Dissecting the molecular mechanisms of lncRNA function in X chromosome inactivation across mammalian gestation evolution

Ioannis Tsagakis

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

Faculty of Biological Sciences

School of Molecular and Cellular Biology

October 2021

The candidate confirms that the work submitted is his own and that appropriate credit has been given where reference has been made to the work of others.

“The proper function of man is to live, not to e-XIST.
I shall not waste my days in trying to prolong them.
I shall use my time.”

— Jack London (1876–1916)

Acknowledgements

This thesis and the whole PhD project by extension would not have been possible without help from my supervisory team. I feel gratitude towards Dr Julie Aspden for taking a chance on me and giving me the opportunity to train in her lab not only as a technician but also as a PhD student with a limited knowledge of biochemistry. I would like to thank her for her patience, guidance and for always finding time to answer my questions.

I would like to thank Dr Niamh Forde for her overall advice on how to be successful in research, teaching me what it takes to achieve the objective in sight but also to take time to destress and be with loved ones. I am grateful for her patience in the early stages of my journey when I would enthusiastically bombard her with questions. Her advice on scientific writing will also stay with me.

Dr Mary O' Connell's role in my PhD has been instrumental both in experimental design and data interpretation when it came to engaging in computational biology, to which field I am a newcomer. She has meticulously spent hours to help me get to grips with concepts about evolutionary biology, which made me develop an appreciation for it. All of the people above have graciously tolerated my failures and equally celebrated my successes, and for that I consider them as great examples to turn to for leadership.

Special thanks to the Aspden group, past and present, for making the PhD a pleasant experience and for being a helpful bunch, ready to support anyone in time of need, with matters of science and otherwise. More specifically, I would like to thank Michaela Agapiou and Isabel Birds for having several discussions about science and using research tools. They also provided fantastic colleagues and set the standards for the team high, which made for well-intended competition. Thanks to Eleanor Walton and Tayah Hopes for showing me around the lab and refamiliarising me with RT-qPCR and western blotting. Special thanks to Tayah Hopes for setting a great example of a meticulous scientist to aspire to, which has on occasion provided motivation to strive to be better at research. I would like to also extend my thanks to Andreas for being a good lab citizen and for coming to the

rescue when things got tricky. Although Mark Handley was only with the group for a year, his incisive comments at journal club together with a sharp insight about experimental design and troubleshooting was particularly beneficial for my training as a researcher. Last but not least, Karl Norris was very supportive throughout the ups and downs of my PhD journey, always offering valuable perspective regardless of whether I was able to tell at the time. For that and for his many hours spent having fruitful discussions with me about science, I am grateful, as he was the closest thing to a mentor and a friend.

I would be amiss if I did not mention and recognise Ali Taylor, Peter Mulhair and David Orr for their significant contributions to my understanding of evolutionary biology and computational analyses. Their assistance in comprehending how to utilise available tools and interpret data was valuable in my PhD training. Thanks should be extended to Irene Malo Estepa, Annika Geijer-Simpson and Haidee Tinning for their contributions to my project, but also for being excellent colleagues who have many a time come to the rescue at difficult times. Irene Malo Estepa and Rachel Quilang have provided valuable perspective at difficult times as well as mental health support and for that I would like to thank them.

Special thanks to Laura Najera Cortazar for her contributions to my knowledge of the placental mammal phylogeny and for her support throughout ups and downs in my PhD. Her friendship made the PhD a fun experience.

Katerina Douka and Iosifina Sampson proved to be more than colleagues after having countless conversations about troubleshooting experiments, discussing future career plans and trajectories as well as exchanging life advice. I am particularly indebted to them for encouragement with my studies and always being available for mental health support and coffee sessions, and for that I consider them my extended ‘lab-related’ family.

The PhD journey would not have been as a pleasant experience in its early days without emotional support and mindfulness from Stella. I am grateful for the valuable life lessons you taught me. Your actions provided me with additional motivation to

work harder and more efficiently whilst also aiming to relish and take full advantage of free time.

Perhaps the cornerstone of my success, rooting for me and encouraging me to complete my studies for over 12 years are my family, who deserve a special mention here. Without you, this might not have been possible and you deserve this accreditation as much as me.

Abstract

Long non-coding RNAs (lncRNAs) are >200-nucleotide transcripts that do not encode proteins, but interact with proteins and regulate gene expression. Whilst full-length sequence conservation is rare for lncRNAs, short regions of higher conservation can exist across species. The *XIST* lncRNA mediates dosage compensation via X chromosome inactivation (XCI) of a single X chromosome in females. Maintaining gene dosage across sexes is vital for placental mammal prenatal development, failure of which is embryonic lethal. Despite its presence throughout placental mammals, most studies have focused on mouse *Xist*. It is yet to be determined whether mouse *Xist*-protein interactions are shared across placental mammals where the timing and nature of XCI differ. Here, we aimed to dissect *XIST*'s interactors in placental mammals with different implantation strategies.

Spen, Hnrnpk, Ciz1, Rbm15 and Wtap proteins were previously identified as mouse *Xist* functional interactors. Their average amino acid identity is >70% across human, mouse, cow and pig. RT-qPCR and western blotting revealed coordinate expression of *XIST* and putative protein partners in endometrial tissues/cells from those species. RNA immunoprecipitations showed SPEN, hnRNPK, WTAP and CIZ1 proteins bind human *XIST* but interactions could not be robustly assessed in cow. RNA pulldowns revealed bovine CIZ1 interacts with *XIST* E-repeat in cow and RBM15 interacts with human *XIST* A-repeat, as occurs in mouse. Proteomic analyses indicated hnRNPU and TOP1 bound bovine *XIST* A-repeat. Bovine *XIST* A-repeat bound MATR3, RALYL and YY1 in human lysates. Selective pressure variation analyses identified residues under positive selection in a subset of these proteins, but could not explain their differential binding to human and bovine *XIST*. Altogether, the *XIST* interactome was characterised for the first time in endometrial-derived bovine cells, revealing cow-specific and conserved interactors across placental mammals. Bovine *XIST* interactors from human cells reported here may contribute to our understanding of lncRNA-protein partner co-evolution.

Table of Contents

List of Figures

List of Tables

List of abbreviations

BEB, Bayes empirical Bayes
CLIP-Seq, Cross-linking immunoprecipitation coupled to sequencing
ChIP-Seq, Chromatin immunoprecipitation coupled to sequencing
DMS-Seq, Dimethylsulfate sequencing
dN, non-synonymous substitutions per nonsynonymous site
dS, synonymous substitution per synonymous site
EGA, embryonic genome activation
EMSA, electrophoretic mobility assays
gBGC, GC-biased conversion
H3K27me3, tri-methylation of lysine 27 on histone H3 protein
hESC, human embryonic stem cell
irCLIP, Infrared crosslinking immunoprecipitation followed by sequencing
KH, K Homology domain
LC-MS, liquid chromatography-mass spectrometry
lncRNA, long non-coding RNA
LRT, likelihood ratio test
MEF, mouse embryonic fibroblast
mESC, mouse embryonic stem cell
Ne, effective population
NES, nuclear export signal
NLS, nuclear localisation signal
PAML, Phylogenetic Analysis by Maximum Likelihood
Pol II, RNA polymerase II
PP, posterior probability
PRC1, Polycomb Repressive Complex 1
PRC2, Polycomb Repressive Complex 2
RG/RGG, arginine–glycine–glycine repeat domain
RID, nuclear receptor interaction domain
RIP, RNA immunoprecipitation
RMM, RNA recognition motif domain
RNA FISH, RNA fluorescent in situ hybridisation
RNA-seq, RNA-sequencing

RNP, RNA protein complex

RT-qPCR, reverse transcriptase quantitative PCR

SAP, SAF-A/B, Acinus and PIAS domain

SHAPE, Selective 2’ Hydroxyl Acylation analyzed by Primer Extension

SPOC, SPEN paralogue/orthologue C-terminal domain

SPRY, SPla and the RYanodine Receptor domain

TE, Transposable element

XCI, X chromosome inactivation

ZF, zinc finger domain

ω, ratio of dN/dS

Main Introduction

1.1. Discovery and features of long non-coding RNAs (lncRNAs)

The complete draft of the Human Genome Project originally identified 20,000-25,000 protein-coding genes and did not aim to characterise non-protein-coding genes (International Human Genome Sequencing, 2004). The ENCODE project found 20,687 protein-coding genes, representing 2.94% of the genome (Dunham et al., 2012). In addition, deep sequencing of the transcriptome revealed that >80% of the genome participates in a biochemical RNA or chromatin-associated reaction (Dunham et al., 2012). The cascade of RNA sequencing experiments that succeeded the Human Genome Project have identified that the human genome is pervasively transcribed (Carninci et al., 2005, Derrien et al., 2012). One class of transcripts identified are long non-coding RNAs (lncRNAs) which are defined as RNAs > 200 nt long that do not encode for proteins. Although currently this description does not reflect a single class of transcripts given it is defined by exclusion criteria, it has remained an easy classification benchmark adopted by many researchers (Gil and Ulitsky, 2020, Seal et al., 2020). Despite the growing numbers of these transcripts being detected, the proportion of them shown to be functional remains extremely low. For instance, the current catalogue of long non-coding RNA transcripts, according to LNCipedia v5.2, comes up to >100,000 (Volders et al., 2019) but only around 100 have an experimentally validated function. According to GENCODE v38, in human, there are 17,944 lncRNA genes generating 48,752 transcripts whereas there are 19,955 protein-coding genes, giving rise to 86,757 transcripts.

LncRNAs comprise a heterogeneous pool of transcripts since several transcripts fit into that category. In order to provide some structure into this extensive catalogue of transcripts, lncRNAs can be further sub-categorised based on their genomic organisation. The majority of lncRNAs are transcribed from distinct, well-annotated genomic loci similar to protein coding genes, also termed long intergenic (or intervening) non-coding RNAs (or lincRNAs; Figure 1A). There are lncRNAs that can be transcribed in the antisense orientation to a protein-coding gene, i.e. natural antisense transcripts (NATs; Figure 1B and 1C) (9). Bi-directional promoters can

direct transcription of genes in the sense and antisense orientation at the same time. Therefore, depending on how lncRNA genes are positioned, this can give rise to antisense lncRNAs (Figure 1D) (10). Moreover, lncRNA transcription start sites can overlap exons or introns of protein-coding genes (see Figure 1E and F) and might even be harboured entirely inside an intron of a protein-coding gene (1). Additionally, a protein-coding or another non-coding RNA (ncRNA) gene may reside within a larger spanning lncRNA locus (Figure 1G). Finally, lncRNA have additionally been detected to originate from enhancers (eRNAs and elncRNAs; as in Figure 1H), upstream to and from promoters (PROMPTS and plncRNAs; Figure 1I) as well as from ultra-conserved elements (11, 12). Overall, lncRNAs can arise from multiple genomic loci, frequently overlapping other non-coding RNA or protein-coding genes which complicates the study of their biological significance. Beyond this, transcripts within the lncRNA category can further vary depending on their subcellular localisation (see below), conservation (Section 1.2) and function (Section 1.3).

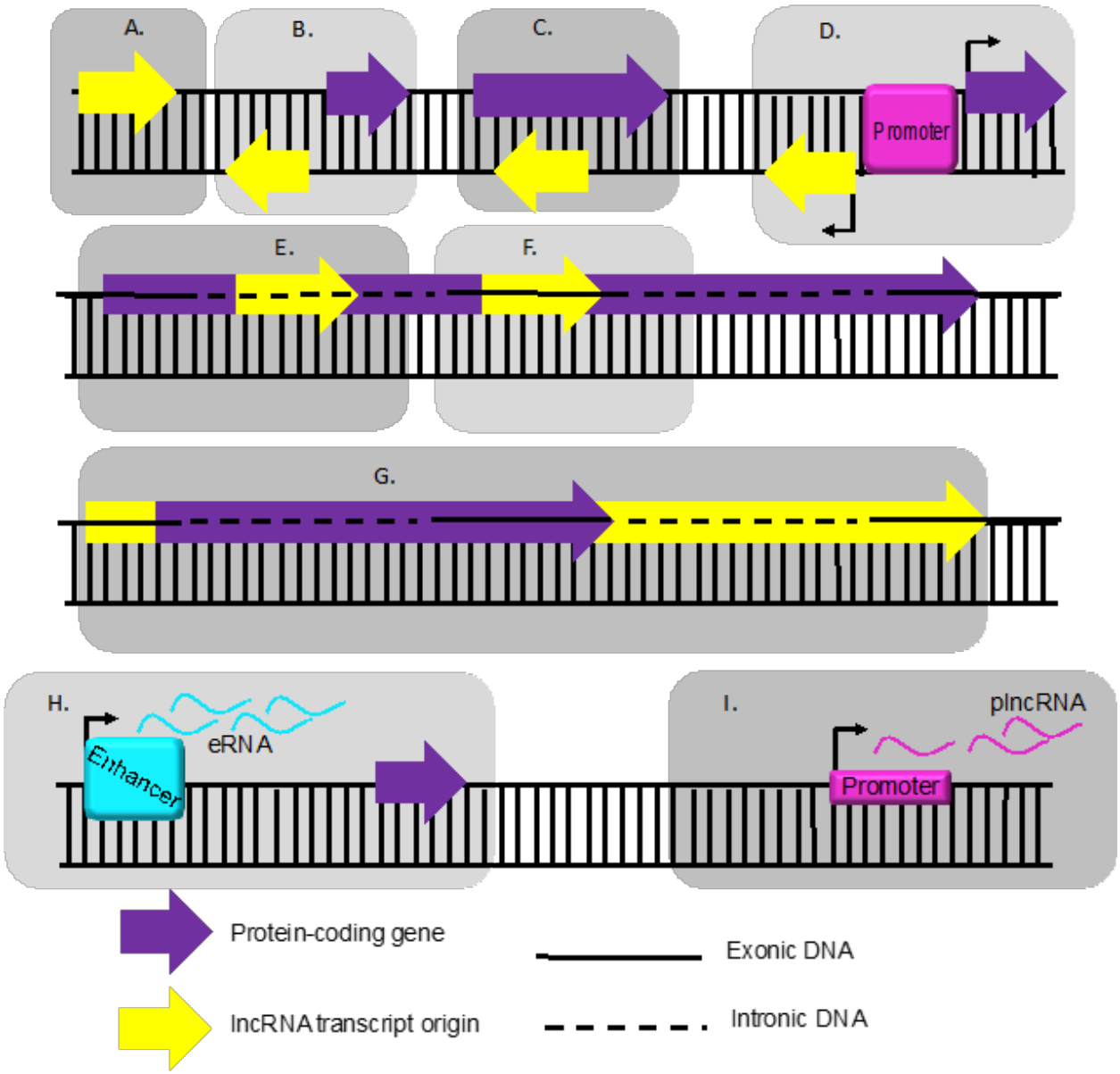


Figure 0.1. LncRNAs stem from diverse genomic loci and be transcribed in various orientations.

