles in R-loop biology (Britton *et al.*, 2014; Herrera-Moyano *et al.*, 2014; Alfano *et al.*, 2019; Liang *et al.*, 2019; Prendergast *et al.*, 2020). In fact, there was highly significant overlap of the different interactomes and both SA1 and SA2 specifically co-IP with R-loops. RNase H1 overexpression is a routinely used method to modulate R-loop levels in growing cells, however, in the HCT116 RmAC OsTIR1 cells it had little effect on enrichment of proteins with s9.6 IP. A possible reason for this is activity of RNase H1 predominantly in the nucleolus, as observed in Hela cells previously (Shen *et al.*, 2017). The requirement of extensive digestion of purified chromatin with the NEB RNaseH alternatively suggests that R-loops are particularly stabilised in these cells. Dr. Yang Li confirmed interaction of SA and R-loops by IF and further identified a feedback loop between the abundance of R-loops in individual cells and SA levels on chromatin (Porter *et al.*, 2021). As both ribosome biogenesis and mRNA processing factors are enriched in the SA1 interactome, this regulatory feedback loop could be at play with R-loops in both the nucleolus and nucleoplasm of cells. If cohesin ring binding also plays a role in the suppression of R-loops, reduced levels of RAD21 compared to un-tagged isogenic cells may account for the seemingly stabilised R-loops observed.

It remains unclear if the interaction of SA and R-loops is mediated by association with R-loop binding proteins or by binding of the SA proteins to a specific nucleic acid component of the R-loop. Interaction of cohesin with RNA-binding proteins has previously been shown to stabilize interaction with CTCF specifically at the *IGH2/H19* locus, suggesting an importance for the protein interactions in the final stabilization of cohesin binding, at least (Yao *et al.*, 2010). In collaboration with Professor Richard Jenners lab it was determined that both SA1 and SA2 can bind to RNA. Binding of the SA proteins to RNA and R-loop structures *in vitro* has recently been shown by an additional group, validating the results shown in this

thesis (Pan *et al.*, 2020). Importantly, SA was shown to interact with endogenous R-loops in this thesis, confirming the recombinant work shown in this paper. Nucleic acid binding domains of SA could recognize the RNA:DNA hybrid itself, the displaces ssDNA, or RNA that is resolved from the R-loop for processing to mRNA, or a combination of these elements. To investigate whether the SA proteins interact with R-loops via a specific component of the R-loop, specific RNA or DNA binding domains would need to be identified within the SA proteins, which could then be mutated and tested for effect on R-loop binding.

As discussed in the introduction Chapter 1, cohesin plays a role in DNA damage repair, MCM2-7 localises loading of cohesin during S-phase, and R-loops are enriched at sites of head-on transcription and replication collision (Birkenbihl and Subramani, 1992; Bauerschmidt *et al.*, 2009; Hamperl *et al.*, 2017; Zheng *et al.*, 2018). Together these different studies might suggest that MCM2-7 and the presence of ssDNA may help target cohesin to R-loops to help correct the DNA damage ensued by collision of transcription and replication machinery. This thesis shows that, alternatively, SA proteins can directly detect R-loop structures, allowing it to localise to all R-loop structures in the nucleus. As R-loop structures occur at sites of transcription – of both mRNA and rDNA – replication, and DNA damage, binding of SA to R-loops could help to localise cohesin to a wide variety of genomic locations throughout the nucleoplasm and nucleolus (El Hage *et al.*, 2010; Chakraborty and Grosse, 2011; Skourti-Stathaki, Proudfoot and Gromak, 2011; Sollier *et al.*, 2014; Yang *et al.*, 2014; Schwab *et al.*, 2015; Salas‐Armenteros *et al.*, 2017; Yasuhara *et al.*, 2018).

Supercoiling and chromatin structure can influence transcription elongation rates, which in turn impacts on co-transcriptional RNA processing (Bentley, 2014; Ma and Wang, 2016). Hence, loading of cohesin at R-loops may regulate gene expression by slowing transcription by increasing chromatin contacts downstream of RNA polymerase or by speeding up transcription by recruiting topoisomerase to resolve topological tangles and coils (Uusküla-Reimand *et al.*, 2016; Orlandini, Marenduzzo and Michieletto, 2019). As SA1 can regulate R-loop levels, and R-loop themselves are regulated by RNA processing, a complex nexus of feedback loops are at play at sites of transcription, likely allowing tighter regulation. As discussed in the introduction section 1.5.5, compound protein

interactions and DNA structure are required for cohesin loading during S phase. Hence, I hypothesised that R-loops represent archetypal cohesin loading sights and investigated the role of the SA proteins and R-loops in cohesin loading in Chapter 5.

The diverse nucleic acid binding abilities and protein interactors of the SA proteins described here may help to explain conflicting mechanisms proposed for SA-mutation-mediated tumourigenesis. The conflicting mechanisms range from aneuploidy to gene expression misregulation. SA2 mutation in GBM and colon cancer cell lines and knock-out of SA1 in mouse fibroblasts result in aneuploidy, suggesting that the role of the SA proteins in cohesion is most important for tumourigenesis (Solomon *et al.*, 2011; Remeseiro, Cuadrado, Carretero, *et al.*, 2012). Solomon et al. (2011) further identified similar expression profiles between the SA2-proficient and deficient paired GBM cell lines, suggesting that interphase roles in gene expression regulation were not involved in tumourigenesis. In contrast, analysis of 17 patient bladder tumours found that loss-of-function SA2 mutation was predominantly in genomically stable samples (Balbás-Martínez *et al.*, 2013). Similarly, of 19 AML patient tumours carrying cohesin mutations, only 1 showed aneuploidy (Welch *et al.*, 2012).

These opposing findings may reflect the limitations of *in vitro* techniques, in which, lack of interplay with the tumour microenvironment *in vivo* can lead to misinterpretation of tumourigensis. Alternatively, distinct mutations may each affect a particular aspect of SA function (i.e. interaction with the cohesin ring, DNA-binding, CES-binding protein interaction), resulting in diverse tumorigenesis mechanisms and conflicting reports of tumourigenesis. Congruently, Kim et al. (2016) determined that not all nonsense tumour-derived SA2 mutations effect interaction with the cohesin ring proteins, yet all nonsense mutations result in defective cohesion. Hence, SA-mediated tumourigenesis is defined by more than just interactions with the cohesin complex and the full range of SA activities, cohesin-dependent and -independent, must be fully understood to determine their roles in cancer.

5

SA can load cohesin independently from the NIPBL-MAU2 loader complex

5.1 Introduction

As discussed in section 1.4.2, the crystal structure of NIPBL and SA are strikingly similar, in that both are highly bent, HEAT-repeat proteins (Hara *et al.*, 2014; Kikuchi *et al.*, 2016; Chao *et al.*, 2017). As the structure of proteins designates their function, I hypothesised that, like NIPBL, the SA proteins may also be capable of inducing conformational change in the cohesin ring structure and inducing loading onto chromatin. Additional evidence in the literature supported the idea that SA proteins contribute to cohesin's association with chromatin. In yeast, interaction of the SA orthologue with the loader complex is required for efficient association of the cohesin ring with DNA and subsequent ATPase activation (Murayama and Uhlmann, 2014; Orgil *et al.*, 2015). Separating interactions into SA-loader and cohesin ring-loader subcomplexes still impairs cohesin loading, indicating that SA functions as more than just a bridge protein between the cohesin ring and NIPBL (Orgil *et al.*, 2015).

To investigate the role of the SA proteins in loading of cohesin, I devised an experiment to test loading of cohesin in the presence or absence of the loader complex and the SA proteins. Importantly, using HCT116 RmAC OsTIR1 cells, auxin-mediated knockdown of RAD21 is reversible and by simply washing the cells with PBS and incubating them in new auxin-free media, cohesin levels on chromatin can be recovered. Hence, proteins of interest can be knocked down in the cells using siRNA prior to auxin treatment and removal, allowing investigation of their contribution to RAD21 loading onto chromatin (Figure 64). In this way,

cohesin loading was assessed in the absence of the cohesin loader, the SA proteins, and the cohesin loader plus the SA proteins, allowing quantification of the contribution of these proteins to cohesin loading.

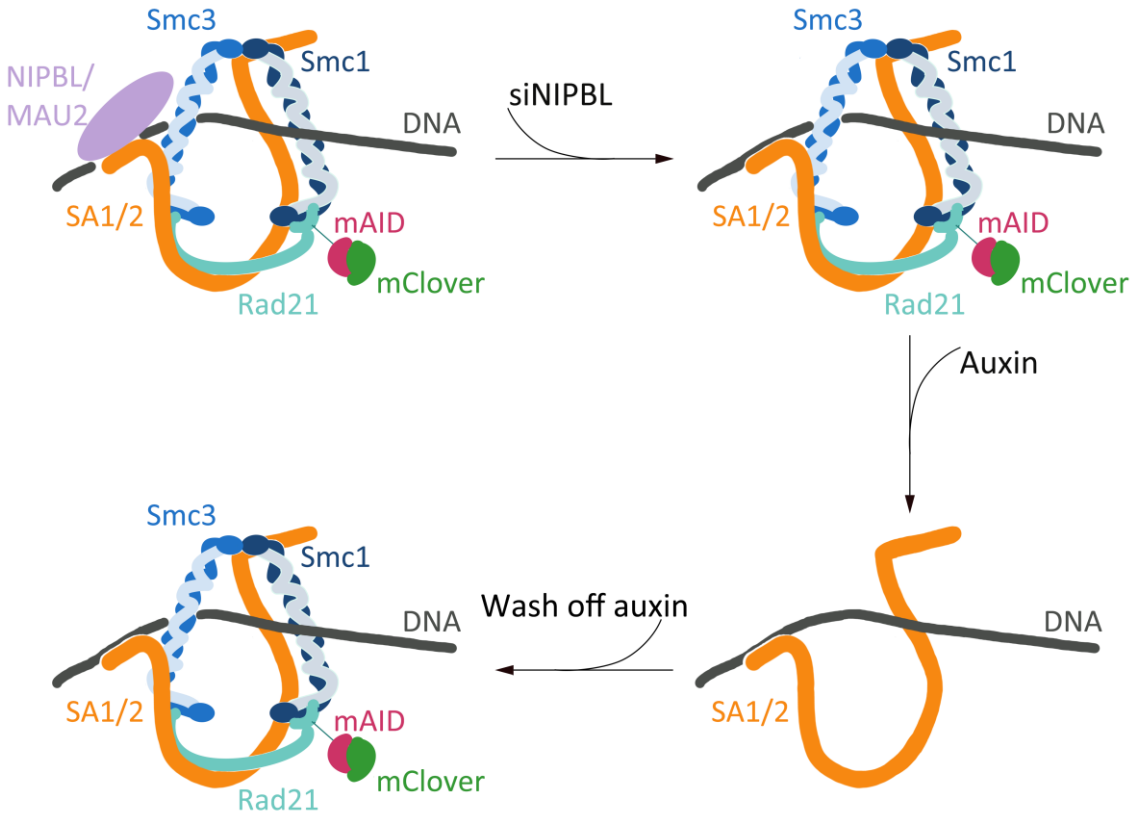


Figure 64: Schematic of reloading experiment to test for role of SA proteins in association of cohesin ring with chromatin. HCT116 RmAC OsTIR1 cells are treated with siRNA to NIPBL for 72 hrs to knockdown the NIPBL/MAU2 loader complex. Already loaded cohesin may remain on chromatin at this point. However the cells are treated with auxin at the 64 hr timepoint, 4 hrs afterwhich the cohesin ring will be depleted from chromatin except for the SA proteins. The auxin-containing media can be washed off and replaced with fresh media. Under these conditions the cohesin complex can reform and in the presence of NIPBL/MAU2 can reload onto chromatin (Rao *et al.*, 2017a; Yesbolatova *et al.*, 2019). Here the experiment will ask if such reloading can occur in the absence of the loader complex.

The results described in Chapter 4 indicate that SA proteins can interact with a variety of proteins, R-loops, and RNA, in the presence or absence of RAD21. Cohesin loading during S-phase is mediated by interaction with the MCM2-7 complex and capture of single-strand DNA at the replication fork (Hara *et al.*, 2014; Kikuchi *et al.*, 2016; Murayama *et al.*, 2018; Zheng *et al.*, 2018). Taking all these results together, I hypothesised that SA proteins may induce loading of cohesin at R-loops, which could act as a structural alternative to the replication fork. Hence, R-loop nuclease and helicase proteins were knocked down using

siRNAs to increase R-loops levels in cells and assess impact on SA proteins and cohesin loading in the absence of the canonical cohesin loader.

5.2 Results

5.2.1 Optimising detection of NIPBL by western blot and investigating the role of NIPBL and MAU2 in cohesin loading

To assess a role for the SA proteins in cohesin loading the canonical NIPBL-MAU2 loader complex first needed to be depleted from cells. NIPBL is a large 320 kDa protein that is difficult to detect by conventional immunoblotting and immunofluorescent techniques. To detect NIPBL, tris-acetate gels were used as they are specialized for the separation of high molecular weight proteins during gel electrophoresis. By maintaining a pH of 8.1, Tris-acetate gels help to prevent protein modification and ensure sharp bands. In addition, transfer conditions were altered to maximize protein transfer to the membrane and prevent precipitation of large proteins. This work was started in Hela and HCT116 cells before the HCT116 RmAC OsTIR1 cells were obtained.

Hela cells were treated with non-targeting control pool siRNA (siCon), 10nM of siNIPBL and siMAU2, or 50nM of siNIPBL and siMAU2 and chromatin samples were run on a Tris-Acetate gel (Figure 65A). Transfer and detection of NIPBL was successful and showed that treatment with 10nM siNIPBL and siMAU2 reduced NIPBL levels and treatment with 50nM siNIPBL and siMAU2 removed all detectable NIPBL. To confirm NIPBL depletion was affecting cohesin loading, the blot was stripped and re-probed for SMC3. SMC3 levels were reduced with 10nM siRNA treatment compare to control, however, much greater loss was observed with 50nM siRNA treatment. The residual SMC3 protein observed may represent a fraction of stably bound cohesin or indicate that some cohesin loading is still occurring in these cells, either by very low amounts of residual NIPBL or by another protein capable of inducing cohesin loading onto chromatin.

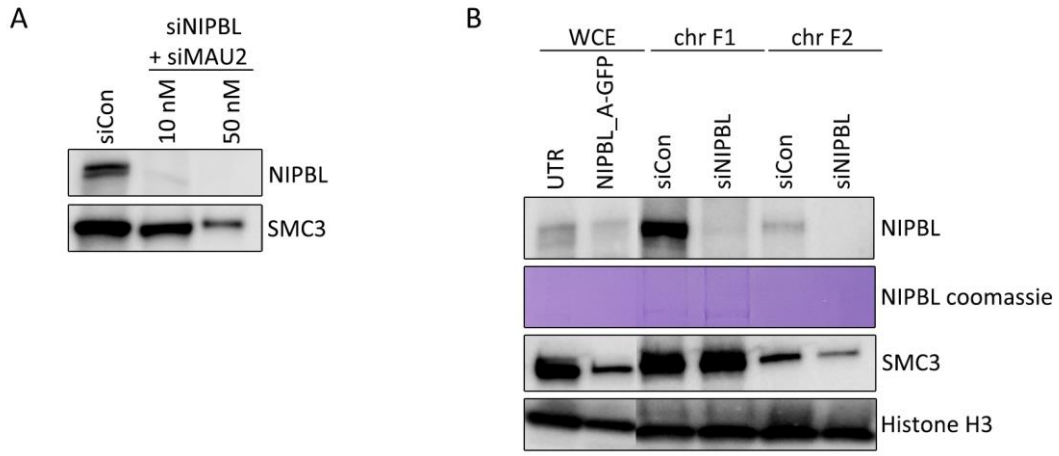


Figure 65: Detection of NIPBL knockdown by western blot analysis. (A) Detection of NIPBL on chromatin in Hela cells. Using a NuPAGE™ 3-8% Tris-Acetate gel, NIPBL could be detected by western blot following O/N wet transfer. Detection of NIPBL was confirmed by treatment with 10 or 50nM siNIPBL + siMAU2. SMC3 was also probed to assess effect of loader complex knockdown on cohesin levels on chromatin. (B) Assessment of the salt conditions required to solubilise NIPBL from chromatin in Hela cells. NIPBL levels in siCon and siNIPBL samples were tested from chromatin solubilised in two rounds. First chromatin was solubilised in low-salt buffer (200mM) and benzonase treatment (chr F1). Insoluble material from this fraction was further solubilised in high-salt buffer (500mM) and sonication (chr F2). WCE from cells that were UTR or over-expressing GFP-tagged NIPBL isoform A (NIPBL_A-GFP) were included as a positive control for the presence of NIPBL. Histone H3 was blotted as a loading control. Extraneous lanes are cropped from the images (between WCE and chr F1 lanes).

To determine if treatment with siNIPBL alone could also result in depletion of NIPBL and removal of SMC3 from the chromatin, Hela cells were treated with 50nM siCon or 50nM siNIPBL. Western blotting was repeated as above with running of the gel shortened from 2hrs 30mins to 1hr 45mins, to keep NIPBL in a lower percentage portion of the gel, from which transfer should occur more readily. Whole cell extract (WCE) from Hela cells and from Hela cells transfected with NIPBL tagged with GFP (obtained from Lena Strom (Bot *et al.*, 2017)) were included as positive controls for the presence of NIPBL. The chromatin sample was sequentially produced in two fractions; Fraction 1 (F1), isolated by low salt buffer and benzonase digestion and Fraction 2 (F2), isolated by high salt buffer and sonication.

NIPBL levels were strongest in the low salt chromatin fraction, suggesting a relatively weak association with chromatin (Figure 65B). 50nM siNIPBL treatment reduced NIPBL levels on chromatin in both fractions, with only a very faint NIPBL band left in the siNIPBL F1 sample. Coomassie staining of the NIPBL gel following transfer revealed that transfer was predominantly efficient with only a band slightly below NIPBL still visible from the siCon and siNIPBL F1 samples.

This band was slightly stronger in the siNIPBL sample compared to the siCon sample, however, the difference in transfer efficiency between the two lanes did not account for the difference in antibody-detected signal observed. Immunoblotting of SMC3 showed loss of SMC3 from chromatin only from the F2 fraction (Figure 65B). This suggests that two pools of cohesin exist on chromatin – a NIPBL-dependent pool that associates more strongly with chromatin and a NIPBL-independent pool that associates more weakly with chromatin. To observe correlation more easily between NIPBL and SMC3 levels this split fractionation was not used again. Instead, chromatin was collected in one high salt fraction, in which, all proteins released by the lower salt buffer should be captured anyway.

Knockdown of NIPBL, MAU2, and NIPBL and MAU2 was tested in HCT116 cells to determine the effect of knockdown of each of the loader complex subunits alone or in combination. Nuclear soluble material was run on the western blot alongside chromatin material to determine if any effects of knockdown could also be detected in this fraction of the nucleus. However, no transfer of proteins larger than 250 kDa was evident for the nuclear soluble fraction making assessment of NIPBL levels impossible from this fraction. On chromatin, NIPBL levels were reduced compared to Hela cells, however, they were still detectable by switching antibody from the Abbiotec rat monoclonal NIPBL antibody to the Bethyl rabbit polyclonal NIPBL antibody (Figure 66A). NIPBL was lost in all three knockdown conditions. Coomassie staining of the corresponding NIPBL gel after transfer showed that all protein was successfully transferred to the blot from all the lanes. Similarly, MAU2 levels in both the nuclear soluble and chromatin fractions were reduced in all three knockdown conditions. The nuclear soluble UTR samples had to be covered to avoid strong ladder signal overwhelming the image. This reciprocal loss of NIPBL and MAU2 supports previous publications that demonstrate co-stabilisation of the loader complex subunits (Watrin *et al.*, 2006; Hinshaw *et al.*, 2015). Hence, in future experiments MAU2 was used as an alternative protein to assess NIPBL levels in the absence of reliable NIPBL signal.

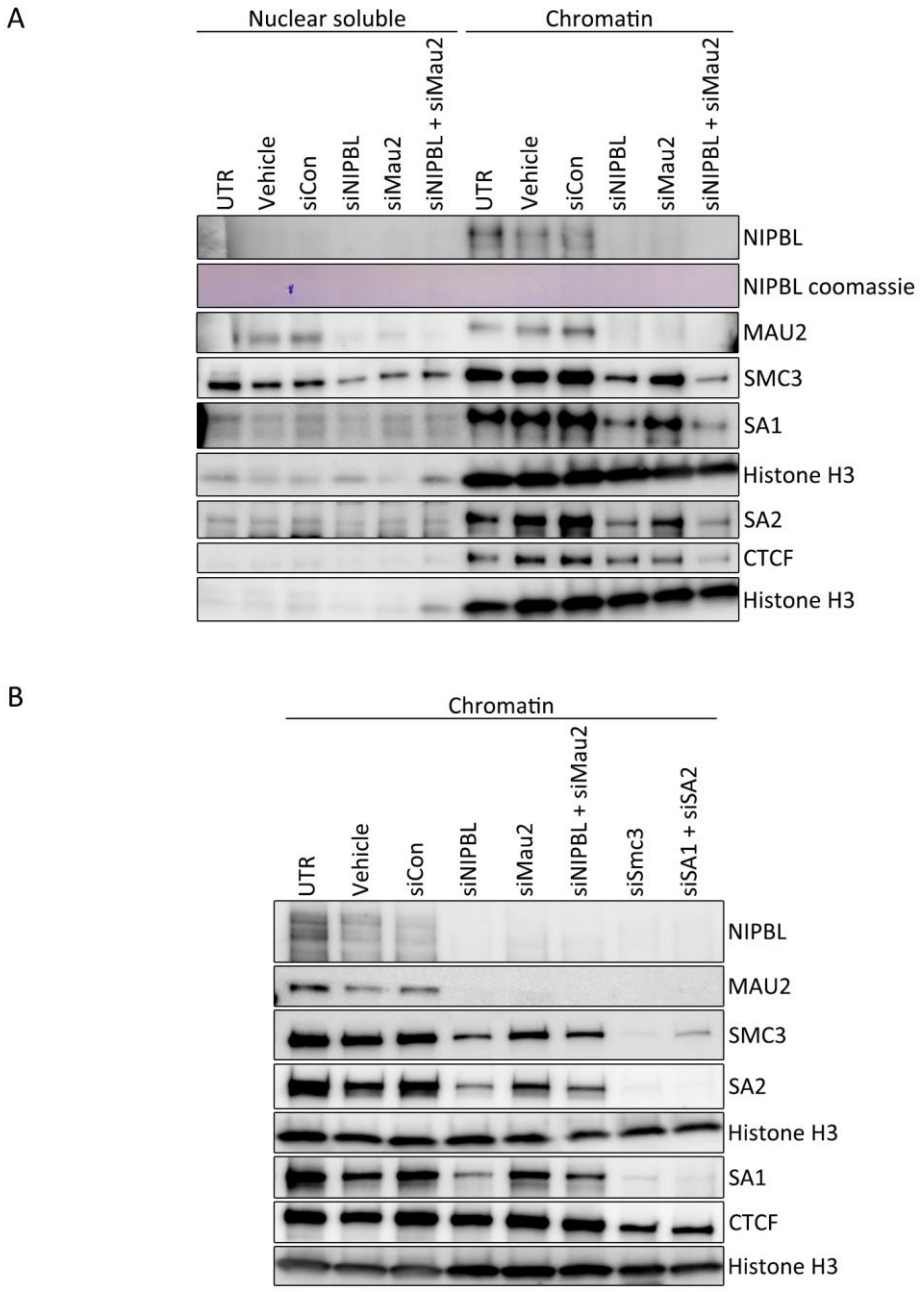


Figure 66: NIPBL and MAU2 knockdown differentially effect cohesin levels on chromatin. (A) Assessment of NIPBL, MAU2, and NIPBL + MAU2 knockdown with 50nM siRNA in HCT116 cells. Scramble siRNA (siCon), lipofectamine-only (Vehicle), and UTR samples were included as negative controls. Nuclear soluble and chromatin fractions are shown. Histone H3 was blotted as a loading control. (B) Biological repeat of (A) with siSMC3, and siSA1 + siSA2 samples included. WCE = Whole cell extract.

Interestingly, despite equal loss of NIPBL, cohesin levels on chromatin were higher in the siMAU2 sample than in the siNIPBL and siNIPBL + siMAU2 samples (Figure 66A). This was true for SMC3, SA1, and SA2 components of cohesin. CTCF levels were also reduced with knockdown of the cohesin loader complex, an unexpected result given CTCFs intrinsic ability to bind to DNA via its zinc finger repeat structure. In mouse erythroleukemia cells CTCF has been pulled down

with NIPBL, in fact with more efficiency than SMC3, suggesting a functional interaction between the two proteins (Lee *et al.*, 2017). SMC3 was the only cohesin component that was robustly detectable in the nuclear soluble fraction. As on chromatin, reduction of SMC3 in the nuclear soluble fraction was also seen in all three knockdown conditions. Thus, cohesin may interact with the loader complex before association with chromatin. Alternatively, knockdown of NIPBL and MAU2 may affect the overall level of proteins in the cells due to stress or cell cycle effects. WCE samples and cell cycle analyses would need to be generated to assess such phenomenon.

The experiment was repeated to determine reproducibility of these results (Figure 66B). In the biological repeat, CTCF levels were not as strongly reduced with knockdown of the loader complex, however, SMC3 and SA levels were still differentially affected by siMAU2 treatment compared to siNIPBL treatment. siSMC3 and siSA1 + siSA2 samples were also included in this experiment to assess the effects of reduction of each of these components of the cohesin complex. Reciprocal loss of SMC3 and the SA proteins was evident, as well as loss of NIPBL and MAU2 from chromatin. However, these samples had high amounts of cell death and showed abnormal cell physiologies indicating that cell stress, death, and cycle effects may have confounded the results from these two samples.

Cohesin levels on chromatin were equally affected by knocking down NIPBL alone as knocking down NIPBL and MAU2 together. Hence, for the reloading experiments, the canonical NIPBL-MAU2 loader was knocked down using siNIPBL only in order to minimise the amount of siRNA added to the cells. Moreover, retention of cohesin on chromatin with MAU2 knockdown suggested that, while it plays an important role in stabilisation of NIPBL, MAU2 loss has a confounding effect on association of cohesin with chromatin. In the absence of MAU2, de-stabilised NIPBL may still be able to catalyse cohesin loading and thus, it is most important to ensure no NIPBL is present in the cells. Alternatively, this retention may also be mediated by an unknown protein that associates with NIPBL-MAU2 but is differentially affected by single knockdown of the two proteins.

Lower NIPBL levels in HCT116 cells compared to Hela cells meant that higher amounts of protein needed to be run on the western blots to detect it. While the Bethyl NIPBL antibody was slightly better at detecting lower amounts of NIPBL, samples could sometimes be too dilute to fit the amount of protein needed for robust signal into the gel wells. To mitigate this issue, three changes were made to try to maximise signal while reducing protein concentration: 1) sample buffer was switched from Bio-Rad Laemmli Sample Buffer to NuPAGE™ LDS Sample Buffer, which is the recommended sample buffer for use with the NuPAGE™ Tris-acetate gels. While both sample buffers contain anionic detergents, namely, SDS and LDS, respectively, LDS may be better suited to the NIPBL transfer set-up as it does not precipitate at low temperature; 2) NuPAGE™ Sample Reducing Agent was added to the protein samples alongside the LDS sample buffer, this agent helps to keep reduced proteins from reoxidising during electrophoresis; and 3) NuPAGE™ Antioxidant was added during electrophoresis and transfer buffers to maintain proteins in a reduced state and prevent reoxidation and to enhance transfer of proteins to the membrane. Altogether these changes helped to enhance NIPBL signal on western blots from HCT116 cells, meaning that detection could now be achieved using either antibody. As the Bethyl antibody gave a double band that could sometime be quite smeary, the Abbiotec NIPBL antibody was used where possible.

5.2.2 HCT116 RmAC OsTIR1 siRNA optimisation experiments

To safeguard cell viability, the lowest possible concentration of siRNA-mediated knockdown of NIPBL and SA1 and SA2 was optimised in the polyclonal HCT116 RmAC-OsTIR1 cells. To determine optimal NIPBL siRNA conditions, cells were treated with 5, 10, 25, or 50nM of non-targeting siControl or siNIPBL and collected at 72 or 96hrs post-transfection (Figure 67A). Two 10cm plates were set up for the 50nM siNIPBL sample, and one 10cm plate set up for all other samples. In addition, two extra plates were treated with 25nM siControl or siNIPBL, incubated for 24hrs, and then treated with 25nM siControl or siNIPBL again. These two samples were then incubated for 48hrs and collected alongside the 72hr samples. A sample treated with lipofectamine only (Vehicle) was set up and collected 72hrs post-transfection to act as a control for treatment with lipofectamine for 72hrs. A

vehicle sample was not set up for the 96hr time point as this had been included in previous experiments and found not to differ from untreated (Figure 66 A & B).

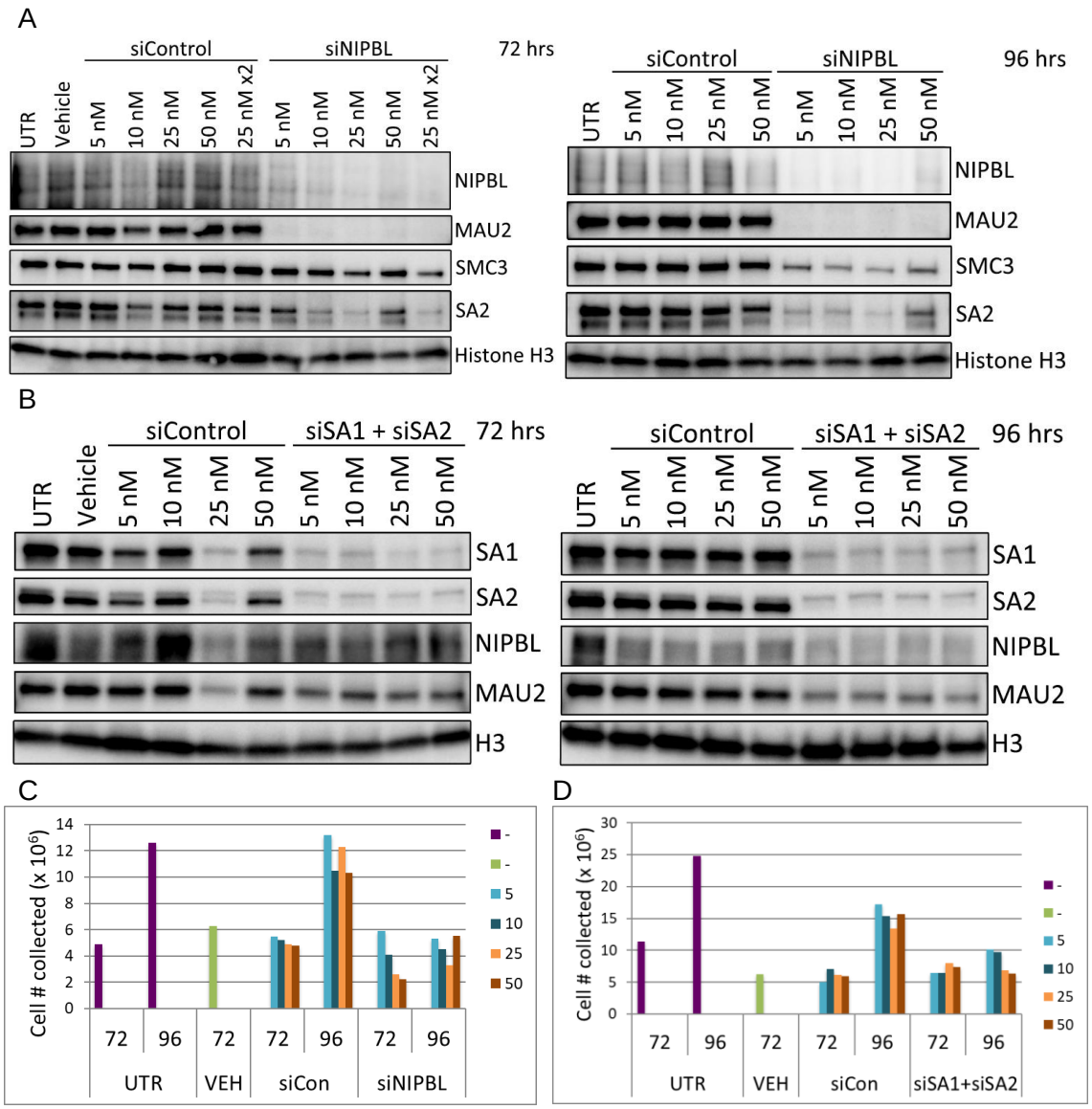


Figure 67: Optimisation of siRNA transfection conditions. (A) HCT116 cells were treated with a titration of scramble siRNA (siControl) or siNIPBL for 72hrs (left) or 96hrs (right). Additional samples were treated with 25nM siControl or siNIPBL, incubated for 24hrs, and then treated with 25nM siControl or siNIPBL again (25 nM x 2). Lipofectamine-only (Vehicle) and UTR samples were included as controls. (B) Repeat of (A) with siSA1 + siSA2 tested in place of siNIPBL. (C) Bar graph of the number of cells collected per sample in (A). The colour legend indicates the nM siRNA treated. VEH = Vehicle. (D) As in (C) but showing the number of cells collect in (B). Histone H3 was blotted as a loading control. UTR = Untreated; H3 = Histone H3.