A) LncRNA (or lincRNA) gene in the sense direction. **B&C**) lncRNA (or lincRNA) gene in the anti-sense orientation. **D**) Bi-directional promoters can drive simultaneous expression of both sense and anti-sense transcripts. **E&F**) LncRNAs can reside in exons or introns of protein-coding genes. **G**) Protein-coding genes can be part of a larger spanning non-coding RNA gene. **H**) Enhancer or **I**) promoter regions are capable of generating lncRNA transcripts. Coloured arrows pointing to the right indicate sense whereas arrows pointing to the left indicate anti-sense transcription orientation.

LncRNAs share similarities with mRNAs in that they are both transcribed by RNA polymerase II (Pol II) and some lncRNAs can be 5'-capped and polyadenylated (Cabili et al., 2011, Derrien et al., 2012). Some lncRNAs also contain multiple exons and can be spliced, albeit at a lower efficiency than mRNAs (Tilgner et al., 2012, Melé et al., 2017), coincident with weaker lncRNA binding by SR proteins (Krchňáková et al., 2019). The median half-life of lncRNAs is also comparable to mRNAs (3.5 vs 5.1 hrs) (Clark et al., 2012), with *MALAT1* and *NEAT1* lncRNAs having a longer half-life due to the presence of a triple helix motif, instead of a poly-A tail (although that is likely to vary with cell type)(Brown et al., 2012, Wilusz, 2016). Despite their name, lncRNAs tend to be shorter than mRNAs, with a median length of 592 nt compared to 2453 nt, respectively (Derrien et al., 2012). LncRNAs are also less abundantly expressed than protein-coding genes and, cell-type and tissue-specificity restriction, indicative of more specialised expression patterns (Mercer et al., 2008, Derrien et al., 2012, Pauli et al., 2012). This may explain why they were only detected after extensive RNA-Seq of a variety of biological samples, such as 15 different human cell lines (Djebali et al., 2012), zebrafish embryos (Pauli et al., 2012), mouse brain (Ponjavic et al., 2009) and human and mouse blood cells (Roux et al., 2017), to name a few.

The function of a lncRNA is likely to be related to its subcellular localisation (Carlevaro-Fita and Johnson, 2019). LncRNAs can be localised both in the nucleus and the cytoplasm, despite original estimates suggesting a bias in nuclear enrichment (Ulitsky and Bartel, 2013). The precise mechanistic details which would lead to a predictive power of knowing where each lncRNA should localise based on its sequence are currently lacking. Several parameters have been characterised to influence lncRNA localisation though, all of which highlight the importance of lncRNA-protein interactions. Certain lncRNAs can utilise the same nuclear export machinery route via nuclear RNA export factor 1 (NXF1) and the transcription-export (TREX) complex, as mRNAs do (Zuckerman et al., 2020). LncRNAs that associate strongly with splicing factors, are efficiently spliced or contain more than one exon are more likely to be granted cytoplasmic entry, although lncRNAs tend to display weaker interactions with splicing proteins (Krchňáková et al., 2019). Despite that, lncRNAs with less efficient splicing could be exported via splicing-independent factors such as translocated promoter region (TPR) (Lee et al., 2019). Inefficient

splicing and intron retention are two mechanisms via which cytoplasmic export can be evaded, instead promoting nuclear retention (Ntini et al., 2018). The U1 small nuclear ribonucleoprotein (snRNP) complex is also thought to be involved in nuclear retention as knockdown of SNRPA, SNRNP70, SNRPD2 proteins (part of the snRNP complex) induced cytoplasmic localisation of the *MEG3* lncRNA in a reporter system (Azam et al., 2019). U1 recognition sites have been found in both exonic and intronic regions of lncRNAs and the binding of the U1 snRNP to these sites can impede polyadenylation and facilitate chromatin tethering of RNAs (Yin et al., 2020). However, even intronless lncRNAs can end up in the cytoplasm, implicating sequence-specific motifs (and as an extension lncRNA-protein interactions) as instrumental in lncRNA localisation (Krchňáková et al., 2019, Khan et al., 2021). Motifs typically represent binding sites for proteins and there are different motifs for cytoplasmic export and nuclear retention (Carlevaro-Fita and Johnson, 2019). For instance, a C-rich motif was recently found to be bound by the heterogeneous nuclear ribonucleoprotein K (hnRNPK) protein and direct nuclear localisation of both mRNAs and lncRNAs (Lubelsky and Ulitsky, 2018). Finally, there are instances of dynamically localised lncRNA, such as ubiquitin C-terminal hydrolase L1 antisense (*UCHL1-AS*) lncRNA, which shuttles from the nucleus to the cytoplasm upon rapamycin treatment (Carrieri et al., 2012). Overall, lncRNA-protein interactions are important for correct lncRNA processing and localisation, which enable lncRNAs to perform their function.

Although very few lncRNA have been ascribed a function, the functions that have been characterised involve gene expression regulation and play crucial roles in health and homeostasis. For example, lncRNA differentiation antagonizing non-protein coding RNA (*DANCR*) has been shown to play a role in preventing the expression of differentiation genes, thereby maintaining the developmental potential (potency) of human epidermal progenitor keratinocytes (Kretz et al., 2012). In contrast, yin yang lncRNA (*yylncT*) regulates gene expression of human embryonic stem cells (hESCs) to promote differentiation and mesoderm lineage specification (Frank et al., 2019). Promoter of CDKN1A antisense DNA damage activated RNA (*PANDAR*) lncRNA was demonstrated to inhibit apoptosis in response to DNA damage, instead promoting cell cycle arrest in human foetal lung fibroblasts (Hung et al., 2011). Conversely, there are several lncRNAs that contribute to disease

pathogenesis, too. Beta secretase 1 antisense (*BACE1-AS*) lncRNA levels are upregulated in the brains of Alzheimer disease patients, where it stabilises *BACE1* mRNA, promoting β-amyloid production and hence senile plaques, contributing to disease progression (Faghihi et al., 2008, Faghihi et al., 2010). Expression of the maternally expressed 3 (*MEG3*) lncRNA was revealed to be upregulated in hepatocytes of a type 2 diabetes mouse model fed with a high fat diet which was responsible for an aberrant build-up of triglycerides and glucagon, facilitating hyperglycaemia (Zhu et al., 2016). Metastasis associated lung adenocarcinoma transcript 1 (*MALAT1*) was originally reported to be highly expressed in lung cancer and enriched in metastatic tumours (Ji et al., 2003) and has since then been linked to cell proliferation and migration in myeloma and oesophageal squamous cell carcinoma (Amodio et al., 2018). Although disease association for many lncRNAs has been discovered there are only few examples where lncRNAs actually drive disease onset. In fact, despite an array of lncRNAs, which have been found dysregulated in cancer (21,064 results of the 'lncRNA and cancer' term in NCBI; 24th Sept 2021), there are no known lncRNA 'drivers' of cancer (Lanzós et al., 2017, Carlevaro-Fita et al., 2020). Driver genes are those which will be mutated early in tumourigenesis and mutations might be positively selected whereas 'downstream' genes are those which can contribute to cancer progression, via non-genetic induced changes in their expression levels, localisation pattern or partner interactions (Carlevaro-Fita et al., 2020). Silencing or overexpressing driver lncRNA genes should theoretically be enough to cause tumour formation in mouse. Overall, lncRNAs are a recently emerged class of transcripts with a few representative lncRNA having an established importance in both physiological and diseased states. A common feature among them that will be explored in this thesis is how these functions are linked to their interactions with protein partners.

1.2. Evolution and conservation of lncRNAs

An important question the scientific community is currently trying to address is distinguishing which lncRNAs are functional and which are the by-product of pervasive genome transcription (non-functional)(Struhl, 2007). Traditionally, one approach that has been taken to tackle this is comparative genomics, which may not inform on whether a lncRNA is functional but it may hint at its importance.
The level of protein-coding gene conservation is higher than that of lncRNAs and spans larger regions than that seen for lncRNAs (Ulitsky et al., 2011). Whilst conservation across the whole length of lncRNAs has been established to be weak (Ulitsky et al., 2011, Hezroni et al., 2015), smaller sections of a lncRNA have been found to exhibit low rates of sequence evolution. As an example, conservation of lncRNA exons could range from 30% to 4% between mouse and chicken orthologs (Chodroff et al., 2010) whereas according to a different study, conservation between a typical lncRNA across human and mouse could be as high as 20%, which would drop to 5% when comparing human to fish (Hezroni et al., 2015). A study of lncRNAs across human and mouse innate immune responses pointed towards a degree of sequence conservation as low as 1%, but with multiple short regions (12-50 nt) displaying a higher conservation (Roux et al., 2017). This would suggest the bulk of the genic lncRNA sequence is evolving under neutral or weak purifying selection (Ulitsky, 2016).

Whether lncRNA structures show conservation has been a topic of intense scientific debate (Tavares et al., 2019, Rivas and Eddy, 2020). A diversity of plausible and complex structural arrangements can be computed from the same sequence of similar length and free energy. Combined with the fact that even random sequences in solution will adopt some form of structure, the issue lies in distinguishing the ‘real’ structure (with some confidence statistic) and whether structures formed are biologically relevant. Recent software that can analyse multiple RNA sequence alignments and estimate statistical confidence for an evolutionarily conserved RNA secondary structure have failed to identify support for conserved lncRNA structures (Rivas et al., 2017, Rivas et al., 2020). It is important to note that covariation analyses cannot distinguish between lncRNAs lacking evidence of covariation because they are too similar or too variable (Rivas, 2021).

Out of the lncRNAs that have been well-researched, there are few examples where function is linked to structure. One of these examples is the RNA on X 1 (*roX1*) and *rox2* lncRNAs which mediate dosage compensation by X chromosome upregulation in Drosophilids. A low sequence conservation has been described for orthologs of these genes when comparing to other fruit fly subspecies or even outgroups (Park et al., 2007, Quinn et al., 2016). However, secondary structures such as stem loops in *roX1* are functionally important, and present across Drosophilids albeit the number of stem loops present in roX gene orthologs varies (Quinn et al., 2016). roX1 orthologs with fewer stem loops display a weaker ability to occupy the X chromosome as a result to engage in protein interactions with the Male-Specific Lethal complex (MSL) (Ilik et al., 2013). A *roX*-null *D. melanogaster* phenotype is lethal for males, but introduction of a region containing stem loops from wild-type *D. virilis* roX1 modestly rescued males by 18% vs the *roX*-null. Grafting the *D. melanogaster* stem loops onto the *D. virilis* roX1, resulted into a significantly higher rescue of males (43%), compared to the *roX*-null (Quinn et al., 2016). It is also important to remember that the presence of a structure does not guarantee the functionality of a transcript (Rivas, 2021).

Given the rapid sequence divergence of lncRNAs, sequence similarity could be challenging to find across distantly-related species. However, genomic position conservation (synteny) could exist, whereby a specific lncRNA could occupy the same locus across two species, which can be measured in relation to a conserved protein-coding gene. This is a particularly useful aspect of lncRNA conservation to examine given function could be maintained from short sequence motifs across the lncRNA sequence, allowing the rest of the sequence to diverge rapidly (Hezroni et al., 2015), as seen for roX1/roX2 lncRNAs in fruit flies.

Conversely, local stretches of conservation within a locus could also reflect DNA elements acting as promoters/enhancers of distal sites. Notably, some loci harbouring a lncRNA have been shown to elicit a function that's uncoupled from the lncRNA transcript. In one study, terminating *linc1536* transcription in mESCs via the insertion of an early polyadenylation signal in its first intron (to prevent its transcription), did not affect the expression of the neighbouring *Bend4* gene

(Engreitz et al., 2016). This demonstrates that if the lncRNA transcript played a role in regulating gene expression of the BEN domain containing 4 (Bend4) gene, abolishing the lncRNA transcript would have impacted on *Bend4* gene expression. However, deleting the promoter of linc1536 decreased *Bend4* expression roughly by half (Engreitz et al., 2016). This suggested that enhancer or other regulatory elements in the promoter region of the lncRNA could have been evolutionarily co-opted for the activation of downstream protein-coding gene whilst the lncRNA transcript generated from that region could be a by-product. In another study, knockdown of *linc1405* in mouse embryonic stem cells (mESCs) had no effect in the neighbouring Eomesodermin (*Eomes*) gene expression whereas deleting the lncRNA locus resulted in *Eomes* downregulation, decoupling the actual lncRNA transcript as necessary for function (Tuck et al., 2018). Examples like this highlight the act of transcription, instead of the lncRNA transcript, as critical for the upregulation or downregulation of nearby genes, presumably through mechanisms involving transcriptional interference (MacDonald and Mann, 2020) or chromatin remodelling (Navarro et al., 2005, Ard et al., 2017). All in all, due to a rapid sequence divergence of lncRNAs, functional lncRNAs could be conserved in one or a small group of species.

1.3. Mechanisms of action of lncRNAs involving protein interactions and their potential origin via transposable element insertions

A lack of sequence conservation need not always be equivalent to an absence of function (Pang et al., 2006). The very elements that could contribute to lncRNA sequence divergence could also contribute to lncRNAs gaining a function. A central and widespread role in lncRNA gene birth, lncRNA sequence divergence or lncRNA loss of function is the insertion, expansion or elimination of transposable or retroviral elements. Transposable elements (TEs) can become part of a transcript by a process called exonization whereby TEs retrotranspose in intronic regions and can then be alternatively spliced (Sela et al., 2010). In fact, RNA-seq from 28 different human tissues and cell lines, revealed that 83% of lncRNAs that are expressed from a locus not overlapping protein-coding genes contain at least one TE (Kelley and Rinn, 2012). TEs and retroviral elements are known to interact with a wide variety of proteins and can also engage in base-pairing with other RNAs (as reviewed in (Johnson and Guigo, 2014). Therefore, the insertion of such an element into a lncRNA, would enable it to assume these interactions based on TEs harboured in its sequence. Furthermore, TEs can exhibit differences from one clade to another, such as for example being expanded (longer in length of repeat or have more copies of repeat monomer) in one species or shrunk in another due to selective pressure. Such events can influence the evolution of lncRNAs and depending on the sequences inserted, guide future function of a lncRNA. One example where all these concepts have come together to manifest into a functional, TE-driven lncRNA is antisense non-coding RNA in the INK4 locus (*ANRIL*). This lncRNA first emerged in a common ancestor of placental mammals where over time TEs accumulated in ANRIL's exons, especially in simians (He et al., 2013). Rabbit and rodent ANRIL have been found to have fewer exons compared to other placental mammals, and interestingly, the exons that acquired TEs became more conserved (He et al., 2013). Human ANRIL contains TEs from the Alu family. Alu elements on ANRIL are thought to contribute to ANRIL-dependent recruitment of repressive chromatin remodelers and downstream gene expression regulation of target genes. ANRIL was shown to interact with Polycomb Repressive Complex 1 (PRC1), a chromatin remodelling complex, via the chromobox 7 (CBX7) subunit (Yap et al., 2010). Overexpression of *ANRIL* isoforms lacking Alu elements was demonstrated to reverse the upregulation

of TSC22 Domain Family Member 3 (TSC22D3) and the downregulation of Collagen Type III Alpha 1 Chain (COL3A1) genes seen with full-length *ANRIL* which resulted in stunted cell growth and increased apoptosis (Holdt et al., 2013). This highlights the potential of lncRNA-protein interactions (mediated by TEs in this case) playing a key role in *ANRIL*'s function. Hence, depending on the sequences inserted into lncRNAs and the ensuing evolutionary constraints imposed, lncRNAs can evolve to interact with a variety of partners.

RNA binding proteins have well-defined roles in RNA processing/metabolism and post-transcriptional regulation of gene expression (Gerstberger et al., 2014b). Genetic alterations leading to reduced expression levels, a loss of RNA binding capacity or deletion of RNA binding proteins that interact with coding and/or non-coding transcripts can lead to disease (Gerstberger et al., 2014a). There are over 1,400 RNA binding proteins expressed (Mallam et al., 2019), each of which can bind 3-8 nt long sites on RNAs (Mitchell and Parker, 2014). From birth till death, mRNAs are bound by proteins (Singh et al., 2015), and there are no known lncRNA exceptions to date. A meta-analysis of cross-linking immunoprecipitation coupled to sequencing (CLIP-Seq) data for 65 RNA binding proteins revealed that 56.8% of 12,255 human lncRNAs are bound by at least one protein, with 16 lncRNAs demonstrating binding sites for more than 30 RNA binding proteins (Li et al., 2015). In fact, even though lncRNAs have been reported to engage in RNA-DNA and RNA-RNA pairing, the most common interactions observed for well-documented lncRNAs involves RNA-protein interactions. This activity fits well with the pattern of low sequence conservation typically seen across the entire length of a lncRNA transcript albeit with high conservation of local short stretches and an overlap with tandem repeats. This is because protein binding sites could be harboured within tandem repeats, which would comprise short regions of high conservation, if the RNA-protein interaction is important. Such activity combined with distinct features of the lncRNA class of transcripts (**Section 1.1**) has made them highly specialised regulators of gene expression and they can achieve this by a number of mechanisms listed below.

LncRNA-protein interactions typically serve to bridge two or more proteins which cannot interact directly or nucleating the recruitment of several proteins, acting as scaffolds. Namely, lncRNA nuclear enriched abundant transcript 1 (*NEAT1*) (also

known as MENε/β) has been recognised to serve as a docking platform for p54, splicing factor proline and glutamine rich (SFPQ) and polymerase suppressor protein 1 (PSP1) and nucleate the recruitment of more proteins in the formation of the paraspeckle, a membraneless organelle in the nucleus (Sasaki et al., 2009). To date more than 40 proteins have been shown to localise to paraspeckles associating these structures with multiple crucial functions such as nuclear retention of A-to-I edited transcripts and sequestering proteins to modulate a transcriptional response to hypoxia and cell differentiation (Pisani and Baron, 2019).

LncRNAs can guide proteins to a specific cellular location. For instance, lncRNA antisense to SPHK1 (*Khps1*) recruits the p300/CREB-binding protein (p300/CBP) histone acetyltransferase complex to the promoter of the sphingosine kinase 1 (SPHK1) gene in human cells (Postepska-Igielska et al., 2015). Consequently, this allows for chromatin remodeling into a ‘permissive’ environment where gene transcription of the SPHK1 gene can initiate. p300/CBP can bind RNA that is proximally transcribed, in a sequence-independent manner (Bose et al., 2017), highlighting an example of a lncRNA-protein interaction that is not mediated by base-pair recognition. In summary, lncRNAs can employ a wide range of interactions with any component in a cell to fine-tune gene expression regulation across specific temporal and spatial contexts.

1.4. X chromosome inactivation and the effector lncRNA, *XIST*

The X-inactive specific transcript, (*Xist*) (Brockdorff et al., 1992, Brown et al., 1992, Penny et al., 1996), is a lncRNA that orchestrates dosage compensation by X chromosome inactivation (XCI), which involves the transcriptional silencing of one of a female's two X chromosomes (Lyon, 1962). Dosage compensation refers to the act of equalising gene expression across different sets of sex chromosomes between females and males of the same species. In placental mammals, males have one copy of an X chromosome and one copy of a Y chromosome (XY) in contrast to females who bear two copies of the X chromosome (XX). However, not all genes on the X chromosome have a homolog on the much smaller and currently degenerating Y chromosome (Deng et al., 2014). Maintaining gene dosage across sexes is vital since gene dosage imbalance is deleterious for proper mouse embryo development (Takagi and Abe, 1990).

More specifically, once *Xist* is expressed, it recruits a multitude of protein partners to orchestrate the repression of active genes on the X inactive chromosome elect (Jegu et al., 2017). This occurs at different points in development in difference species (**see Section 1.7**). Additionally, *Xist* remodels the inactive X (Xi) by promoting heterochromatin formation, essentially condensing it to a structure termed Barr body or sex chromatin (Lyon, 1962). However, female mouse fibroblasts lacking Xist proliferated as normal and the proportion of cells expressing two X-linked genes, phosphoglycerate kinase 1 (*Pgk1*) and hypoxanthine-guanine phosphoribosyltransferase (*Hprt*), did not differ before and after Cre-loxP deletion of XIST, demonstrating that Xist deletion does not result in transcriptional upregulation of X-linked genes (Csankovszki et al., 1999). *In vivo*, even female mice without the Xist locus that survive to term, all die by weaning age at a median of 18 ± 10.4 days (Yang et al., 2016). RNA fluorescent in situ hybridisation (RNA FISH) in primary tail-tip fibroblasts of mice on the first day after birth demonstrated that some genes (e.g. alpha-thalassemia/mental retardation, X-linked - *Atrx*) did not have altered expression levels when comparing Xist-null mice to control heterozygotes. On the other hand, there were genes whose bi-allelic expression increased from 4.2% of cells in controls to 30% in mutant cells (Yang et al., 2016). Taken together, it is likely that an inherent dosage compensation mechanism exists *in vivo* to achieve partial

dosage compensation in the absence of *Xist*. Such a mechanism is not sufficient for long term survival however, as observed from the eventual death of mice lacking *Xist*.

1.5. *XIST* features and conservation in placental mammals

XIST arose in a common ancestor of eutheria and not in earlier mammals, according to an analysis of 14 mammalian genomes including monotremes, marsupials and placental mammals (Duret et al., 2006). *XIST* is an atypical lncRNA given it is characterised by both a positional conservation (synteny) and being present across eutheria (placental mammals), contrary to most lncRNAs (**Section 1.2**). *XIST* is one of the longest lncRNAs consisting of at least 7 exons, with its (spliced) size varying depending on species, e.g. 19.2 kb in human (Howe et al., 2021), ~17.9 kb in mouse (Howe et al., 2021), 25 kb in pig (Hwang et al., 2013) and ~32-35 kb in cow (based on homology estimates). XIST also exhibits a peculiar genomic structure, with the first and last exons comprised of nearly 10 kbp, with exons in between being up to hundreds of bases (Brockdorff et al., 1992, Brown et al., 1992). More specifically, XIST arose from a combination of several events, starting with the pseudogenisation of ligand of numb-protein x 3 (Lnx3), a protein-coding gene still present in chickens (Duret et al., 2006). Chicken Lnx3 exons 4 and 11 were shown to share homology with human XIST exons 4 and 6 (Duret et al., 2006). Mouse and human XIST exon 4 also share 77% sequence similarity and this region is predicted to form a stem-loop structure (Caparros et al., 2002). Deletion of mouse Xist exon 4 resulted in a reduction in *Xist* expression levels, which was shown not to be a result of altered stability but perhaps a processing or transcription related defect (Caparros et al., 2002). The stem loop formed by mouse Xist exon 4 was confirmed by *in vivo* Dimethylsulfate mutational chemical probing with sequencing (DMS-Seq) and 17/27 bases in the region responsible for the structure were maintained in human, pig and mouse (Fang et al., 2015). Additional homology also exists between XIST's promoter and exons 1 and 2 of Lnx3 (Elisaphenko et al., 2008). More surprisingly for a lncRNA, chicken Lnx3 exons 3 and 5 were up to 60% conserved with introns 3 and 4 of human XIST (rodent Xist exon 5 had a 65% similarity to Lnx3 exon 5)(Elisaphenko et al., 2008).

Perhaps the second event in the evolution of XIST that aided in its (neo)functionalisation involved the gain of transposable elements, as evidenced by the presence of both retroviral and DNA transposons across its exons, originally described in mouse *Xist* (Elisaphenko et al., 2008). These regions which are mostly

repetitive are found primarily in the first and last exons of XIST, display positional and sequence conservation and have been characterised in several placental mammals, including mouse, vole, rat, mole, cow, pig, dog and human (Yen et al., 2007, Hwang et al., 2013, Fang et al., 2015)(see **Figure 1.1**). These tandem repeat regions are designated from A to F. In addition to repeats A and B, two more species-specific repeats have been characterised in porcine *XIST* (Hwang et al., 2013) (**Figure 1.1**). Not all repetitive regions found in human or mouse have been found in cow or pig, two examples being the F and C repeats (Hwang et al., 2013, Yen et al., 2007). Repetitive regions in XIST differ across from placental mammals (**Figure 1.1**), both in length of a repeat monomer (e.g. *XIST* A repeat in mouse is a 26-mer) and in the number of times a monomer repeats in tandem (e.g. 7.5 and 8.5 copies in mouse and human, respectively)(Duszczyk et al., 2011, Brockdorff, 2018). Notably, a complete XIST RNA sequence is missing for cow whereas the pig XIST RNA sequence has not been extensively characterised as seen in human and mouse (Hwang et al., 2013). Functional experiments to test various isoforms that are generated in cow and pig are also lacking. Due to that, the precise size of bovine and porcine XIST escape the research community, and the exact location of repetitive sequences in these species is elusive.

Selective 2' hydroxyl acylation analyzed by primer extension (SHAPE) has been employed to reveal higher-order structures that mouse Xist assumes both *ex vivo* and *in vivo*. *Ex vivo* SHAPE data from mouse *Xist* indicated structural variability in repeat A (high Shannon entropy), with a single hairpin containing an AU-rich loop and GC-rich stem linking repeat 3 and 4 (out of seven)(Smola et al., 2016). A high Shannon entropy predicts several possible structures whereas low entropy suggests structure constraint, usually indicating a single structure. Of the 14 nucleotides participating in the formation of the repeat A hairpin, 11 were conserved with cow and 12 with human, suggesting >78% conservation (Smola et al., 2016). The A repeat could also contact other nearby regions and form a pseudoknot. The structural flexibility of the *Xist* A repeat could permit dynamic interactions with partners and perhaps imply accessibility to proteins that bind to regions with loose or no secondary structure. A different study employing DMS-Seq identified compensatory mutations which maintain the secondary structure of XIST repeat A hairpins (Fang et al., 2015). One example is the shift from a U:A base pair in repeat

5 for rodents to a CG pair in other placental mammals, including human and pig, presumably stabilising the structure. Another example in the A repeat is the G:C pair between repeats 2 and 4 mouse, which is found as G:U in primates and A:U in rabbits. Previous studies have clearly shown that *Xist* silencing capacity is lost upon deletion of a region encompassing the A repeat (Wutz et al., 2002). A linear relationship between number of repeat A monomers present and the degree of gene silencing exhibited by *XIST* was demonstrated by repression of EGFP inserted into an autosome downstream from an artificial *XIST* construct harbouring 2, 3, 4, 5, 6 and 9 repeat A monomers (Minks et al., 2013). The presence of 2 repeat A monomers achieved ~30-20% EGFP repression whereas this increased to 80 with the 9-mer in a HT1080 male fibrosarcoma cell line.

Repeat F is characterised by a 16 nt monomer repeating twice in both human and mouse (Nesterova et al., 2001). A link between this repeat and XIST function has not been established to date, neither has the structure of this repeat been examined (to the best of my knowledge). Repeat B is a 7 nt cytidine-rich monomer present in 29 copies in human and 32 in mouse (Brockdorff, 2018). The structure of this repeat has not been studied due to its highly repetitive GC nature.

Repeat C comprises a single 115 nt monomer in humans, which repeats 14 times in mouse, consistent with an expanded C repeat in mice (Nesterova et al., 2001). This repeat was found to fold into four consecutive hairpins, forming a multibranch loop (Fang et al., 2015). Locked nucleic acid probes (nucleic acid analogs containing a methylene bridge between the 2′ oxygen and the 4′ carbon) targeted at repeat C were previously shown to transiently displace Xist from the Xi in mouse embryonic fibroblasts (MEFs), highlighting its significance in *Xist* localisation and potentially implicating secondary structure as functionally important (Sarma et al., 2010).

The size of the repeat D monomer is 290 nt and repeats 10 times in mice and 26 times in humans (Nesterova et al., 2001). No consensus sequence has been reported in the literature. Repeat D of mouse Xist exhibited structural variability between *ex vivo* and *in vivo* SHAPE data, with a reduced SHAPE reactivity in cells (Smola et al., 2016). Such discrepancies between *ex vivo* and *in vivo* SHAPE data could be owed to factors endogenous in cells, such as protein binding (Smola et al.,

2016). This would suggest that protein binding is not restricted to repeats A, B, C, F and E of *Xist* and could include the largely understudied repeat D.