Treatment with 10nM or more siNIPBL removed NIPBL from chromatin by 72hrs (Figure 67A). However, 96hrs of siRNA knockdown was required to see efficient reduction of cohesin (SMC3) from chromatin, potentially due to residual stable cohesin that can remain bound to chromatin. Confounding cell cycle effects may also have contributed to the reduction in SMC3 levels at 96hrs, as cell number

did not double between the 72 and 96hrs siNIPBL samples, as is did for siControl samples (Figure 67C).

The same experiment set up was also used to assess optimal siRNA conditions for simultaneous knockdown of SA1 and SA2 (Figure 67B). In this case, treatment with a minimum of 5nM siRNA was required to observe loss of SA1 and SA2 from chromatin by 72hrs. Thus, 10nM was chosen as the optimal siRNA concentration for all knockdown experiments. As for siNIPBL, treatment with siSA1 and siSA2 reduced cell number by 96hrs compared to siControl-treated samples (Figure 67D). This reduction may be due to increased cell death or cell cycle arrest, as mentioned above. NIPBL and MAU2 levels were reduced in siSA1 + siSA2 treated cells compared to UTR cells, at both 72 and 96 hrs post-transfection. However, variable NIPBL levels in siCon samples make this reduction difficult to interpret as a direct effect.

Given that effect on cohesin levels following knockdown of NIPBL were not observed until 96hrs post-transfection with siRNA, knockdown of NIPBL followed by auxin-mediated degradation of RAD21 at 72 and 96hrs post-siRNA-transfection was tested to determine how both treatments together would affect cell survival and cohesin levels on chromatin (Figure 68). NIPBL was efficiently knocked down at 72 and 96hrs post-transfection. Auxin treatment further reduced RAD21 levels on chromatin over siNIPBL treatment alone, however, RAD21 was not completely lost in any of the samples. Similarly, SMC3 was also observed on chromatin after only 72hrs of siNIPBL treatment. At this point, the cells were FACS sorted for better RAD21 loss with auxin treatment, as discussed in section 3.2.1. In the interim, due to worries of altered cell cycle at the 96hr time point, the 72hr timepoint was chosen and the following experiments were tested first with reduced RAD21 in the polyclonal cells line and subsequently with complete loss of RAD21 in the sorted clones. siNIPBL + auxin was used to test: i) the effect of NIPBL knockdown on CTCF-SA interaction in the absence of RAD21 (section 5.2.3) and ii) effect of NIPBL knockdown on cohesin reloading on to chromatin following withdrawal of auxin-mediated depletion (section 5.2.4).

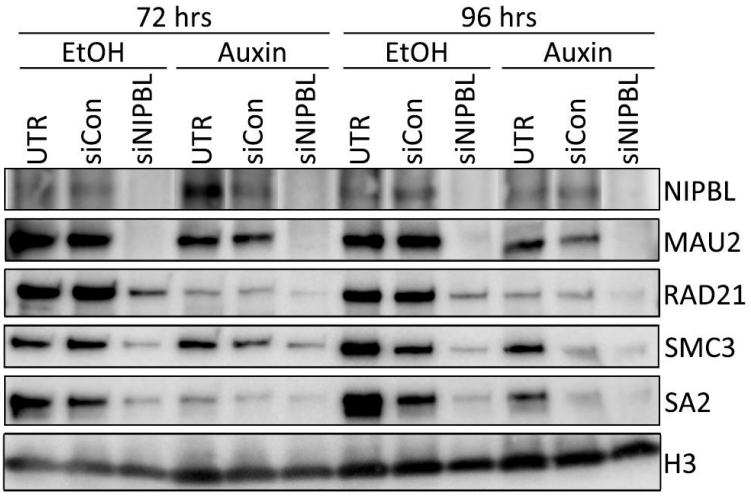


Figure 68: Assessment of siNIPBL and auxin co-treatment. HCT116 RmAC OsTIR1 cells were treated with the indicated siRNA for 72 or 96hrs. Prior to collection, ethanol or auxin was added to media of the indicated samples for 4 hrs. Effect on the levels of the loader complex and cohesin members on cohesin was probed. N.B. Lanes in the NIPBL blot are cut and pasted together to match the loading order of the other blots. The full NIPBL blot is shown in Supplemental Figure 6 A. Histone H3 was blotted as a loading control. EtOH = ethanol.

5.2.3 Interaction of SA1 with CTCF in the absence of RAD21 is not dependent on NIPBL

To assess if the SA proteins required NIPBL to interact with CTCF in RAD21 depleted cells, effect of NIPBL knockdown on CTCF-SA interaction in auxin-treated cells was tested. HCT116 RmAC OsTIR1 cells were treated with either siControl or siNIPBL for 72hrs as optimised above and prior to collection cells were treated with ethanol or auxin for 4hrs.

Two biological replicate IP experiments were run, the first in the polyclonal RmAC OsTIR1 cells (Figure 69A) and the second in the H2 FACS sorted clone (Figure 69B). Chromatin from the polyclonal cells was solubilised for IP using the 6U benzonase conditions as in section 3.2.3 and chromatin from the H2 clonal cells was solubilised using the 85U benzonase condition as in section 3.2.7. RAD21 and NIPBL loss was confirmed by loss of input signal in auxin- and siNIPBL-treated samples, respectively. RAD21 depletion could not be assessed from the polyclonal cells due to technical issues with the western blot, however, assuming depletion similar to other experiments in the polyclonal cells, RAD21 should have been reduced if not completely ablated. NIPBL knockdown was confirmed only in the SA1 IP samples from the polyclonal cells, again due to technical issues with the CTCF western blot.

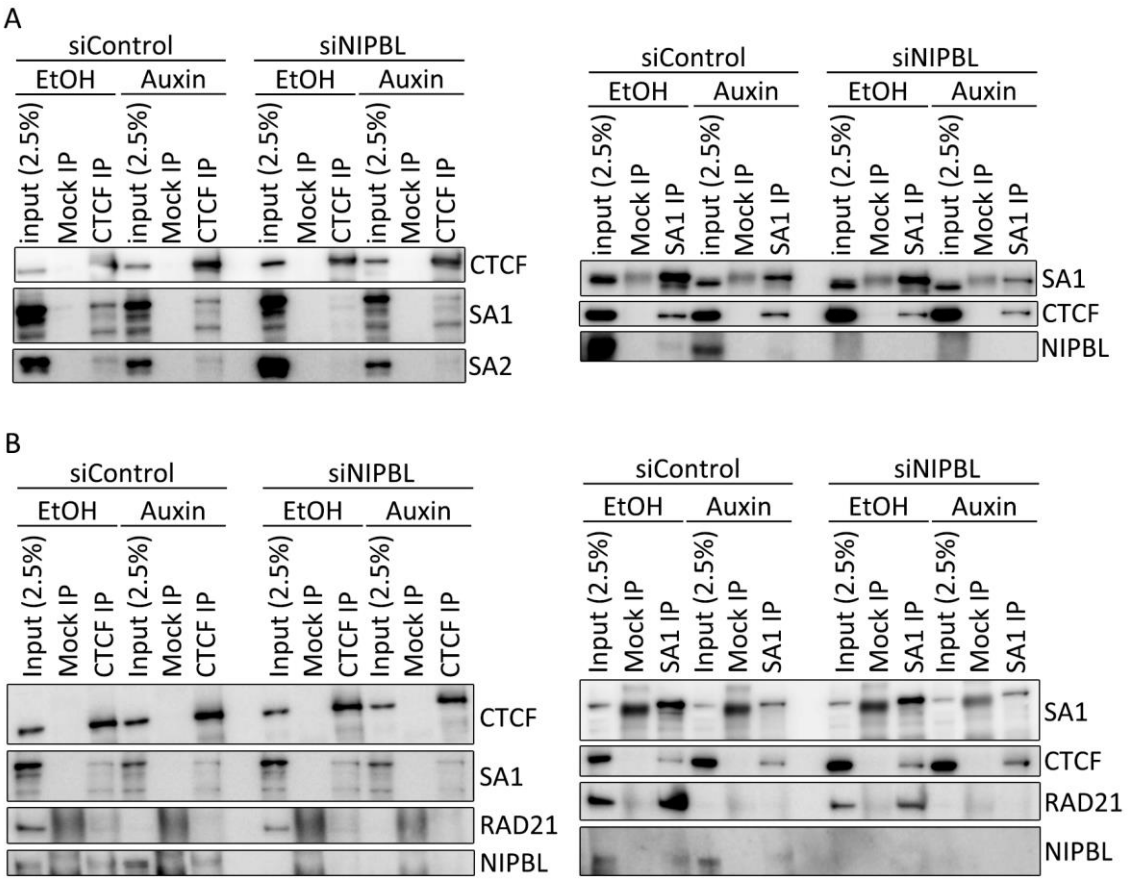


Figure 69: SA1 can interact with chromatin and CTCF independently of NIPBL. HCT116 RmAC OsTIR1 polyclonal (A) or H2 (B) cells were treated with scramble siRNA (siControl) or siNIPBL for 72hrs. Prior to collection cells were treated with ethanol or auxin for 4hrs. Chromatin was solubilised as for previous IPs with 6U benzonase (A) or 85U benzonase (B) per 100 x 10^6 cells. Chromatin was IP'd with endogenous antibodies for CTCF (left) or SA1 (right) or corresponding species matched IgG (Mock). NIPBL was immunoblotted on membranes from separately run gels optimised for high molecular weight proteins as described in section 5.2.1.

CTCF IP appeared even across the different treatments in both experiments. Although evaluation from the polyclonal experiment was difficult as the siControl ethanol CTCF IP sample badly burned the membrane and both siNIPBL CTCF IP bands slightly greyed, indicating substrate depletion due to an excess of HRP in the bands. SA1 was immunoblotted for co-IP with CTCF in both experiments and was detected in all IP samples except for the siNIPBL ethanol CTCF IP from the polyclonal cells, perhaps due to an unidentified technical issue. SA1 co-IP in siControl ethanol and auxin conditions was as expected with the 6U benzonase used for the polyclonal cells, however, co-IP was not increased as expected with the 85U benzonase used for the clonal cells. Despite low enrichment, co-IP levels of SA1 was similar in siControl and siNIPBL auxin CTCF IPs, indicating that NIPBL is not essential for interaction of SA1 with CTCF in the absence of RAD21. SA2 was immunoblotted only in the CTCF IP from the polyclonal cells. Only faint

co-IP signal was observed as expected with the 6U benzonase. This low signal made it impossible to assess complete necessity of NIPBL for CTCF-SA2 interaction, although there may have been a slight decrease in SA2 co-IP in the siNIPBL sample. Interestingly, NIPBL was detected in the siControl auxin IP with 85U benzonase digestion, indicating that NIPBL can interact with CTCF and SA in the absence of RAD21 even if it is not required for the interaction.

Reciprocal co-IP of CTCF with SA1 confirmed independence of CTCF-SA1 interaction in the absence of NIPBL. In both experiments, enrichment of SA1 over input was low compared to previous experiments. Enrichment was low in both siControl and siNIPBL samples indicating an unknown effect of the siRNA transfection that results in reduced IP of SA1. Auxin treatment reduced SA1 input and IP levels and, as expected, this reduction was increased in the clonal cells, wherein RAD21 is more efficiently depleted from the cells. In both experiments, CTCF was equally enriched over input for all samples, regardless of differences in SA1 IP levels. Thus, CTCF was still enriched with SA1 in the absence of RAD21 and the CTCF-SA1 interaction does not require NIPBL. Furthermore, SA1 input levels in auxin treated samples was not change in siNIPBL samples compared to siControl, indicating that SA1s ability to interact with chromatin in the absence of RAD21 does not require NIPBL-mediated loading. Altogether these experiments suggested that SA1 can interact with chromatin and CTCF in complex with or independently from NIPBL.

5.2.4 Reloading experiment optimisation

Using the polyclonal HCT116 RmAC OsTIR1 cell line, a timecourse of RAD21 reloading was assessed (Figure 70A). Cells were treated with either siControl or siNIPBL for 72hrs. Prior to collection, cohesin was rapidly removed from chromatin by auxin-mediated degradation of degron-tagged RAD21. Auxin-containing media was then washed off the cells and replaced with auxin-free growth media for the withdrawal times indicated (0, 1, 2, 4, or 20hrs). Successful knockdown of the NIPBL-MAU2 loader complex was observed for all siNIPBL-treated samples, meaning that, any reloading of RAD21 observed in these samples should have occurred independently of the canonical loader complex.

Reassociation of RAD21 with chromatin was observed as quickly as 1hr post-auxin removal, with similar levels of reassociation observed in siControl- and siNIPBL-treated samples. Thus, this experiment suggested cohesin can associate with chromatin independently of NIPBL-MAU2. In the siCon samples, MAU2 and SA2 levels mimic the loss and reassociation of RAD21 with chromatin observed for 0 and 4hr samples, respectively. This suggests that the SA2 and the loader complex require RAD21 to associate with chromatin, possibly as the cohesin complex and loader complex all interact together before association with the chromatin or perhaps suggesting continued interaction of the loader complex with cohesin following loading (Rhodes *et al.*, 2017).

The experiment was repeated to confirm reloading of RAD21 and to test RAD21 levels on chromatin with just 30mins withdrawal of auxin. As the reloading was observed by as little as 1hr in the previous experiment and prolonged knockdown of multiple cohesin components would add stress to the cells, the 20hr timepoint was disregarded. As for the first experiment, all samples could not be run on a single gel so ethanol- and auxin-treated samples were split across two gels, with a few control samples also run on the auxin sample gel to allow comparison of protein quantities. RAD21 degradation was more pronounced in this repeat experiment and reassociation of RAD21 with the chromatin was diminished compared to the first experiment, even in the siCon sample (Figure 70B). In fact, for the siNIPBL samples there was little difference in RAD21 abundance between the 0 and 4hr withdrawal samples. Blotting for MAU2 confirmed that the loader complex was depleted from the cells. As only minor RAD21 reloading was observed no NIPBL gel was run to double check the MAU2 result. An optimisation of the cell number plated was tested for effect on RAD21 reloading in siCon conditions, however, no effect was seen by increasing or decreasing the cell number plated by 50%. Residual RAD21 in the 0hr auxin samples was concerning for interpreting RAD21 signal in auxin withdrawal samples and alongside the issues discussed in section 3.2.1 led to the decision to FACS sort the polyclonal HCT116 cells for the population of highest auxin responding cells.

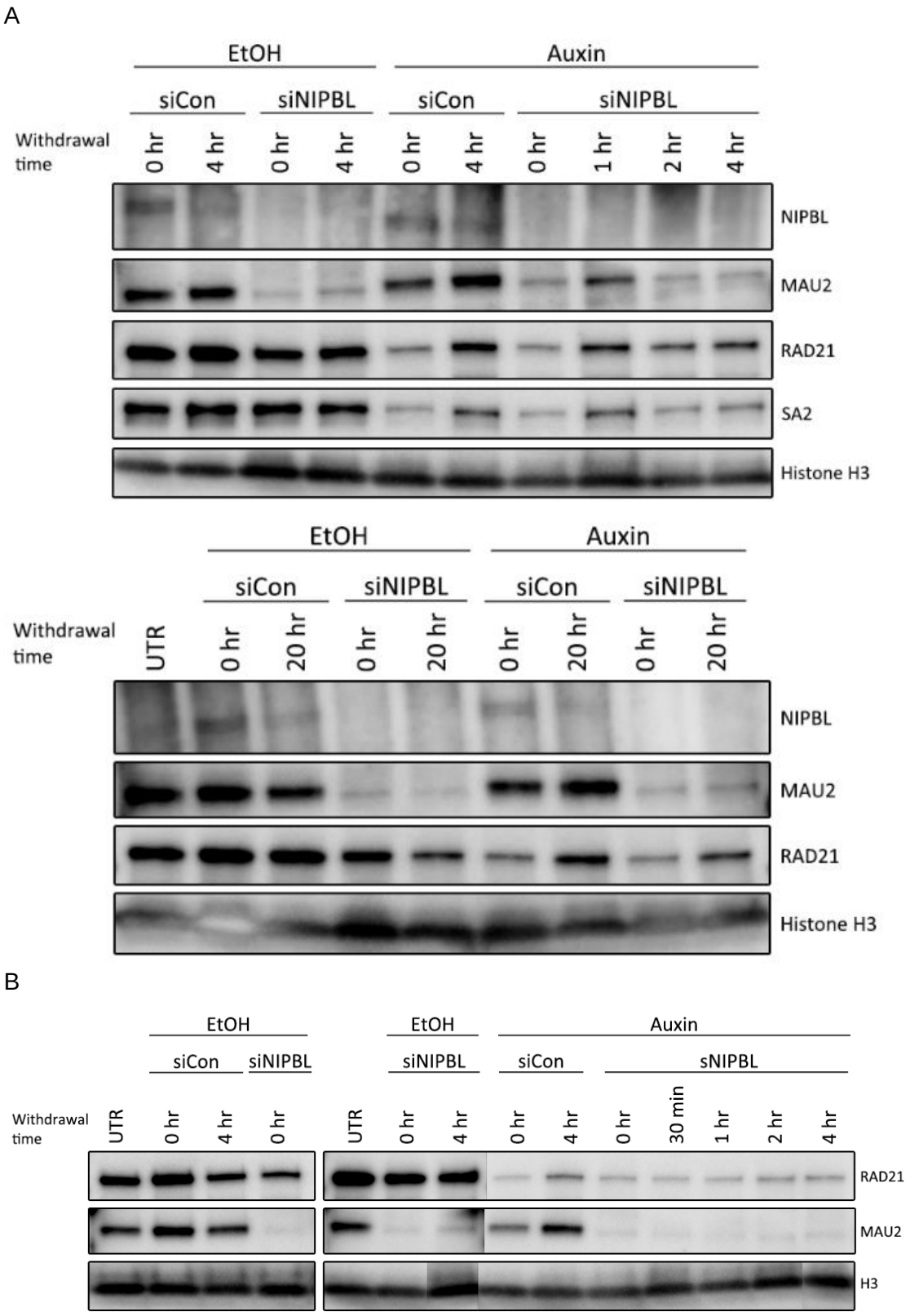


Figure 70: Re-assocation of RAD21 with chromatin in the absence of NIPBL in polyclonal HCT116 RmAC OsTIR1 cells. (A) HCT116 RmAC OsTIR1 cells were treated with siCon or siNIPBL for 72 hrs. Prior to collection, ethanol or auxin was added to the media of the indicated samples for 4hrs. The media was washed off and replaced with fresh, unmodified media for the indicated withdrawal times. 0-4hrs withdrawal times (top) and UTR and 0-20hr withdrawal times (bottom) were run on separate gels. Histone H3 was blotted as a loading control (B) Biological repeat of (A) with slightly altered timecourse of ethanol/auxin withdrawal. N.B. Additional bands are cropped from the RAD21 and MAU2 blots and the H3 lanes are rearranged to match those above. EtOH = ethanol.

Reloading was first tested in the FACS sorted H11 clone, at the median auxin-removal timepoint of 4hrs and, in case reduced reloading might occur, the longer timepoint of 22hrs. As in the polyclonal cell line, 10nM siNIPBL treatment for 72hrs efficiently removed NIPBL and MAU2 from the chromatin (Figure 71A). RAD21 levels were even across untreated and ethanol-treated samples, except for the 22hr siNIPBL sample, where loss of RAD21 from chromatin was observed, perhaps due to over-stressing of the cells (Figure 71 A & B). In the sorted cells, treating the cells with auxin for 4hrs completely degraded RAD21 from the chromatin. With the complete loss of RAD21, modest reloading was captured with 4hrs of auxin-removal. The quantity of reloaded RAD21 was similar in siCon and siNIPBL condition, and in fact, as RAD21 was slightly reduced in the 0hr siNIPBL signal, the fold change of reloading was increased in the absence of the NIPBL loader (Figure 71B). Extending the recovery time to 22hrs did not increase reloading levels, perhaps due to increased stress with the extended time in NIPBL-knockdown conditions.

RAD21 reassociation with chromatin was also tested in the H2 clone, however the efficiency of reloading was reduced in this clone (Supplemental Figure 6B). The reason for this discrepancy is unclear, as the two clones performed similarly in IP experiments (Figure 14 A & C). Further investigation would be required to test if this difference was reproducible and was caused by differences in cohesin or its regulators between the two cell clones.

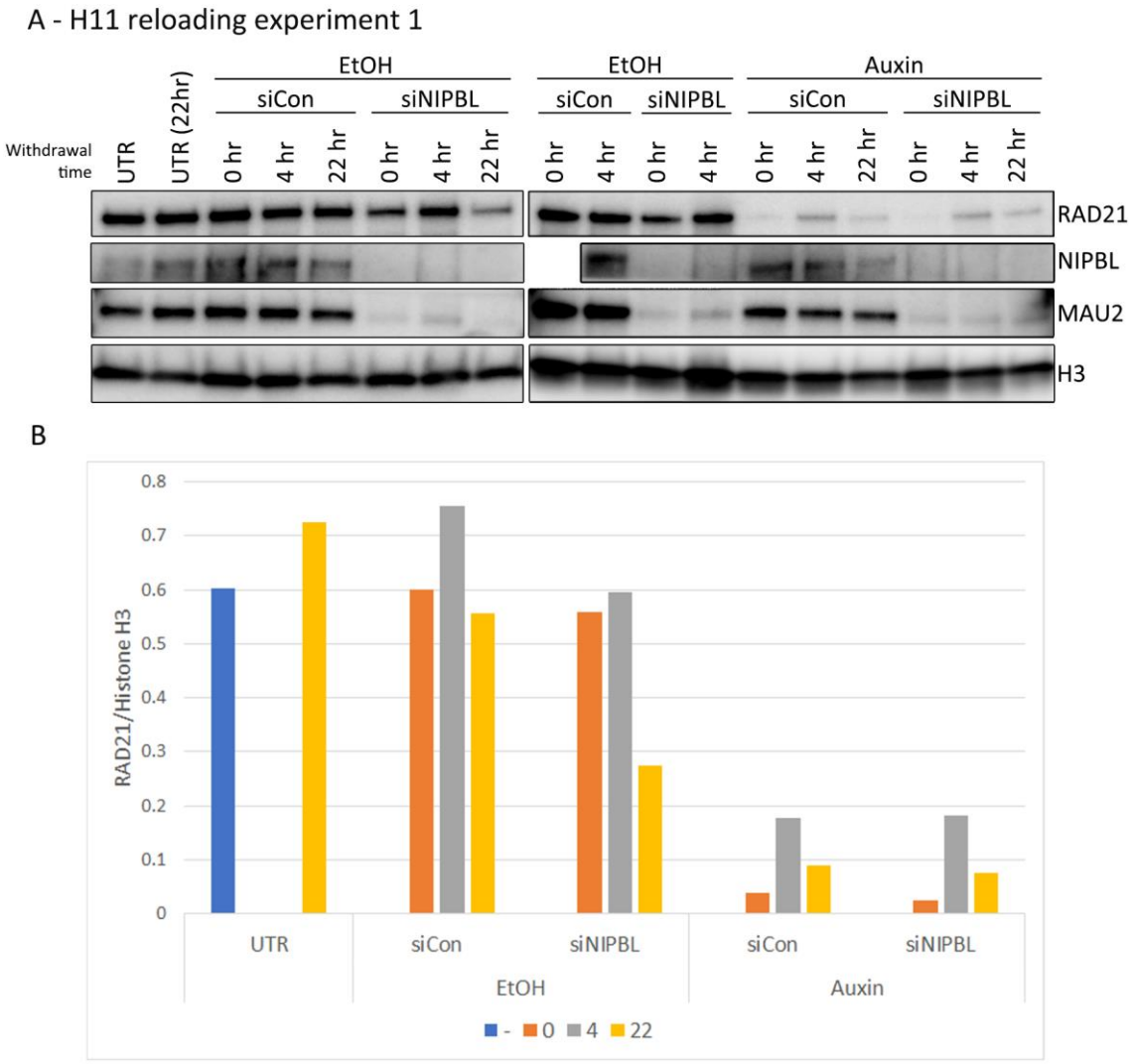


Figure 71: Re-assocaition of RAD21 with chromatin in the absence of NIPBL in H11 HCT116 RmAC OsTIR1 cells. (A) Biological repeat of Figure 70 (A) performed on the FACs sorted H11 clone of HCT116 RmAC OsTIR1 cells. (B) Quantification of RAD21 levels normalised to Histone H3 from (A). The colour legend indicates the withdrawal timepoints from ethanol/auxin.

The role of SA1 and SA2 in cohesin loading was evaluated in the H11 clone using the same experiment set-up and the median auxin-removal timepoint of 4hrs in two biological (replicate 2: Figure 72 A & C, replicate 3: Figure 72 B & D). In the replicate 2 experiment an addition triple knockdown sample was included with cell number plated increased from 1 to 1.25 million. This was done in case cell survival was considerably lower in these cells in hopes of collecting samples with similar confluence. The H11 cells tolerated all the siRNA knockdowns well and this sample was not required for future experiments. MAU2 was probed as a proxy for the loader complex in both experiments and confirmed knockdown in both experiments. SA2 was also efficiently reduced in all knockdown samples. In contrast, low levels of SA1 appeared refractory to siRNA treatment. This retention

of SA1 was consistent across experiments so any affect from it should be constant. RAD21 reassociation with chromatin was observed in all samples, albeit with varying magnitudes. For the two experiments, RAD21 reloading levels in the absence of NIPBL were ~80% of control, whereas, in the absence of SA, and SA and NIPBL, reloading levels were ~40% of control.

These results revealed two key findings; i) RAD21 can efficiently reassociate with chromatin in the absence of the loader complex, and ii) SA1 and SA2 are required for efficient reassociation of cohesin with chromatin. The SA proteins may act upstream of the loader complex since knock down of SA1, SA2, and NIPBL results in similar levels of reloading as knock down of SA1 and SA2. Such a pathway is also suggested by MAU2 densitometry. Quantification of MAU2 normalised to H3 in 7 biological replicate experiments and 1 technical repeat revealed reduction of MAU2 levels on chromatin in siSA samples, at 0 and 4hrs of auxin removal (Figure 73A). To account for variation in protein levels across the different experiments the data was also plotted as the normalised MAU2 relative to corresponding siCon. With normalisation to the control sample, MAU2 levels still showed reduction in siSA samples, indicating a reliance for the loader complex on chromatin with SA1 and SA2 (Figure 73B). Co-reduction of MAU2 may contribute to reduced levels of reloading observed in siSA conditions, however, samples with higher levels of residual NIPBL/MAU2 did not show the highest levels of reloading, illustrating that this is not the only contributing factor.

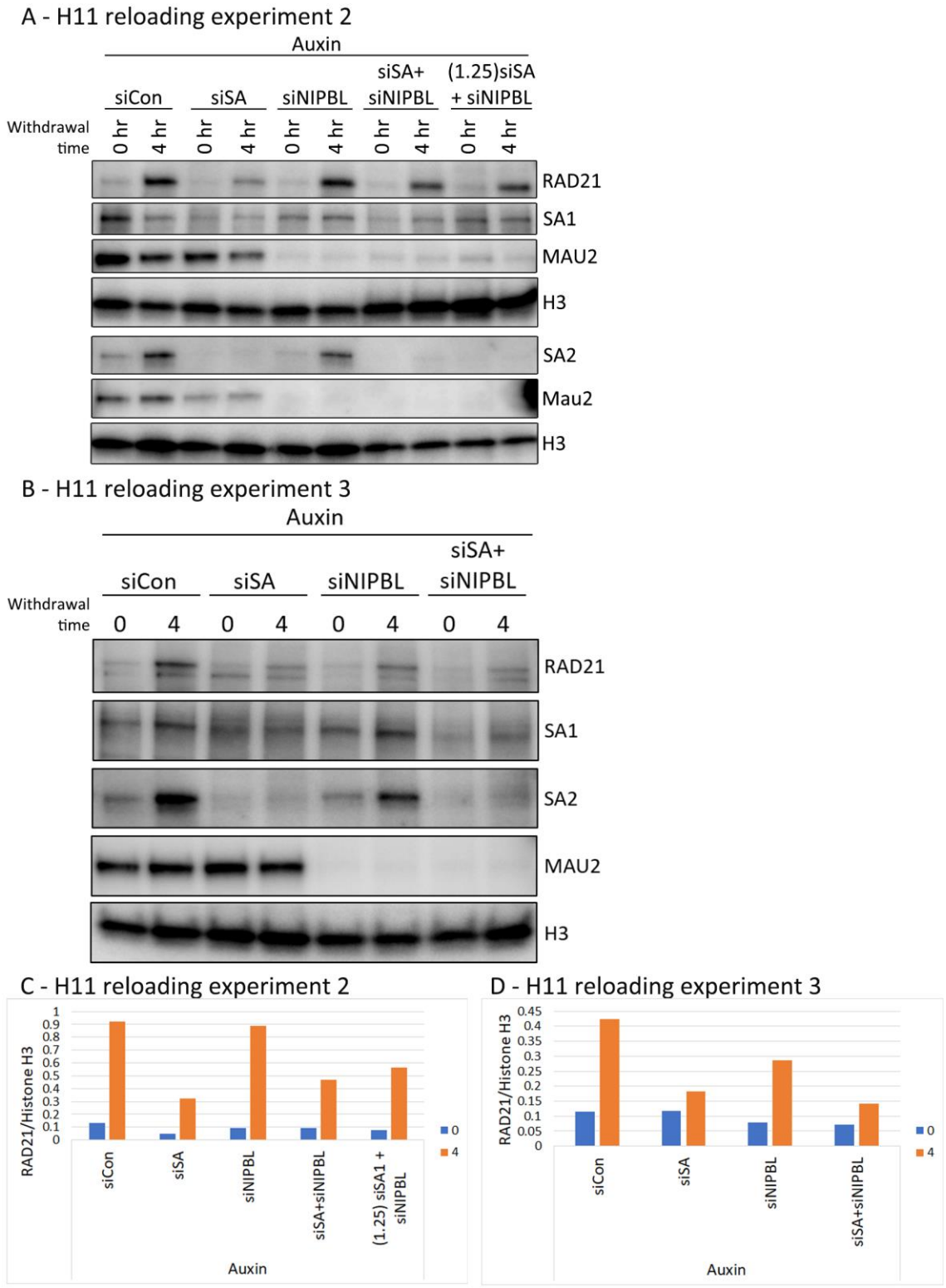


Figure 72: The SA proteins contribute to cohesin loading in H11 HCT116 RmAC OsTIR1 cells. (A) H11 cells were treated with siCon, siSA, siNIPBL, or siSA + siNIPBL for 72hrs. Additional siSA + siNIPBL samples with an increase cell number was also included ((1.25) siSA + siNIPBL). Prior to collect, auxin was added to the cell media for 4hrs and then washed off and replaced with fresh, unmodified media for the timepoints indicated. Histone H3 was blotted as a loading control. (B) Biological repeat of (A) without the additional siSA + siNIPBL sample. (C) Quantification of RAD21 levels normalised to Histone H3 from (A). The colour legend indicates the withdrawal timepoints from auxin. (D) Quantification of RAD21 levels normalised to Histone H3 from (B). The colour legend indicates the withdrawal timepoints from auxin. siSA = siSA1 + siSA2.