Repeat E is comprised of U-rich 20-28 nt long monomers (Nesterova et al., 2001). In mouse, monmers are 25 nt long and repeat 50 times whereas in human, monomers are 28 nt long and repeat 25 times. This repeat appeared unstructured from *ex vivo* SHAPE data but *in vivo*, it adopted a strong secondary structure with a combination of hairpin and internal loops (Smola et al., 2016). The region downstream from repeat E (14-17.9 kbp in mouse *Xist*) was also shown to be largely unstructured *ex vivo* but displayed a constrained structure *in vivo.* Deleting this downstream region was shown to decrease the half-life of mouse *Xist* by three-fold (Smola et al., 2016), compared to a full-length counterpart, perhaps indicating that structure alone or in combination with unknown binding partners contribute to evasion from 3’ end decay.

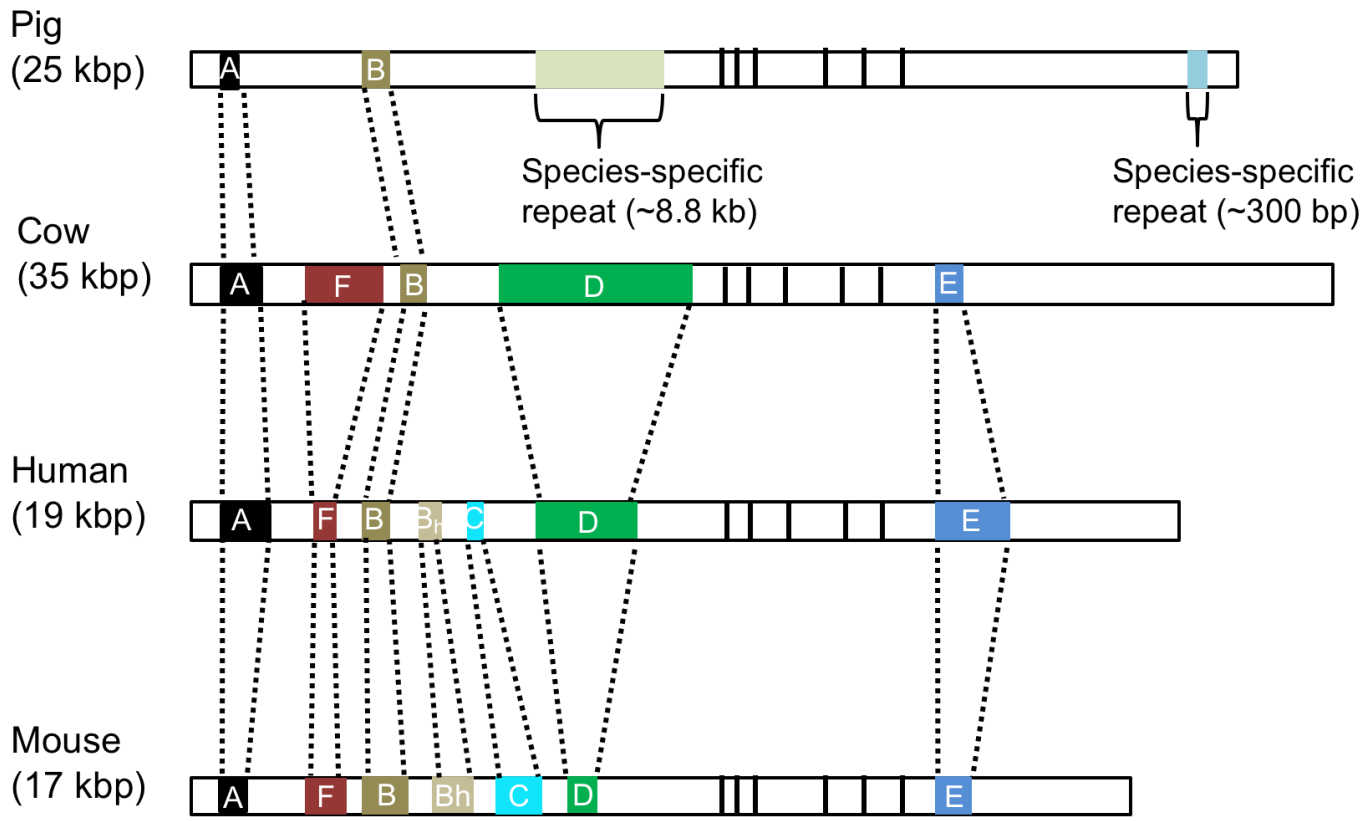


Figure 0.2. XIST transcript organisation shared across placental mammals.

XIST spliced RNA transcripts (longest isoforms) depicted as bars with intervening vertical black lines denoting exon boundaries. Coloured boxes indicate repetitive regions. Note different repeats have emerged in different species and the span of conserved repeats varies across species. Adapted from (Yen et al., 2007). Not to scale.

Despite the fact that *XIST* is 5'-capped, spliced and polyadenylated, it is predominantly localised in the nucleus (Brockdorff et al., 1992, Brown et al., 1992). According to Ensemble v104 (Howe et al., 2021), there are 30 isoforms of human *XIST*, although not much work has been carried out to characterise the localisation or function of alternative human *XIST* transcripts. In contrast, two mouse *Xist* isoforms have been described. Both are transcribed from the same promoter, have been shown to be nuclear and are capable of eliciting XCI, despite the shorter isoform of the two having a smaller part of the last exon and possibly a shorter poly-A tail (Memili et al., 2001, Yue and Ogawa, 2017). Localisation of *XIST* is actively maintained by mechanisms of nuclear retention and inefficient export. Contrary to most mRNAs, *XIST* exhibits a reduced association with the export factor TAP/NXF1 in human embryonic kidney cells (HEK293) (Cohen and Panning, 2007, Viphakone et al., 2019). Nevertheless, XIST can still bind the adaptor protein, ALYREF export factor (Viphakone et al., 2019), which typically recruits TAP/NXF1. Recently, a 57-nt nuclear localisation element was found on the *XIST* lncRNA that occurred at 18 sites, some of which were overlapping repetitive regions C and D (Shukla et al., 2018). Moreover, *XIST* harbours binding sites for several proteins that tether it not only to the Xi but also anchor it to the nuclear matrix and nuclear periphery (such as LBR, CIZ1, MATR3, CELF1, YY1 and hnRNPU; **details in Section 1.6.2**)(Pandya-Jones et al., 2020). Moreover, redundancy is likely to have been acquired for this localisation pattern in a cell-type-dependent manner, since in the absence of heterogeneous nuclear ribonucleoprotein U (hnRNPU), hnRNPU-like 1 (hnRNPUL1) could partly recapitulate the *XIST* localisation pattern seen with hnRNPU in mouse Neuro2A cells (Sakaguchi et al., 2016, Creamer and Lawrence, 2017). All in all, XIST is a lncRNA that evolved after acquiring several repetitive elements, which presumably contributed to its capacity to interact with a diverse set of proteins and elicit its functions. The presence of repetitive elements in XIST from other placental mammals hints at their importance in the function of XIST.

1.6.1. Stages of XCI: Choice

In mice, XCI occurs in two waves, with the first one initiating from the 4-cell stage whereby the paternal X chromosome carries an imprint and is always inactivated preferentially (**Figure 1.2**)(Vallot et al., 2016). Whilst the inactive paternal X is maintained in extra-embryonic tissues, it is reactivated transiently in cells of the inner cell mass in the blastocyst. Within 24 hours, a second wave of random XCI occurs in the epiblast whereby the paternal and maternal X chromosomes have an equal chance of being inactivated (**Figure 1.2**)(Vallot et al., 2016). The choice of which X chromosome to be silenced has been shown to be random in human and rabbit embryos and occur in a single step and there is no imprinted XCI stage (Okamoto et al., 2011). Despite a lack of imprinted XCI in human pre-implantation embryos (Petropoulos et al., 2016), RNA-seq in placenta samples exhibited skewed XCI from either the maternal or paternal X chromosomes, depending on what patch of the placenta was sampled (Phung et al., 2021). This was in contrast to adult tissues, whereby 90% of cells exhibited mosaic XCI. No imprinted XCI has been detected in pigs when assessing *XIST* expression across parthenogenic and *in vitro* fertilized female blastocysts via reverse transcriptase quantitative PCR (RT-qPCR) (Hwang et al., 2015). RNA-sequencing (RNA-seq) performed on mammary gland biopsies in cows demonstrated no preference for the expression of maternal or paternal alleles from genes located outside of pseudo-autosomal regions on the X chromosome (Couldrey et al., 2017), hinting at random XCI. Evidence is conflicting when it comes to XCI in the bovine placenta however. The expression of the X-linked gene monoamine oxidase type A (MAOA) was measured via RT-qPCR in placentas from three female calves born from natural reproduction and revealed maternal-specific expression, implying a skew towards paternal XCI (Xue et al., 2002). RT-qPCR in placentas from female *Bos gaurus/Bos taurus* hybrid foetuses found paternal-specific *XIST* expression, arguing in favour of imprinted XCI (Dindot et al., 2004). A more recent study employing bisulfite sequencing detected a hypomethylated XIST promoter and repeat A region in the trophectoderm cells of blastocysts, supporting a lack of imprinted XCI (Mendonca et al., 2019). Finally, RNA-seq and allele-specific pyrosequencing employed in day 33 mule conceptuses and horse day 33 chorionic girdle (placenta-associated structure) samples displayed random XCI (Wang et al., 2012). Taken together, XCI can occur in different ways across placental mammals

with patterns of XCI across human, cow and pig being more similar between them than the mouse.

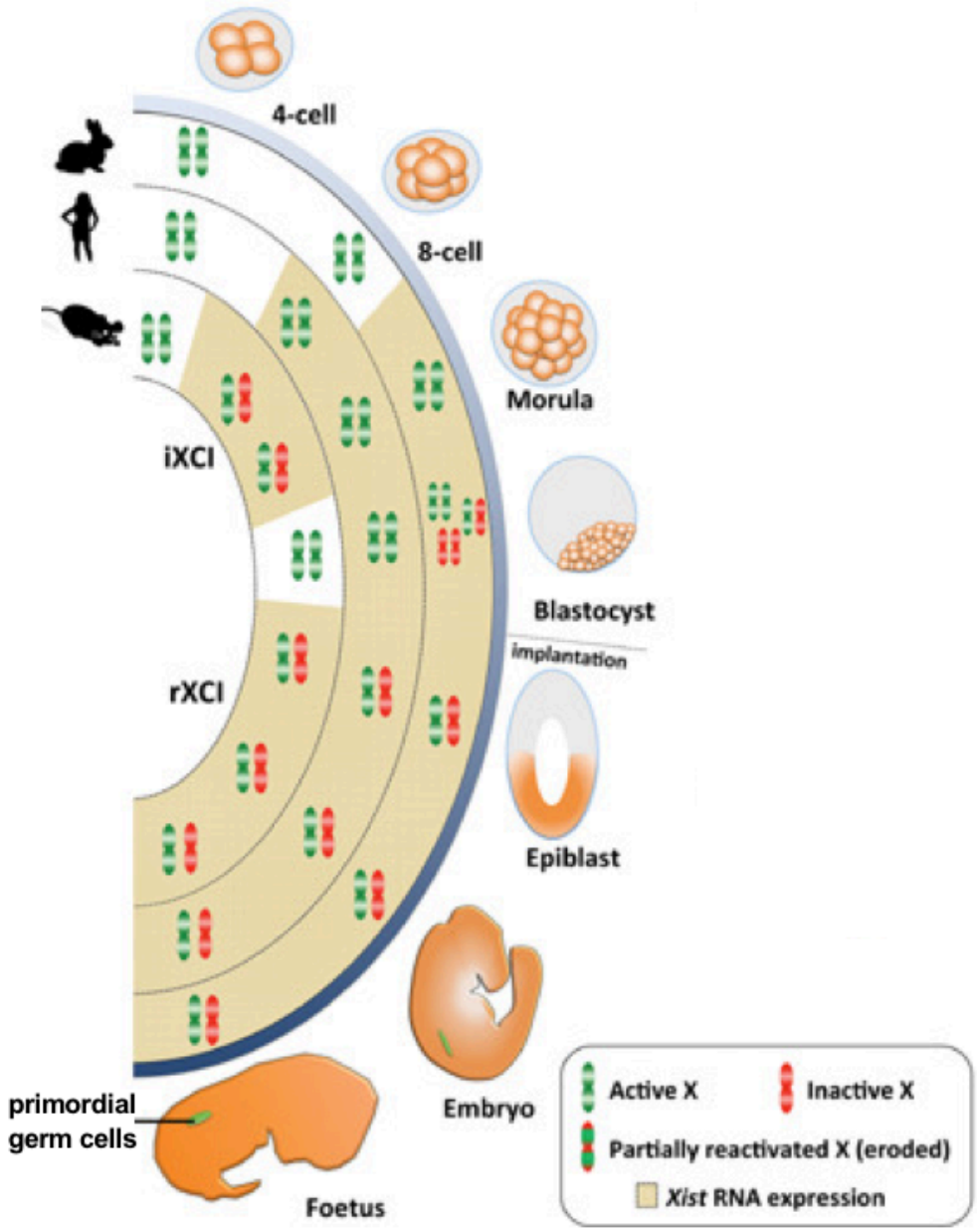


Figure 0.3. XCI onset is tightly coupled to embryonic development.

Embryonic genome activation in the mouse is quickly succeeded with increasing *Xist* RNA levels aimed to preferentially inactivate the paternal X chromosome (imprinted XCI). The inactive paternal X is retained in the trophoectoderm and future extra-embryonic tissues (such as the placenta). In the epiblast cells of the blastocyst the paternal X is reactivated prior to implantation, following decreasing *Xist* levels. Random XCI is completed 24 hours afterwards in embryonic tissues. iXCI, imprinted XCI; rXCI, random XCI; *Xist*, X specific inactive transcript. Adapted from (Vallot et al., 2016).

1.6.2. Stages of XCI: Initiation

The onset of XCI is characterised by the molecular co-operation of *XIST* and its protein partners. As many as 81 protein partners in the mouse aid with *Xist* localisation, spreading along the X chromosome, X-linked gene silencing, X chromosome compaction and finally depositing repressive chromatin marks to ensure long-lasting XCI (Chu et al., 2015b). Some of these protein partners have been elucidated in human and mouse (Table 1.1), however, the protein interactome of *XIST* from other placental mammals such as cow, pig or rabbit, has yet to be characterised.