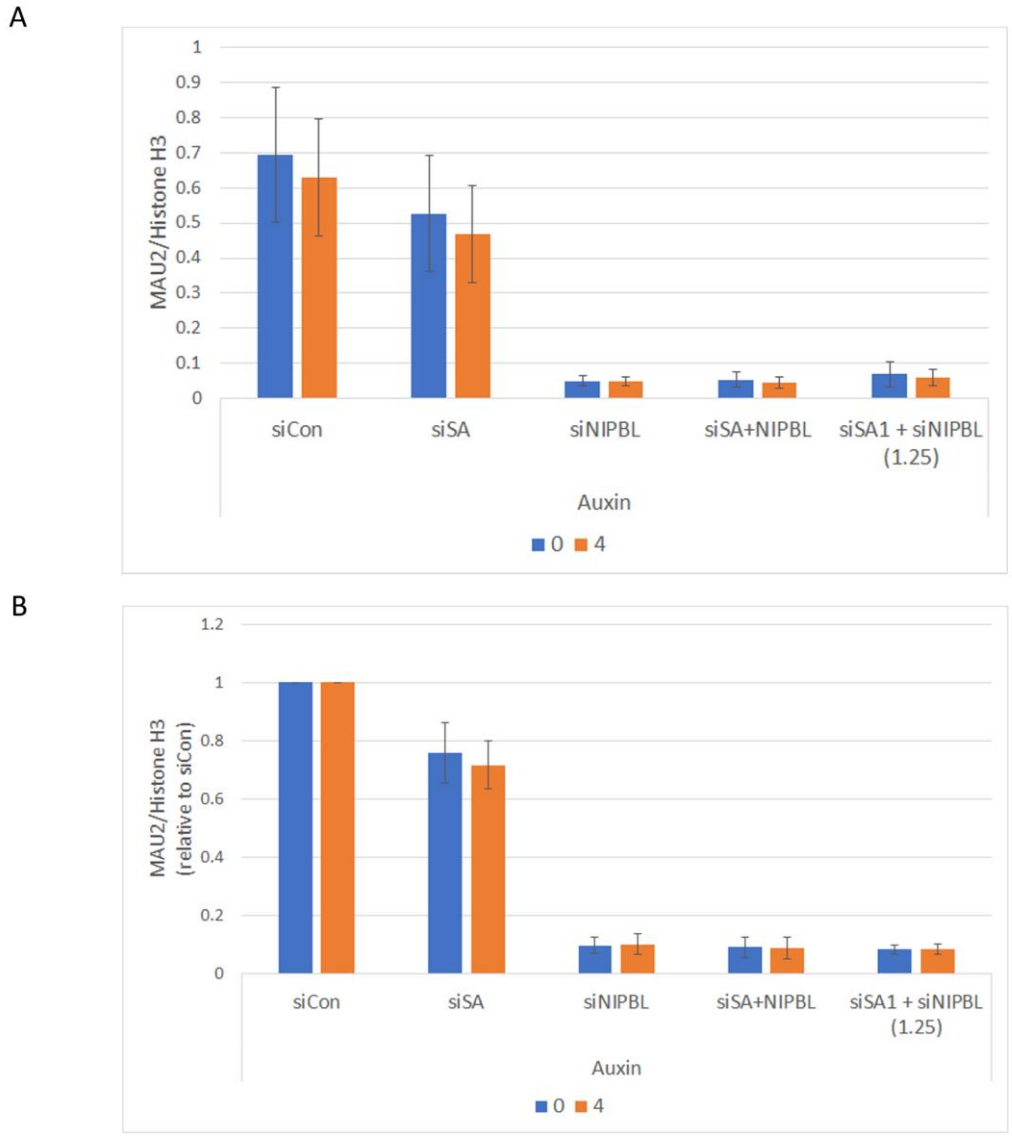


Figure 73: MAU2 levels on chromatin are reduced with knockdown of SA1 and SA2. (A) Quantification of MAU2 levels on chromatin normalised to Histone H3 from 4 biological replicates and 2 technical replicates (including Figure 72 A & B). The colour legend indicates the withdrawal timepoints from auxin. (B) Values from (A) now normalised to siCon.

To confirm the observations above were auxin-specific, RAD21 association with chromatin was assessed in both ethanol- and auxin-treated cells following knockdown of SA1 and SA2, NIPBL, or SA1 and SA2 and NIPBL for 72hrs. Control ethanol-treated chromatin samples were loaded onto three gels; i) a pre-cast NuPAGE™ 3-8% Tris-Acetate gel to probe for NIPBL, ii) a Bio-Rad 4–20% Mini-PROTEAN® TGX™ precast protein gel to probe for RAD21, SA2, MAU2, and Histone H3, and iii) a Bio-Rad 4–20% Mini-PROTEAN® TGX™ precast protein gel to probe for SA1, MAU2 (repeat), and Histone H3. Auxin-treated

samples were run in the same set-up, but with siCon ethanol-treated samples also loaded on the gel to allow matching of exposure of each protein across the separate gels. Ethanol-treated samples confirmed the levels of each protein after the respective siRNA treatments and no difference was observed between 0 and 4hr removal from ethanol-treatment (Figure 74A). NIPBL and MAU2 were both knocked down in cells transfected with siNIPBL and were reduced in siSA compared to siCon. SA2 was completely lost in siSA-treated samples, whereas low levels of SA1 were retained on chromatin. RAD21 levels were reduced by just over 50% in siNIPBL, ~70% in siSA, and ~80% in siSA + siNIPBL (Quantification shown in Figure 74C). RAD21 was not completely lost from chromatin in the siSA + siNIPBL sample, as would be expected if the loss of SA and NIPBL was additive. The residual signal observed may represent a pool of very stably bound cohesin or may indicate a positive antagonistic epistatic relation between SA and NIPBL (the effect of loss of both SA and NIBPL is reduced compared to the theoretical additive loss of both). This would suggest that SA and NIPBL work within the same pathway for loading of cohesin.

The same effects of siRNA treatment were observed in auxin-treated samples, except here RAD21, SA1, and SA2 were lost from chromatin in 0hr withdrawal samples (Figure 74A). Hence, their signal in 4hr withdrawal samples represents re-association with chromatin. To better visualise differences in reloading of RAD21 in the different siRNA-treated samples, the ethanol siCon samples were covered and exposure increased for RAD21, SA1, and SA2 (Figure 74B). RAD21 levels after 4hrs withdrawal from auxin were highest in siCon-treated cells, followed by siNIPBL, siSA, and lowest in siSA + siNIPBL.

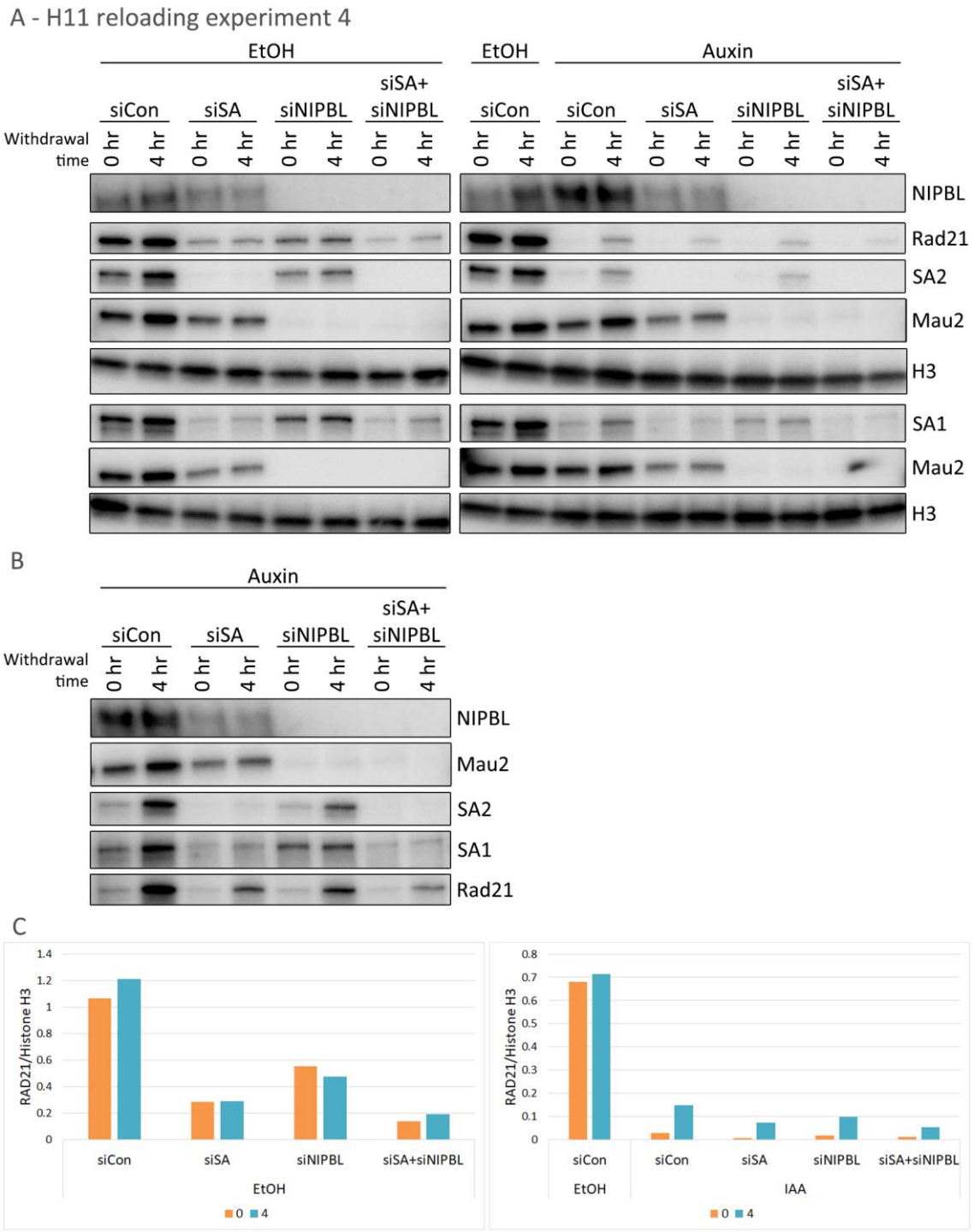


Figure 74: Cohesin can re-associate with chromatin in the absence of NIPBL – representative WB. (A) A representative western blot of the cohesin reloading experiment including ethanol-treated controls. Samples were treated as in Figure 71. The siCon and ethanol treated samples were run on both blots to allow matching of exposure levels and comparison of the two blots. Histone H3 was blotted as a loading control. (B) Auxin samples from (A) with increased exposure. (C) Quantification of RAD21 levels normalised to Histone H3 from (A) with EtOH blot on left and EtOH/IAA blot on right. The colour legend indicates the withdrawal timepoints from ethanol/auxin. EtOH = ethanol.

Given the different 'base' quantities of RAD21 on chromatin in the different siRNA transfections, a standardized measure of reloading of RAD21 was calculated as

fold change of the 4hr withdrawal signal over 0hr withdrawal signal. The mean fold change of reloading for each treatment is shown in Figure 75A and the individual fold change values for each experiment are shown in Figure 75B. Variation of the fold change values was evident among auxin-treated samples and was highest in siSA and siNIPBL conditions. Such variability may be driven by fluctuating cell state at the onset of treatment, variability in siRNA-mediated knockdown, or signify that without its loaders, RAD21 reloading occurs in a stochastic manner. The average fold change revealed that reloading was most similarly efficient in siSA and siNIPBL samples, due to the variability between experiments, this increase was not significant. Triple knockdown of all three proteins significantly reduced reloading efficiency compared to siCon. Reloading was also reduced compared to siSA or siNIPBL alone, at amounts close to, but not quite significant (0.0930 and 0.0736, respectively). Hence, both NIPBL and the SA proteins play a role in the reassociation of cohesin with chromatin and the SA proteins may compensate for loss of the loader complex in the siNIPBL samples.

To try and consider the differing levels of base RAD21 with siSA and siNIPBL treatments, the same data was also calculated as fold change compared to siCon 0hr (either ethanol or IAA treated; Figure 76). It is evident that RAD21 levels are decreased with siNIPBL, siSA, and siSA + siNIPBL in ethanol and auxin treated samples. With this calculation it is possible to observe the importance of the SA proteins to reassociation of cohesin with chromatin. In the absence of the loader complex alone (siNIPBL) the mean level of reassociation is reduced compared to siCon, but the range of reloading observed is somewhat similar. In contrast, siSA and siSA + siNIPBL samples have a much lower mean level of reassociation. Reassociation of in siSA 4hr is significantly reduced compared to siCon 4h, with a p-value of 0.05.

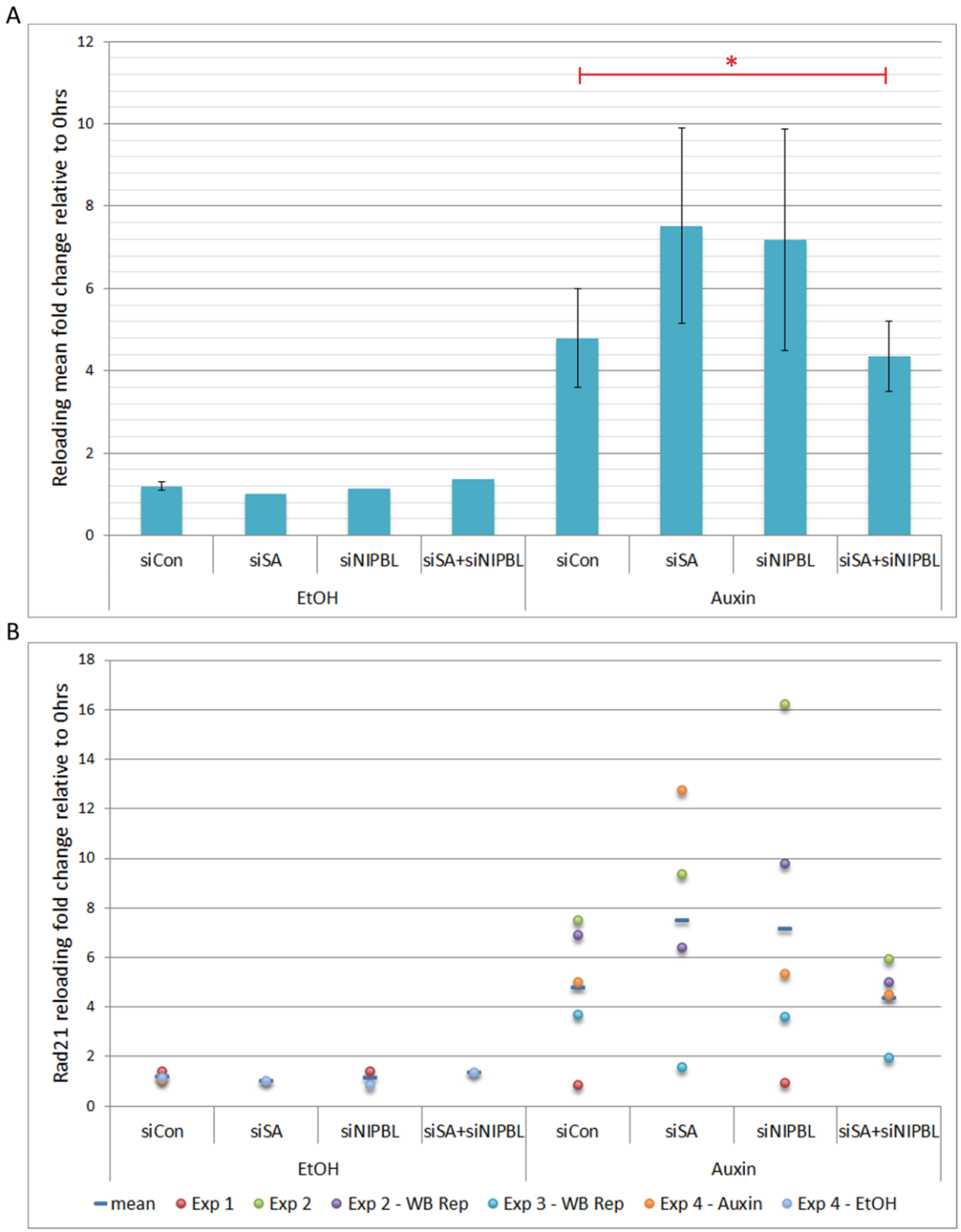


Figure 75: Cohesin can re-associate with chromatin in the absence of NIPBL – densitometry. Reloading efficiency was calculated by the fold change of the 4hr withdrawal RAD21 quantification levels normalised to H3 over the 0hr withdrawal RAD21 quantification levels normalised to H3. Mean values are shown in (A) (n=4) and mean and individual values are shown in (B). * indicates statistically significant change between siCon and sisiSA1 + siNIPBL, calculated by t-test (pvalue < 0.05).

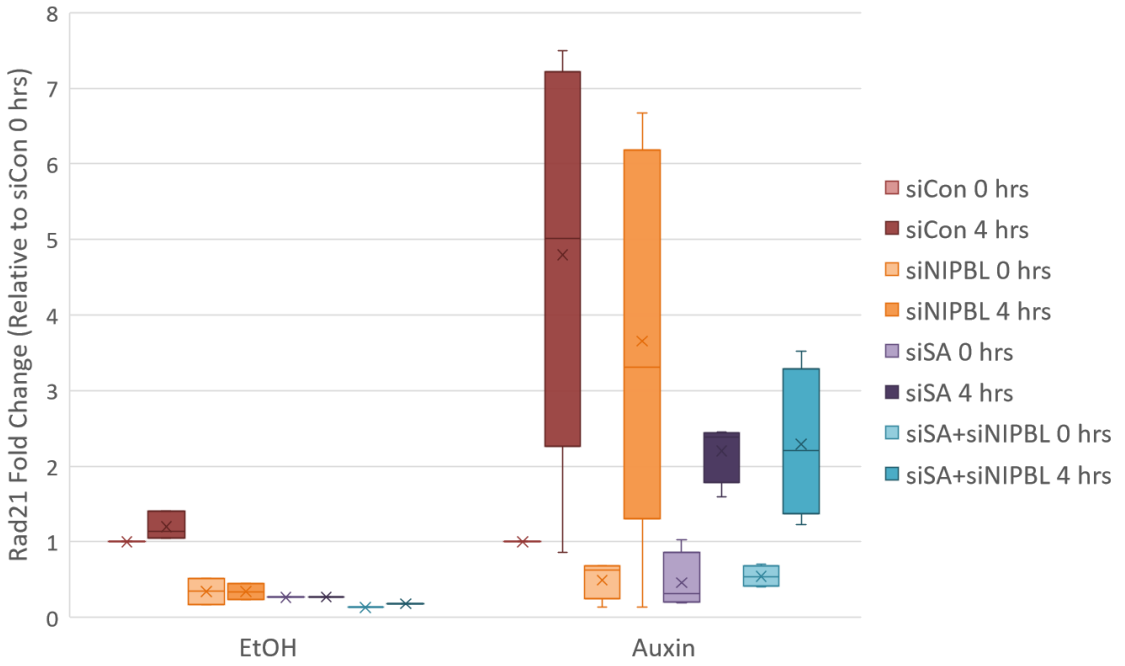


Figure 76: Cohesin can re-associate with chromatin in the absence of NIPBL – alternative densitometry. An alternative normalisation technique used to calculate the mean reloading efficacy using the same data as Figure 41. Here fold change of RAD21 quantification levels normalised to H3 were calculated for each sample relative to the corresponding siCon 0hr withdrawal sample. Data is represented as a box and whisker plot. There is a statistically significant change between Auxin siCon 4hrs and siSA 4hrs, as calculated by t-test (pvalue = 0.05).

5.2.5 Influence of R-loop structures on RAD21 reloading

As discussed in section 4.2.4, loading of cohesin occurs at specific nucleic acid structures during replication (Murayama *et al.*, 2018; Zheng *et al.*, 2018). We hypothesised that SA1 may co-opt this loading behaviour to load cohesin onto chromatin at R-loops during interphase. Following on from the novel findings thus far that SA1 interacts with a myriad of R-loop proteins, is enriched in R-loop IPs, and may account for reassociation of cohesin in NIPBL knockdown cells, the influence of altering R-loops levels on cohesin reloading in NIPBL knockdown conditions was assessed.

For this experiment R-loop levels needed to be altered in cell culture, however, as outlined in section 4.2.5.1, overexpression of RNase H1 had little effect on R-loop levels in our cell system. Overexpression of RNase H1 is the most widely use technique to modulate R-loop levels in cell culture in the literature, although alternative methods have also been used. As discussed in section 4.3, RNase H1 may be localised predominantly to the nucleolus in human cells. A second

RNase H enzyme is RNase H2, which is more strongly expressed in human cells than RNase H1 and has been shown to localised more globally in the nucleus (Bubeck *et al.*, 2011; Choi, Hwang and Ahn, 2018). Hence, this enzyme may represent a stronger manipulator of R-loops in our cells. RNase H2 is composed of three subunits, termed H2A, H2B, and H2C. As all three subunits are required to form the full enzyme, overexpression of the enzyme to reduce R-loop levels was considered too complicated to test. Instead, knockdown of the H2A subunit with siRNA was chosen as this has previously been shown to destabilise the enzyme and increase nucleolar and nuclear s9.6 signal (Chon *et al.*, 2009; Choi, Hwang and Ahn, 2018). Knockdown of the RNA:DNA hybrid nuclease Aquarius (AQR) was also chosen as a strategy to modulate R-loops, as this has been shown to increase R-loop levels in HEK, Hela, and HCT116 cells (Sollier *et al.*, 2014; Nguyen *et al.*, 2017; Sakasai *et al.*, 2017). In Hela cells, Sollier et al. (2014), showed by immunofluorescence that this increase occurred in the nucleus. Hence, reloading of cohesin following auxin removal was tested in cells in the presence and absence of the loader complex, in the presence of control or increased levels of R-loops.

5.2.5.1 Optimisation of siAQR and siRNASEH2A

To optimise the new method of R-loop modulation, knockdown of AQR and RNASEH2A was tested for 72hrs with 10 and 20 nM siRNA. A double knockdown of AQR and RNASEH2A together, each at 10nM was also assessed. Efficiency of knockdown was considered by the level of AQR and RNASEH2A on chromatin and by the quantity of RNA:DNA hybrids detected in the chromatin lysate by dot blot. SA1 was also blotted for to check for alterations to its levels on chromatin. Cells treated with 20nM RNASEH2A or 10 or 20nM AQR had enlarged cell bodies and irregular edges, indicating an underlying change to the cell health or biology was induced in these conditions. An example of the morphology observed is shown in Figure 77B. AQR was successfully knockdown with 10 and 20nM siRNA treatments, whereas, RNASEH2A levels were increased rather than decreased with its siRNA treatment (Figure 77A). SA1 levels were also decreased in siAQR samples, however, it was unclear if this was a consequence of loss of the helicase or indirectly from the changes occurring to the cell morphology. RNA:DNA hybrids levels were assessed by dot blot of the chromatin lysates. A titration of the UTR sample is shown on the left and the dots for each sample below the

corresponding western blot bands. siAQR and siRNASEH2A single siRNA treatments slightly reduced hybrid signals compared to their corresponding siCon controls. Double knockdown of AQR and RNASEH2A reduced the hybrid levels further, however, this is the opposite phenotype to that expected with loss of the proteins. Reduction of R-loops with loss of repressive regulators may stem from senescence of the cells and halting of transcription or a feedback loop that in fact increased RNASEH2A levels on chromatin.

A second optimisation experiment was run to test 5nM siAQR treatment for 40 and 65hrs to specifically reduce the morphological changes. Co-transfection with 10nM of either siNIPBL or siRNASEH2A were also evaluated for effect on cell health and RNA:DNA hybrid levels. AQR was knocked down after 40 or 65hrs, however, after 65hrs most of the cells had an enlarged irregular morphology and MAU2 and SA1 levels on chromatin were reduced. Altogether this signalled that the cells likely had altered health or biology. Reducing the transfection incubation to 40hrs reduced the proportion of irregular cells within the population. Thus, for the reloading experiment, both the amount and timepoint of siRNA treatment needed to be reduced. Blotting for MAU2 confirmed that the loader complex was still efficiently depleted from chromatin with reduction to 40hrs (Figure 77C). R-loop levels were again assessed by dot blot of the chromatin lysate. A pipette tip insert was no longer used to create a grid on the membrane and instead the lysate was ejected from the pipette slowly to form a small bubble on the tip of the pipette that could then be wicked onto the membrane. Variability in R-loop levels between the two amounts was apparent, however, a general increase in R-loops was observed between siCon and siAQR samples. This increase in R-loops was retained with co-transfection of siAQR and siNIPBL. There also seemed to be a substantial increase in R-loops in the siAQR, siRNASEH2A double transfection sample, however, comparison to H3 levels suggest that this is simply due to overloading of this sample. As for the first optimisation experiment, siRNASEH2A treatment had no observable effect on the protein levels under the conditions tested.

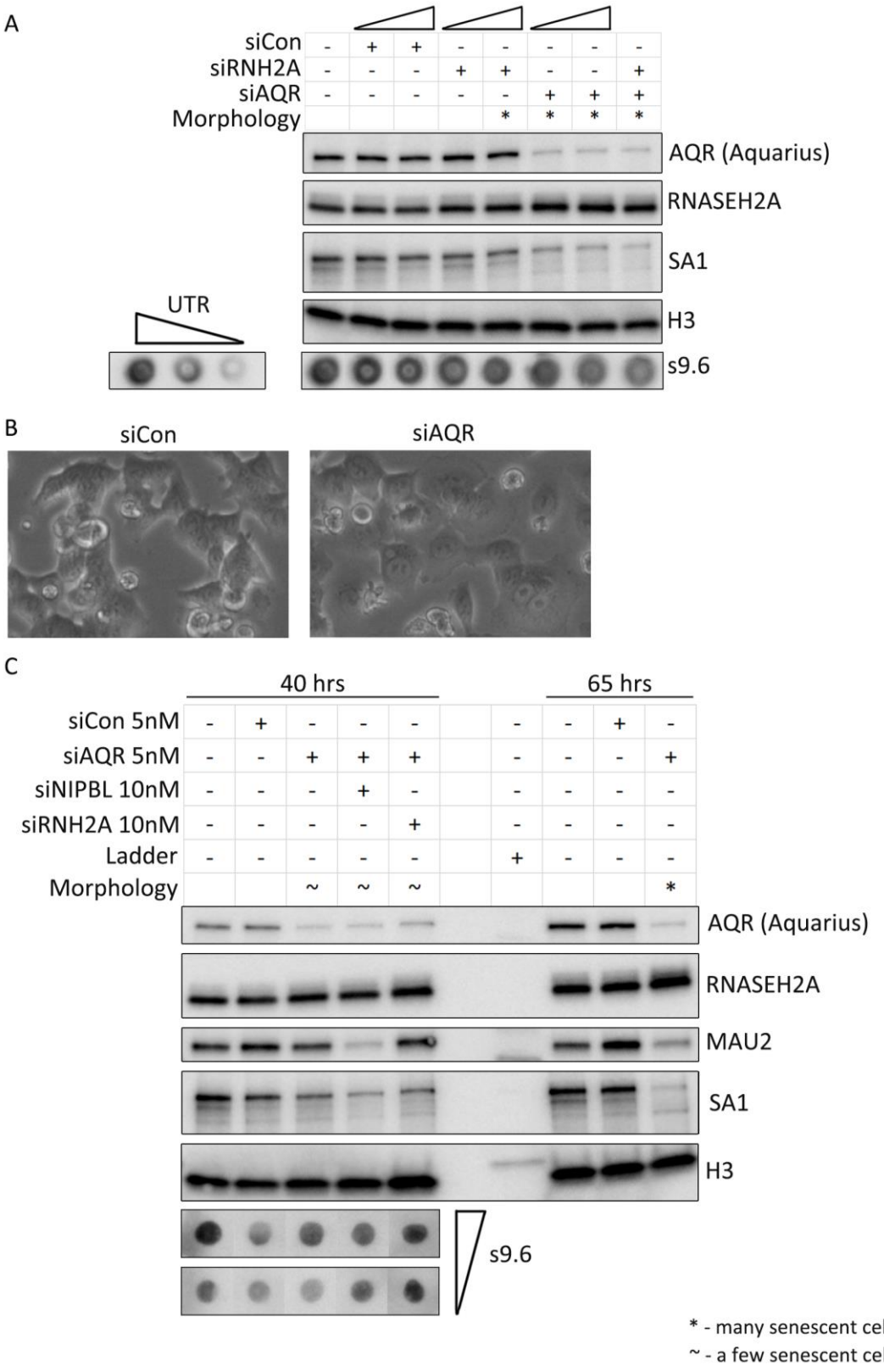


Figure 77: Optimisation of siAQR and siRNASH2A to increase R-loop levels. (A) H11 cells were treated with 10 or 20nM of siAQR or siRNASEH2A, or 10nm siAQR + siRNASEH2A, for 72hrs. Cells were fractionated to isolate chromatin-bound proteins, without benzonase treatment. H3 was blotted as a loading control. Dot blot of s9.6 signal is shown at the bottom with a titration of 7.5, 3.75, 1.87 ug of the UTR sample on the left and 7.5ug of each sample shown beneath the corresponding western blot lane. (B) Example of the morphological changes observed with siAQR transfection. The image was captured using the phase channel of a Zeiss brightfield inverted bench-top microscope. (C) Treatment of cells with siCon, siAQR, siAQR + siNIPBL, or siAQR + siRNASEH2A at the indicated concentrations for 40 or 65hrs. Dot blot of s9.6 for the 40 hrs transfection samples is shown for 7.5 and 3.75 ug of chromatin extract. RNH2A = RNASEH2A.

5.2.5.2 Reloading at modulated R-loops

Using the 5nM siAQR condition optimised above a full set of ethanol and auxin-treated cells were tested for reloading with removal of auxin in siCon, siNIPBL, siAQR, and siAQR + siNIPBL conditions. Even though there were no clear changes to the protein levels of RNASEH2A in the optimisation experiments, RAD21 reloading was also tested with RNASEH2A and RNASEH2A + NIPBL knockdown to test for any effect on R-loops and RAD21 levels after auxin treatment and withdrawal. The samples were split across three blots and ethanol- and auxin-treated siNIPBL samples were run on each blot to allow matching of exposure. Summary figures for each blot are shown on the left with normalised RAD21, SA1, and s9.6 densitometry values on the right ethanol- and auxin-treated siCon and siNIPBL samples were run on blot 1 (Figure 78A). RAD21, SA1, and MAU2 behaviour in the samples was in line with the previous reloading experiments in section 5.2.3. AQR levels were reduced on chromatin with the 4hr auxin treatment. Like MAU2 and SA1, AQR showed a recovery with the 4hr removal of auxin, this suggests that cohesin itself plays a role either in targeting of AQR to the chromatin or the levels of R-loops. Yang Li, a post-doctoral researcher in the lab carried out IF of s9.6, cohesin, and SA1 in control and siSA1 cells and observed an increase in s9.6 signal with loss of SA1, implicating SA1 as a suppressor of R-loops. RNASEH2A levels were generally unchanged across the treatments. s9.6 levels were similar between the ethanol- and auxin-treated samples although they was slightly more variation in ethanol conditions (Figure 78A and Figure 80B).

Ethanol-treated siNIPBL, siAQR, siAQR + siNIPBL, siRNASEH2A, and siRNASEH2A + siNIPBL samples were run on blot 2 (Figure 78B). In these control conditions, RAD21 levels were relatively similar between siNIPBL and siAQR samples. Increased RAD21 was recorded in siRNASEH2A samples. In parallel, s9.6 levels were highest in the 4hr ethanol withdrawal siRNASEH2A sample (Figure 78B and Figure 80B). MAU2 and AQR displayed a similar increase to RAD21 in these samples, even following normalisation to H3. RNASEH2A levels themselves appeared refractory to the siRNA treatment. S9.6 was increased in all siAQR-treated samples and were not increased in any of the other siRNASEH2A-treated samples.