One of the first proteins described to interact with mouse *Xist* and be critical in X-linked gene silencing was Split ends (Spen). Spen is a ~400 kDa protein containing four RNA recognition motif (RMM) domains, a nuclear receptor interaction domain (RID) and a SPEN paralogue/orthologue C-terminal (SPOC) domain. Spen depletion via RNA interference in female differentiating mESCs could upregulate expression of otherwise repressed X-linked genes, such as glypican 4 (*Gpc4*) and *Atrx* (McHugh et al., 2015), protein kinase G (*Pkg1*)*, methyl CpG binding protein 2 (Mecp2)* and *Rnf12* in mESCs (Chu et al., 2015a) and ubiquitin specific peptidase 9 X-Linked (*Usp9x), ubiquitin like modifier activating enzyme 1 (Uba1), Hprt, HECT, UBA And WWE Domain Containing E3 Ubiquitin Protein Ligase 1 (Huwe1)* in HATX3 ESCs (Monfort et al., 2015). Moreover, coupling RNA-FISH to immunofluorescence in Spen-depleted mESCs, it was shown that *Xist* and Pol II could occupy the same DNA regions, unlike the situation in wild type cells (McHugh et al., 2015), suggesting a role for Spen in excluding Pol II from X-linked genes. Spen can recruit chromatin remodelers such as the nuclear receptor co-repressor 2 (NCOR2/SMRT) complex, the nucleosome remodeling deacetylase (NuRD) complex and histone deacetylase 3 (HDAC3), to synergistically achieve gene silencing, given depletion of NCOR2/SMRT or HDAC3 in mESCs mirrored the inability of Spen-depleted cells to effectively silence the X-linked genes assayed (McHugh et al., 2015, Dossin et al., 2020). Recruitment of these protein complexes is likely mediated by the SPOC domain of Spen (Ariyoshi and Schwabe, 2003) whereas interactions with the *Xist* A-repeat were demonstrated to occur via its RRM domains in electrophoretic mobility assays (EMSA) (Monfort et al., 2015)(**Figure 1.3**). Infrared

crosslinking immunoprecipitation followed by sequencing (irCLIP) of flag-tagged RRM2-4 SPEN domains also highlighted specific interactions with *Xist* repeat A in mESCs (Carter et al., 2020). The importance of RRM2-4 domains of SPEN was also highlighted by overexpressing different truncated Spen cDNA mutants in an mESC line after depletion of endogenous Spen via an auxin-inducible degron (Dossin et al., 2020). This study also found that RRM1 and RID domains of Spen did not contribute to Xist binding as measured by Spen recruitment to the Xi and these domains were not essential for X-linked gene silencing (Dossin et al., 2020). A modified MS2 tag system coupled with luciferase detection using human XIST repeat A also identified SPEN as an interactor in HEK293T (Graindorge et al., 2019).

Table 1.1. XIST protein partners known prior to this study.

XIST protein interactors		
Protein symbol	Mouse	Human
SPEN/ SHARP	(Chu et al., 2015b, McHugh et al., 2015, Moindrot et al., 2015, Monfort et al., 2015)	
RBM15	(Chu et al., 2015b, Moindrot et al., 2015)	(Patil et al., 2016)
WTAP	(Moindrot et al., 2015)	(Patil et al., 2016)
LBR	(McHugh et al., 2015, Chen et al., 2016b)	
HNRNPK	(Chu et al., 2015b, Pintacuda et al., 2017a)	
SAF-A/HNRNPU	(Hasegawa et al., 2010, Chu et al., 2015b, McHugh et al., 2015)	(Minks, 2012)
HNRNPC	(Chu et al., 2015b, McHugh et al., 2015)	
HNRNPM	(Chu et al., 2015b, McHugh et al., 2015)	
RALY	(Chu et al., 2015b, McHugh et al., 2015)	
MYEF2	(Chu et al., 2015b, McHugh et al., 2015)	
CIZ1	(Chu et al., 2015b, Ridings-Figueroa et al., 2017, Sunwoo et al., 2017)	(Sunwoo et al., 2017)
PTBP1	(Chu et al., 2015b, Moindrot et al., 2015, Vuong et al., 2016)	
MATR3	(Chu et al., 2015b, McHugh et al., 2015)	
CELF1	(Chu et al., 2015b)	
YY1	(Jeon and Lee, 2011a)	(Minks, 2012)
RYBP	(Chu et al., 2015b, McHugh et al., 2015)	

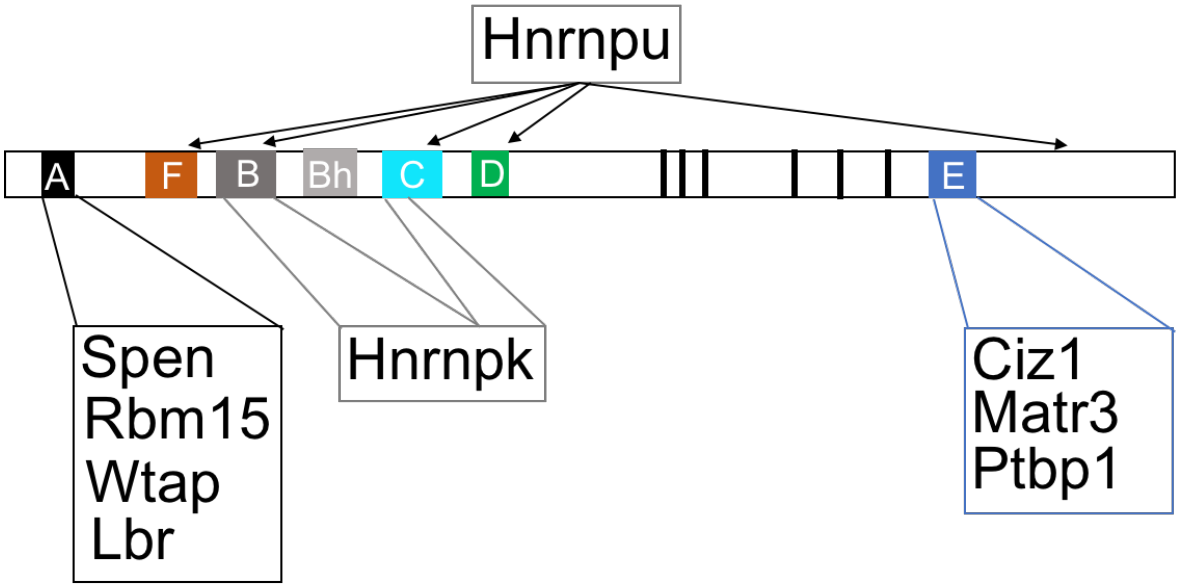


Figure 0.4. Overview of key mouse Xist protein partners and their binding sites.

Schematic depicts the spliced mouse Xist transcript with coloured boxes representing repeat elements characterised. Boxes under repeat elements highlight key protein partners reported to interact with Xist via these repetitive elements. Hnrnpu has not been found to interact with a specific region, rather exhibits a broad binding profile. References for mouse Xist and protein partner interactions in text.

RNA-binding protein 15 (RBM15) and WT1 Associated Protein (WTAP) proteins were also among the first described interactors of human (Patil et al., 2016) and mouse *Xist* (Chu et al., 2015b, McHugh et al., 2015), showing a preference for the A repeat.

RBM15 is comprised of three RRM domains and a SPOC domain. The SPOC domain of Rbm15 and that of the Spen protein in mouse both originated from the spen gene in Drosophila and are 35% similar (Ma et al., 2007). WTAP contains a WTAP/Mum2 domain as does its yeast ortholog, Mum2, which has been shown to interact with methyltransferase-like 3/14 (METTL3/14) (the catalytic complex for m^6A methylation)(Ping et al., 2014). Besides the Mum2 domain, no other RNA-binding or catalytic domains have been mapped onto WTAP. RBM15 has been shown to be the driver of 78 m^6A methylation modifications upon binding on the human *XIST* lncRNA and recruiting the METTL3/14 complex via associations with WTAP in HEK293T cells (Patil et al., 2016). Knock-down of RBM15 caused a drop in the levels of *Xist* which was m^6A methylated, as detected by chromatin immunoprecipitation with an m^6A-specific antibody, establishing a link between RBM15 and WTAP-mediated m^6A methylation (Patil et al., 2016). Whereas knockdown of RBM15 in human has been shown to reduce X-linked gene silencing (Patil et al., 2016), the effect of mouse Wtap in X-linked gene gene silencing appears to be weak (not significantly affected upon Wtap depletion) in the context of XCI (Nesterova et al., 2019). This could be consistent with a role of Wtap bridging Mettl3 to mouse *Xist*, instead of a central one affecting gene silencing or m6A installation (Nesterova et al., 2019).

Lamin B receptor (Lbr) was identified as a bona fide interacting partner of mouse *Xist*, mediating proper localisation of the Xi in the cell's nuclear lamina (Chen et al., 2016b). Lbr interacts with *Xist* specifically via its repeat A region (McHugh et al., 2015, Chen et al., 2016b, Lu et al., 2020b). Depletion of Lbr in female mESCs resulted in loss of *Xist* co-localisation with active X-linked genes (Chen et al., 2016b). A mutant of the Lbr protein lacking its arginine-serine (RS) domain displayed 97% decreased association with *Xist* in mESCs and did not contribute to X-linked gene silencing (Chen et al., 2016b). In contrast, a mutant of the Lbr protein lacking most of its transmembrane domains did not exhibit any of these defects, highlighting the role of the RS motif in mediating the *Xist*-Lbr interaction (Chen et al., 2016b).

Furthermore, by measuring the distance between *Xist* and Lamin B1 (a marker of the nuclear lamina) in LBR knockdown and wild-type mESCs, the localisation of the Xi to the nuclear lamina was shown to be dependent on the association of *Xist* with LBR (Chen et al., 2016b). The effect of Lbr on gene silencing seems to vary, similar to Ciz1, with some studies finding a prominent effect on upregulating X-linked genes upon Lbr depletion in differentiating female mESCs (McHugh et al., 2015, Chen et al., 2016b), and others reporting a negligible effect, if any at al,l in mESCs (Nesterova et al., 2019).

Heterogeneous nuclear ribonucleoprotein K (hnRNPK) has been shown to bind repeat B and C of mouse *Xist* (Almeida et al., 2017, Pintacuda et al., 2017a, Bousard et al., 2019) and human *XIST* (Lu et al., 2020). The hnRNPK protein encompasses three K homology (KH) domains and an arginine–glycine–glycine repeat (RG/RGG) domain. Both KH1 and KH2 domains as well as the RG/RGG domain are required for a high affinity interaction of an RNA with mouse hnRNPK, whereas the KH3 domain is dispensable for interactions with RNAs longer than 10 nt (Nakamoto et al., 2020). Mouse hnRNPK binding is selective for two neighbouring cytosine repeats (each with three or more C nucleotides), with highest affinity interactions when these stretches of cytosine are within internal loops, demonstrating that hnRNPK binding can be affected by the presence of structure (Nakamoto et al., 2020). Following mESC differentiation, hnRNPK can recruit the PRC1 and PRC2 complexes, via its RG/RGG domain, to deposit repressive chromatin modifications (H2AK119ub1 and H3K27me2/3) on active X-linked genes (Chu et al., 2015b, Pintacuda et al., 2017a), a critical step in ensuring long term silencing (Csankovszki et al., 1999).

Hnrnpu was first shown to bind *Xist* in mouse Neuro2A cells and play a role in XCI by tethering *Xist* on the X chromosome (Hasegawa et al., 2010). The hnRNPU protein possesses a SAF-A/B, Acinus and PIAS (SAP) domain, a SPla and the RYanodine Receptor (SPRY) domain and an RGG (NTPase) domain. More specifically, the working model put forward is that mouse Hnrnpu is able to non-competitively bind mouse *Xist* RNA (via its RGG domain) and chromosomal DNA (presumably via the SAP domain), bridging the two together (Hasegawa et al., 2010). Loss of either of those two domains decreased the accumulation of *Xist* on the Xi (Hasegawa et al., 2010). Ultraviolet radiation (UV) coupled to RNA

immunoprecipitation (UV-RIP) analyses in HEK293T cells demonstrated an interaction between hnRNPU and *XIST* at exons 1 and 7 (outside of repeat E) (Yamada et al., 2015). More recent studies have demarcated that hnRNPU exhibits broad binding across the whole length of human *XIST*, with multiple enhanced CLIP (eCLIP) peaks spanning exon 1 (repeats B, C, D and F) and exon 6 (outside of repeat E) (Lu et al., 2020).

A few studies have established a role for mouse Ciz1 as key for proper *Xist* anchoring to the Xi, an essential aspect of XCI which facilitates gene silencing (Ridings-Figueroa et al., 2017, Sunwoo et al., 2017, Stewart et al., 2019). *Xist* repeat E interacts with Ciz1 (Ridings-Figueroa et al., 2017), and is responsible for Ciz1 localisation to the Xi (Sunwoo et al., 2017). Transient transfections of the C- and N-terminal domains of Ciz1 showed that only the C-terminus construct could co-localise with *Xist* (Ridings-Figueroa et al., 2017). The C-terminus of Ciz1 contains C2H2-type zinc fingers and Matrin3-type RNA-binding zinc finger domains. Loss of Ciz1 function in female mouse embryonic fibroblasts lead to *Xist* delocalization from the Xi and dispersal in the nucleoplasm with one study reporting only 28 genes aberrantly upregulated (Stewart et al., 2019) and another describing a marked loss of the repressive chromatin mark tri-methylation of lysine 27 on histone H3 protein (H3K27me3) on the Xi (Sunwoo et al., 2017).

Matrin 3 (Matr3) has been described as a mouse *Xist* interactor in both pulldown studies (Chu et al., 2015b) and genetic screens (Moindrot et al., 2015). MATR3 was also found to be interacting with human *XIST* repeat A (Graindorge et al., 2019). Nevertheless, in a more recent study, eCLIP of Matr3 and polypyrimidine tract binding protein 1 (Ptbp1) revealed binding to mouse *Xist* repeat E (Pandya-Jones et al., 2020). Ptbp1 has been shown to bind mouse *Xist* previously by pulldown studies (Chu et al., 2015a, McHugh et al., 2015), genetic screens (Moindrot et al., 2015) and eCLIP had found binding peaks on repeat E (Vuong et al., 2016). Ptbp1 binding to repeat E was confirmed by EMSA (Pandya-Jones et al., 2020). Depletion of Matr3 or Ptbp1 in mESCs caused a loss of *Xist* localisation from the Xi as well as a reduction in repressive chromatin marks from the Xi (H3K27me3), albeit with negligible changes in *Xist* levels (Pandya-Jones et al., 2020). Chromatin immunoprecipitation coupled to sequencing (ChIP-Seq) also revealed Ptbp1 binding to genomic *Xist*

repeat E, which could suggest co-transcriptional loading of Ptbp1(Pandya-Jones et al., 2020). This would be consistent with a role of Ptbp1 in mouse *Xist* splicing at the onset of XCI (Stork et al., 2019). MATR3 has a nuclear export signal (NES) and two matrin/zinc-finger type domains flanking two RRM domains and a nuclear localisation signal (NLS). Ptbp1 is comprised of a NES, NLS and four RRM domains that recognise CU motifs (Coelho et al., 2015). Ptbp1 and Matr3 are known to also interact directly with one another (Coelho et al., 2015). Introducing point mutations in Matr3 to disrupt binding to Ptbp1, could not rescue *Xist* localisation to the Xi or repress X-linked genes, defects noticed after deleting *Xist* repeat E, despite partially recovering H3K27me3 modifications on the Xi (Pandya-Jones et al., 2020). Reciprocally, the same effects were noted with point mutations disrupting the Ptbp1 interaction with Matr3, underscoring the synergy between the two proteins for efficient XCI.

In summary, *XIST* requires the recruitment of a multitude of proteins to establish XCI across placental mammals, highlighting the importance of lncRNA-protein interactions in this process. Given XCI and *XIST* are present in all placental mammals, XIST is an important model for understanding the co-evolution of functional lncRNA-protein interactions when the lncRNA sequence is changing rapidly compared to protein interactors.

1.6.3. Stages of XCI: Maintenance

The XCI is a dynamic process that employs many layers of regulation across space and time to ensure robust and faithful dosage compensation consistently. Indeed, there is a specific developmental time window when *XIST* initiates XCI, after which *XIST* expression becomes dispensable for maintenance of X-linked gene repression (**Section 1.4**).

In a similar manner, the Xi adopts a distinct localisation within the nucleus, which is separate to the one seen for the active X chromosome (Comings, 1968). The position where the Xi is observed depends on the stage of the cell cycle that cells are found at, with the majority of cells displaying a perinuclear (near the nuclear envelope / nuclear lamina) Xi in interphase (Dyer et al., 1989). Some of the aforementioned proteins (**Section 1.6.2**) are mediators of *XIST*-induced Xi tethering to the nuclear lamina, given *XIST* has been observed both on the Xi and the nuclear lamina (Jonkers et al., 2008). Nuclear lamina localisation is thought to be concordant with gene expression repression (Shevelyov et al., 2009). If this was the case, depleting SPEN (a key protein previously described to silence active X-linked genes) would not be enough to abolish X-linked gene activity, since localisation to the nuclear lamina mediated by other proteins would contribute to gene silencing. However, X-linked gene transcription was still observed in the nuclear lamina following SPEN depletion and *XIST* could still access those genes (Chen et al., 2016b). On the other hand, upon depletion of LBR, one of the proteins involved in *XIST* tethering to the nuclear lamina, the ability of *XIST* to access X-linked genes was lost, implicating a link between nuclear lamina tethering and *XIST* spreading (Chen et al., 2016b).

Towards the late S phase of the cell cycle, most cells shift their Xi to a perinucleolar (adjacent to the nucleolus) position, presumably to spatially separate heterochromatic regions that are replicated last (Zhang et al., 2007). It is likely this compartmentalisation ensures the carry-over of the epigenetically inactive state of the X to the next generation, ensuring long-lasting XCI maintenance. It is also at this stage when *XIST* expression does not co-localise with the Xi and instead displays a diffuse signal in the nucleus (Jonkers et al., 2008).

Given the consensus that *XIST* expression is not necessary for the maintenance stage of XCI, most studies have focused on characterising the protein interactome of *XIST* at the onset of XCI in stem cells that can be induced to differentiate. Therefore, it is likely that protein interactions identified at the onset of XCI, might not persist throughout the maintenance stage of XCI. In fact, since 2017 (the beginning of this project), there were no studies to characterise *XIST* protein partners in differentiated cells. Two recent studies have examined the identity of proteins bound to XIST from HEK293 kidney (Graindorge et al., 2019), K562 myeloid and GM12878 B-cell (Yu et al., 2021), which are all differentiated cells. In fact the latter study showed an overlap of ~57% between protein partners of *XIST* at the onset vs the maintenance stage of XCI (Yu et al., 2021). Why would a dispensable lncRNA for XCI maintenance be dynamically exchanging protein partners? More importantly, how do the functions of those proteins differ and what implications does that hold for XCI?

1.7. Placental mammal differences in development and reproductive morphologies

In the previous sections (**Sections 1.4-1.6**), XIST was shown to be present and orchestrate XCI across placental mammals, often employing similar effector proteins between human and mouse. Clear differences pertaining to *XIST* features as well as details of the XCI process were also laid out. Despite that, most of the knowledge about *XIST*'s role in XCI and *XIST*'s protein partners is derived from studies of the mouse and mouse and human stem cells. However, the mouse is not representative of all mammals.

In fact, differences between placental mammals are evident from their gestation period, even prior to embryonic development. During very early embryonic development, the timing of embryonic genome activation (EGA) varies across different placental mammals. In mice EGA begins at the 2-cell stage, in humans and pigs at the 4- to 8-cell stage whereas it's from the 8- to 16-cell stage in cows and sheep (Telford et al., 1990). Variation in EGA could affect the timing of lineage specification commitment. For example, cell fate commitment of the trophectoderm lineage (crucial for blastocyst formation) occurs much faster in mice compared to pigs (Wei et al., 2018) and cattle (Berg et al., 2011, Wei et al., 2017). *Xist* expression

is also seen at different developmental windows in different placental mammal species, not always equating the onset of XCI. *Xist* expression can be detected from the late 2-cell to 4-cell stage in mice (Deuve et al., 2015), 8-cell stage (van den Berg et al., 2009) or later in humans (Petropoulos et al., 2016), 8-cell stage or earlier in cows (De La Fuente et al., 1999) and 16-cell stage in pigs (Park et al., 2011). Due to the imprinted nature of the first XCI wave in mice, expression of *Xist* will trigger XCI whereas as seen for human (Okamoto et al., 2011, Petropoulos et al., 2016, Sahakyan et al., 2017) and cow (Yu et al., 2020), *XIST* upregulation on an X chromosome is not equivalent to gene silencing early on in systems where XCI occurs randomly in a single step. Furthermore, implantation timing varies across species with the mouse and human embryo implanting in the endometrium five and seven days after fertilisation, respectively whereas pigs and cattle could take up to two to three weeks, respectively (Berg et al., 2010, Bou et al., 2017). Humans more closely resemble cows during early development, since factors for cell lineage specification and cell signalling (e.g. octamer-binding protein 4 - OCT4) are required earlier in embryo development compared to the mouse (Fogarty et al., 2017, Daigneault et al., 2018).

As the name placental mammals suggest, all species in this clade have a placenta. However, there are differences in the structure the placenta adopts during pregnancy with respect to its shape and its interface with the uterine wall (**Figure 1.4**). The placental morphology of cows and sheep (‘cotyledonary’) is different to that of pig (‘diffuse’) or human and mouse (‘discoid’). Cotyledonary refers to the presence of ‘cotyledons’ or spot-like structures which can connect to the uterine wall for resource exchange. Diffuse means that there is an extended surface area allowing for a greater nutrient supply whereas discoid, resembling a disc, has a small surface area (Mess, 2014, Gundling and Wildman, 2015, Roberts et al., 2016). Thus, although part of the same clade, placental mammals have diverged sufficiently to display distinct developmental and XCI-related characteristics. Hence, the aforementioned differences in placental mammals could have provided grounds for *XIST* to evolve different protein partners, in a species-specific way, so that XCI can be achieved in the context of divergent placental mammal developmental. Knowledge about placental mammal divergence times as well as their evolutionary relationships would thus be conducive to an assessment of interacting partner conservation.

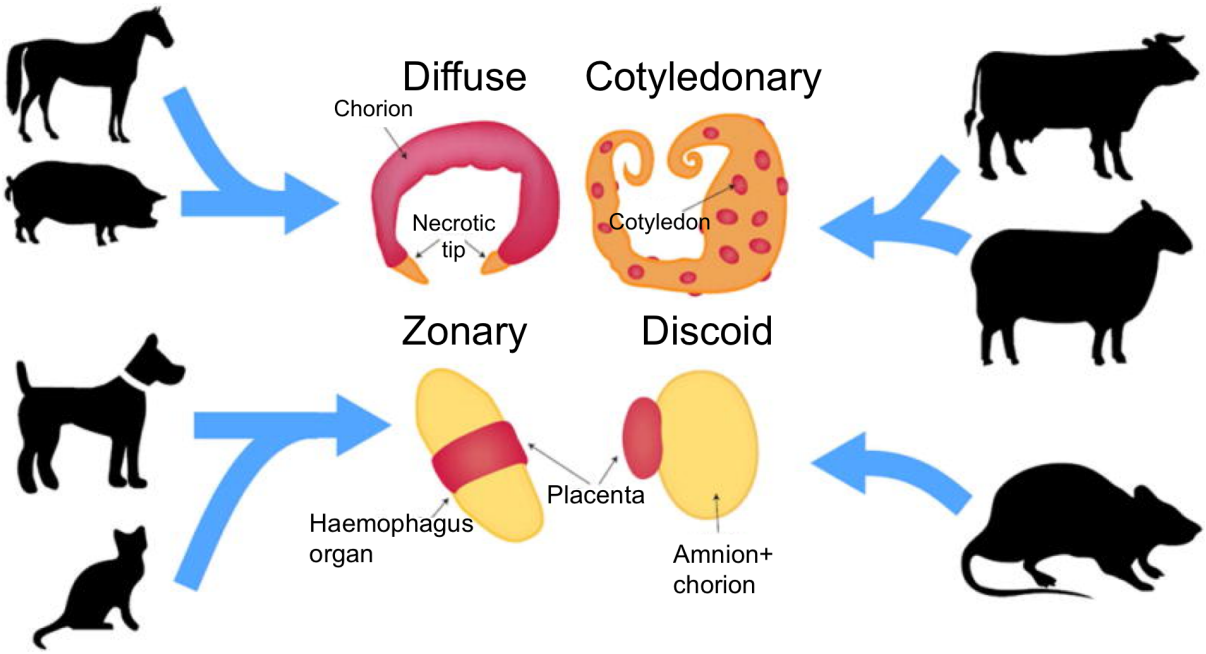


Figure 0.5. Diversity of placental morphology across eutheria.

The placenta of horses and pigs develops to a flat and thin surface for the embryo to be as close as possible, ensuring an efficient exchange of small molecules between the foetal and maternal circulation via micro-cotyledons (top left). Micro-cotyledons are groups of highly vascularized projections which extend into cavities in the endometrium and serve as the interface between the foetus and the parent. These structures probably developed in size in the placenta of cows and sheep (top right). Placentas of dogs and cats surround the embryo in a band of tissue, rich in vasculature (bottom left). In contrast, the placenta adopts a discoid shape mice and humans (bottom right). Adopted from (Roberts et al., 2016).

1.8. Project aims

The *XIST* lncRNA is a master regulator of the XCI process which is indispensable for embryonic development. *XIST* is expressed across all placental mammals and all placental mammals employ XCI to achieve dosage compensation. However, most of our knowledge about XCI mechanisms derives from studies of mouse *Xist*. Recent studies into human *XIST* have revealed its protein interactome can be cell-type and XCI-stage dependent. The mechanisms of how *XIST* functions in placental mammals other than mouse and human is unclear. XCI in placental mammals only occurs in females and the expression of *XIST* is higher in female reproductive tissues. Moreover, prominent differences between placental mammals relate to their divergent reproductive morphologies and early pregnancy events. Despite those, *XIST* is modestly conserved across mouse, human, cow and pig, which is in contrast to the lncRNA class of molecules, generally displaying low to no sequence conservation. Therefore, it's possible that a eutherian-wide *XIST* protein partner interactome has been maintained across all placental mammals given the common endpoint of XCI. Given the lack of a high *XIST* sequence conservation, this would either require short regions of conservation or compensatory mutations on its protein partners for an interaction to be maintained. LncRNA-protein partner co-evolution has not been investigated previously given the trend of lncRNAs lacking sequence conservation. Alternatively, aforementioned differences could have provided grounds for XIST protein partners to diverge accordingly in each placental mammal whilst still achieving XCI. To date, no protein partners of XIST from placental mammals outside of mouse or human have been elucidated and none have been investigated in the context of reproductive tissues/cells.

This project aims to dissect the *XIST* lncRNA protein partners across placental mammals with different early pregnancy events by addressing 3 main objectives:

1) To examine whether already characterized mouse *Xist* protein partners are conserved and present in uterine-derived tissues/cells from human, cow and pig. These species represent mammals with different implantation strategies and early pregnancy events. To this end, Clustal-ω will be used for conservation score estimates and RT-qPCR together with western blotting will

assay for a concomitant expression of *XIST* and its putative protein partners in uterine-derived tissues/cells.

2) To determine whether putative protein partners of *XIST* interact with *XIST* from human and cow in uterine-derived tissues/cells. This will be addressed via pulldown assays coupled to western blotting and mass spectrometry. Based on results from data analysis of the previous objective, candidate protein partners will be selected and their interaction with *XIST* validated and their roles in XCI and reproduction further investigated.

3) To assess for a potential functional shift in a subset of XIST protein partners across mouse, human, cow and pig lineages as a result of positive selection. Selective pressure variation analyses will be performed via codon-based models of evolution using the PAML suite of tools, available from the VESPA pipeline.

2. Chapter 2: Characterisation of *XIST* and protein partner conservation and expression across human, mouse, cow and pig

2.1 Introduction

Maintaining gene dosage across sexes is vital since an imbalance would be deleterious for embryo developmental competency and failure to do so would result in embryonic lethality. All placental mammals employ *XIST* for XCI to achieve dosage compensation. *XIST* is a lncRNA, the length of which varies depending on placental mammal species, from 17 to 35 kb lncRNA (human *XIST*: 19.2 kb; cow *XIST*: 35 kb). Across *XIST*'s first and last exons, repetitive regions discovered serve as its 'functional domains' (Brockdorff, 2018). Namely, repeats A, B, C and E contain protein binding sites for as many as 81 proteins in mouse (Chu et al., 2015b) and can also fold into complex secondary structures (Pintacuda et al., 2017b). Once *XIST* is expressed in a cell, it recruits a multitude of protein partners to orchestrate the repression of active genes on the X chromosome to achieve dosage compensation (Jegu et al., 2017).