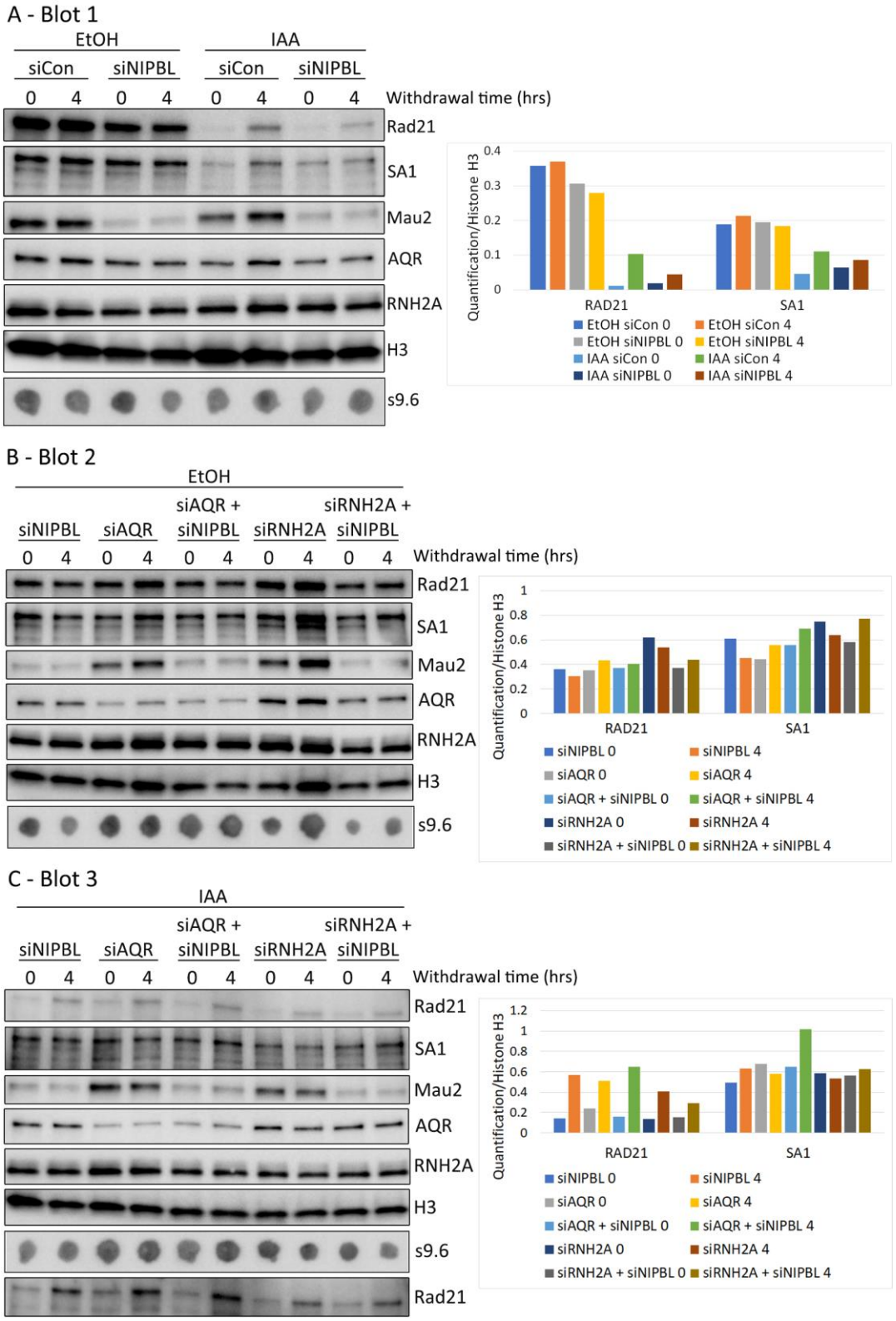


Figure 78: RAD21 reassociation with chromatin is increased with knockdown of AQR – replicate 1. H11 cells were treated with non-targeting siRNA (siCon) or siRNA targeting NIPBL, AQR, RNASEH2A, in the indicated combinations for 40hrs. Prior to collection, ethanol or auxin was added to the cell media as indicated for 4hrs, then washed off and replaced with fresh media for the indicated withdrawal time. Samples were fractionated to purify chromatin-bound proteins in the absence of benzonase treatment and blotted on three membranes. Western and dot blot results are shown on the left and quantification relative to H3 for RAD21 and SA1 is shown on the right. (A) siCon and siNIPBL samples were run on blot 1. (B) Ethanol-treated siNIPBL, siAQR, siAQR + siNIPBL, siRNASEH2A, and siRNASEH2A + siNIPBL samples were run on blot 2. (C) Ethanol-treated siNIPBL, siAQR, siAQR + siNIPBL, siRNASEH2A, and siRNASEH2A + siNIPBL samples were run on blot 3. EtOH = Ethanol; IAA = auxin; RNH2A = RNASHEH2A.

Auxin-treated siNIPBL, siAQR, siAQR + siNIPBL, siRNASEH2A, and siRNASEH2A + siNIPBL samples were run on blot 3 (Figure 78C). RAD21 levels on chromatin were increased in all 4hr withdrawal samples, compared to their corresponding 0hr samples. Hence, reloading of RAD21 occurred in all condition tested. The levels of reassociated RAD21 were similar between siNIPBL and siAQR samples, however the highest amounts were recorded in the double transfected siAQR + siNIPBL sample. This increase corresponded with a spike in SA1 and s9.6 levels (Figure 78C and Figure 80B). Corresponding spike in these proteins perhaps suggests a mechanism whereby knockdown of AQR facilitates retention of R-loop structures and binding of SA1 to these structures facilitates the loading of cohesin in the absence of the loader complex. Similar spikes of SA1 and s9.6 were not observed in the single siAQR knockdown, this may be due to a difference in the state of the cells or point to an importance for the loss of NIPBL to observe this pathway. It may be important to note that MAU2 and AQR were also increased in the 4hr siAQR + siNIPBL auxin withdrawal sample compared to the 0hr siAQR + siNIPBL auxin withdrawal sample, however even with the increase the levels of both proteins was very low, especially compared to the increase observed in SA1 (Supplemental Figure 7A).

The experiment was repeated to determine reproducibility of the above results. Again the samples were run on three gels, but for the repeat experiment siCon was included on the same gel as siNIPBL and siAQR to allow better comparison of RAD21 reloading and the RNASEH2A samples were split across the three gels to allow exposure-matching. Here ethanol-treated siCon, siNIPBL, siAQR, siAQR + siNIPBL, and siRNASEH2A + siNIPBL samples were run on blot 1 (Figure 79 A). RAD21, SA1, and RNASEH2A levels were relatively even across all of the samples. Lack of reduction of RAD21 and SA1 levels in siNIPBL-treated samples suggests that the transfection may not have been optimal for this experiment. MAU2 signal suggested a loss of the loader complex from chromatin, however there was unusual background signal on the blot, perhaps from the use of old antibody solution. The AQR knockdown was successful and was probably more efficient than for replicate 1 as AQR levels in control samples were higher in this experiment and still AQR levels in knockdown samples were lower than replicate

1. In this replicate s9.6 levels were increased in all siAQR- and siRNASEH2A-treated samples (Figure 79 and Figure 80).

Corresponding auxin-treated samples were run on gel 2 (Figure 79B). Again, reloading of RAD21 was observed for all the samples tested. As in replicate 1, the levels of reloaded RAD21 were similar across the conditions tested, however, in this case the siAQR + siNIPBL sample showed the lowest levels of reloaded RAD21. No spike in SA1 or s9.6 levels were observed in the siAQR + siNIPBL 4hr auxin withdrawal sample, suggesting that R-loop increase with siAQR treatment is somewhat variable, perhaps depending on cell state and initial levels of R-loops in cells. Finally, blot 3 contained the RNASEH2A knockdown samples. RNASEH2A was refractory to change, but s9.6 levels were still increased. The amount of RAD21 reloaded on chromatin in the siRNASEH2A + siNIPBL auxin withdrawal sample was similar to siNIPBL alone.

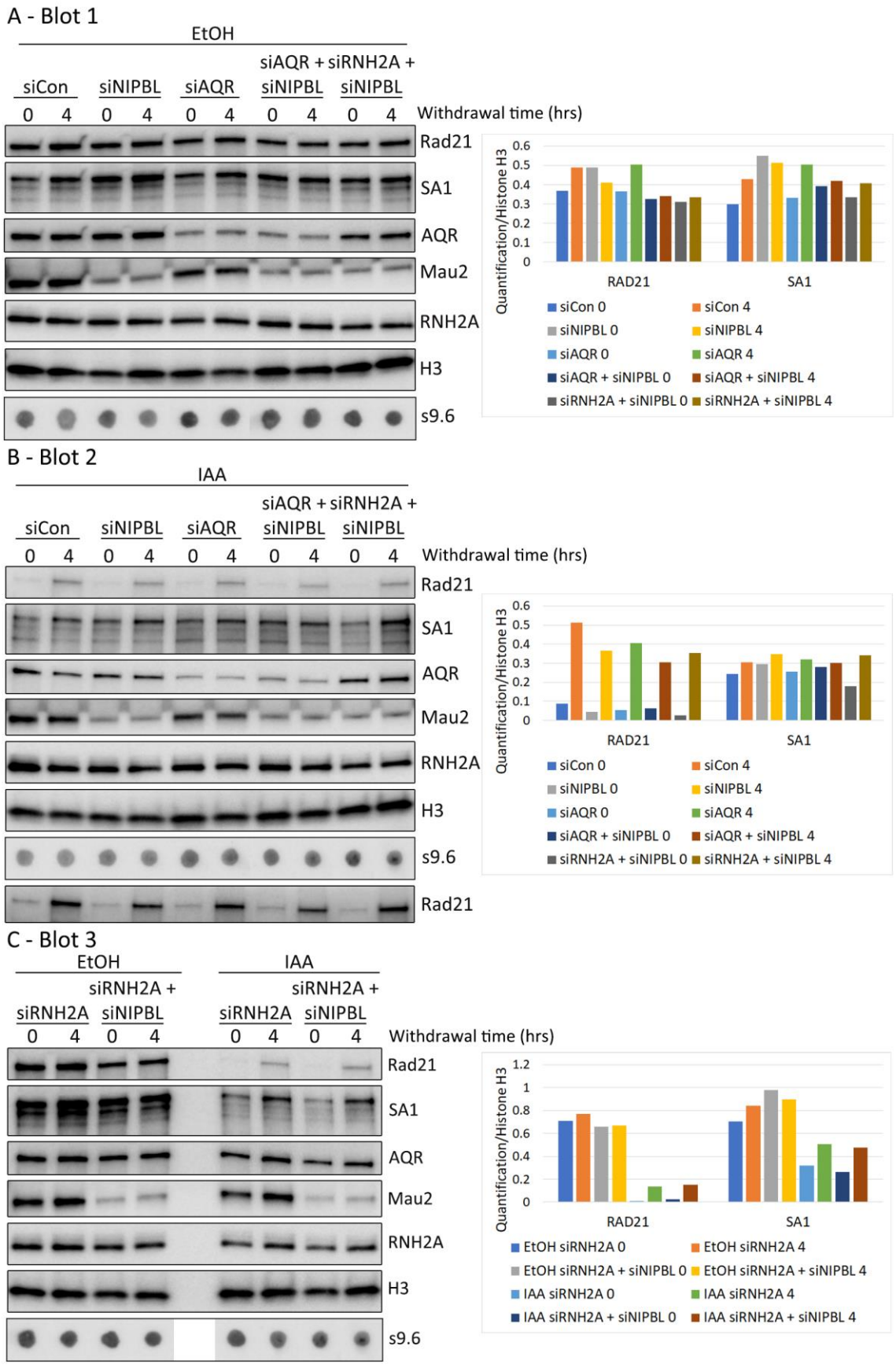


Figure 79: RAD21 reassociation with chromatin is increased with knockdown of AQR – replicate 2. Biological replicate of Figure 78. (A) Ethanol-treated siCon, siNIPBL, siAQR, siAQR + siNIPBL, and siRNASEH2A + siNIPBL samples were run on blot 1. (B) Auxin-treated siCon, siNIPBL, siAQR, siAQR + siNIPBL, and siRNASEH2A + siNIPBL samples were run on blot 2. (C) siRNASEH2A and siRNASEH2A + siNIPBL samples were run on blot 3. EtOH = Ethanol; IAA = auxin; RNH2A = RNASHEH2A; H3 = Histone H3.

To better compare replicates 1 and 2, the fold change of reloading was calculated for each 0 and 4hr withdrawal sample pair and the resulting reloading efficiency values graphs for replicate 1, replicate 2, and the mean value of the two experiments (Figure 80A). For all ethanol-treated samples, the reloading efficiency was approximately 1 as there was no loss of RAD21 and subsequent reassociation with chromatin in these cells. Comparison of RAD21 fold change values across the conditions and experiments revealed 3 key details about the differences. Firstly, the signal context of the blot itself resulted in changes in the reloading efficiency value. For example, the same siNIPBL IAA sample was run on blot 1 and blot 3 in experiment 1 and gave reloading values of 2.4 and 3.95, respectively. This difference is mostly likely caused by the level of background signal and the strength of the strongest signal on the blot, both of which would affect quantification of lower strength IAA bands. This means that comparison of reloading efficiency values across different blots is not possible without a further level of normalisation, such as scaling by the difference between the double loaded samples.

Secondly, the efficiency of loading in siNIPBL and siAQR samples was decreased compared to siCon in experiment 1 but increased compared to siCon in experiment 2. The reason for the increased reloading in siNIPBL is currently unclear as there is no large change in any of the proteins blotted for across the two experiments. Variation of siNIPBL reloading efficiency above and below the efficacy in siCon samples was also observed in earlier experiments and perhaps represents some underlying variation in the population. In experiment 1 reloading efficiency in siAQR + siNIPBL was about equal with the siNIPBL sample run on the same gel. The siNIPBL sample that was run on the same gel as siCon showed decreased reloading efficiency (~3-fold). This suggests that in experiment 1, reloading in the siAQR + siNIPBL was reduced compared to siCon ~3-fold. In comparison, siAQR + siNIPBL reloading was only ~1.2-fold reduced compared to siCon in experiment 2. Hence, reloading efficiency in the siAQR + siNIPBL experiment 2 was likely not reduced compared to experiment 1, but the efficiency of reloading in the other experiment 2 samples was increased. Thirdly, RAD21 reloading efficiency in siRNASEH2A + siNIPBL was reduced compared to siNIPBL in experiment 1 but increased compared to siNIPBL in experiment 2.

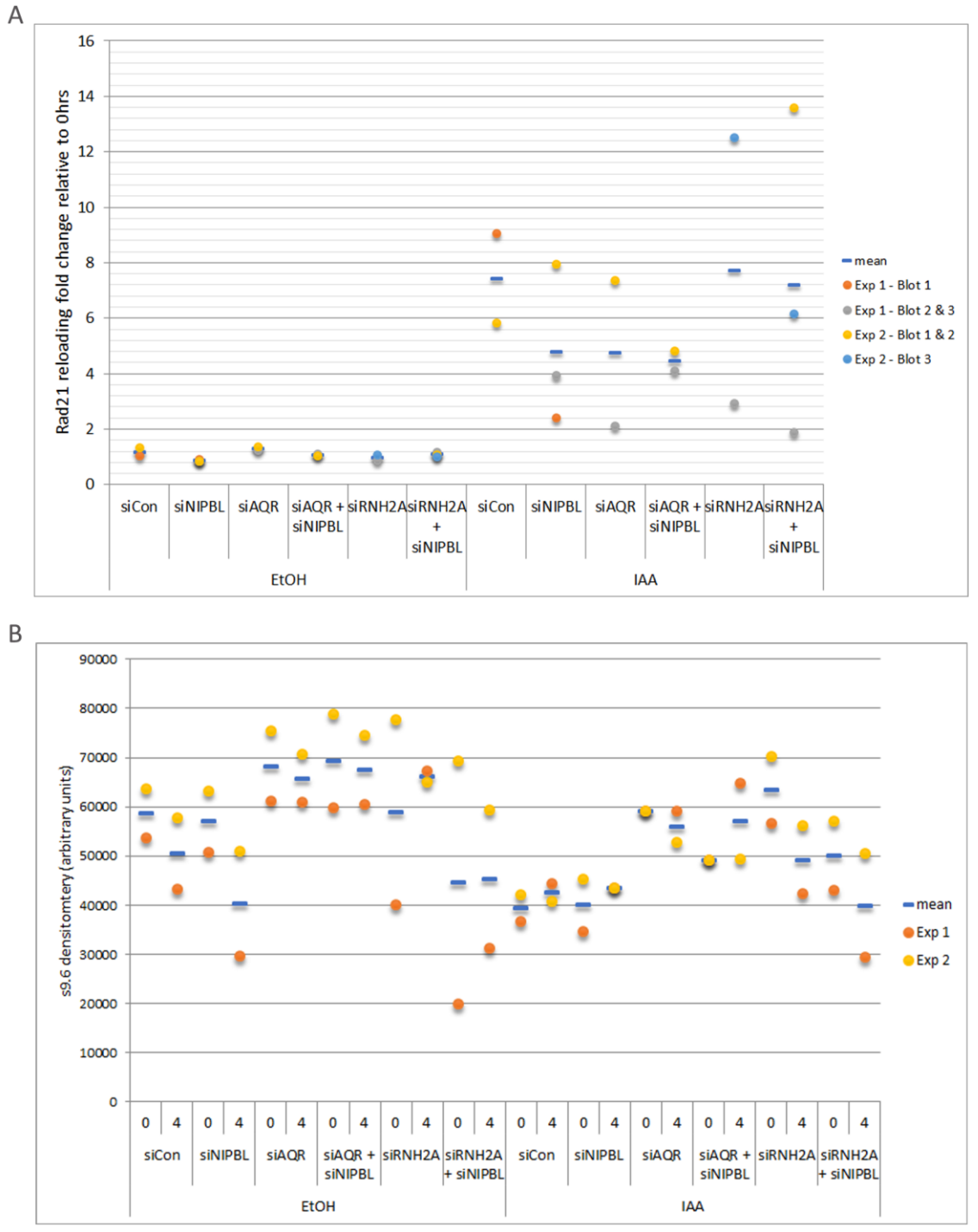


Figure 80: RAD21 reassociation with chromatin is increased with increase in R-loops. (A) RAD21 reloading efficiency was calculated as the fold change of Histone H3 normalised 4hr withdrawal over Histone H3 normalised 0hr withdrawal. Mean and individual values from Figure 44 and Figure 45 are plotted. (B) Raw densitomtery values from s9.6 dot blots from Figure 78 and Figure 79. Mean and individual values are shown.

To compare with the RAD21 reloading efficiency graph discussed above, the raw s9.6 densitometry values from experiment 1 and 2 were plotted, alongside the mean of the two datasets (Figure 80B). Raw densitometry values were used as s9.6 was blotted on a separate membrane to the western blot membranes and

so normalisation to H3 from the western blots could skew the data by any differences in pipetting error. These values showed an increase in R-loop levels with siAQR or siRNASEH2A transfection, except for a few of the siRNASEH2A-treated samples in experiment 1. This demonstrates that, as expected, knockdown of the R-loop repressor proteins increased R-loop abundance in the cells. Of note, the s9.6 signal in the 4hr auxin withdrawal siAQR + siNIPBL experiment 1 sample was more increased compared to the 4hr auxin withdrawal siAQR + siNIPBL sample from experiment 2. RAD21 reloading efficiency in this experiment 1 sample was also increased compared to the other experiment 1 samples, suggesting that in a condition of increased R-loops, reloading of cohesin on chromatin is increased. SA1 levels were also increased in this sample. Similarly, s9.6 levels were increased in siRNASEH2A samples in experiment 2 concomitant with an increase in RAD21 reloading efficiency. Thus, overall these experiments demonstrated that variability exists in RAD21 reassociation with chromatin in the absence of its canonical loader complex and in R-loop levels, but, when a condition of increased R-loops is observed, RAD21 reloading efficiency is also increased.

5.2.6 CTCF and SA colocalise at long-range contacts

Rao *et al.* (2017) determined that long-range chromosomal contacts are enriched upon auxin-mediated depletion of RAD21 from HCT116 RmAC OsTIR1 cells and that 41 of these 64 long-range contacts corresponded to the locations of super-enhancers. Dr. Christopher Barrington developed a script to identify long-range contact hotspots in the Hi-C data from this same paper (Supplemental Figure 8). To assess the distance between hotspots from control and auxin datasets, a density distribution plot of the distance between the midpoints of the hotspots was generated (Figure 81). Interactions at TAD-level distances (100 kb to 1 Mb) account for the majority of the control hotspots, but only a small proportion of the auxin hotspots. Between 1 Mb and 10 Mb, a similar proportion of hotspots was observed for both conditions. Very long-range contacts were enriched in the auxin condition, whereas, hotspots of this length only make up a small proportion of the control hotspots. Therefore, chromosomal contacts show a shift from TAD-level

distances to long-range distances with the loss of RAD21. NIPBL is established to localise at promoters engaged in long-range contacts (Seitan *et al.*, 2013; Muto *et al.*, 2014). When the long-range contact hotspots were subset for overlap with NIPBL peaks from control and auxin ChIP-seq datasets generated by Rao *et al.* (2017), the same density distribution as all hotspots was observed. Similarly, when the hotspots were subset for overlap with CTCF-SA1 and CTCF-SA2 colocalised ChIP peaks, a shift from TAD-level distance to long-range distance was observed. 40-56% of the long-range contact intervals overlap with CTCF-SA1 and CTCF-SA2 peaks from auxin-treated cells and 8-9% of CTCF-SA1 and CTCF-SA2 peaks from auxin-treated cells overlap with the long-range contact hotspots.

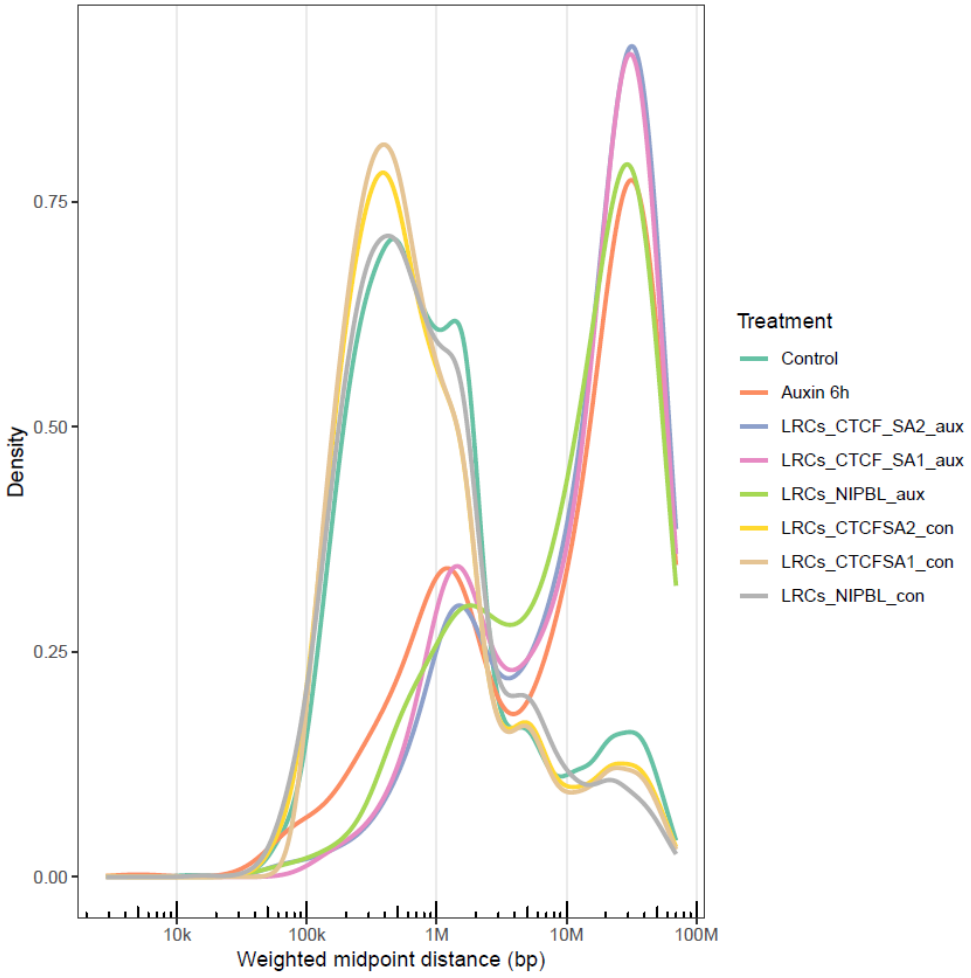


Figure 81: SA proteins localise to long-range chromatin contacts in the absence of cohesin. Hi-C data describing long-range contact (LRC) hotspots observed in HCT116 RmAC OsTIR1 cells upon depletion of cohesin was obtain from Rao et al. (2017) and processed by Dr. Christopher Barrington. Density of contact hotspot in control and auxin-treated cells. LRC intervals that overlap with CTCF and SA1/2 or NIPBL (ChIP data obtained from Rao et al. (2017)) are also plotted for control and auxin-treated datasets.

ChromHMM is software for the characterisation of chromatin states (Ernst and Kellis, 2017). By splitting the genome into sections (or 'bins') of a set size, ChromHMM will map the presence or absence of a mark along the genome to generate a pattern of signal. Patterns for multiple ChIP-seq datasets are compared and based on a set number of output chromatin states, the marks will be group according to the similarity of their signal pattern. Alongside ENCODE transcription factor and histone modification datasets, ChromHMM analysis was carried out to characterise the chromatin state of CTCF and SA sites in the presence or absence of RAD21. In control and auxin conditions, CTCF, SA1, and SA2, were group with control SMC3 and RAD21 datasets (Figure 82, left). In contrast, SMC3 and RAD21 auxin datasets had a lower likelihood of being in the same chromatin states (as indicated by the level of blueness). YY1, CBX3, SIN3A, RNA polymerase II, NIPBL, H3K27ac, and H3K4me3 also grouped with CTCF and SA, indicating a pattern of binding at promoter and transcription start sites. ChromHMM also computes enrichment of external annotations for each state to facilitate biological interpretation of each state (Figure 82, right). State 6 represents binding to CpG islands, exons and TSSs. Interestingly, R-loops are enriched at CpG islands, perhaps accounting for this enrichment.

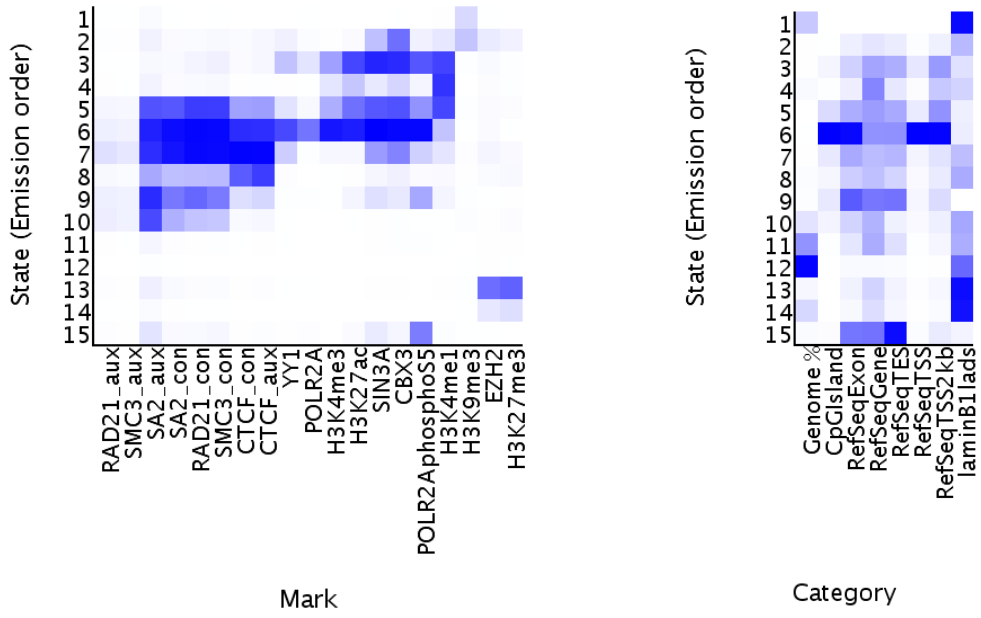


Figure 82: SA proteins localise to long-range chromatin contacts at genic regions in the absence of cohesin - ChromHMM. ChromHMM evaluation of 15 chromatin states in the indicated 'Mark' datasets (ChIP-Seq from Section 3.2.4 and ENCODE datasets see methods section 2.11). Enrichment of the category data (right) in the different chromatin states is automatically calculated by ChromHMM.

5.3 Discussion

Preliminary siRNA experiments in Hela cells determined the conditions required to detect NIPBL by western blot and provided insight into the dynamics of NIPBL vs MAU2 knockdown. Structural studies have reported extensive interaction of disordered regions of NIPBL and MAU2, and hence, the two proteins have been shown to stabilise one another (Watrin *et al.*, 2006; Hinshaw *et al.*, 2015; Chao *et al.*, 2017). Reciprocal knockdown of the loader complex proteins was observed in Hela cells, however, despite the loss of both proteins, siNIPBL and siMAU2 treatments differentially effected cohesin levels on chromatin. Cohesin levels were ~1/2 of control levels in siMAU2 samples and ~1/4 of control levels in siNIPBL. It is unclear how this difference arises as both proteins are lost in both conditions, so the influence of either protein should be lost regardless of siRNA treatment.

To minimise effect on cell cycle dynamics and cell health, siRNA knockdown of NIPBL and SA was tested across a range of concentrations at 72 and 96 hrs. These experiments revealed that the loader complex was efficiently depleted by 72hrs, meaning that no loading of cohesin mediated by NIPBL-MAU2 was occurring at this timepoint. Correspondingly, there was a decrease in cohesin levels on chromatin, however, this decrease was more severe by the 96 hrs timepoint, suggesting either the retention of residual stable cohesin on chromatin at 72 hrs or induction of cell cycle defects by 96 hrs. Interestingly, siSA treatment reduced NIPBL and MAU2 levels on chromatin following 72 or 96 hrs of knockdown, indicating a novel role of the SA proteins in stabilisation of the loader complex on chromatin. Fluorescent recovery after photobleaching (FRAP) and single-molecule tracking experiments in yeast and Hela cells indicate a much shorter residence time for NIPBL on chromatin compared to cohesin, suggesting that NIPBL only dynamically associate with chromatin and cohesin during loading (Hu *et al.*, 2011; Rhodes *et al.*, 2017). However, in additional studies in yeast and Drosophila cells, NIPBL orthologs have increased residency half-life on chromatin that is similar to cohesin (McNairn and Gerton, 2009; Gause *et al.*, 2010). Additionally, under conditions that stabilise cohesin residence on chromatin in Hela cells, increased abundance of NIPBL on chromatin has been observed, validating the idea that NIPBL occupancy on chromatin is influence by

cohesin (Rhodes *et al.*, 2017). The experiments described in this chapter further suggest that interaction with SA proteins plays a role in these chromatin-binding dynamics of NIPBL. It is not clear if SA proteins contribute to stabilisation of NIPBL during the loading reaction or continued binding of NIPBL thereafter, or both. Importantly, co-treatment with siNIPBL and auxin removed residual RAD21 and SMC3 from chromatin even with 72hrs of siRNA incubation, meaning that reloading experiments could be undertaken without question of the possible cell cycle effects observed at 96hrs. A flow cytometry experiment was attempted to more specifically address cell cycle in siNIPBL conditions however, staining was unsuccessful and time constraints of the PhD did not allow optimisation of the experiment.

Pilot reloading experiments in the polyclonal HCT116 RmAC OsTIR1 cells resulted in varying levels of RAD21 loss and consequently, varying magnitudes of RAD21 reassociation with chromatin following removal of auxin from the cell media. However, these experiments did suggest that reloading of cohesin could occur in the absence of the loader complex. Repetition in the FACS sorted H11 cell population confirmed that 4hrs-post auxin removal, RAD21 could be detected on chromatin. While reassociation was not seen to the endogenous levels of ethanol control samples, it was similar to that of auxin-treated siCon samples. Increasing the incubation withdrawal timepoint to allow more reformation of the cohesin complex did not increase association, likely due to over-stressing of the cells with the extended time in siRNA conditions. The diminished cohesin levels may also have been due to residual auxin in cells. Enhanced RAD21 expression following auxin wash-off has been recorded by treatment of cells with the OsTIR1 inhibitor auxinole (Yesbolatova *et al.*, 2019). Repetition of these experiments with inclusion of auxinole in the recovery media may increase the reassociation levels closer to endogenous levels.

Knockdown of the SA proteins alone or in combination with NIPBL was carried out to assess a potential role in the retained reloading ability. These samples determined that the SA proteins were even more essential for reassociation of RAD21 with chromatin than NIPBL and that, at least some, of the retained reloading ability was reliant on their presence in the cell. Reloading efficiency varied across the experiments however quantification of 6 replicates confirmed

the difference in reloading with loss of NIPBL and the SA proteins. Evidence exists in the literature suggesting that alongside the canonical NIPBL-MAU2 loading complex SA proteins contribute to cohesin's association with chromatin. In yeast, interaction of the SA orthologue with the loader complex is required for efficient association of the cohesin ring with DNA and subsequent ATPase activation (Murayama and Uhlmann, 2014; Orgil *et al.*, 2015). Separating interactions into SA-loader and cohesin ring-loader subcomplexes still impairs cohesin loading, indicating that SA functions as more than just a bridge protein (Orgil *et al.*, 2015). The experiments described here further indicate that SA proteins can induce cohesin's association with chromatin independently of NIPBL-MAU2.

Given the aforementioned reduction of NIPBL and MAU2 in siSA samples and the confounding effect of MAU2 loss on cohesin levels on chromatin, variation in MAU2 levels was assessed alongside the variation in reloading efficiency. It was determined that siSA samples with the highest levels of MAU2 did not correspond to the samples with the highest levels of cohesin reloading. Thus, loss of the loader complex in siSA samples likely contributes to reduced association of RAD21 with chromatin, but it is not the only contributing factor, indicating a role for the SA proteins themselves in cohesin loading.

Quantification of RAD21 levels in the four replicates with the highest levels of reassociation in their siCon sample, revealed that reassociation of cohesin with chromatin was equally efficient in siNIPBL and siSA samples. Only in the triple knockdown siNIPBL + siSA samples was a significant reduction in reloading efficiency observed compared to siCon. Thus, all three of these proteins play an important role in the loading of cohesin onto chromatin and it appears that the loader complex and SA proteins are able to compensate for each other activities, at least to some extent. Recent cryo-EM studies of cohesin in complex with its loader suggest that NIPBL and SA both wrap around the cohesin ring and DNA to position and further entrap DNA as it engaged by the cohesin ring, implying a role in the initial recruitment of cohesin to DNA alongside NIPBL (Higashi *et al.*, 2020; Shi *et al.*, 2020). The reloading experiments presented in this thesis suggest that NIPBL and SA can also induce such conformational changes separately. Functional independence of SA from NIPBL was shown by the

relative efficiency of RAD21 reassociation with chromatin in siNIPBL and siSA samples and interaction of SA1 with chromatin and CTCF in the absence of RAD21 either in complex with or independently of NIPBL. Hence, it is possible that in the absence of RAD21, SA1, and perhaps SA2, can interact with chromatin at sites determined by interaction with numerous binding partners, such as CTCF, and here contribute to the loading of cohesin in the presence or absence of NIPBL. NIPBL and SA have strikingly similar crystal structures (Hara *et al.*, 2014; Kikuchi *et al.*, 2016; Chao *et al.*, 2017) and thus may be capable of inducing similar conformational changes in cohesin and DNA. Structural analysis of the cohesin ring loading onto DNA in the presence and absence of SA and/or NIPBL-MAU2 would be required to investigate such activity.

It is not clear if the cohesin reloaded in the absence of NIPBL represents stably loaded cohesin or a pool of rapidly associating and dissociating cohesin. Recombinant cohesin has been shown to have an intrinsic ability to bind to DNA molecules in the absence of its loader complex *in vitro*, however, binding is less efficient and less stable than in the presence of the loader (Murayama and Uhlmann, 2014; Stigler *et al.*, 2016; Davidson *et al.*, 2019). This raises the question of whether cohesin can bind to chromatin in a physiologically relevant manner in the absence of NIPBL-MAU2. Murayama and Uhlmann (2014) observed a further loss of DNA-binding efficiency in the absence of the yeast SA ortholog, suggesting that SA proteins specifically mediate this intrinsic ability. The majority of studies investigating intrinsic activity of cohesin consider SA proteins as core complex members, rather than regulators of the cohesin ring, so this specificity of the activity can be overlooked. The reloading experiments presented here indicate that in dividing cells SA can indeed contribute to association of cohesin with chromatin and that the extent of loading is similar to NIPBL-MAU2 alone. Hence, it is important to understand this activity more fully. For example, could the SA and NIPBL compensation described in this thesis vary in human populations and play a role in disease severity in developmental diseases and cancers driven by loss-of-function of SA1, SA2, or NIPBL.

It is also possible that, without its canonical loader, cohesin randomly associates with the chromatin and thus is not localised to properly mediate organisation of the chromatin. Rao *et al.* (2017) report restoration of chromosome contacts

following 3 hrs of withdrawal from auxin. This suggests that at least for the siCon samples here, the levels of RAD21 restored on chromatin are sufficient to mediate chromosome organisation. As similar levels of RAD21 were restored in the absence of the cohesin loader, SA-mediated loading may be sufficient to maintain chromosome organisation. Stigler *et al.,* (2016) observed a strong preference for cohesin loading at A/T-rich DNA *in vitro*, potentially mediated by affinity of the cohesin loader complex for such sequences. SA1 can also bind to A/T-rich DNA via an AT-hook region in its N-terminus, perhaps suggesting that *in vivo*, both the loader complex or SA could direct cohesin to such locations (Bisht, Daniloski and Smith, 2013; Lin *et al.*, 2016). The work present in Chapter 4 suggests that SA could localise cohesin via interaction with specific proteins or nucleic acid structures. Hence, either via affinity for A/T-rich DNA or specific interactions, SA-mediated loading may localise cohesin to the required genomic locations.

In the absence of the NIPBL-MAU2 loader complex, increasing R-loop levels by siRNA-mediated knockdown of R-loop regulatory proteins increased loading of RAD21 on chromatin. Chromatin-bound SA1 was increased in the sample with highest R-loop levels and can interact with both R-loop interactome members and the R-loops themselves suggesting that SA1 may play a role in this reloading. These results support the importance of DNA structure in cohesin loading that has previously been shown on ds- and ss-DNA plasmids *in vitro* and at the replication fork in Hela cells (Murayama *et al.*, 2018; Zheng *et al.*, 2018). As discussed in Chapter 4 R-loop structures occur at sites of transcription – of both mRNA and rDNA – replication, and DNA damage (El Hage *et al.*, 2010; Chakraborty and Grosse, 2011; Skourti-Stathaki, Proudfoot and Gromak, 2011; Sollier *et al.*, 2014; Yang *et al.*, 2014; Schwab *et al.*, 2015; Salas‐Armenteros *et al.*, 2017; Yasuhara *et al.*, 2018). Therefore, binding of SA to R-loops could help to localise cohesin to a wide variety of genomic locations throughout the nucleoplasm and nucleolus and once again SA-mediated loading may localise cohesin to important genomic locations.

Structural analysis of chromosome organisation in HCT116 RmAC OsTIR1 cells depleted of cohesin has previously revealed enrichment of long-range cis contacts, and clustering of superenhancers thereby (Rao *et al.*, 2017a). I further

show here that these long-range hotspots overlap with CTCF-SA1 and CTCF-SA2 binding sites. Perhaps suggesting that even in the absence of cohesin, the nucleic acid- and protein-binding abilities of SA can mediate structural organisation in cells. Genome compartments are preserved in the ethanol and auxin Hi-C maps generated by Rao *et al.* (2017), indicating that compartments are not lost with RAD21 depletion. Simulation of DNA contact maps produced by phase separation are reminiscent of compartmentalizastion maps. As SA protein have intriniscially disordered regions and can interact with both DNA and RNA, this raises the question of whether the SA observed at these long-range contacts may contribute to condensation of compartments of chromatin even in the absence of the cohesin ring.

Characterisation of the chromatin ‘state’ was carried out in ChromHMM by comparison with ENCODE ChIP-Seq datasets. CTCF and SA sites in the absence of cohesin were characteristic of CpG islands, exons, and TSSs. As R-loops are enriched at CpG islands, due to their G-rich nature, this is again suggestive of the importance of SA targeting to such sites. Together these analyses suggest that in cells depleted for RAD21, SA proteins remain bound to chromatin at active transcription sites that retain long-range chromosomal contacts. The reloading experiments further suggest that SA proteins at such site may mediate loading of cohesin on to chromatin, in the presence or absence of NIPBL. Such clustering of cohesin loading sites in the nucleus would allow easier regulation of loading in areas of specific metabolites, regulatory proteins, and feedback loop activity.

6

Conclusions & Future Perspectives

From yeast to humans, cohesin ring components have always been found in complex with an SA protein. As such, SA proteins are considered as core members of the complex and their function is studied in the context of cohesin activity. Yet, we still do not understand the full contribution of the SA proteins to cohesin function and cell identity or how the divergence of multiple paralogs across evolution differentially contributes to these processes. Over the past 10 years, mutation of SA2, and increasingly of SA1, has been identified in numerous cancers, including, bladder cancer, Ewing sarcoma, glioblastoma multiform (GBM), and acute myeloid leukaemia (Rocquain *et al.*, 2010; Solomon *et al.*, 2011; Balbás-Martínez *et al.*, 2013; Guo *et al.*, 2013; Romero-Pérez *et al.*, 2019). In fact, SA2 is one of only twelve genes that contains statistically significant somatic point mutations in four or more cancer types (Lawrence *et al.*, 2014). As such, understanding the function of SA1 and SA2 is imperative to help us to understand mechanisms of chromosomal organisation and the aetiology of these cancers. Using acute depletion of Rad21 this thesis shows that SA proteins remain on chromatin and in complex with a wide variety of proteins in the absence of the cohesin ring proteins. This represents a heretofore undiscovered role for the SA proteins and an important new activity to understand in the context of development and disease.

Optimisation of a co-IP protocol that allows detection of the cohesin proteins in complex under endogenous conditions allowed determination of physiologically relevant conditions of interaction. Sufficient cell number, a salt concentration of 200mM KCl and digestion of nucleic acids were required for most efficient interaction of the SA proteins with CTCF. Fragmentation of DNA and RNA to a size range of 25 – 1000bp was optimal for CTCF-SA co-IP, however this still

represents quite a broad size range. In ChIP-seq experiments sonication is used to digest DNA to fragments in a tighter size range. More specific digestion of DNA and RNA, such as in a ChIP-seq experiments, would help to determine more exactly the nucleic acid molecules mediating this interaction.

In the presence and absence of cohesin, CTCF was strikingly enriched with SA1 compared to SA2. The use of endogenous conditions and the opposite enrichment of RAD21 indicate that this illustrates varied stability between the two interactions. Investigation of naturally occurring variants suggested that a basic, C-terminal domain contributes to this difference, however, it was not possible to confirm this due to inconclusive western blot of the SA2 variants. siRNAs specifically targeting the different variants and co-IP with CTCF may help to confirm the importance of this exon. Alternatively, CRISPR-Cas9-mediated deletion of exon 31 in SA1 could help confirm its importance for interaction with CTCF. However, disruption of an essential domain of SA1 that is required for interaction with CTCF, nucleic acids, or other unknown functions could have adverse effects on cell viability, making clonal selection of the deletion difficult. Mutation mapping of recombinant version of the proteins have already been published and the varied N- and C-terminal dependencies observed between different studies suggest that the interaction is multiplex in nature (Xiao, Wallace and Felsenfeld, 2011; Li *et al.*, 2020; Nishana *et al.*, 2020; Pugacheva *et al.*, 2020).

ChIP-seq was used as an orthogonal approach to investigate CTCF and SA interaction and confirmed overlapping distribution of SA1 and SA2 with CTCF in the absence of the cohesin ring proteins. The population level nature of ChIP-seq means that it cannot be used to test for specific interactions on chromatin. As such, Re-ChIP methods have been developed that allow detection of multiple proteins bound to a single DNA sequence by sequential IP of the proteins of interest prior to library preparation and sequencing (Geisberg and Struhl, 2005; Truax and Greer, 2012). However, loss of material is seen with each sequential IP and subsequently, libraries may have high PCR duplication levels and low complexity. Importantly, Re-ChIP only determines that proteins localise to the same region of DNA, no direct interaction can be established. The joint BiFC-ChIP method developed here represents a promising method to detect the

chromatin localisation of proteins directly in complex with one another. This method could be used to confirm the genomic locations of CTCF-SA1 and CTCF-SA2 complexes in future. While the literature suggests that SA2 is more commonly located to genic regions than SA1, the majority of SA sites were overlapping in this study and both were found at sites of transcription, genes, and CpG islands. The postdoctoral researchers Dr. Stanimir Dulev and Dr. Yang Li confirmed interaction and co-localisation of SA and CTCF in the absence of cohesin using siRNA knockdown in a separate cell line (U2OS/Hela) and immunofluorescence, respectively. Dr. Stanimir Dulev also confirmed interaction of SA and CTCF following siRNA-mediated knockdown of SMC3, confirming independence from cohesin (Porter *et al.*, 2021).

Mass Spectrometry revealed the SA1 interactome in the presence and absence of the cohesin ring. This work determined that SA1 interacts with proteins involved in chromosome organisation, transcription, RNA processing, ribosome biogenesis, translation, DNA replication, and DNA repair. Interaction with the proteins involved in chromosome organisation, transcription, RNA processing, ribosome biogenesis, and translation were maintained in the absence of the cohesin ring. Furthermore, there was a significant enrichment of proteins involved in ribosome biogenesis and RNA processing with RAD21 depletion. IP from an unrelated chromatin-binding protein in ethanol and auxin conditions would help to confirm specificity of the proteins identified. Banded mass spectrometry of CTCF and SA2 suggest that CTCF enriches a very similar set of proteins to SA1, while SA2 shows similar and distinct interactions. Chromatin-binding and -regulatory, RNA binding, DNA binding, translational, and cytoskeletal proteins were major protein classes enriched by all three proteins. Whereas gene-specific transcriptional regulators were only enriched with CTCF and SA1 and a large class of metabolite interconversion enzymes were only enriched with SA2. This similarity of SA1 with CTCF compared to SA2 correlates well with the different efficacies of co-IP between the proteins. Full lane samples would be required to confirm these results, and especially the lack of transcriptional regulators observed in the SA2 IP, as ChIP-seq experiments suggest it is highly enriched to genic regions (Faure *et al.*, 2012; Kojic *et al.*, 2018; Cuadrado *et al.*, 2019; Viny *et al.*, 2019; Casa *et al.*, 2020). However, there was compelling similarity between the enriched protein groups from the SA1 banded and full lane mass spec

samples, suggesting that the banded mass spec protein lists are indicative of the IP proteins interactome.

FGF-motif proteins have been shown to interact with SA2 via its CES (Hara *et al.*, 2014; Li *et al.*, 2020). Here interaction of FGF-like motif proteins with SA1 and SA2 was confirmed in human cells. Furthermore, increased enrichment with SA1 compared to SA2 was observed, suggesting that the interaction is maintained by more than just the CES. Mutation of the FGF-like motif in candidate proteins may help to confirm the specificity of the interaction. In addition, a full lane SA2 IP would help to assess all FGF-like motif proteins that interact with SA2. It would be interesting to assess by IP and western blot if any of the FGF-like motif proteins are enriched with SA2 compared to SA1 in human cells. For example, sororin has been shown to co-IP SA2 *in vitro* (Zhang and Pati, 2015) and was not detected in the SA1 IP-MS experiments. If SA2 does in fact interact with Sororin, it would be interesting to assess if mutation of the FGF-motif in sororin disrupts interaction with SA2 and does this have any downstream effect on sororin-mediated stabilisation of cohesin. Downstream effects could be tested via a range of experiments, including, i) analysis of chromosome spreads to assess integrity of sister chromatid cohesion, ii) FRAP of cohesin to assess changes to cohesin stabilisation, and iii) quantification of WAPL co-IP with cohesin to assess potential WAPL anagonsim. As Sororin interacts with PDS5 in a manner dependent on its FGF-motif (Nishiyama *et al.*, 2010), sororin-PDS5 interaction would also need to be assess compared to sororin-SA2 interaction. Artificial fusion of sororin-PDS5 compared to artificial fusion of sororin-SA2 may help to differentiate which interaction prevents WAPL-mediated release of cohesin. It is also possible that both interactions are important, although the presence of only one FGF-motif in sororin suggests is should only interact with one protein at a time via this motif.

Further to the striking enrichment of RNA-binding and processing proteins in the SA1 interactome, interaction of SA1 and SA2 with R-loops was confirmed. Whether interaction with RNA-binding and processing factors localises the SA proteins to R-loops or interaction with the R-loop localises the SA proteins to the RNA processing factors remains unclear. Digestion of R-loops and IP-MS of the SA1 interactome was initially trialled to help answer such a question however R-loop digestion was more difficult than anticipated and further time would be

required to achieve the replicates required for such an experiment. In collaboration with Professor Richard Jenners lab interaction of SA1 and SA2 with RNA was also confirmed. Binding to RNA could also contribute to SA association with R-loops or localisation at sites of active transcription. FGF-like motif proteins are predicted to interact with the SA proteins via their CES domain (Li *et al.*, 2020). However, co-IP of the endogenous proteins in HCT116 cells showed increase co-IP with SA1 compared to SA2, thus, indicating that the interaction is mediated by more than just the CES domain. It is possible that RNA might interact more strongly with SA1, due to inclusion of the basic exon 31 in its C-terminus, and contribute to the different stabilisation of protein interactions observed. Although there was not an extreme difference in RNA enrichment in the SA1 and SA2 CLIP. DNA or alterative domains in the SA proteins may also contribute to stabilisation of the interactions. For example, SA1 also contains an AT-hook in its N-terminus that promotes interaction with AT-rich DNA (Bisht, Daniloski and Smith, 2013; Lin *et al.*, 2016). The SA proteins are most divergent in their terminal ends, perhaps as mutations in these regions were selected for effect on stabilising different proteins and nucleic acid interactions between SA paralogs. CRISPR-Cas9-mediated tagging of the SA1 AT-hook domain into SA2's N-terminus could help to assess if it is sufficient to mediate SA1-specific activities, however, it would be extremely important and likely very difficult to ensure proper tertiary structure was retained following such modification.

The tertiary structure of SA and NIPBL proteins is strikingly similar and so potential overlapping function in cohesin loading was investigated. This thesis uncovered NIPBL-independent loading of cohesin in HCT116 cells and illuminated a role for the SA proteins in this activity. Very recent crystallisation studies in yeast and human cells indicate a role for SA1, at least, in bending DNA and cohesin to induce loading, in combination with NIPBL (Higashi *et al.*, 2020; Shi *et al.*, 2020). Accordingly, only when NIPBL, SA1, and SA2 were depleted from cells was cohesin association with chromatin significantly reduced compared to control levels. Confirmation of these results using the dCas9-KRAB repression system to prevent expression of NIPBL would corroborate these results and ensure that artefacts of siRNA transfection did not impact the output. Guide RNAs for the NIPBL promoter and expression of three different dCas9-KRAB plasmids have been evaluated. Unfortunately, generation of the dCas9-

Krab cell line was not possible within the timeline of this thesis. Modulation of R-loop levels by repression of RNase H2 and AQR confirmed that in conditions of enriched R-loops, increased association of cohesin with chromatin occurs, in a NIPBL-independent manner. It would be ideal to confirm the importance of the SA proteins for this increased re-loading, however, treatment of the cells with additional siRNAs would likely render them inviable as the conditions already had to be altered to preserve cell growth/health. However specific increase of SA1 alongside R-loops suggests that SA1 is present at the new cohesin loading sites, at least. Repetition of this experiment to confirm the results and allow statistical analysis of the different protein dynamics is required.

Single-molecule tracking or FRAP experiments may help to assess if the reloaded cohesin is stabilised on chromatin or represents a stochastic pool of cohesin that can rapidly associate and dissociate from chromatin. A more simple alternative to assess stability of chromatin association may be to determine the salt fraction that reloaded cohesin occupies in siSA and siNIPBL chromatin samples, and whether there is a difference in the severity of the conditions required to solubilise it. Additional experiments to assess the functional capacity of the reloaded cohesin include ChIP-seq analysis of distribution at canonical cohesin binding sites and Hi-C analysis of chromosome contacts in the absence of the SA and NIPBL proteins.

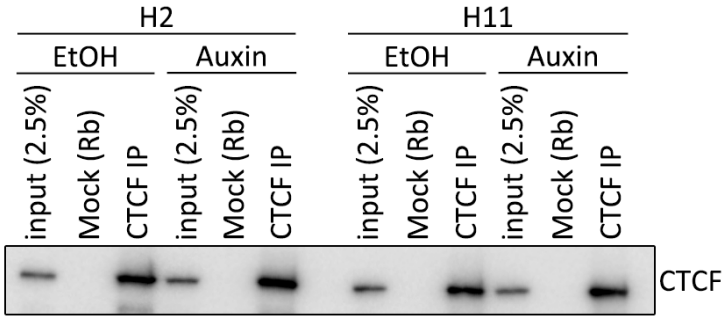
Long-range contacts and ‘hubs’ of superenhancers are observed in the absence of RAD21 (Rao *et al.,* 2017). I determined that CTCF-SA1 and CTCF-SA2 overlap with this long-range hospots, suggesting that CTCF and SA may contribute to the interactions observed. ChIP-qPCR of RAD21 at a range of these sites following ethanol and auxin withdrawal may indicate if these sites represent sites of cohesin reloading with different combination of SA and NIPBL-MAU2 proteins. Comparison with random sites and CTCF-SA sites that do no overlap with the long-range hotspots would need to be run for comparison. Atomic force microscopy (AFM) of CTCF, SA, DNA, and RNA could be used to help investigate whether the clusters observed represent condensates. Mutation of the CES and exon31/32 in the SA protein may further help to determine if these domains contribute to potential condensates via protein or nucleic acid interactions,

respectively. AFM of cohesin +/- SA1 and SA2 would also be of interest to determine the contribute of the individual SA proteins to condensation of the DNA.

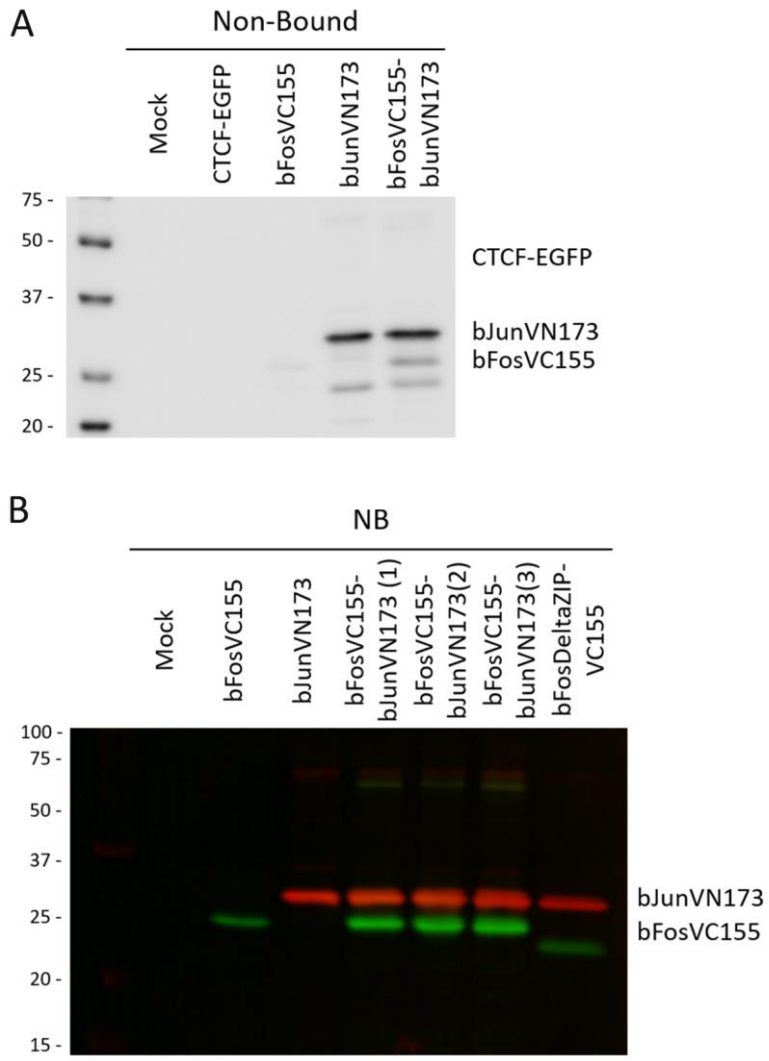
Overall, this thesis demonstrates new understanding into the function of SA proteins, differences between their activity, and the molecular mechanisms of cohesin biology.

7

Supplemental Figures



Supplemental Figure 1: Replicate CTCF IP in H2 and H11 clones. HCT116 RmAC OsTIR1 H2 and H11 clones were treated with ethanol or auxin for 4hrs, as indicated. Cells were collected and fractionated from chromatin. Chromatin-bound proteins were IP'd using endogenous antibodies to IgG (Mock) or CTCF. IP elutes were run on a gel and immunoblotted for CTCF.



Supplemental Figure 2: (A) Non-bound IP samples corresponding to Figure 24C. (B) Non-bound IP samples corresponding to Figure 25A.

A

```
SA1 isoform 1   1     MITS-ELPVLQDSTNETTAHSDAGSELEETEVKGKRKRGRPGRPPSTNKKPRKSPGEKSRIEAGIRGAGR-----GRANG  74
SA2 isoform 2   1     MIAAPEIPTDFNLLQESETHFSSDTDFEDIE--GKNQKQGKGK---TCKKGKKGPAEKGK---GGNGGGKPPSGPNRMNG  72

SA1 isoform 1   75    HPQQNGEGEPVTLFEVVKLGKSAMQSVVDDWIESYKQDRDIALLDLINFFIQCSGCRGTVRIEMFRNMQNAEIIRKMTEE  154
SA2 isoform 2   73    HHQQNGV-ENMMLFEVVKMGKSAMQSVVDDWIESYKHDRDIALLDLINFFIQCSGCKGVVTAEMFRHMQNSEIIRKMTEE  151

SA1 isoform 1   155   FDEDSGDYPLTMPGPQWKKFRSNFCEFIGVLIRQCQYSIIYDEYMMDTVISLLTGLSDSQVRAFRHTSTLAAMKLMTALV  234
SA2 isoform 2   152   FDEDSGDYPLTMAGPQWKKFKSSFCEFIGVLVRQCQYSIIYDEYMMDTVISLLTGLSDSQVRAFRHTSTLAAMKLMTALV  231

SA1 isoform 1   235   NVALNLSIHQDNTQRQYEAERNKMIGKRANERLELLLQKRKELQENQDEIENMMNSIFKGIFVHRYRDAIAEIRAICIEE  314
SA2 isoform 2   232   NVALNLSINMDNTQRQYEAERNKMIGKRANERLELLLQKRKELQENQDEIENMMNAIFKGVFVHRYRDAIAEIRAICIEE  311

SA1 isoform 1   315   IGVWMKMYSDAFLNDSYLKYVGWTLHDRQGEVRLKCLKALQSLYTNRELFPKLELFTNRFKDRIVSMTLDKEYDVAVEAI  394
SA2 isoform 2   312   IGIWMKMYSDAFLNDSYLKYVGWTMHDKQGEVRLKCLTALQGLYYNKELNSKLELFTSRFKDRIVSMTLDKEYDVAVQAI  391

SA1 isoform 1   395   RLVTLILHGSEEALSNEDCENVYHLVYSAHRPVAVAAGEFLHKKLFSRHDPQAEEALAKRRGRNSPNGNLIRMLVLFFLE  474
SA2 isoform 2   392   KLLTLVLQSSEEVLTAEDCENVYHLVYSAHRPVAVAAGEFLYKKLFSRRDPE-EDGMMKRRGRQGPNANLVKTLVFFFLE  470

SA1 isoform 1   475   SELHEHAAYLVDSLWESSQELLKDWECMTELLLEEPVQGEEAMSDRQESALIELMVCTIRQAAEAHPPVGRGTGKRVLTA  554
SA2 isoform 2   471   SELHEHAAYLVDSMWDCATELLKDWECMNSLLLEEPLSGEEALTDRQESALIEIMLCTIRQAAECHPPVGRGTGKRVLTA  550

SA1 isoform 1   555   KERKTQIDDRNKLTEHFIITLPMLLSKYSADAEKVANLLQIPQYFDLEIYSTGRMEKHLDALLKQIKFVVEKHVESDVLE  634
SA2 isoform 2   551   KEKKTQLDDRTKITELFAVALPQLLAKYSVDAEKVTNLLQLPQYFDLEIYTTGRLEKHLDALLRQIRNIVEKHTDTDVLE  630

SA1 isoform 1   635   ACSKTYSILCSEEYTIQNRVDIARSQLIDEFVDRFNHSVEDLLQEGEEADDDDIYNVLSTLKRLTSFHNAHDLTKWDLFG  714
SA2 isoform 2   631   ACSKTYHALCNEEFTIFNRVDISRSQLIDELADKFNRLLEDFLQEGEEPDEDDAYQVLSTLKRITAFHNAHDLSKWDLFA  710

SA1 isoform 1   715   NCYRLLKTGIEHGAMPEQIVVQALQCSHYSILWQLVKITDGSPSKEDLLVLRKTVKSFLAVCQQCLSNVNTPVKEQAFML  794
SA2 isoform 2   711   CNYKLLKTGIENGDMPEQIVIHALQCTHYVILWQLAKITESSSTKEDLLRLKKQMRVFCQICQHYLTNVNTTVKEQAFTI  790

SA1 isoform 1   795   LCDLLMIFSHQLMTGGREGLQPLVFNPDTGLQSELLSFVMDHVFIDQDEENQSMEGDEEDEANKIEALHKRRNLLAAFSK  874
SA2 isoform 2   791   LCDILMIFSHQIMSGGRDMLEPLVYTPDSSLQSELLSFILDHVFIEQDDDNNSADGQQEDEASKIEALHKRRNLLAAFCK  870

SA1 isoform 1   875   LIIYDIVDMHAAADIFKHYMKYYNDYGDIIKETLSKTRQIDKIQCAKTLILSLQQLFNELVQEQGPNLDRTSAHVSGIKE  954
SA2 isoform 2   871   LIVYTVVEMNTAADIFKQYMKYYNDYGDIIKETMSKTRQIDKIQCAKTLILSLQQLFNEMIQENGYNFDRSSSTFSGIKE  950

SA1 isoform 1   955   LARRFALTFGLDQIKTREAVATLHKDGIEFAFKYQNQKGQEYPPPNLAFLEVLSEFSSKLLRQDKKTVHSYLEKFLTEQM  1034
SA2 isoform 2   951   LARRFALTFGLDQLKTREAIAMLHKDGIEFAFKEPNPQGESHPPLNLAFLDILSEFSSKLLRQDKRTVYVYLEKFMTFQM  1030

SA1 isoform 1   1035  MERREDVWLPLISYRNSLVTGGEDDRMSVNSGSSSSKTSSVRNKKGRPPLHKKRV----------EDESLDNTWLNRTDT  1104
SA2 isoform 2   1031  SLRREDVWLPLMSYRNSLLAGGDDDTMSVISGI-SSRGSTVRSKKSKPSTGKRKVVEGMQLSLTEESSSSDSMWLSREQT  1109

SA1 isoform 1   1105  MIQTPGPLPAPQLTSTVLRENSRPMGDQ----IQEPESEHGSEP-DFLHNPQMQISW-LGQPKLEDLNRK-DRTGMNYMK  1177
SA2 isoform 2   1110  L-HTPVMMQTPQLTSTIMREPKRLRPEDSFMSVYPMQTEHHQTPLDY----NTQVTWMLAQRQQEEARQQQERAAMSYVK  1184

SA1 isoform 1   1178  VRTGVRHAVR---GLMEEDAEPIFEDVMMSSRSQLEDMNEEFE-DTMVIDLPPSRNRRERAELRPDFFDSAAIIEDDSGF  1253
SA2 isoform 2   1185  LRTNLQHAIRRGTSLMEDDEEPIVEDVMMSSEGRIEDLNEGMDFDTMDIDLPPSKNRRERTELKPDFFDPASIM-DESVL  1263

SA1 isoform 1   1254  GMPMF  1258
SA2 isoform 2   1264  GVSMF  1268
```