Some of those partners have been elucidated in mouse and human (details in **Section 1.7.4**). Briefly, Spen was shown to interact with the A-repeat of *Xist* in mouse embryonic stem cells (Monfort et al., 2015, Chen et al., 2016b, Lu et al., 2016, Lu et al., 2020b) and in human embryonic kidney cells (HEK293T) (Graindorge et al., 2019). SPEN mediates the silencing of active genes on an X chromosome via associating with the NCOR2/SMRT complex, the NuRD complex and HDAC3, remodelling chromatin and excluding RNAPII from active genes of the Xi elect chromosome in mice (Dossin et al., 2020). Rbm15 and Wtap proteins were found to interact with the A-repeat of mouse *Xist* (Chu et al., 2015b, McHugh et al., 2015) and play a role in X-linked gene silencing. In particular, RBM15 was also shown to recruit the m6A methylation machinery (METTL3/4) to *XIST* via associations with WTAP in HEK293T cells (Patil et al., 2016). hnRNPK is able to bind repeat B and C of mouse *Xist* (Almeida et al., 2017, Pintacuda et al., 2017a, Bousard et al., 2019) and interact with human *XIST* (Lu et al., 2020). hnRNPK links PRC1/PRC2-mediated repressive chromatin modification (H2AK119ub1 and H3K27me2/3) installation on active X-linked genes (Chu et al., 2015b, Pintacuda et al., 2017a). Hnrnpu binds *Xist* in mouse Neuro2A (Hasegawa et al., 2010) and HEK293T cells (Lu et al., 2020a) and

plays a role in XCI by localising *Xist* on the X chromosome by bridging mouse *Xist* RNA and chromosomal DNA using different domains. Ciz1 binds *Xist* repeat E and is involved in anchoring *Xist* to the nuclear matrix (nuclear periphery) and to the Xi, facilitating gene silencing (Ridings-Figueroa et al., 2017, Sunwoo et al., 2017, Stewart et al., 2019). Lbr was identified as a bona fide interacting partner of mouse *Xist*, mediating proper. Lbr interacts with *Xist* specifically via its repeat A region (McHugh et al., 2015, Chen et al., 2016b, Lu et al., 2020b) and tethers *Xist* in the cell's nuclear lamina (Chen et al., 2016b). Altogether, effector proteins are necessary for Xist to commence and establish XCI, although a detailed list of those is currently lacking outside of human and mouse.

Contrary to most members of the lncRNA class which exhibit rapid evolution and thus lack conservation (Pang et al., 2006), *XIST* is an atypical lncRNA considering its moderate conservation among the eutherian clade (Yen et al., 2007). Yet our current knowledge regarding the role of *XIST* and its interaction partners in XCI comes from either the mouse model or human and mouse stem cells which are unlikely to be representative of the entire eutherian clade. For instance, certain XCI milestones such as temporal *XIST* expression and X-linked gene silencing are achieved at different time-points in distinct placental mammal species (Okamoto et al., 2011, Vallot et al., 2016). The timing of implantation differs among placental mammals with the mouse and human embryo implanting in the uterus five and seven days after fertilisation, respectively whereas pigs and cattle could take up to two to three weeks, respectively (Berg et al., 2010, Bou et al., 2017). There is also substantial variation in development and reproductive morphologies across the placental mammal clade. Furthermore, the nature of XCI varies (imprinted vs random) along with the extent of XCI escape across placental mammals. During mouse embryo development, there is a wave of imprinted XCI where the paternal X is preferentially briefly inactivated from the 4-cell stage to the blastocyst until the paternal X is reactivated transiently in preparation for the second wave of random XCI (reviewed in (Vallot et al., 2016). However, in humans (Petropoulos et al., 2016) and pigs (Zou et al., 2019), a single wave of random XCI has been described so far whereas the picture is unclear in cow (Xue et al., 2002, Dindot et al., 2004, Chen et al., 2016c, Couldrey et al., 2017).

Therefore, under the prism that eutherian mammals have diverged in terms of their early development, XCI timing and reproductive morphologies, it is possible that their mechanisms of XCI have diverged as well, perhaps co-opting new protein players. The aim of this chapter is to investigate whether the *XIST* interactome identified in the mouse is shared across placental mammals or if species-specific *XIST* protein partners have evolved. The first step to identify protein partners of *XIST* in placental mammals was to examine whether the previously discovered proteins found to interact with mouse *Xist* are shared. This chapter will focus on work done to determine the conservation of *XIST* putative protein partners and assessing whether they are present in reproductive tissues from placental mammals with different implantation strategies.

2.2 Materials and Methods

2.2.1 Sequence conservation analysis of *XIST* and putative protein partners

The RNA sequences of *XIST* and its mouse protein partners were sourced from Ensemble v90. More specifically, the following RNA sequences were used for *XIST* (GeneBank identifiers for each species in brackets; accessed 19 April 2018): human (NR_001564.2), mouse variant 1 (NR_001463.3), pig (KC753465.1) and predicted bovine variant 1 (XR_001495594.1). The predicted NCBI Nucleotide entry was created from RNA seq data with the NCBI Eukaryotic Genome Annotation Pipeline. The following amino acid sequences were retrieved from UniProt (https://www.uniprot.org/) for mouse, human, pig and cow, respectively (Uniprot accession numbers for each species in brackets; accessed 19 April 2018): SPEN (Q96T58, Q62504, A0A287BPC2, F1MRK2), WTAP (Q15007, Q9ER69, F1SB61, F1MN80), RBM15 (Q96T37, Q6PHZ5, F1S619, E1BFX6) and CIZ1 (Q9ULV3, Q8VEH2, F1RRW9, F1MZB8).

To compare the sequence identities of *XIST* RNA sequences across the four species and the amino acid sequences coding for protein partners of mouse *Xist*, the multiple sequence alignment software Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) was used with the default settings to derive percent identity matrix scores. The default transition matrix is Gonnet, gap opening penalty is 6 bits, gap extension is 1 bit.

To consider whether % similarity scores observed for amino acid sequences of putative XIST protein partners spanned a protein's functional domains, the InterPro website was used (https://www.ebi.ac.uk/interpro/) (Blum et al., 2021) to predict the location and identity of domains using the human homolog of these proteins (given its superior annotation). Where available, predictions were confirmed by browsing for the domains described for the human homolog in Uniprot as well. The start/end location of domains along with their identity were overlaid on the multiple sequence alignment performed.

2.2.2 Cell culture and tissue handling

Unless stated otherwise, all consumables were sourced from Sigma-Aldrich (UK).

ISHIKAWA cells were a kind gift from Prof. Aplin (Manchester University) and Dr Karen Forbes (University of Leeds)(European Collection of Authenticated Cell Cultures: Cat. No. 99040201). These were grown at 37^{o}C (5% CO_2) in a 50-50 mix of two different media DMEM and F12 B-Nut mix (ThermoFisher) supplemented with 10% (v/v) heat-inactivated Fetal Bovine Serum and 5% (v/v) Penicillin-Streptomycin-Glutamine. To passage, cells were washed with 1x PBS (-$CaCl_2$, -$MgCl_2$) and detached with 0.025% trypsin in PBS for 5 minutes. Cell suspensions were centrifuged at 1000 xg for 5 minutes and pellets resuspended in complete media before splitting at 1:4 every two days.

Endometrial tissue was collected from the ipsilateral uterine horn at the late luteal stage of the estrous cycle from female *Bos taurus* cows and *Sus scrofa* pigs (local slaughterhouse). Mouse uteri (n=3) were collected from unknown age female NFAT-Luc ApoE *Mus musculus* following a schedule one procedure. All tissues were snap-frozen in LN_2 and stored at -80^{o}C.

2.2.3 Generation of cell and tissue lysates for RT-qPCR expression profiling

For cytoplasmic/nuclear fractionation, a pool of ~30 million ISHIKAWA cells were lysed in a hypotonic lysis buffer [10 mM HEPES (ThermoFisher), 1.5 mM $MgCl_2$, 1 mM KCl, 0.5 mM DTT, supplemented with EDTA-free protease inhibitor cocktail] for 40 minutes on ice. Lysates were centrifuged at 800 xg for 8 minutes and the supernatant (cytoplasmic fraction) was transferred into a fresh tube and centrifuged twice more whereas the pellet (nuclear fraction) was resuspended in 1 ml of ice-cold hypotonic lysis buffer. The resuspended pellet was kept on ice for 5 minutes before centrifuging at 800 xg for 8 minutes and then resuspending in 500 µl of PBS.

Bovine endometrial and mouse uterine tissue pieces were homogenised via a mechanical rotor (Heidolph SilentCrusher S) in 600 µl of RNA lysis buffer (mirVana miRNA isolation). Up to four 20-second pulses were used to create single cell

suspensions. Total RNA was isolated (mirVana miRNA isolation) by adding 1/10 volume (60 µl) miRNA Homogenate Additive to each homogenate and vortexing for 20 seconds. The homogenates were incubated on ice for 10 minutes, after which 1 volume (600 µl) of acidified phenol:chloroform was added to each sample and these were vortexed for 30 seconds. The samples were then centrifuged at 10,000 xg for 5 minutes at room temperature and the aqueous phase was collected. Then, 1.25 volumes of 100% ethanol were added to the aqueous phase, mixed and added to a filter column which was centrifuged at 10,000 xg for 15 seconds. The filter column was washed with 700 µl miRNA Wash Solution 1 and centrifuged at 10,000 xg for 20 seconds, followed by the addition of 500 µl Wash Solution 2/3 and a second centrifugation at 10,000 xg for 20 seconds. The filter cartridge was dried by centrifuging at 16,000 xg for 1 minute. RNA was eluted into a fresh collection tube using 50 µl of nuclease-free water pre-heated at 95°C. Tubes were centrifuged at 10,000 xg for 30 seconds. Samples were quantified using the NanoDrop N1000 (Thermo Fisher Scientific, USA) and stored at -80°C.

2.2.4 RT-qPCR expression profiling of *XIST* and putative protein partners

To synthesize cDNA, 1 µg of RNA was used per sample (High-Capacity cDNA Reverse Transcription, ThermoFisher), according to the manufacturer's instructions, in a Veriti PCR thermocycler (Applied biosystems, UK). Briefly, in the final reaction, 1 µg of RNA was mixed with 4 mM dNTPs, 1x random primers and MultiScribe™ Reverse Transcriptase in ddH2O. Cycling conditions were at 25^{o}C for 10 minutes, 37^{o}C for 2 hours and 85^{o}C for 5 minutes. RT-qPCR was carried out using the SYBR Green dye (Roche Diagnostics LightCycler 480 SYBR Green I) and LightCycler 480 II (Roche, UK). All samples were diluted accordingly so that 10 ng of cDNA were loaded per 10 µl RT-qPCR reaction and primers were used at 0.5 µM final concentration per reaction. Serial dilutions were prepared from a 1:10 diluted pool of cDNA derived from all cDNA samples assayed. The thermocycling program used was as follows: once at 95 for 5 minutes, 45 times at 95^{o}C for 10 seconds, 60^{o}C for 10 seconds and 72^{o}C for 10 seconds. Melt curves were estimated by heating for 5 seconds at 95^{o}C, 65^{o}C for 1 minute followed by a continuous increase to 97oC while acquiring fluorescence readings (5/^{o}C). Expression levels obtained for the various

genes of interest assayed (using primers in **Table 2.1**) in the different systems were normalised to β-actin (Actb) from each species. To calculate expression relative to Actb, the following formula was used: $2^{\wedge-(Ct_{Xist}- Ct_{Actb})}$.

Table 2.1. List of primers used for RT-qPCR assessment of transcript abundance for XIST and the mRNA of its mouse protein partners.

		Primer	
Species	Gene	Orientation	Sequence (5'-3')
Human	*XIST*	Forward	GGCTCCTCTTGGACATTCTGAG
		Reverse	AGCTTGGCCAGATTCTCAAAG
	SPEN	Forward	GAAGGATGACGGTGGAGACAGA
		Reverse	CTTGAGGGACTCGGTCTGGC
	WTAP	Forward	CTAAAGCAACAACAGCAGGAGTC
		Reverse	GGTACTGGATTTGAGTAGTACACTCT
	RBM15	Forward	GTCTTCTTGTGGAGGGTTCAACT
		Reverse	CCCTGCTACTTTGATGCGTC
	LBR	Forward	AGGAGTACCTGGTGTGTTTCTCAT
		Reverse	CTGGCAAAGGAGGAGGGAA
	CIZ1	Forward	GAGATGCCAGGGGTATGGG
		Reverse	TGGAGGAGACGGAGTCACTGG
	hnRNPK	Forward	GCGTCCCATGCCTCCATCTAGAAG
		Reverse	CTGAAACCAACCATGCCGTC
	hnRNPU	Forward	GCTATCCATACCCTCGTGCC
		Reverse	CGTCCTCTGAAGTTCTGGTTGT
	ACTB	Forward	CTTCCTGGGCATGGAGTCC
		Reverse	TGATCTTGATCTTCATTGTGCTGG

Species	Gene	Primer	
		Orientation	Sequence (5'-3')
Mouse	*Xist*	Forward	AGTGGAAATTGGCTGGATTCAG
		Reverse	CTTGGTCTTGGGGATAGAAGGA
	Spen	Forward	CTCCAATCAGCCTGCCTACG
		Reverse	GTTCAGAGCCTCACACCGAG
	Wtap	Forward	ACCACTCAAATCCAGTACCTCAAG
		Reverse	TTGGGCTTGTTCCAGTTTGTC
	Rbm15	Forward	ACCGATTTGGCACCATTCG
		Reverse	CCTGAGGCGACGATCTGG
	Lbr	Forward	CAGGAGAGAAGAGGTCAAAGCC
		Reverse	ATGAGGACCGCACCAGGTACT
	Ciz1	Forward	CAAGCAGGTGAAGCCGAG
		Reverse	TTTGACAGACATAGCCCATCACT
	hnRNPK	Forward	TCCGTACAGACTACAATGCCAG
		Reverse	GCCCTCTTCCAAGGTAGG
	hnRNPU	Forward	AGAGGACCGAGTTAGAGGACC
		Reverse	CCTGCCACCATCATCTTGTC
	Actb	Forward	GGCACCAGGGTGTGATGG
		Reverse	TCCATGTCGTCCCAGTTGG