B

NW Score	Identities	Positives	Gaps
-55	9/62(15%)	16/62(25%)	19/62(30%)

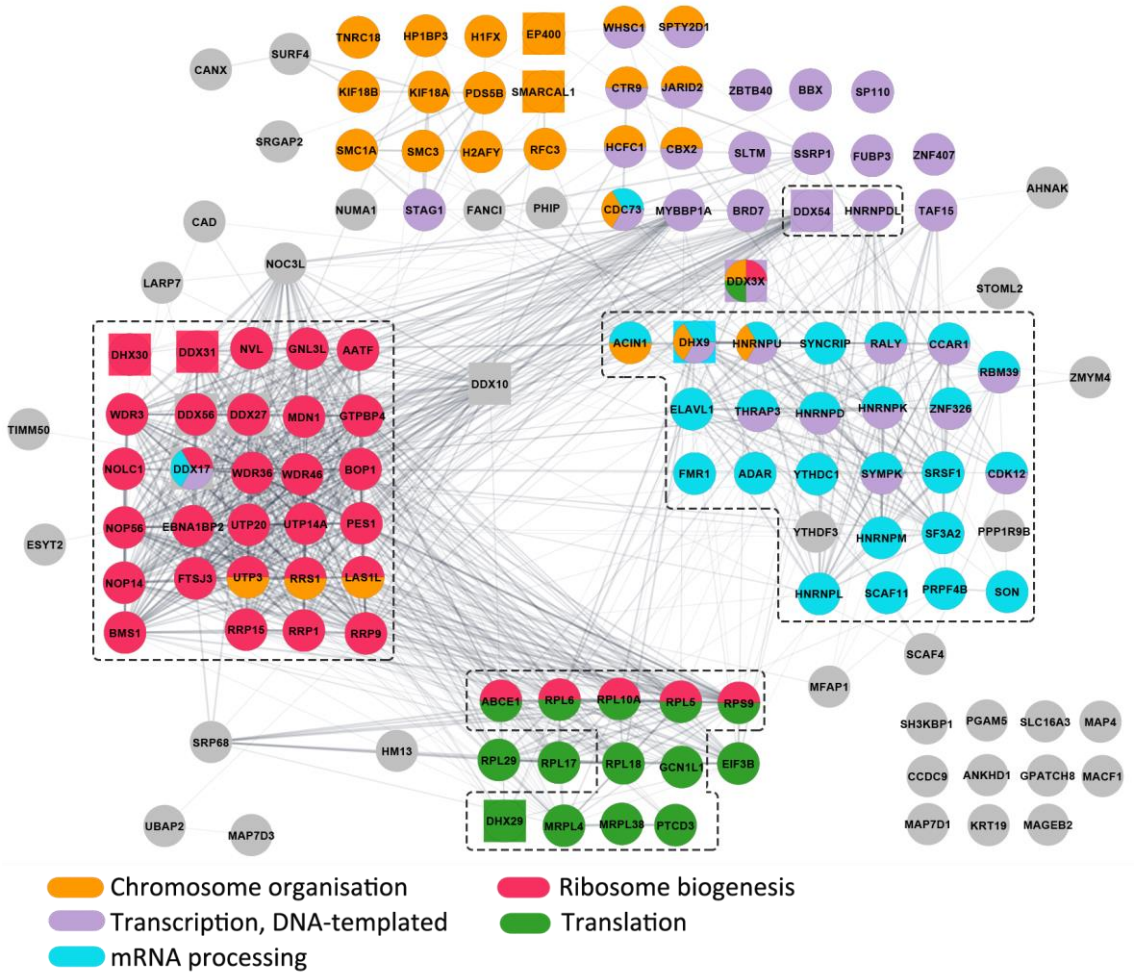
```
SA1 exon31  1    PQMQIS---WLGQ-----------PKLEDLNRKD----RTGMNYMKVRTGVRHAV
                  +   S   WL +           P+L     ++    R   ++M V     H
SA2 exon31  1    EESSSSDSMWLSREQTLHTPVMMQTPQLTSTIMREPKRLRPEDSFMSVYMQTEHHQTPLD

SA1 exon31  61   YN  62
```

C

NW Score	Identities	Positives	Gaps
-22	6/50(12%)	13/50(26%)	22/50(44%

```
SA1 exon31  1    PQ-------------MQISWLGQPKLEDLNRKDRTGMNYMKVRTGVRHAV  37
                  +               +    + ++EDLN  D   M+           +
SA2 exon33  1    RRGTSLMEDDEEPIVEDVMMSSEGRIEDLNGMDFDTMDI---------DL  41
```

Supplemental Figure 3: (A) NCBI Constraint-based Multiple Alignment Tool (COBALT) global alignment of SA1 +exon31 and SA2 +exon32. A 3 bit conservation setting was applied. Red residues are conserved, blue residues are less conserved with no gaps, and gray residues are less conserved and bridge gaps. Blast® Needleman-Wunsch local alignment results for compaison of the protein sequence of SA1 exon31 and SA2 exon31 (B) or SA2 exon33 (C). Summary statistics for the alignment are shown at the top of the table. Sequences for each of the exons were obtained from the UCSC genome browser Human GRCh38 track. Alignment of the two sequnces is shown in the middle line of sequence; a letter indicates an identical match, + indicates a positive match of similarity, and white space indicates mismatch.

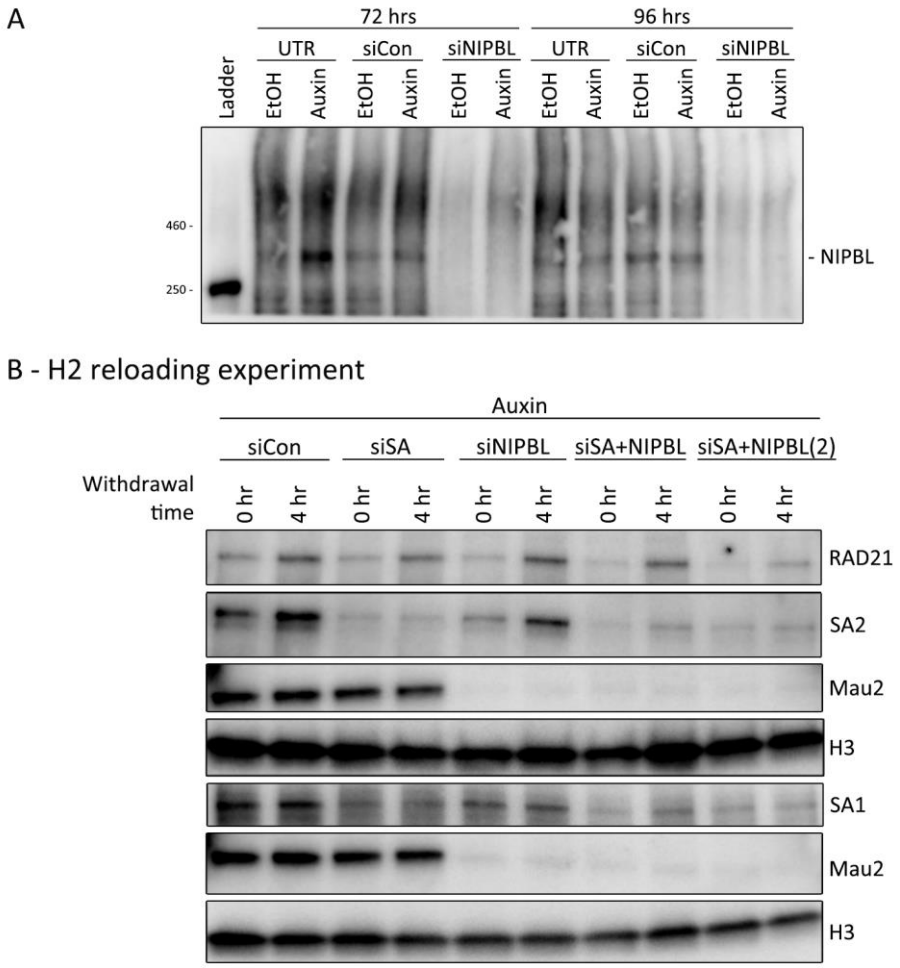


Supplemental Figure 4: Full network of proteins co-purified with SA1 and with altered abundance in IAA conditions compared to UTR conditions. A subset of this network is shown in Figure 48 A. Proteins were considered to have altered abundance for log2FC ≥ 0.581 and ≤ -0.58 and a pvalue of < 0.05. Protein interactions and GO term enrichment analysis was generated using STRING. Node colours denote the major enriched categories compare to whole genome background, with square nodes signifying helicase proteins. Dotted lines encompass the processes within the network that are enriched within the SA1 interactome itself with IAA treatment, compared to the UTR SA1 interactome.

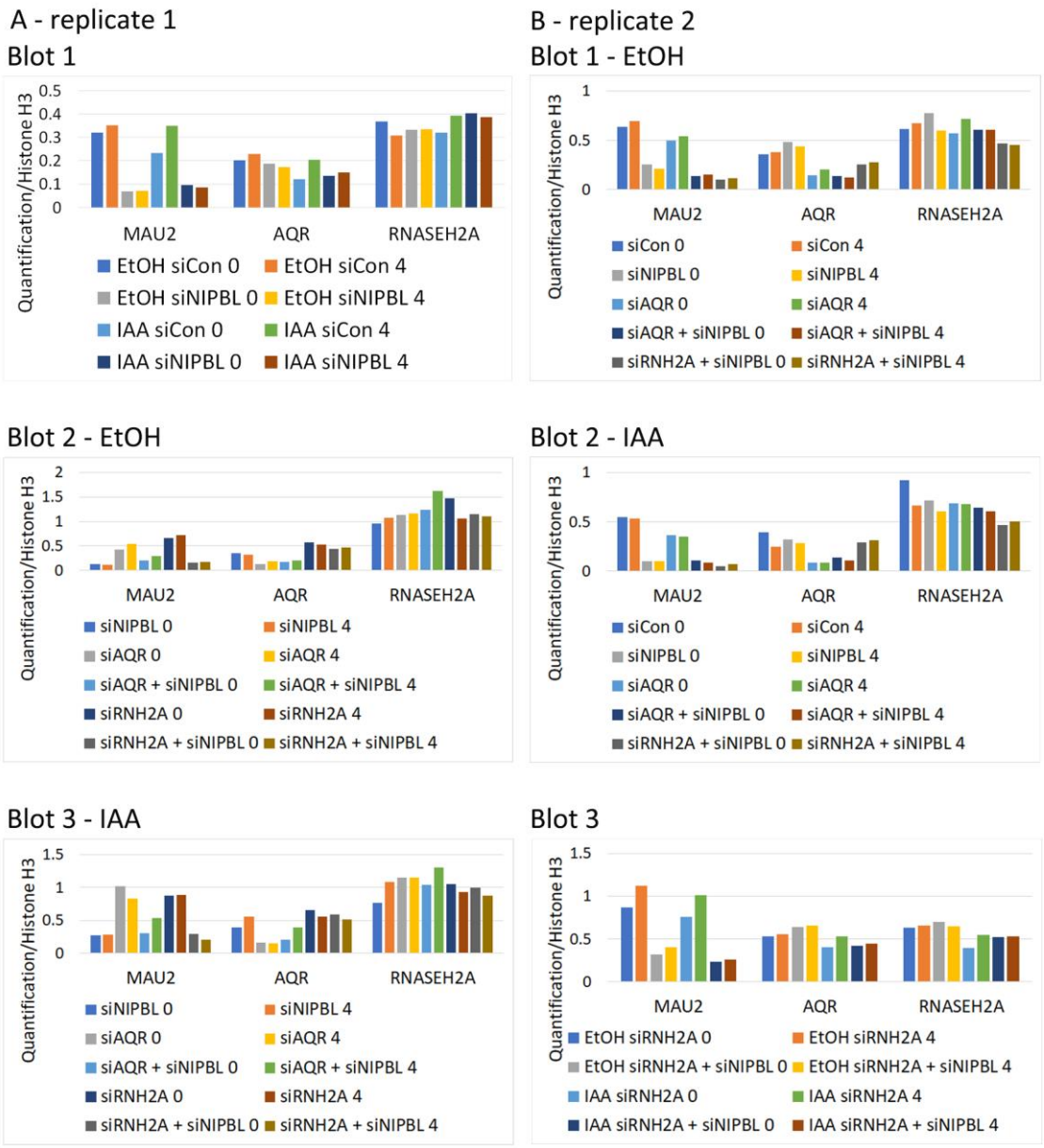
ProteinAcc	ProteinName	GeneName	Hit
A0FGR8	Extended synaptotagmin-2	ESYT2	sqrrsAFGFDDgnfpg
A5PLL1	Ankyrin repeat domain-containing protein 34B	ANKRD34B	eteltLFGFKDLElagsn
O43525	Potassium voltage-gated channel subfamily KQT member 3	KCNQ3	idkvsPYGFFAHDpvnlp
O43776	Asparagine--tRNA ligase, cytoplasmic	NARS	kedgtFYEFGEDipeap
P04629	High affinity nerve growth factor receptor	NTRK1	afmdnPFEFNPEdpipv
P12111	Collagen alpha-3(VI) chain	COL6A3	sgpveAFDFDEyqpem
P17302	Gap junction alpha-1 protein	GJA1	nshaqPFDFPDDnqnsk
P25205	DNA replication licensing factor MCM3	MCM3	gdsydPYDFSDTEeempq
P49711	Transcriptional repressor CTCF	CTCF	dvdvsVYDFEEEqqegl
Q01433	AMP deaminase 2	AMPD2	elrsaPYEFPEEspieq
Q13164	Mitogen-activated protein kinase 7	MAPK7	pdcapPFDFAFDrealt
Q13347	Eukaryotic translation initiation factor 3 subunit I	EIF3I	yfdpqYFEFEFEa
Q14676	Mediator of DNA damage checkpoint protein 1	MDC1	leraqPFGFIDSDtdaee
Q1KMD3	Heterogeneous nuclear ribonucleoprotein U-like protein 2	HNRNPUL2	ehgraYYEFREEayhsr
Q53T59	HCLS1-binding protein 3	HS1BP3	gndeeAFDFFEEqdqva
Q5FBB7	Shugoshin 1	SGO1	vssndAYNFNLEegvhl
Q5JTC6	APC membrane recruitment protein 1	AMER1	ysgdaLYEFYEPDdslen
Q5TCZ1	SH3 and PX domain-containing protein 2A	SH3PXD2A	eydipAFGFDSEpelse
Q5XG87	Non-canonical poly(A) RNA polymerase PAPD7	PAPD7	grggaFFNFADgapsa
Q7Z3K3	Pogo transposable element with ZNF domain	POGZ	setesFYGFEEADldlme
Q7Z5K2	Wings apart-like protein homolog	WAPL	estgdPFGFDSDdeslp
Q7Z5K2	Wings apart-like protein homolog	WAPL	dvkleFFGFEDHEtggde
Q7Z5K2	Wings apart-like protein homolog	WAPL	nykikYFGFDDlsese
Q8IZU3	Synaptonemal complex protein 3	SYCP3	dqftrAYDFETEdkkdl
Q8N5A5	Zinc finger CCCH-type with G patch domain-containing protein	ZGPAT	paprnVFDFLNEklqgq
Q8TD26	Chromodomain-helicase-DNA-binding protein 6	CHD6	qkhrrPYEFEVErdaka
Q96MT3	Prickle-like protein 1	PRICKLE1	rtrrrVYNFEErgsrs
Q96N16	Janus kinase and microtubule-interacting protein 1	JAKMIP1	hvvetFFGFDEEsvdse
Q96NW7	Leucine-rich repeat-containing protein 7	LRRC7	qrmtvAFEFEDkkedd
Q96PQ7	Kelch-like protein 5	KLHL5	tsevpAFEFTAEDcggah
Q99961	Endophilin-A2	SH3GL1	psckaLYDFEPEndgel
Q99962	Endophilin-A1	SH3GL2	pccraLYDFEPEnegel
Q99963	Endophilin-A3	SH3GL3	pccrgLYDFEPEnqgel
Q9BPY3	Protein FAM118B	FAM118B	sdideIFGFFNDgeppt

Q9HCE6	Rho guanine nucleotide exchange factor 10-like protein	ARHGEF10L	ddpgeAFEFDDSDdeedt
Q9NY33	Dipeptidyl peptidase 3	DPP3	qdekgAFNFDQEtvinp
Q9Y6N6	Laminin subunit gamma-3	LAMC3	egrpsAYNFEEspglq

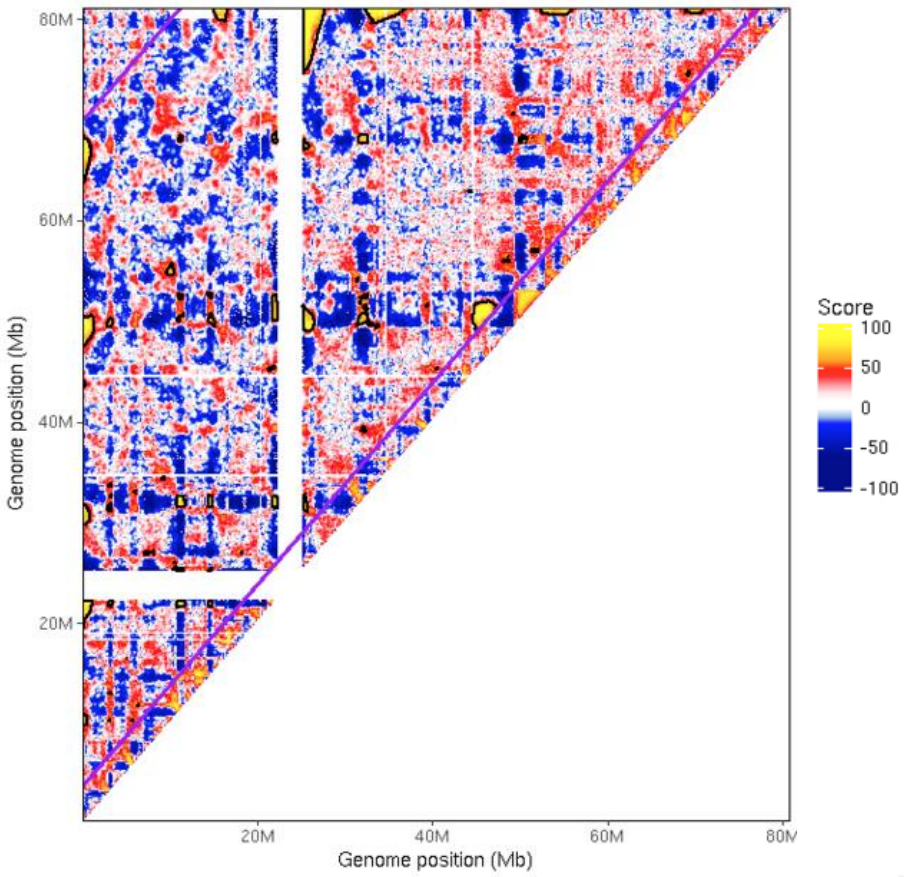
Supplemental Figure 5: Slimsearch results for the FGF-like motif predicted to interact with the SA CES domain (Li *et al.*, 2020) in the human proteome. Protein accession (ProteinAcc), protein name, gene name and FGF-like motif are shown.



Supplemental Figure 6: (A) Full NIPBL blot corresponding to Figure 67 E. (B) Reloading of RAD21 on chromatin in the H2 clone of HCT116 RmAC OsTIR1 cells. H2 cells were treated with siCon, siSA, siNIPBL, or siSA + siNIPBL for 72hrs. Additional siSA + siNIPBL samples with an increase cell number was also included ((2) siSA + siNIPBL). Prior to collect, auxin was added to the cell media for 4hrs and then washed off and replaced with fresh, unmodified media for the timepoints indicated. Histone H3 was blotted as a loading control. Note the reduced levels of RAD21 reassociation with chromatin following 4 hr withdrawal from auxin in siCon-treated cells – RAD21 levels are low enough that very low level signal in the 0 hr samples can be observed.



Supplemental Figure 7: (A) Quantification relative to Histone H3 for MAU2, AQR, and RNASEH2A for Blot 1 (top), Blot 2 (middle), and Blot 3 (bottom) from Figure 78. (B) Quantification relative to Histone H3 for MAU2, AQR, and RNASEH2A for Blot 1 (top), Blot 2 (middle), and Blot 3 (bottom) from Figure 79.



Supplemental Figure 8: Example of long-range contact (LRC) hotspots (yellow) observed in HCT116 cells upon depletion of cohesin. Hi-C data was obtain from Rao et al. (2017) and processed by Dr. Christopher Barrington.

8

References

Abakir, A. *et al.* (2020) ‘N 6-methyladenosine regulates the stability of RNA:DNA hybrids in human cells’, *Nature Genetics*. doi: 10.1038/s41588-019-0549-x.

Alfano, L. *et al.* (2019) ‘Depletion of the RNA binding protein HNRNPD impairs homologous recombination by inhibiting DNA-end resection and inducing R-loop accumulation’, *Nucleic Acids Research*. doi: 10.1093/nar/gkz076.

Alipour, E. and Marko, J. F. (2012) ‘Self-organization of domain structures by DNA-loop-extruding enzymes’, *Nucleic Acids Research*, 40(22), pp. 11202–11212. doi: 10.1093/nar/gks925.

Alomer, R. M. *et al.* (2017) ‘Esco1 and Esco2 regulate distinct cohesin functions during cell cycle progression’, *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1708291114.

Arab, K. *et al.* (2019) ‘GADD45A binds R-loops and recruits TET1 to CpG island promoters’, *Nature Genetics*. doi: 10.1038/s41588-018-0306-6.

Arumugam, P. *et al.* (2003) ‘ATP Hydrolysis Is Required for Cohesin’s Association with Chromosomes’, *Current Biology*. doi: 10.1016/j.cub.2003.10.036.

Azvolinsky, A. *et al.* (2009) ‘Highly Transcribed RNA Polymerase II Genes Are Impediments to Replication Fork Progression in Saccharomyces cerevisiae’, *Molecular Cell*. doi: 10.1016/j.molcel.2009.05.022.

Balbás-Martínez, C. *et al.* (2013) ‘Recurrent inactivation of STAG2 in bladder

cancer is not associated with aneuploidy', *Nature Genetics*. doi: 10.1038/ng.2799.

Banani, S. F. *et al.* (2017) 'Biomolecular condensates: Organizers of cellular biochemistry', *Nature Reviews Molecular Cell Biology*. doi: 10.1038/nrm.2017.7.

Barrington, C. *et al.* (2019) 'Enhancer accessibility and CTCF occupancy underlie asymmetric TAD architecture and cell type specific genome topology', *Nature Communications*. doi: 10.1038/s41467-019-10725-9.

Bartkuhn, M. *et al.* (2009) 'Active promoters and insulators are marked by the centrosomal protein 190', *EMBO Journal*. doi: 10.1038/emboj.2009.34.

Batlle, C. *et al.* (2020) 'hnRNPDL Phase Separation Is Regulated by Alternative Splicing and Disease-Causing Mutations Accelerate Its Aggregation', *Cell Reports*. doi: 10.1016/j.celrep.2019.12.080.

Baude, A. *et al.* (2015) 'Hepatoma-derived growth factor-related protein 2 promotes DNA repair by homologous recombination', *Nucleic Acids Research*. doi: 10.1093/nar/gkv1526.

Bauerschmidt, C. *et al.* (2009) 'Cohesin promotes the repair of ionizing radiation-induced DNA double-strand breaks in replicated chromatin', *Nucleic Acids Research*, 38, pp. 477–487. doi: 10.1093/nar/gkp976.

Beckouët, F. *et al.* (2016) 'Releasing Activity Disengages Cohesin's Smc3/Scc1 Interface in a Process Blocked by Acetylation', *Molecular Cell*. doi: 10.1016/j.molcel.2016.01.026.

Beier, S. *et al.* (2017) 'MISA-web: a web server for microsatellite prediction', *Bioinformatics (Oxford, England)*. doi: 10.1093/bioinformatics/btx198.

Bell, A. C. and Felsenfeld, G. (2000) 'Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene', *Nature*. doi: 10.1038/35013100.

Bell, A. C., West, A. G. and Felsenfeld, G. (1999) ‘The protein CTCF is required for the enhancer blocking activity of vertebrate insulators’, *Cell*. doi: 10.1016/S0092-8674(00)81967-4.

Ben-Shahar, T. R. *et al.* (2008) ‘Eco1-dependent cohesin acetylation during establishment of sister chromatid cohesion’, *Science*. doi: 10.1126/science.1157774.

Benedict, B. *et al.* (2020) ‘WAPL-Dependent Repair of Damaged DNA Replication Forks Underlies Oncogene-Induced Loss of Sister Chromatid Cohesion’, *Developmental Cell*. doi: 10.1016/j.devcel.2020.01.024.

Bentley, D. L. (2014) ‘Coupling mRNA processing with transcription in time and space’, *Nature Reviews Genetics*. doi: 10.1038/nrg3662.

van den Berg, D. L. C. *et al.* (2017) ‘Nipbl Interacts with Zfp609 and the Integrator Complex to Regulate Cortical Neuron Migration’, *Neuron*, 93(2), pp. 348–361. doi: 10.1016/j.neuron.2016.11.047.

Bessat, M. and Ersfeld, K. (2009) ‘Functional characterization of cohesin SMC3 and separase and their roles in the segregation of large and minichromosomes in Trypanosoma brucei’, *Molecular Microbiology*. doi: 10.1111/j.1365-2958.2009.06611.x.

Bhattacharyya, N. P. *et al.* (1994) ‘Mutator phenotypes in human colorectal carcinoma cell lines’, *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.91.14.6319.

Birkenbihl, R. P. and Subramani, S. (1992) ‘Cloning and characterization of rad21 an essential gene of Schizosaccharomyces pombe involved in DNA double-strand-break repair’, *Nucleic Acids Research*, pp. 6605–6611. doi: 10.1093/nar/20.24.6605.

Bisht, K. K., Daniloski, Z. and Smith, S. (2013) ‘SA1 binds directly to DNA through its unique AT-hook to promote sister chromatid cohesion at telomeres’,

Journal of Cell Science. doi: 10.1242/jcs.130872.

Boque-Sastre, R. *et al.* (2015) 'Head-to-head antisense transcription and R-loop formation promotes transcriptional activation', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1421197112.

Bose, T. *et al.* (2012) 'Cohesin proteins promote ribosomal RNA production and protein translation in yeast and human cells', *PLoS Genetics*. doi: 10.1371/journal.pgen.1002749.

Bot, C. *et al.* (2017) 'Independent mechanisms recruit the cohesin loader protein NIPBL to sites of DNA damage', *Journal of Cell Science*, 130(6), pp. 1134–114636. doi: 10.1242/jcs.197236.

Britton, S. *et al.* (2014) 'DNA damage triggers SAF-A and RNA biogenesis factors exclusion from chromatin coupled to R-loops removal', *Nucleic Acids Research*. doi: 10.1093/nar/gku601.

Bubeck, D. *et al.* (2011) 'PCNA directs type 2 RNase H activity on DNA replication and repair substrates', *Nucleic Acids Research*. doi: 10.1093/nar/gkq980.

Buheitel, J. and Stemmann, O. (2013) 'Prophase pathway-dependent removal of cohesin from human chromosomes requires opening of the Smc3-Scc1 gate', *EMBO Journal*. doi: 10.1038/emboj.2013.7.

Buratowski, S. *et al.* (1989) 'Five intermediate complexes in transcription initiation by RNA polymerase II', *Cell*. doi: 10.1016/0092-8674(89)90578-3.

Busslinger, G. A. *et al.* (2017) 'Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl', *Nature*, 544(7651), pp. 503–507. doi: 10.1038/nature22063.

Byun, T. S. *et al.* (2005) 'Functional uncoupling of MCM helicase and DNA

polymerase activities activates the ATR-dependent checkpoint', *Genes and Development*. doi: 10.1101/gad.1301205.

Canudas, S. and Smith, S. (2009) 'Differential regulation of telomere and centromere cohesion by the Scc3 homologues SA1 and SA2, respectively, in human cells', *Journal of Cell Biology*. doi: 10.1083/jcb.200903096.

Cao, L. *et al.* (2015) 'Negative regulation of p21Waf1/Cip1 by human INO80 chromatin remodeling complex is implicated in cell cycle phase G2/M arrest and abnormal chromosome stability', *PLoS ONE*. doi: 10.1371/journal.pone.0137411.

Carramolino, L. *et al.* (1997) 'SA-1, a nuclear protein encoded by one member of a novel gene family: Molecular cloning and detection in hemopoietic organs', *Gene*. doi: 10.1016/S0378-1119(97)00121-2.

Casa, V. *et al.* (2020) 'Redundant and specific roles of cohesin STAG subunits in chromatin looping and transcriptional control', *Genome Research*. doi: 10.1101/gr.253211.119.

Castronovo, P. *et al.* (2009) 'Premature chromatid separation is not a useful diagnostic marker for Cornelia de Lange syndrome', *Chromosome Research*. doi: 10.1007/s10577-009-9066-6.

Chakraborty, P. and Grosse, F. (2011) 'Human DHX9 helicase preferentially unwinds RNA-containing displacement loops (R-loops) and G-quadruplexes', *DNA Repair*. doi: 10.1016/j.dnarep.2011.04.013.

Chan, K.-L. *et al.* (2013) 'Pds5 promotes and protects cohesin acetylation', *Proceedings of the National Academy of Sciences*, 110(32), pp. 13020–13025. doi: 10.1073/pnas.1306900110.

Chan, K. L. *et al.* (2012) 'Cohesin's DNA exit gate is distinct from its entrance gate and is regulated by acetylation', *Cell*. doi: 10.1016/j.cell.2012.07.028.

Chao, W. C. H. *et al.* (2015) ‘Structural Studies Reveal the Functional Modularity of the Scc2-Scc4 Cohesin Loader’, *Cell Reports*, 12(5), pp. 719–725. doi: 10.1016/j.celrep.2015.06.071.

Chao, W. C. H. *et al.* (2017) ‘Structure of the cohesin loader Scc2’, *Nature Communications*, 8. doi: 10.1038/ncomms13952.

Chen, H. *et al.* (2012) ‘Comprehensive identification and annotation of cell type-specific and ubiquitous CTCF-binding sites in the human genome’, *PLoS ONE*. doi: 10.1371/journal.pone.0041374.

Chen, L. *et al.* (2017) ‘R-ChIP Using Inactive RNase H Reveals Dynamic Coupling of R-loops with Transcriptional Pausing at Gene Promoters’, *Molecular Cell*. doi: 10.1016/j.molcel.2017.10.008.

Cheng, H. *et al.* (2006) ‘Human mRNA Export Machinery Recruited to the 5′ End of mRNA’, *Cell*. doi: 10.1016/j.cell.2006.10.044.

Chien, R. *et al.* (2011) ‘Cohesin mediates chromatin interactions that regulate mammalian β-globin expression’, *Journal of Biological Chemistry*. doi: 10.1074/jbc.M110.207365.

Choi, J., Hwang, S. Y. and Ahn, K. (2018) ‘Interplay between RNASEH2 and MOV10 controls LINE-1 retrotransposition’, *Nucleic Acids Research*. doi: 10.1093/nar/gkx1312.

Chon, H. *et al.* (2009) ‘Contributions of the two accessory subunits, RNASEH2B and RNASEH2C, to the activity and properties of the human RNase H2 complex’, *Nucleic Acids Research*. doi: 10.1093/nar/gkn913.

Cimprich, K. A. and Cortez, D. (2008) ‘ATR: An essential regulator of genome integrity’, *Nature Reviews Molecular Cell Biology*. doi: 10.1038/nrm2450.

Ciosk, R. *et al.* (2000) ‘Cohesin’s binding to chromosomes depends on a separate complex consisting of Scc2 and Scc4 proteins.’, *Molecular cell*, 5(2),

pp. 243–54. doi: 10.1016/S1097-2765(00)80420-7.

Clarkson, C. T. *et al.* (2019) ‘CTCF-dependent chromatin boundaries formed by asymmetric nucleosome arrays with decreased linker length’, *Nucleic acids research*. doi: 10.1093/nar/gkz908.

Conrad, T. *et al.* (2016) ‘Serial interactome capture of the human cell nucleus’, *Nature Communications*. doi: 10.1038/ncomms11212.

Coster, G. and Goldberg, M. (2010) ‘The cellular response to dna damage: A focus on MDC1 and its interacting proteins’, *Nucleus*. doi: 10.4161/nucl.11176.

Countryman, P. *et al.* (2018) ‘Cohesin SA2 is a sequence-independent DNA-binding protein that recognizes DNA replication and repair intermediates’, *Journal of Biological Chemistry*. doi: 10.1074/jbc.M117.806406.

Cox, J. and Mann, M. (2008) ‘MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification’, *Nature Biotechnology*. doi: 10.1038/nbt.1511.

Cremer, M. *et al.* (2020) ‘Cohesin depleted cells rebuild functional nuclear compartments after endomitosis’, *Nature Communications*. doi: 10.1038/s41467-020-19876-6.

Cristini, A. *et al.* (2018) ‘RNA/DNA Hybrid Interactome Identifies DXH9 as a Molecular Player in Transcriptional Termination and R-Loop-Associated DNA Damage’, *Cell Reports*. doi: 10.1016/j.celrep.2018.04.025.

Cuadrado, A. *et al.* (2012) ‘The specific contributions of cohesin-SA1 to cohesion and gene expression: Implications for cancer and development’, *Cell Cycle*. doi: 10.4161/cc.20318.

Cuadrado, A. *et al.* (2019) ‘Specific Contributions of Cohesin-SA1 and Cohesin-SA2 to TADs and Polycomb Domains in Embryonic Stem Cells’, *Cell Reports*. doi: 10.1016/j.celrep.2019.05.078.

Cuddapah, S. *et al.* (2009) ‘Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains’, *Genome Research*. doi: 10.1101/gr.082800.108.

Darwiche, N., Freeman, L. A. and Strunnikov, A. (1999) ‘Characterization of the components of the putative mammalian sister chromatid cohesion complex’, *Gene*. doi: 10.1016/S0378-1119(99)00160-2.

Davidson, I. F. *et al.* (2019) ‘DNA loop extrusion by human cohesin’, *Science*. doi: 10.1126/science.aaz3418.

Deardorff, M. A. *et al.* (2007) ‘Mutations in cohesin complex members SMC3 and SMC1A cause a mild variant of Cornelia de Lange syndrome with predominant mental retardation’, *American Journal of Human Genetics*. doi: 10.1086/511888.

Dixon, J. R. *et al.* (2012) ‘Topological domains in mammalian genomes identified by analysis of chromatin interactions’, *Nature*, 485(7398), pp. 376–380. doi: 10.1038/nature11082.

Donze, D. *et al.* (1999) ‘The boundaries of the silenced HMR domain in Saccharomyces cerevisiae’, *Genes and Development*. doi: 10.1101/gad.13.6.698.

Dorsett, D. *et al.* (2005) ‘Effects of sister chromatid cohesion proteins on cut gene expression during wing development in Drosophila’, *Development*. doi: 10.1242/dev.02064.

Duc, T. N. *et al.* (2012) ‘Nanobody-based chromatin immunoprecipitation’, *Methods in Molecular Biology*, 911, pp. 491–505. doi: 10.1007/978-1-61779-968-6-31.

Dunham, I. *et al.* (2012) ‘An integrated encyclopedia of DNA elements in the human genome’, *Nature*. doi: 10.1038/nature11247.

Dunn, K. and Griffith, J. D. (1980) ‘The presence of RNA in a double helix inhibits its interaction with histone protein’, *Nucleic Acids Research*. doi: 10.1093/nar/8.3.555.

Eichinger, C. S. *et al.* (2013) ‘Disengaging the Smc3/kleisin interface releases cohesin from Drosophila chromosomes during interphase and mitosis’, *EMBO Journal*. doi: 10.1038/emboj.2012.346.

Eissenberg, J. C. *et al.* (1992) ‘The heterochromatin-associated protein HP-1 is an essential protein in Drosophila with dosage-dependent effects on position-effect variegation’, *Genetics*. doi: 10.1093/genetics/131.2.345.

Ernst, J. and Kellis, M. (2017) ‘Chromatin-state discovery and genome annotation with ChromHMM’, *Nature Protocols*. doi: 10.1038/nprot.2017.124.

Fan, H. *et al.* (2018) ‘The nuclear matrix protein HNRNPU maintains 3D genome architecture globally in mouse hepatocytes’, *Genome Research*. doi: 10.1101/gr.224576.117.

Faure, A. J. *et al.* (2012) ‘Cohesin regulates tissue-specific expression by stabilizing highly occupied cis-regulatory modules’, *Genome Research*. doi: 10.1101/gr.136507.111.

Filippova, G. N. *et al.* (1996) ‘An exceptionally conserved transcriptional repressor, CTCF, employs different combinations of zinc fingers to bind diverged promoter sequences of avian and mammalian c-myc oncogenes.’, *Molecular and Cellular Biology*. doi: 10.1128/mcb.16.6.2802.

Fisher, J. B. *et al.* (2017) ‘The cohesin subunit Rad21 is a negative regulator of hematopoietic self-renewal through epigenetic repression of Hoxa7 and Hoxa9’, *Leukemia*. doi: 10.1038/leu.2016.240.

Fu, Y. *et al.* (2008) ‘The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome’, *PLoS Genetics*. doi: 10.1371/journal.pgen.1000138.

Fudenberg, G. *et al.* (2016a) ‘Formation of Chromosomal Domains by Loop Extrusion’, *Cell Reports*, 15(9), pp. 2038–2049. doi: 10.1016/j.celrep.2016.04.085.

Fudenberg, G. *et al.* (2016b) ‘Formation of Chromosomal Domains by Loop Extrusion’, *Cell Reports*. doi: 10.1016/j.celrep.2016.04.085.

Furuya, K., Takahashi, K. and Yanagida, M. (1998) ‘Faithful anaphase is ensured by Mis4, a sister chromatid cohesion molecule required in S phase and not destroyed in G1 phase’, *Genes and Development*. doi: 10.1101/gad.12.21.3408.

Garcia-Luis, J. *et al.* (2019) ‘FACT mediates cohesin function on chromatin’, *Nature Structural and Molecular Biology*. doi: 10.1038/s41594-019-0307-x.

Gause, M. *et al.* (2010) ‘Dosage-Sensitive Regulation of Cohesin Chromosome Binding and Dynamics by Nipped-B, Pds5, and Wapl’, *Molecular and Cellular Biology*. doi: 10.1128/mcb.00642-10.

Geisberg, J. V. and Struhl, K. (2005) ‘Analysis of Protein Co-Occupancy by Quantitative Sequential Chromatin Immunoprecipitation’, in *Current Protocols in Molecular Biology*. doi: 10.1002/0471142727.mb2108s70.

Gerlich, D. *et al.* (2006) ‘Live-Cell Imaging Reveals a Stable Cohesin-Chromatin Interaction after but Not before DNA Replication’, *Current Biology*. doi: 10.1016/j.cub.2006.06.068.

Gertz, J. *et al.* (2013) ‘Distinct properties of cell-type-specific and shared transcription factor binding sites’, *Molecular Cell*. doi: 10.1016/j.molcel.2013.08.037.

Ghirlando, R. and Felsenfeld, G. (2016) ‘CTCF: Making the right connections’, *Genes and Development*. doi: 10.1101/gad.277863.116.

Ghosh, I., Hamilton, A. D. and Regan, L. (2000) ‘Antiparallel leucine zipper-

directed protein reassembly: Application to the green fluorescent protein', *Journal of the American Chemical Society*, 122. doi: 10.1021/ja994421w.

Gillespie, P. J. and Hirano, T. (2004) 'Scc2 couples replication licensing to sister chromatid cohesion in Xenopus egg extracts', *Current Biology*. doi: 10.1016/j.cub.2004.07.053.

Ginno, P. A. *et al.* (2012) 'R-Loop Formation Is a Distinctive Characteristic of Unmethylated Human CpG Island Promoters', *Molecular Cell*. doi: 10.1016/j.molcel.2012.01.017.

Gligoris, T. G. *et al.* (2014) 'Closing the cohesin ring: Structure and function of its Smc3-kleisin interface', *Science*. doi: 10.1126/science.1256917.

Gnatt, A. L. *et al.* (2001) 'Structural basis of transcription: An RNA polymerase II elongation complex at 3.3 Å resolution', *Science*. doi: 10.1126/science.1059495.

Gómez-González, B. *et al.* (2011) 'Genome-wide function of THO/TREX in active genes prevents R-loop-dependent replication obstacles', *EMBO Journal*. doi: 10.1038/emboj.2011.206.

Grosselin, K. *et al.* (2019) 'High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer', *Nature Genetics*. doi: 10.1038/s41588-019-0424-9.

Gruber, S., Haering, C. H. and Nasmyth, K. (2003) 'Chromosomal cohesin forms a ring', *Cell*, 112(6), pp. 765–777. doi: 10.1016/S0092-8674(03)00162-4.

Grunseich, C. *et al.* (2018) 'Senataxin Mutation Reveals How R-Loops Promote Transcription by Blocking DNA Methylation at Gene Promoters', *Molecular Cell*. doi: 10.1016/j.molcel.2017.12.030.

Guastafierro, T. *et al.* (2008) 'CCCTC-binding factor activates PARP-1 affecting DNA methylation machinery', *Journal of Biological Chemistry*. doi:

10.1074/jbc.M801170200.

Gullerova, M. and Proudfoot, N. J. (2008) ‘Cohesin Complex Promotes Transcriptional Termination between Convergent Genes in S. pombe’, *Cell*. doi: 10.1016/j.cell.2008.02.040.

Guo, G. *et al.* (2013) ‘Whole-genome and whole-exome sequencing of bladder cancer identifies frequent alterations in genes involved in sister chromatid cohesion and segregation’, *Nature Genetics*. doi: 10.1038/ng.2798.

Haarhuis, J. H. I. *et al.* (2017) ‘The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension’, *Cell*, 169(4), pp. 693-707.e14. doi: 10.1016/j.cell.2017.04.013.

Hadjur, S. *et al.* (2009) ‘Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus’, *Nature*. doi: 10.1038/nature08079.

Haering, C. H. *et al.* (2002) ‘Molecular architecture of SMC proteins and the yeast cohesin complex’, *Molecular Cell*, 9(4), pp. 773–788. doi: 10.1016/S1097-2765(02)00515-4.

Haering, C. H. *et al.* (2008) ‘The cohesin ring concatenates sister DNA molecules’, *Nature*. doi: 10.1038/nature07098.

El Hage, A. *et al.* (2010) ‘Loss of Topoisomerase I leads to R-loop-mediated transcriptional blocks during ribosomal RNA synthesis’, *Genes and Development*. doi: 10.1101/gad.573310.

Hakimi, M. A. *et al.* (2002) ‘A chromatin remodelling complex that loads cohesin onto human chromosomes’, *Nature*. doi: 10.1038/nature01024.

Halász, L. *et al.* (2017) ‘RNA-DNA hybrid (R-loop) immunoprecipitation mapping: An analytical workflow to evaluate inherent biases’, *Genome Research*. doi: 10.1101/gr.219394.116.

Hamperl, S. *et al.* (2017) ‘Transcription-Replication Conflict Orientation Modulates R-Loop Levels and Activates Distinct DNA Damage Responses’, *Cell*. doi: 10.1016/j.cell.2017.07.043.

Han, X. *et al.* (2015) ‘Phosphorylation of minichromosome maintenance 3 (MCM3) by checkpoint kinase 1 (Chk1) negatively regulates DNA replication and checkpoint activation’, *Journal of Biological Chemistry*. doi: 10.1074/jbc.M114.621532.

Hansen, A. S. *et al.* (2017) ‘CTCF and cohesin regulate chromatin loop stability with distinct dynamics’, *eLife*. doi: 10.7554/elife.25776.

Hansen, A. S. *et al.* (2019) ‘Distinct Classes of Chromatin Loops Revealed by Deletion of an RNA-Binding Region in CTCF’, *Molecular Cell*. doi: 10.1016/j.molcel.2019.07.039.

Hara, K. *et al.* (2014) ‘Structure of cohesin subcomplex pinpoints direct shugoshin-Wapl antagonism in centromeric cohesion’, *Nature Structural and Molecular Biology*, 21(10), pp. 864–870. doi: 10.1038/nsmb.2880.

Hark, A. T. *et al.* (2000) ‘CTCF mediates methylation-sensitive enhancer-blocking activity at the H19/Igf2 locus’, *Nature*. doi: 10.1038/35013106.

Harris, B. *et al.* (2014) ‘Cohesion promotes nucleolar structure and function’, *Molecular Biology of the Cell*. doi: 10.1091/mbc.E13-07-0377.

Hashimoto, H. *et al.* (2017) ‘Structural Basis for the Versatile and Methylation-Dependent Binding of CTCF to DNA’, *Molecular Cell*. doi: 10.1016/j.molcel.2017.05.004.

Hauf, S. *et al.* (2005) ‘Dissociation of cohesin from chromosome arms and loss of arm cohesion during early mitosis depends on phosphorylation of SA2’, in *PLoS Biology*. doi: 10.1371/journal.pbio.0030069.

Hauf, S., Waizenegger, I. C. and Peters, J. M. (2001) ‘Cohesin cleavage by

separase required for anaphase and cytokinesis in human cells', *Science*. doi: 10.1126/science.1061376.

He, Y. *et al.* (2016) 'Near-atomic resolution visualization of human transcription promoter opening', *Nature*. doi: 10.1038/nature17970.

Heger, P. *et al.* (2012) 'The chromatin insulator CTCF and the emergence of metazoan diversity', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1111941109.

Herrera-Moyano, E. *et al.* (2014) 'The yeast and human FACT chromatinreorganizing complexes solve R-loopmediated transcription-replication conflicts', *Genes and Development*. doi: 10.1101/gad.234070.113.

Higashi, T. L. *et al.* (2020) 'A Structure-Based Mechanism for DNA Entry into the Cohesin Ring', *Molecular Cell*. doi: 10.1016/j.molcel.2020.07.013.

Hilmi, K. *et al.* (2017) 'CTCF facilitates DNA double-strand break repair by enhancing homologous recombination repair', *Science Advances*. doi: 10.1126/sciadv.1601898.

Hinshaw, S. M. *et al.* (2015) 'Structural evidence for Scc4-dependent localization of cohesin loading', *eLife*, 4, p. e06057. doi: 10.7554/eLife.06057.

Hinshaw, S. M. *et al.* (2017) 'The Kinetochore Receptor for the Cohesin Loading Complex', *Cell*. doi: 10.1016/j.cell.2017.08.017.

Hirano, M. and Hirano, T. (2002) 'Hinge-mediated dimerization of SMC protein is essential for its dynamic interaction with DNA', *The EMBO Journal*, 21, pp. 5733–5744. doi: 10.1093/emboj/cdf575.

Holzmann, J. *et al.* (2019) 'Absolute quantification of cohesin, CTCF and their regulators in human cells', *eLife*. doi: 10.7554/elife.46269.

Horsfield, J. A. *et al.* (2007) 'Cohesin-dependent regulation of Runx genes',

Development. doi: 10.1242/dev.002485.

Hou, F. and Zou, H. (2005) 'Two human orthologues of Eco1/Ctf7 acetyltransferases are both required for proper sister-chromatid cohesion', *Molecular Biology of the Cell*. doi: 10.1091/mbc.E04-12-1063.

Hu, B. *et al.* (2011) 'ATP hydrolysis is required for relocating cohesin from sites occupied by its Scc2/4 loading complex', *Current Biology*. doi: 10.1016/j.cub.2010.12.004.

Hu, C. D., Chinenov, Y. and Kerppola, T. K. (2002) 'Visualization of interactions among bZIP and Rel family proteins in living cells using bimolecular fluorescence complementation', *Molecular Cell*, 9(4), pp. 789–798. doi: 10.1016/S1097-2765(02)00496-3.

Huis In't Veld, P. J. *et al.* (2014) 'Characterization of a DNA exit gate in the human cohesin ring', *Science*. doi: 10.1126/science.1256904.

Hyle, J. *et al.* (2019) 'Acute depletion of CTCF directly affects MYC regulation through loss of enhancer-promoter looping', *Nucleic Acids Research*. doi: 10.1093/nar/gkz462.

Ivanov, D. and Nasmyth, K. (2005) 'A topological interaction between cohesin rings and a circular minichromosome', *Cell*. doi: 10.1016/j.cell.2005.07.018.

Jiang, H. *et al.* (2017) 'Long noncoding RNA CRNDE stabilized by hnRNPUL2 accelerates cell proliferation and migration in colorectal carcinoma via activating Ras/MAPK signaling pathways', *Cell death & disease*. doi: 10.1038/cddis.2017.258.

Kagey, M. H. *et al.* (2010) 'Mediator and cohesin connect gene expression and chromatin architecture', *Nature*, 467(7314), pp. 430–435. doi: 10.1038/nature09380.

Kanduri, C. *et al.* (2000) 'Functional association of CTCF with the insulator

upstream of the H19 gene is parent of origin-specific and methylation-sensitive', *Current Biology*. doi: 10.1016/S0960-9822(00)00597-2.

Kanduri, M. *et al.* (2002) 'Multiple Nucleosome Positioning Sites Regulate the CTCF-Mediated Insulator Function of the H19 Imprinting Control Region†', *Molecular and Cellular Biology*. doi: 10.1128/mcb.22.10.3339-3344.2002.

Kanke, M. *et al.* (2016) 'Cohesin acetylation and Wapl-Pds5 oppositely regulate translocation of cohesin along DNA', *The EMBO Journal*, 35(24), pp. 2686–2698. doi: 10.15252/embj.201695756.

Kargapolova, Y. *et al.* (2020) 'Overarching control of autophagy and DNA damage response by CHD6 revealed by modeling a rare human pathology', *bioRxiv*. doi: 10.1101/2020.01.27.921171.

Kaur, M. *et al.* (2005) 'Precocious sister chromatid separation (PSCS) in Cornelia de Lange syndrome', *American Journal of Medical Genetics*. doi: 10.1002/ajmg.a.30919.

Kikuchi, S. *et al.* (2016) 'Crystal structure of the cohesin loader Scc2 and insight into cohesinopathy', *Proceedings of the National Academy of Sciences*, 113(44), pp. 12444–12449. doi: 10.1073/pnas.1611333113.

Kim *et al.* (2019) 'Systematic proteomics of endogenous human cohesin reveals an interaction with diverse splicing factors and RNA-binding proteins required for mitotic progression', *Journal of Biological Chemistry*. doi: 10.1074/jbc.RA119.007832.

Kim, T. H. *et al.* (2007) 'Analysis of the Vertebrate Insulator Protein CTCF-Binding Sites in the Human Genome', *Cell*. doi: 10.1016/j.cell.2006.12.048.

Kim, T. K., Ebright, R. H. and Reinberg, D. (2000) 'Mechanism of ATP-dependent promoter melting by transcription factor IIH', *Science*. doi: 10.1126/science.288.5470.1418.

Kim, Y. *et al.* (2019) ‘Human cohesin compacts DNA by loop extrusion’, *Science*. doi: 10.1126/science.aaz4475.

Kitajima, T. S. *et al.* (2006) ‘Shugoshin collaborates with protein phosphatase 2A to protect cohesin’, *Nature*. doi: 10.1038/nature04663.

Kodama, Y. and Hu, C. D. (2010) ‘An improved bimolecular fluorescence complementation assay with a high signal-to-noise ratio’, *BioTechniques*, 49(5), pp. 793–803. doi: 10.2144/000113519.

Kojic, A. *et al.* (2018) ‘Distinct roles of cohesin-SA1 and cohesin-SA2 in 3D chromosome organization’, *Nature Structural and Molecular Biology*. doi: 10.1038/s41594-018-0070-4.

Kong, X. *et al.* (2009) ‘Cohesin associates with spindle poles in a mitosis-specific manner and functions in spindle assembly in vertebrate cells’, *Molecular Biology of the Cell*. doi: 10.1091/mbc.E08-04-0419.

Kong, X. *et al.* (2014) ‘Distinct Functions of Human Cohesin-SA1 and Cohesin-SA2 in Double-Strand Break Repair’, *Molecular and Cellular Biology*. doi: 10.1128/mcb.01503-13.

Kosicki, M., Tomberg, K. and Bradley, A. (2018) ‘Repair of double-strand breaks induced by CRISPR–Cas9 leads to large deletions and complex rearrangements’, *Nature Biotechnology*. doi: 10.1038/nbt.4192.

Krantz, I. D. *et al.* (2004) ‘Cornelia de Lange syndrome is caused by mutations in NIPBL, the human homolog of Drosophila melanogaster Nipped-B’, *Nature Genetics*. doi: 10.1038/ng1364.

Kueng, S. *et al.* (2006) ‘Wapl Controls the Dynamic Association of Cohesin with Chromatin’, *Cell*. doi: 10.1016/j.cell.2006.09.040.

Kugel, J. F. and Goodrich, J. A. (1998) ‘Promoter escape limits the rate of RNA polymerase II transcription and is enhanced by TFIIE, TFIIH, and ATP on

negatively supercoiled DNA', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.95.16.9232.

Kulemzina, I. *et al.* (2012) 'Cohesin Rings Devoid of Scc3 and Pds5 Maintain Their Stable Association with the DNA', *PLoS Genetics*. doi: 10.1371/journal.pgen.1002856.

Kurukuti, S. *et al.* (2006) 'CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.0600326103.

Ladurner, R. *et al.* (2014) 'Cohesin's ATPase activity couples cohesin loading onto DNA with Smc3 acetylation', *Current Biology*. doi: 10.1016/j.cub.2014.08.011.

Ladurner, R. *et al.* (2016) 'Sororin actively maintains sister chromatid cohesion', *The EMBO Journal*. doi: 10.15252/embj.201592532.

Larson, A. G. *et al.* (2017) 'Liquid droplet formation by HP1α suggests a role for phase separation in heterochromatin', *Nature*. doi: 10.1038/nature22822.

Lawrence, M. S. *et al.* (2014) 'Discovery and saturation analysis of cancer genes across 21 tumour types', *Nature*. doi: 10.1038/nature12912.

Lee, B. G. *et al.* (2016) 'Crystal Structure of the Cohesin Gatekeeper Pds5 and in Complex with Kleisin Scc1', *Cell Reports*, 14(9), pp. 2108–2115. doi: 10.1016/j.celrep.2016.02.020.

Lee, C. G. and Hurwitz, J. (1992) 'A new RNA helicase isolated from HeLa cells that catalytically translocates in the 3' to 5' direction', *Journal of Biological Chemistry*. doi: 10.1016/s0021-9258(18)42849-9.

Lee, J. *et al.* (2017) 'The LDB1 Complex Co-opts CTCF for Erythroid Lineage-Specific Long-Range Enhancer Interactions', *Cell Reports*. doi:

10.1016/j.celrep.2017.05.072.

Leighton, P. A. *et al.* (1995) ‘An enhancer deletion affects both H19 and Igf2 expression’, *Genes and Development*. doi: 10.1101/gad.9.17.2079.

Lengauer, C., Kinzler, K. W. and Vogelstein, B. (1997) ‘Genetic instability in colorectal cancers’, *Nature*. doi: 10.1038/386623a0.

Lengronne, A., Katou, Y., Mori, S., Yokabayashi, S., *et al.* (2004) ‘Cohesin relocation from sites of chromosomal loading to places of convergent transcription’, *Nature*, 430(6999), pp. 573–578. doi: 10.1038/nature02742.

Lengronne, A., Katou, Y., Mori, S., Yokobayashi, S., *et al.* (2004) ‘Cohesin relocation from sites of chromosomal loading to places of convergent transcription’, *Nature*, 430(6999), pp. 573–578. doi: 10.1038/nature02742.

Lengronne, A. *et al.* (2006) ‘Establishment of Sister Chromatid Cohesion at the S. cerevisiae Replication Fork’, *Molecular Cell*. doi: 10.1016/j.molcel.2006.08.018.

Li, R., Zhu, H. and Luo, Y. (2016) ‘Understanding the functions of long non-coding RNAs through their higher-order structures’, *International Journal of Molecular Sciences*. doi: 10.3390/ijms17050702.

Li, X. and Manley, J. L. (2005) ‘Inactivation of the SR protein splicing factor ASF/SF2 results in genomic instability’, *Cell*. doi: 10.1016/j.cell.2005.06.008.

Li, Y. *et al.* (2018) ‘Structural basis for scc3-dependent cohesin recruitment to chromatin’, *eLife*. doi: 10.7554/eLife.38356.

Li, Y. *et al.* (2020) ‘The structural basis for cohesin–CTCF-anchored loops’, *Nature*. doi: 10.1038/s41586-019-1910-z.

Liang, Z. *et al.* (2019) ‘Binding of FANCI-FANCD2 Complex to RNA and R-Loops Stimulates Robust FANCD2 Monoubiquitination’, *Cell Reports*. doi:

10.1016/j.celrep.2018.12.084.

Lieberman-Aiden, E. *et al.* (2009) ‘Comprehensive mapping of long-range interactions reveals folding principles of the human genome’, *Science*, 326(5950), pp. 289–293. doi: 10.1126/science.1181369.

Lin, J. *et al.* (2016) ‘Functional interplay between SA1 and TRF1 in telomeric DNA binding and DNA-DNA pairing’, *Nucleic Acids Research*. doi: 10.1093/nar/gkw518.

Liu, H. *et al.* (2015) ‘Mitotic Transcription Installs Sgo1 at Centromeres to Coordinate Chromosome Segregation’, *Molecular Cell*. doi: 10.1016/j.molcel.2015.06.018.

Liu, H., Rankin, S. and Yu, H. (2013) ‘Phosphorylation-enabled binding of SGO1-PP2A to cohesin protects sororin and centromeric cohesion during mitosis’, *Nature Cell Biology*. doi: 10.1038/ncb2637.

Liu, J. *et al.* (2009) ‘Transcriptional Dysregulation in NIPBL and Cohesin Mutant Human Cells’, *PLoS Biology*. doi: 10.1371/journal.pbio.1000119.

Liu, J. *et al.* (2021) ‘The RNA m6A reader YTHDC1 silences retrotransposons and guards ES cell identity’, *Nature*, 591(7849), pp. 322–326. doi: 10.1038/s41586-021-03313-9.

Lobanenkov, V. V. *et al.* (1990) ‘A novel sequence-specific DNA binding protein which interacts with three regularly spaced direct repeats of the CCCTC-motif in the 5’-flanking sequence of the chicken c-myc gene’, *Oncogene*.

Lopez-Serra, L. *et al.* (2014) ‘The Scc2-Scc4 complex acts in sister chromatid cohesion and transcriptional regulation by maintaining nucleosome-free regions’, *Nature Genetics*, 46, pp. 1147–1151. doi: 10.1038/ng.3080.

Losada, A. *et al.* (2000) ‘Identification and characterization of SA/Scc3p subunits in the Xenopus and human cohesin complexes’, *Journal of Cell*

Biology, 150(3), pp. 405–416. doi: 10.1083/jcb.150.3.405.

Loukinov, D. I. *et al.* (2002) ‘BORIS, a novel male germ-line-specific protein associated with epigenetic reprogramming events, shares the same 11-zinc-finger domain with CTCF, the insulator protein involved in reading imprinting marks in the soma’, *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.092123699.

Lutz, T., Stöger, R. and Nieto, A. (2006) ‘CHD6 is a DNA-dependent ATPase and localizes at nuclear sites of mRNA synthesis’, *FEBS Letters*. doi: 10.1016/j.febslet.2006.09.049.

Ma, J. and Wang, M. D. (2016) ‘DNA supercoiling during transcription’, *Biophysical Reviews*. doi: 10.1007/s12551-016-0215-9.

Majumder, P. *et al.* (2008) ‘The insulator factor CTCF controls MHC class II gene expression and is required for the formation of long-distance chromatin interactions’, *Journal of Experimental Medicine*. doi: 10.1084/jem.20071843.

Mao, Y. S., Zhang, B. and Spector, D. L. (2011) ‘Biogenesis and function of nuclear bodies’, *Trends in Genetics*. doi: 10.1016/j.tig.2011.05.006.

Masuda, S. *et al.* (2005) ‘Recruitment of the human TREX complex to mRNA during splicing’, *Genes and Development*. doi: 10.1101/gad.1302205.

McNairn, A. J. and Gerton, J. L. (2009) ‘Intersection of ChIP and FLIP, genomic methods to study the dynamics of the cohesin proteins’, *Chromosome Research*. doi: 10.1007/s10577-008-9007-9.

Medvedovic, J. *et al.* (2013) ‘Flexible long-range loops in the VH gene region of the Igh locus facilitate the generation of a diverse antibody repertoire’, *Immunity*. doi: 10.1016/j.immuni.2013.08.011.

Melby, T. E. *et al.* (1998) ‘The symmetrical structure of structural maintenance of chromosomes (SMC) and MukB proteins: Long, antiparallel coiled coils,

folded at a flexible hinge', *Journal of Cell Biology*, 142, pp. 1595–1604. doi: 10.1083/jcb.142.6.1595.

Mendelson Cohen, N. *et al.* (2017) 'SHAMAN: Bin-free randomization, normalization and screening of Hi-C matrices', *bioRxiv*. doi: 10.1101/187203.

Mendez, J. and Stillman, B. (2000) 'Chromatin Association of Human Origin Recognition Complex, Cdc6, and Minichromosome Maintenance Proteins during the Cell Cycle: Assembly of Prereplication Complexes in Late Mitosis', *Molecular and Cellular Biology*. doi: 10.1128/mcb.20.22.8602-8612.2000.

Merkenschlager, M. and Nora, E. P. (2016) 'CTCF and Cohesin in Genome Folding and Transcriptional Gene Regulation', *Annual Review of Genomics and Human Genetics*. doi: 10.1146/annurev-genom-083115-022339.

Michaelis, C., Ciosk, R. and Nasmyth, K. (1997) 'Cohesins: Chromosomal proteins that prevent premature separation of sister chromatids', *Cell*, 91(1), pp. 35–45. doi: 10.1016/S0092-8674(01)80007-6.

Michieletto, D. and Gilbert, N. (2019) 'Role of nuclear RNA in regulating chromatin structure and transcription', *Current Opinion in Cell Biology*. doi: 10.1016/j.ceb.2019.03.007.

Minor, A. *et al.* (2014) 'Two novel RAD21 mutations in patients with mild Cornelia de Lange syndrome-like presentation and report of the first familial case', *Gene*. doi: 10.1016/j.gene.2013.12.045.

Molliex, A. *et al.* (2015) 'Phase Separation by Low Complexity Domains Promotes Stress Granule Assembly and Drives Pathological Fibrillization', *Cell*. doi: 10.1016/j.cell.2015.09.015.

Moon, H. *et al.* (2005) 'CTCF is conserved from Drosophila to humans and confers enhancer blocking of the Fab-8 insulator', *EMBO Reports*. doi: 10.1038/sj.embor.7400334.

Moore, S. *et al.* (2019) ‘The CHD6 chromatin remodeler is an oxidative DNA damage response factor’, *Nature Communications*. doi: 10.1038/s41467-018-08111-y.

Morello, L. G. *et al.* (2011) ‘The human nucleolar protein FTSJ3 associates with NIP7 and functions in pre-rRNA processing’, *PLoS ONE*. doi: 10.1371/journal.pone.0029174.

Muir, K. W. *et al.* (2016) ‘Structure of the Pds5-Scc1 Complex and Implications for Cohesin Function’, *Cell Reports*, 14(9), pp. 2116–2126. doi: 10.1016/j.celrep.2016.01.078.

Muñoz, S. *et al.* (2019) ‘A Role for Chromatin Remodeling in Cohesin Loading onto Chromosomes’, *Molecular Cell*. doi: 10.1016/j.molcel.2019.02.027.

Murakami-Tonami, Y. *et al.* (2016) ‘SGO1 is involved in the DNA damage response in MYCN-Amplified neuroblastoma cells’, *Scientific Reports*. doi: 10.1038/srep31615.

Murayama, Y. *et al.* (2018) ‘Establishment of DNA-DNA Interactions by the Cohesin Ring’, *Cell*. doi: 10.1016/j.cell.2017.12.021.

Murayama, Y. and Uhlmann, F. (2013) ‘Biochemical reconstitution of topological DNA binding by the cohesin ring’, *Nature*, 505(7483), pp. 367–371. doi: 10.1038/nature12867.

Murayama, Y. and Uhlmann, F. (2014) ‘Biochemical reconstitution of topological DNA binding by the cohesin ring’, *Nature*, 505(7483), pp. 367–371. doi: 10.1038/nature12867.

Murayama, Y. and Uhlmann, F. (2015) ‘DNA Entry into and Exit out of the Cohesin Ring by an Interlocking Gate Mechanism’, *Cell*, 163(7), pp. 1628–1640. doi: 10.1016/j.cell.2015.11.030.

Muto, A. *et al.* (2014) ‘Nipbl and Mediator Cooperatively Regulate Gene

Expression to Control Limb Development', *PLoS Genetics*. doi: 10.1371/journal.pgen.1004671.

Nakahashi, H. *et al.* (2013) 'A Genome-wide Map of CTCF Multivalency Redefines the CTCF Code', *Cell Reports*. doi: 10.1016/j.celrep.2013.04.024.

Nasmyth, K. and Haering, C. H. (2005) 'THE STRUCTURE AND FUNCTION OF SMC AND KLEISIN COMPLEXES', *Annual Review of Biochemistry*, 74(1), pp. 595–648. doi: 10.1146/annurev.biochem.74.082803.133219.

Nativio, R. *et al.* (2009) 'Cohesin is required for higher-order chromatin conformation at the imprinted IGF2-H19 locus', *PLoS Genetics*. doi: 10.1371/journal.pgen.1000739.

Natsume, T. *et al.* (2016) 'Rapid Protein Depletion in Human Cells by Auxin-Inducible Degron Tagging with Short Homology Donors', *Cell Reports*. doi: 10.1016/j.celrep.2016.03.001.

Nguyen, H. D. *et al.* (2017) 'Functions of Replication Protein A as a Sensor of R Loops and a Regulator of RNaseH1', *Molecular Cell*. doi: 10.1016/j.molcel.2017.01.029.

Nioi, P. *et al.* (2005) 'The Carboxy-Terminal Neh3 Domain of Nrf2 Is Required for Transcriptional Activation', *Molecular and Cellular Biology*. doi: 10.1128/mcb.25.24.10895-10906.2005.

Nishana, M. *et al.* (2020) 'Defining the relative and combined contribution of CTCF and CTCFL to genomic regulation', *Genome Biology*. doi: 10.1186/s13059-020-02024-0.

Nishiyama, T. *et al.* (2010) 'Sororin mediates sister chromatid cohesion by antagonizing Wapl', *Cell*. doi: 10.1016/j.cell.2010.10.031.

Nishiyama, T. *et al.* (2013) 'Aurora B and Cdk1 mediate Wapl activation and release of acetylated cohesin from chromosomes by phosphorylating Sororin',

Proceedings of the National Academy of Sciences of the United States of America. doi: 10.1073/pnas.1305020110.

Nonaka, N. *et al.* (2002) 'Recruitment of cohesin to heterochromatic regions by Swi6/HP1 in fission yeast', *Nature Cell Biology*. doi: 10.1038/ncb739.

Nora, E. P. *et al.* (2012) 'Spatial partitioning of the regulatory landscape of the X-inactivation centre', *Nature*, 485(7398), pp. 381–385. doi: 10.1038/nature11049.

Nora, E. P. *et al.* (2017) 'Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization', *Cell*. doi: 10.1016/j.cell.2017.05.004.

Nozawa, R. S. *et al.* (2017) 'SAF-A Regulates Interphase Chromosome Structure through Oligomerization with Chromatin-Associated RNAs', *Cell*. doi: 10.1016/j.cell.2017.05.029.

Ohle, C. *et al.* (2016) 'Transient RNA-DNA Hybrids Are Required for Efficient Double-Strand Break Repair', *Cell*. doi: 10.1016/j.cell.2016.10.001.

Ohlsson, R., Renkawitz, R. and Lobanenkov, V. (2001) 'CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease', *Trends in Genetics*. doi: 10.1016/S0168-9525(01)02366-6.

Oldach, P. and Nieduszynski, C. A. (2019) 'Cohesin-mediated genome architecture does not define DNA replication timing domains', *Genes*. doi: 10.3390/genes10030196.

Onn, I. and Koshland, D. (2011) 'In vitro assembly of physiological cohesin/DNA complexes', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1107504108.

Orgil, O. *et al.* (2015) 'A Conserved Domain in the Scc3 Subunit of Cohesin Mediates the Interaction with Both Mcd1 and the Cohesin Loader Complex',

PLoS Genetics. doi: 10.1371/journal.pgen.1005036.

Orlandini, E., Marenduzzo, D. and Michieletto, D. (2019) 'Synergy of topoisomerase and structural-maintenance-of-chromosomes proteins creates a universal pathway to simplify genome topology', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1815394116.

Owens, N. *et al.* (2019) 'CTCF confers local nucleosome resiliency after dna replication and during mitosis', *eLife*. doi: 10.7554/eLife.47898.

Pan, H. *et al.* (2020) 'Cohesin SA1 and SA2 are RNA binding proteins that localize to RNA containing regions on DNA', *Nucleic acids research*. doi: 10.1093/nar/gkaa284.

Pan, X. *et al.* (2013) 'YY1 controls Igκ repertoire and B-cell development, and localizes with condensin on the Igκ locus', *EMBO Journal*. doi: 10.1038/emboj.2013.66.

Panigrahi, A. K. *et al.* (2012) 'A cohesin-RAD21 interactome', *Biochemical Journal*. doi: 10.1042/BJ20111745.

Papamichos-Chronakis, M. *et al.* (2011) 'Global regulation of H2A.Z localization by the INO80 chromatin-remodeling enzyme is essential for genome integrity', *Cell*. doi: 10.1016/j.cell.2010.12.021.

Papamichos-Chronakis, M. and Peterson, C. L. (2008) 'The Ino80 chromatin-remodeling enzyme regulates replisome function and stability', *Nature Structural and Molecular Biology*. doi: 10.1038/nsmb.1413.

Parelho, V. *et al.* (2008) 'Cohesins Functionally Associate with CTCF on Mammalian Chromosome Arms', *Cell*, 132, pp. 422–433. doi: 10.1016/j.cell.2008.01.011.

Pežić, D. *et al.* (2021) 'The cohesin regulator Stag1 promotes cell plasticity

through heterochromatin regulation.', *bioRxiv*.

Phillips-Cremins, J. E. *et al.* (2013) 'Architectural protein subclasses shape 3D organization of genomes during lineage commitment', *Cell*. doi: 10.1016/j.cell.2013.04.053.

Phillips, D. D. *et al.* (2013) 'The sub-nanomolar binding of DNA-RNA hybrids by the single-chain Fv fragment of antibody S9.6', *Journal of Molecular Recognition*. doi: 10.1002/jmr.2284.

Polo, S. E. *et al.* (2012) 'Regulation of DNA-End Resection by hnRNPU-like Proteins Promotes DNA Double-Strand Break Signaling and Repair', *Molecular Cell*. doi: 10.1016/j.molcel.2011.12.035.

Pomerantz, R. T. and O'Donnell, M. (2008) 'The replisome uses mRNA as a primer after colliding with RNA polymerase', *Nature*. doi: 10.1038/nature07527.

Pope, B. D. *et al.* (2014) 'Topologically associating domains are stable units of replication-timing regulation', *Nature*, 515(7527), pp. 402–405. doi: 10.1038/nature13986.

Porter, H. *et al.* (2021) 'STAG proteins promote cohesin ring loading at R-loops', *bioRxiv*, p. 2021.02.20.432055. doi: 10.1101/2021.02.20.432055.

Powell, W. T. *et al.* (2013) 'R-loop formation at Snord116 mediates topotecan inhibition of Ube3a-antisense and allele-specific chromatin decondensation', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1305426110.

Prendergast, L. *et al.* (2020) 'Resolution of R-loops by INO80 promotes DNA replication and maintains cancer cell proliferation and viability', *Nature Communications*. doi: 10.1038/s41467-020-18306-x.

Prieto, I. *et al.* (2002) 'STAG2 and Rad21 mammalian mitotic cohesins are implicated in meiosis', *EMBO Reports*. doi: 10.1093/embo-reports/kvf108.

Pugacheva, E. M. *et al.* (2020) ‘CTCF mediates chromatin looping via N-terminal domain-dependent cohesin retention’, *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.1911708117.

Quitschke, W. W. *et al.* (2000) ‘Differential effect of zinc finger deletions on the binding of CTCF to the promoter of the amyloid precursor protein gene’, *Nucleic Acids Research*. doi: 10.1093/nar/28.17.3370.

Ramírez, F. *et al.* (2016) ‘deepTools2: a next generation web server for deep-sequencing data analysis’, *Nucleic acids research*. doi: 10.1093/nar/gkw257.

Rankin, S., Ayad, N. G. and Kirschner, M. W. (2005) ‘Sororin, a substrate of the anaphase- promoting complex, is required for sister chromatid cohesion in vertebrates’, *Molecular Cell*. doi: 10.1016/j.molcel.2005.03.017.

Rao, S. S. P. *et al.* (2014) ‘A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping’, *Cell*, 159(7), pp. 1665–1680. doi: 10.1016/j.cell.2014.11.021.

Rao, S. S. P. *et al.* (2017a) ‘Cohesin Loss Eliminates All Loop Domains’, *Cell*, 171, pp. 305–320. doi: 10.1016/j.cell.2017.09.026.

Rao, S. S. P. *et al.* (2017b) ‘Cohesin Loss Eliminates All Loop Domains’, *Cell*, 171(2), pp. 305-320.e24. doi: 10.1016/j.cell.2017.09.026.

Rappsilber, J. *et al.* (2002) ‘Large-scale proteomic analysis of the human spliceosome’, *Genome Research*. doi: 10.1101/gr.473902.

Rayner, E. *et al.* (2019) ‘CRISPR-Cas9 Causes Chromosomal Instability and Rearrangements in Cancer Cell Lines, Detectable by Cytogenetic Methods’, *The CRISPR Journal*. doi: 10.1089/crispr.2019.0006.

Recillas-Targa, F. *et al.* (2002) ‘Position-effect protection and enhancer blocking by the chicken β-globin insulator are separable activities’, *Proceedings of the*

National Academy of Sciences of the United States of America. doi: 10.1073/pnas.102179399.

Reid, D. W. and Nicchitta, C. V. (2012) ‘The enduring enigma of nuclear translation’, *Journal of Cell Biology*. doi: 10.1083/jcb.201202140.

Remeseiro, S., Cuadrado, A., Gómez-Lãpez, G., *et al.* (2012) ‘A unique role of cohesin-SA1 in gene regulation and development’, *EMBO Journal*. doi: 10.1038/emboj.2012.60.

Remeseiro, S., Cuadrado, A., Carretero, M., *et al.* (2012) ‘Cohesin-SA1 deficiency drives aneuploidy and tumourigenesis in mice due to impaired replication of telomeres’, *EMBO Journal*. doi: 10.1038/emboj.2012.11.

Remeseiro, S. *et al.* (2013) ‘Reduction of Nipbl impairs cohesin loading locally and affects transcription but not cohesion-dependent functions in a mouse model of Cornelia de Lange Syndrome’, *Biochimica et Biophysica Acta - Molecular Basis of Disease*. doi: 10.1016/j.bbadis.2013.07.020.

Rhodes, J. *et al.* (2017) ‘Scc2/Nipbl hops between chromosomal cohesin rings after loading’, *eLife*. doi: 10.7554/eLife.30000.

Rhodes, J. M. *et al.* (2010) ‘Positive regulation of c-Myc by cohesin is direct, and evolutionarily conserved’, *Developmental Biology*. doi: 10.1016/j.ydbio.2010.05.493.

Richardson, J. P. (1975) ‘Attachment of nascent RNA molecules to superhelical DNA’, *Journal of Molecular Biology*. doi: 10.1016/S0022-2836(75)80087-8.

Riedel, C. G. *et al.* (2006) ‘Protein phosphatase 2A protects centromeric sister chromatid cohesion during meiosis I’, *Nature*. doi: 10.1038/nature04664.

Roberts, R. W. and Crothers, D. M. (1992) ‘Stability and properties of double and triple helices: Dramatic effects of RNA or DNA backbone composition’, *Science*. doi: 10.1126/science.1279808.

Rocquain, J. *et al.* (2010) ‘Alteration of cohesin genes in myeloid diseases’, *American Journal of Hematology*. doi: 10.1002/ajh.21798.

Roig, M. B. *et al.* (2014) ‘Structure and function of cohesin’s Scc3/SA regulatory subunit’, *FEBS Letters*. doi: 10.1016/j.febslet.2014.08.015.

Rollins, R. A. *et al.* (2004) ‘Drosophila Nipped-B Protein Supports Sister Chromatid Cohesion and Opposes the Stromalin/Scc3 Cohesion Factor To Facilitate Long-Range Activation of the cut Gene’, *Molecular and Cellular Biology*. doi: 10.1128/mcb.24.8.3100-3111.2004.

Rollins, R. A., Morcillo, P. and Dorsett, D. (1999) ‘Nipped-B, a Drosophila homologue of chromosomal adherins, participates in activation by remote enhancers in the cut and Ultrabithorax genes’, *Genetics*. doi: 10.1093/genetics/152.2.577.

Romero-Pérez, L. *et al.* (2019) ‘STAG Mutations in Cancer’, *Trends in Cancer*. doi: 10.1016/j.trecan.2019.07.001.

Rotem, A. *et al.* (2015) ‘Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state’, *Nature Biotechnology*. doi: 10.1038/nbt.3383.

Roy, D. *et al.* (2010) ‘Competition between the RNA Transcript and the Nontemplate DNA Strand during R-Loop Formation In Vitro: a Nick Can Serve as a Strong R-Loop Initiation Site’, *Molecular and Cellular Biology*. doi: 10.1128/mcb.00897-09.

Roy, D. and Lieber, M. R. (2009) ‘G Clustering Is Important for the Initiation of Transcription-Induced R-Loops In Vitro, whereas High G Density without Clustering Is Sufficient Thereafter’, *Molecular and Cellular Biology*. doi: 10.1128/mcb.00139-09.

Rubio, E. D. *et al.* (2008) ‘CTCF physically links cohesin to chromatin’, *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.0801273105.

Ryan, V. H. *et al.* (2018) ‘Mechanistic View of hnRNPA2 Low-Complexity Domain Structure, Interactions, and Phase Separation Altered by Mutation and Arginine Methylation’, *Molecular Cell*. doi: 10.1016/j.molcel.2017.12.022.

Ryu, J. K. *et al.* (2021) ‘Bridging-induced phase separation induced by cohesin SMC protein complexes’, *Science Advances*. doi: 10.1126/sciadv.abe5905.

Sakasai, R. *et al.* (2017) ‘Aquarius is required for proper CtIP expression and homologous recombination repair’, *Scientific Reports*. doi: 10.1038/s41598-017-13695-4.

Salas-Armenteros, I. *et al.* (2017) ‘ Human THO –Sin3A interaction reveals new mechanisms to prevent R-loops that cause genome instability ’, *The EMBO Journal*. doi: 10.15252/embj.201797208.

Saldaña-Meyer, R. *et al.* (2014) ‘CTCF regulates the human p53 gene through direct interaction with its natural antisense transcript, Wrap53’, *Genes and Development*. doi: 10.1101/gad.236869.113.

Saldaña-Meyer, R. *et al.* (2019) ‘RNA Interactions Are Essential for CTCF-Mediated Genome Organization’, *Molecular Cell*. doi: 10.1016/j.molcel.2019.08.015.

Sanborn, A. L. *et al.* (2015) ‘Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes’, *Proceedings of the National Academy of Sciences*, 112(47), pp. E6456–E6465. doi: 10.1073/pnas.1518552112.

Sanz, L. A. *et al.* (2016) ‘Prevalent, Dynamic, and Conserved R-Loop Structures Associate with Specific Epigenomic Signatures in Mammals’, *Molecular Cell*. doi: 10.1016/j.molcel.2016.05.032.

Schmidt, D. *et al.* (2010) ‘A CTCF-independent role for cohesin in tissue-specific transcription’, *Genome Research*. doi: 10.1101/gr.100479.109.

Schmitz, J. *et al.* (2007) ‘Sororin Is Required for Stable Binding of Cohesin to Chromatin and for Sister Chromatid Cohesion in Interphase’, *Current Biology*. doi: 10.1016/j.cub.2007.02.029.

Schvartzman, J. M., Thompson, C. B. and Finley, L. W. S. (2018) ‘Metabolic regulation of chromatin modifications and gene expression’, *Journal of Cell Biology*. doi: 10.1083/jcb.201803061.

Schwab, R. A. *et al.* (2015) ‘The Fanconi Anemia Pathway Maintains Genome Stability by Coordinating Replication and Transcription’, *Molecular Cell*. doi: 10.1016/j.molcel.2015.09.012.

Schwarzer, W. *et al.* (2017) ‘Two independent modes of chromatin organization revealed by cohesin removal’, *Nature*, 551(7678), pp. 51–56. doi: 10.1038/nature24281.

Seitan, V. C. *et al.* (2013) ‘Cohesin-Based chromatin interactions enable regulated gene expression within preexisting architectural compartments’, *Genome Research*, *23*(12), pp. 2066–2077. doi: 10.1101/gr.161620.113.

Sexton, T. *et al.* (2012) ‘Three-dimensional folding and functional organization principles of the Drosophila genome’, *Cell*, pp. 458–472. doi: 10.1016/j.cell.2012.01.010.

Shen, W. *et al.* (2017) ‘Dynamic nucleoplasmic and nucleolar localization of mammalian RNase H1 in response to RNAP I transcriptional R-loops’, *Nucleic Acids Research*. doi: 10.1093/nar/gkx710.

Shen, X. *et al.* (2000) ‘A chromatin remodelling complex involved in transcription and DNA processing’, *Nature*. doi: 10.1038/35020123.

Shi, Z. *et al.* (2020) ‘Cryo-EM structure of the human cohesin-NIPBL-DNA complex’, *Science*. doi: 10.1126/science.abb0981.

Shintomi, K. and Hirano, T. (2009) ‘Releasing cohesin from chromosome arms

in early mitosis: Opposing actions of Wapl-Pds5 and Sgo1', *Genes and Development*, 23(18), pp. 2224–2236. doi: 10.1101/gad.1844309.

Shyu, Y. J. *et al.* (2006) 'Identification of new fluorescent protein fragments for bimolecular fluorescence complementation analysis under physiological conditions', *Biotechniques*, 40(1), pp. 61–66. doi: 000112036 [pii].

Skourti-Stathaki, K., Kamieniarz-Gdula, K. and Proudfoot, N. J. (2014) 'R-loops induce repressive chromatin marks over mammalian gene terminators', *Nature*. doi: 10.1038/nature13787.

Skourti-Stathaki, K., Proudfoot, N. J. and Gromak, N. (2011) 'Human Senataxin Resolves RNA/DNA Hybrids Formed at Transcriptional Pause Sites to Promote Xrn2-Dependent Termination', *Molecular Cell*. doi: 10.1016/j.molcel.2011.04.026.

Sofueva, S. *et al.* (2013) 'Cohesin-mediated interactions organize chromosomal domain architecture', *EMBO Journal*, 32(24), pp. 3119–3129. doi: 10.1038/emboj.2013.237.

Sollier, J. *et al.* (2014) 'Transcription-Coupled Nucleotide Excision Repair Factors Promote R-Loop-Induced Genome Instability', *Molecular Cell*. doi: 10.1016/j.molcel.2014.10.020.

Solomon, D. A. *et al.* (2011) 'Mutational inactivation of STAG2 causes aneuploidy in human cancer', *Science*. doi: 10.1126/science.1203619.

Sonoda, E. *et al.* (2001) 'Scc1/Rad21/Mcd1 is required for sister chromatid cohesion and kinetochore function in vertebrate cells.', *Developmental cell*, 1(6), pp. 759–70. doi: 10.1016/S1534-5807(01)00088-0.

Splinter, E. *et al.* (2006) 'CTCF mediates long-range chromatin looping and local histone modification in the β-globin locus', *Genes and Development*. doi: 10.1101/gad.399506.

Stein, H. and Hausen, P. (1969) ‘Enzyme from calf thymus degrading the RNA moiety of DNA-RNA hybrids: Effect on DNA-dependent RNA polymerase’, *Science*. doi: 10.1126/science.166.3903.393.

Stigler, J. *et al.* (2016) ‘Single-Molecule Imaging Reveals a Collapsed Conformational State for DNA-Bound Cohesin’, *Cell Reports*. doi: 10.1016/j.celrep.2016.04.003.

Sumara, I. *et al.* (2000) ‘Characterization of vertebrate cohesin complexes and their regulation in prophase’, *Journal of Cell Biology*, 151(4), pp. 749–761. doi: 10.1083/jcb.151.4.749.

Sutani, T. *et al.* (2009) ‘Budding Yeast Wpl1(Rad61)-Pds5 Complex Counteracts Sister Chromatid Cohesion-Establishing Reaction’, *Current Biology*, 19(6), pp. 492–497. doi: 10.1016/j.cub.2009.01.062.

Szabó, P. E. *et al.* (2000) ‘Maternal-specific footprints at putative CTCF sites in the H19 imprinting control region give evidence for insulator function’, *Current Biology*. doi: 10.1016/S0960-9822(00)00489-9.

Szklarczyk, D. *et al.* (2019) ‘STRING v11: Protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets’, *Nucleic Acids Research*. doi: 10.1093/nar/gky1131.

Tanaka, T. *et al.* (1999) ‘Identification of cohesin association sites at centromeres and along chromosome arms’, *Cell*. doi: 10.1016/S0092-8674(00)81518-4.

Tedeschi, A. *et al.* (2013) ‘Wapl is an essential regulator of chromatin structure and chromosome segregation’, *Nature*. doi: 10.1038/nature12471.

Thorvaldsen, J. L., Duran, K. L. and Bartolomei, M. S. (1998) ‘Deletion of the H19 differentially methylated domain results in loss of imprinted expression of H19 and Igf2’, *Genes and Development*. doi: 10.1101/gad.12.23.3693.

Thurman, R. E. *et al.* (2012) 'The accessible chromatin landscape of the human genome', *Nature*. doi: 10.1038/nature11232.

Tonkin, E. T. *et al.* (2004) 'NIPBL, encoding a homolog of fungal Scc2-type sister chromatid cohesion proteins and fly Nipped-B, is mutated in Cornelia de Lange syndrome', *Nature Genetics*. doi: 10.1038/ng1363.

Tóth, A. *et al.* (1999) 'Yeast cohesin complex requires a conserved protein, Eco1p(Ctf7), to establish cohesion between sister chromatids during DNA replication', *Genes and Development*, 13(3), pp. 320–333. doi: 10.1101/gad.13.3.320.

Truax, A. D. and Greer, S. F. (2012) 'ChIP and Re-ChIP assays: Investigating interactions between regulatory proteins, histone modifications, and the DNA sequences to which they bind', *Methods in Molecular Biology*. doi: 10.1007/978-1-61779-376-9_12.

Tuduri, S. *et al.* (2009) 'Topoisomerase I suppresses genomic instability by preventing interference between replication and transcription', *Nature Cell Biology*. doi: 10.1038/ncb1984.

Udugama, M., Sabri, A. and Bartholomew, B. (2011) 'The INO80 ATP-Dependent Chromatin Remodeling Complex Is a Nucleosome Spacing Factor', *Molecular and Cellular Biology*. doi: 10.1128/mcb.01035-10.

Uhlmann, F. *et al.* (2000) 'Cleavage of cohesin by the CD clan protease separin triggers anaphase in yeast', *Cell*. doi: 10.1016/S0092-8674(00)00130-6.

Uhlmann, F., Lottspelch, F. and Nasmyth, K. (1999) 'Sister-chromatid separation at anaphase onset is promoted by cleavage of the cohesin subunit Scc1', *Nature*. doi: 10.1038/21831.

Uusküla-Reimand, L. *et al.* (2016) 'Topoisomerase II beta interacts with cohesin and CTCF at topological domain borders', *Genome Biology*. doi: 10.1186/s13059-016-1043-8.

Venkatesh, S. and Workman, J. L. (2015) ‘Histone exchange, chromatin structure and the regulation of transcription’, *Nature Reviews Molecular Cell Biology*. doi: 10.1038/nrm3941.

Vian, L. *et al.* (2018) ‘The Energetics and Physiological Impact of Cohesin Extrusion’, *Cell*. doi: 10.1016/j.cell.2018.03.072.

Vietri Rudan, M. *et al.* (2015) ‘Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture’, *Cell Reports*, 10(8), pp. 1297–1309. doi: 10.1016/j.celrep.2015.02.004.

Viny, A. D. *et al.* (2019) ‘Stag1 and Stag2 regulate cell fate decisions in hematopoiesis through non-redundant topological control’, *bioRxiv*. doi: 10.1101/581868.

Wahba, L. *et al.* (2011) ‘RNase H and Multiple RNA Biogenesis Factors Cooperate to Prevent RNA:DNA Hybrids from Generating Genome Instability’, *Molecular Cell*. doi: 10.1016/j.molcel.2011.10.017.

Waizenegger, I. C. *et al.* (2000) ‘Two distinct pathways remove mammalian cohesin from chromosome arms in prophase and from centromeres in anaphase’, *Cell*. doi: 10.1016/S0092-8674(00)00132-X.

Wang, H. *et al.* (2012) ‘Widespread plasticity in CTCF occupancy linked to DNA methylation’, *Genome Research*. doi: 10.1101/gr.136101.111.

Wang, I. X. *et al.* (2018) ‘Human proteins that interact with RNA/DNA hybrids’, *Genome Research*. doi: 10.1101/gr.237362.118.

Wang, S. *et al.* (2016) ‘Spatial organization of chromatin domains and compartments in single chromosomes’, *Science*. doi: 10.1126/science.aaf8084.

Watrin, E. *et al.* (2006) ‘Human Scc4 Is Required for Cohesin Binding to Chromatin, Sister-Chromatid Cohesion, and Mitotic Progression’, *Current Biology*. doi: 10.1016/j.cub.2006.03.049.

Weber, C. M., Henikoff, J. G. and Henikoff, S. (2010) 'H2A.Z nucleosomes enriched over active genes are homotypic', *Nature Structural and Molecular Biology*. doi: 10.1038/nsmb.1926.

Weitzer, S., Lehane, C. and Uhlmann, F. (2003) 'A Model for ATP Hydrolysis-Dependent Binding of Cohesin to DNA', *Current Biology*. doi: 10.1016/j.cub.2003.10.030.

Welch, J. S. *et al.* (2012) 'The origin and evolution of mutations in acute myeloid leukemia', *Cell*. doi: 10.1016/j.cell.2012.06.023.

Wendt, K. S. *et al.* (2008) 'Cohesin mediates transcriptional insulation by CCCTC-binding factor', *Nature*, 451(7180), pp. 796–801. doi: 10.1038/nature06634.

West, S., Gromak, N. and Proudfoot, N. J. (2004) 'Human 5′ → 3′ exonuclease Xm2 promotes transcription termination at co-transcriptional cleavage sites', *Nature*. doi: 10.1038/nature03035.

Wutz, G. *et al.* (2017) 'Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins', *The EMBO Journal*. doi: 10.15252/embj.201798004.

Wutz, G. *et al.* (2020) 'ESCO1 and CTCF enable formation of long chromatin loops by protecting cohesinstag1 from WAPL', *eLife*. doi: 10.7554/eLife.52091.

Xiao, T., Wallace, J. and Felsenfeld, G. (2011) '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', *Molecular and Cellular Biology*. doi: 10.1128/mcb.05093-11.

Xu, C. *et al.* (2014) 'Structural basis for selective binding of m6A RNA by the YTHDC1 YTH domain', *Nature Chemical Biology*. doi: 10.1038/nchembio.1654.

Yaffe, E. and Tanay, A. (2011) 'Probabilistic modeling of Hi-C contact maps

eliminates systematic biases to characterize global chromosomal architecture', *Nature Genetics*. doi: 10.1038/ng.947.

Yang, Y. *et al.* (2014) 'Arginine Methylation Facilitates the Recruitment of TOP3B to Chromatin to Prevent R Loop Accumulation', *Molecular Cell*. doi: 10.1016/j.molcel.2014.01.011.

Yao, H. *et al.* (2010) 'Mediation of CTCF transcriptional insulation by DEAD-box RNA-binding protein p68 and steroid receptor RNA activator SRA', *Genes and Development*. doi: 10.1101/gad.1967810.

Yasuhara, T. *et al.* (2018) 'Human Rad52 Promotes XPG-Mediated R-loop Processing to Initiate Transcription-Associated Homologous Recombination Repair', *Cell*. doi: 10.1016/j.cell.2018.08.056.

Yesbolatova, A. *et al.* (2019) 'Generation of conditional auxin-inducible degron (AID) cells and tight control of degron-fused proteins using the degradation inhibitor auxinole', *Methods*. doi: 10.1016/j.ymeth.2019.04.010.

Yu, E. Y. *et al.* (2007) 'Regulation of Telomere Structure and Functions by Subunits of the INO80 Chromatin Remodeling Complex', *Molecular and Cellular Biology*. doi: 10.1128/mcb.00418-07.

Yuan, B. *et al.* (2015) 'Global transcriptional disturbances underlie Cornelia de Lange syndrome and related phenotypes', *Journal of Clinical Investigation*. doi: 10.1172/JCI77435.

Zampieri, M. *et al.* (2012) 'ADP-ribose polymers localized on Ctcf-Parp1-Dnmt1 complex prevent methylation of Ctcf target sites', *Biochemical Journal*. doi: 10.1042/BJ20111417.

Zhang *et al.* (2008) 'A handcuff model for the cohesin complex', *Journal of Cell Biology*. doi: 10.1083/jcb.200801157.

Zhang, J. *et al.* (2008) 'Acetylation of Smc3 by Eco1 Is Required for S Phase

Sister Chromatid Cohesion in Both Human and Yeast', *Molecular Cell*. doi: 10.1016/j.molcel.2008.06.006.

Zhang, N. *et al.* (2013) 'Characterization of the Interaction between the Cohesin Subunits Rad21 and SA1/2', *PLoS ONE*. doi: 10.1371/journal.pone.0069458.

Zhang, N. and Pati, D. (2015) 'C-terminus of sororin interacts with sa2 and regulates sister chromatid cohesion', *Cell Cycle*. doi: 10.1080/15384101.2014.1000206.

Zhang, Z. Z. *et al.* (2015) 'Complexities due to single-stranded RNA during antibody detection of genomic rna:dna hybrids', *BMC Research Notes*. doi: 10.1186/s13104-015-1092-1.

Zhao, B. S., Roundtree, I. A. and He, C. (2016) 'Post-transcriptional gene regulation by mRNA modifications', *Nature Reviews Molecular Cell Biology*. doi: 10.1038/nrm.2016.132.

Zheng, G. *et al.* (2018) 'MCM2–7-dependent cohesin loading during s phase promotes sister-chromatid cohesion', *eLife*. doi: 10.7554/eLife.33920.

Ziatanova, J. and Caiafa, P. (2009) 'CTCF and its protein partners: Divide and rule?', *Journal of Cell Science*. doi: 10.1242/jcs.039990.

Zilberman, D. *et al.* (2008) 'Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks', *Nature*. doi: 10.1038/nature07324.

Zuin *et al.* (2014) 'Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells', *Proceedings of the National Academy of Sciences*, 111(3), pp. 996–1001. doi: 10.1073/pnas.1317788111.

Zuin, J. *et al.* (2014) 'A Cohesin-Independent Role for NIPBL at Promoters Provides Insights in CdLS', *PLoS Genetics*, 10(2). doi: 10.1371/journal.pgen.1004153.