Species	Gene	Primer	
		Orientation	Sequence (5'-3')
Cow	*XIST*	Forward	AATCGTTTGTGTTGTGTGAGTGG
		Reverse	TACTTAGCACAGTTACCCCTCAG
	SPEN	Forward	CAGTGACAGCACTGATTCCAGC
		Reverse	CGCACTGGAAGATTCTGAACC
	WTAP	Forward	GGAACAAGCCCAAAATGAACTG
		Reverse	GAGATCAGCAATGGTGGACCC
	RBM15	Forward	ACCATACGCACCATTGACTACC
		Reverse	GTCTACTCTAAGGCGACGATCTG
	LBR	Forward	GCTGGTGCTGAAGCCATTTG
		Reverse	CCTGTGTGTGTTTGTGAGGCAT
	CIZ1	Forward	CACCCGAAGACCAGGAACC
		Reverse	GGCGGCTCAGAGGCTTCA
	hnRNPK	Forward	GAATCTTCCTCTTCCACCACC
		Reverse	CTGAAACCAACCATGCCATCA
	hnRNPU	Forward	GGCATTGGCTATCCGTACC
		Reverse	CGTCCTCTGAAGTTCTGGTTGT
	ACTB	Forward	CGCCATGGATGATGATATTGC
		Reverse	AAGCCGGCCTTGCACAT

Species	Gene	Primer	
		Orientation	Sequence (5'-3')
Pig	*XIST*	Forward	AGAAAGGGTGGTGGAATTGGTC
		Reverse	GGTGCTGACTGGCTGAATAGAG
	SPEN	Forward	AGTGACAGAAGAGAAGACCACGG
		Reverse	GAGTCCACTTGTTCAGGCTGTTG
	WTAP	Forward	CTAGCAACCAAGGAGCAAGAG
		Reverse	CGACCATTGTTGATCTCAGTTGG
	RBM15	Forward	CAAAGGTGACAGTTGGGCATAC
		Reverse	AAGTCTACTCTAAGGCGACGATC
	LBR	Forward	GCCTCGGAATGACCTGTC
		Reverse	TAATGACCACCCAGCCAATCA
	CIZ1	Forward	CCAAGACGAGGACCACTTCATC
		Reverse	ATCTCACCTGCTTGCAGAATTCC
	hnRNPK	Forward	AAGGAAGCGACTTTGACTGC
		Reverse	GTCTGAGTGTTCTCCCGAAGTT
	hnRNPU	Forward	AACAGAACAGAAAGGCGGAG
		Reverse	GCGATTTGGCTCTGCTATACT
	ACTB	Forward	GCACGGCATCGTCACCAAC
		Reverse	GTCCAGACGCAGGATGGC

2.2.5 Protein lysate generation for immunoblotting profiling of putative *XIST* protein partners

To generate protein lysates, 10 million ISHIKAWA cells were lysed in 250 µl of RIPA buffer (150 mM NaCl, 1% IGEPAL, 0.5% Sodium deoxycholate, 0.1% SDS, 25 mM Tris (pH 7.4), supplemented with protease inhibitor cocktail in ddH2O). Three biological replicates were performed. Cell lysates were incubated on ice for 10 minutes and passed five times through a 21 G needle and syringe to homogenise. Cell lysates were centrifuged at max speed (17,000 xg) for 1 minute to remove insoluble material and supernatants were frozen at -20^{o}C.

Bovine endometrial and mouse uterine tissue pieces were homogenised via a mechanical rotor (Heidolph SilentCrusher S) in 600 µl of RIPA buffer and protein lysates were generated as described above. Up to four 20-second pulses were used to create single cell suspensions.

To quantify the protein concentration in each lysate, the Pierce™ BCA Protein Assay Kit was used according to the manufacturer's instructions using two technical replicates for each of three biological replicates. Briefly, 25 µL of each standard or unknown sample was pippeted into a 96-well microplate with transparent bottoms, using fresh RIPA buffer or water for the blank control. Then, 200 µL of the working reagent mix (Reagent A:Reagent B, 50:1) were pipetted into each well and the plate was mixed thoroughly on a plate shaker for 30 seconds. The plate was then covered with foil and incubated at 37°C for 30 minutes. The absorbance was measured on a plate reader at 595 nm.

2.2.6. Protein profiling of putative *XIST* protein partners via wester blotting

Protein samples were diluted with in 6x Laemmli buffer (1.2 g SDS, 6 mg bromophenol blue, 4.7 ml glycerol, 1.2 ml Tris 0.5M pH6.8, 0.93 g DTT, 2.1ml ddH2O) and heated at 95^{o}C for 5 minutes. Proteins were separated on a denaturing 10% SDS-polyacrylamide gel at 90 V for 30 minutes and 150 V for 80 minutes using

1x running buffer (1x running buffer, 0.025 M Tris, 0.25 M Glycine and 0.1% SDS in water). Subsequently, proteins were wet-transferred) onto a 0.22 µm PVDF membrane at 200 mA for 1.5 hours using 1x transfer buffer (25 mM Tris and 192 mM Glycine with 20% methanol in water). Membranes were blocked with 5% Marvel milk powder in PBS-T (0.5% Tween) for 1 hour at room temperature while rolling. Blots were incubated with primary antibodies (**Table 2.2**) overnight at 4°C while rolling, washed three times in PBS-T (0.5% Tween) at room temperature for 10 minutes each while rolling and incubated with secondary HRP-conjugated antibodies for 1 hour at room temperature while rolling. Membranes were washed three times in PBS-T (0.5% Tween) at room temperature for 10 minutes each while rolling and developed using ECL (Biological Industries, UK). Chemiluminescent signal was detected using a ChemiDoc XRS+ imaging system (BioRad, UK).

Table 2.2. List of antibodies.

	Protein	Species	Supplier	Dilution	Cat. No. #
Primary	β-tubulin	Mouse	DSHB	1:1000-1:5000	E7-S
	WTAP	Mouse	ProteinTech	1:1000	60881-1-Ig
	WTAP	Rabbit	Abcam	1:1000	ab195380
	RBM15	Rabbit	ProteinTech	1:1000	10587-1-AP
	hnRNPK	Rabbit	Abcam	1:1000	ab52600
	hnRNPU	Rabbit	ProteinTech	1:1000	14599-1-AP
	CIZ1	Rabbit	ThermoFisher	1:1000	PA5-27625
	H3K27me3	Mouse	Abcam	1:1000	ab6002
	SPEN	Rabbit	Abcam	1:1000	ab72266
	Lamin B1	Rabbit	Abcam	1:1000	ab133741
Secondary (HRP-linked)	Mouse IgG	Goat	Cell Signalling Technology	1:5000	7076S
	Rabbit IgG	Goat	Cell Signalling Technology	1:5000	7074S

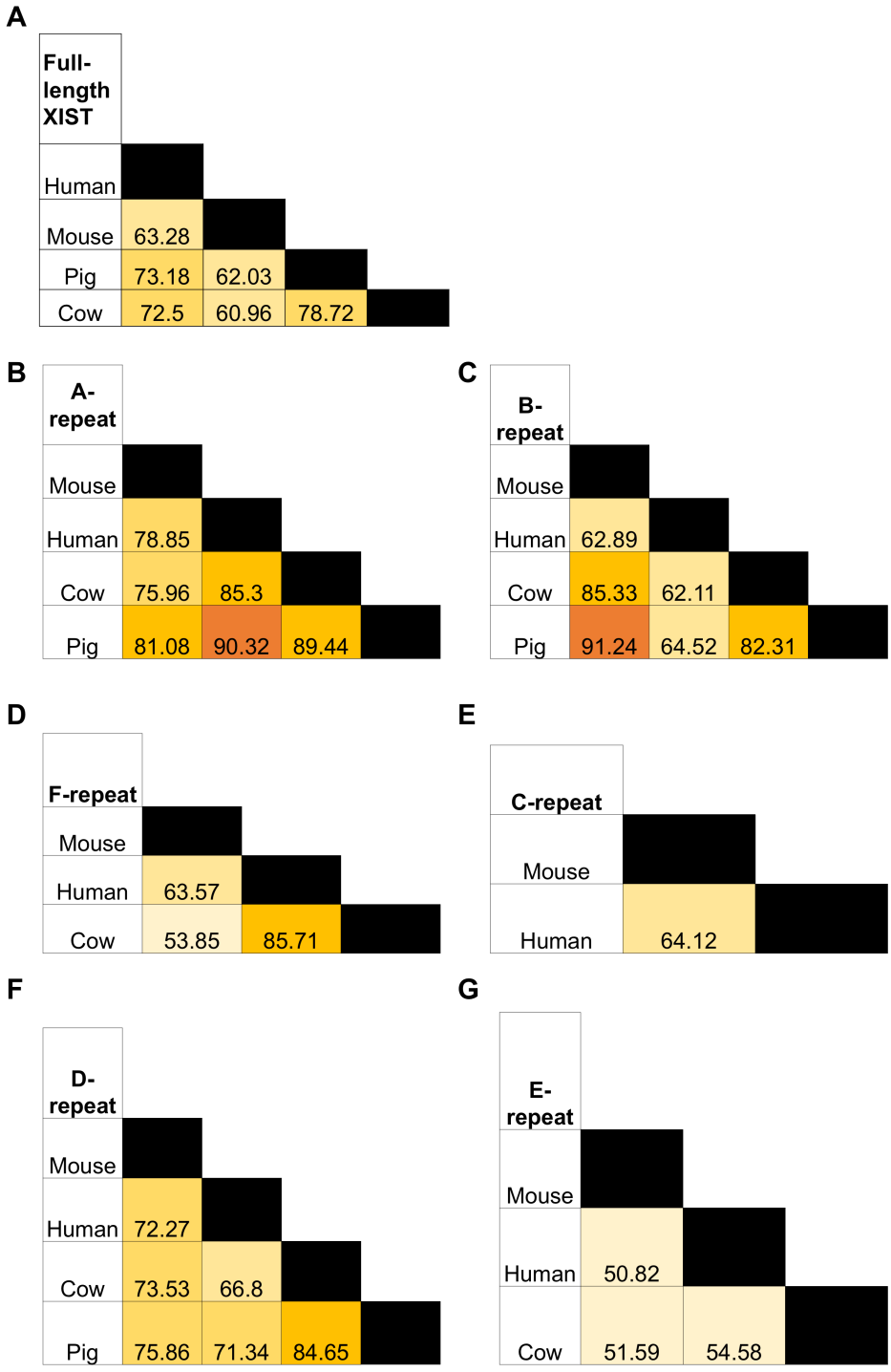
2.3 Results

2.1.1. Estimation of *XIST* RNA conservation across four placental mammals with divergent early pregnancy morphologies

To examine how the sequence of *XIST* varies across placental mammals, human, mouse, cow and pig were selected as they represent species with different implantation strategies. The RNA sequence of full-length *XIST* was available for human, mouse and pig. The complete *XIST* RNA sequence was not available for cow, and so the predicted entry for the longest variant was used (NCBI Reference Sequence: XR_001495594.2). The conservation of the *XIST* RNA sequence in the four species tested was estimated to be ~61-73% (**Table 2.3A**). More specifically, comparing *Xist* from mouse with other placental mammals here, conservation was ~63% at highest whereas *XIST* conservation among human, cow and pig could be as high as ~73%.

Table 2.3. XIST contains regions of high conservation in human, mouse, cow and pig.

Percent identity matrices were generated from aligning full-length *XIST* nucleotide sequences or repetitive regions on the *XIST* RNA from human, mouse, cow and pig with Clustal-ω (default settings; see Methods). Percent identity scores are a proxy for conservation where a high % value shown represents similarity between the two compared sequences (darker colours show higher similarity).



Aligning the sequences from *XIST*'s repeat regions confirmed that some regions within *XIST* are more conserved compared to full-length *XIST* across human, mouse, cow and pig (**Table 2.3B-G**). Among the well-documented repetitive regions characterised in mouse *Xist*, repeats A, B and D displayed the highest conservation. Repeat A displayed the highest % identity score ranging from ~76% to 90% between the four species tested (**Table 2.3B**). As seen from full-length *XIST* comparisons, mouse *Xist* repeat A was at highest 81% similar to pig, whereas human *XIST* repeat A was at highest 90% similar to pig. Repeat B exhibited similarly higher % identity scores than full-length *XIST*, ranging from 62 to 91% depending on the species compared (**Table 2.3C**). In contrast to repeat A, mouse *Xist* repeat B conservation was only ~60% similar to human but ~85 and ~91% similar to cow and pig, respectively. Human *XIST* repeat B ranged from 60-63% in similarity to mouse, cow and pig. Repeat D from mouse was slightly more similar between cow (~74%) and pig (~76%) than human (72%) (**Table 2.3F**). Repeat D from human was more similar to mouse (~72%), than to cow (~67%) and pig (~71%).

Nonetheless, the F and C repeats are moderately conserved in the species they have been mapped to (**Table 2.3D&E**). Mouse repeat F was ~64 and ~54% to human and cow, whereas human repeat F was ~86% similar to cow (**Table 2.3D**). Repeat C is only found in mouse and human in the assayed species where it was ~64% similar (**Table 2.3E**). The E repeat is another example of a repetitive region for which, as of yet, there is no evidence for its presence in the pig. In the three species for which there was evidence of the E repeat, its conservation when considered in isolation, was lower than full-length *XIST*, displaying an identity % score range of 50-54% (**Table 2.3G**). Namely, mouse repeat E was ~51% similar to human and cow whereas human repeat E was ~51 and ~55% similar to mouse and cow, respectively. Overall, previously characterised *XIST* repeats are >50% conserved across human, mouse cow and pig. In particular, repeats A, B and D are more conserved than full-length *XIST* in human, mouse cow and pig. Repeats A and E were more similar across human, cow and pig whereas repeat B and D were more similar across mouse, cow and pig. In all comparisons, repeats from cow and pig were more similar between each other than to human or mouse.

2.1.2. Conservation analysis estimates for protein partners identified in mouse *Xist*

At the amino acid sequence level, all seven proteins examined exhibited high similarity (>70%) between human, mouse, cow and pig (**Table 2.4**).

Specifically, human CIZ1 was ~73% similar with mouse and ~80% similar between cow and pig. Mouse Ciz1 was ~70% and ~69% similar between cow and pig, respectively. Human LBR exhibited an ~80% similarity with mouse, but an ~87% and an 86% similarity with cow and pig, respectively whereas mouse Lbr was 80% and 79% similar to cow and pig LBR, respectively. Human SPEN was ~83% similar to mouse, and 88% similar to cow and pig. In contrast, mouse Spen was ~80% and 81% similar to cow and pig. Human RBM15 demonstrated a 94% similarity to mouse but a ~98 and ~97% similarity towards the cow and pig amino acid sequence. Mouse Rbm15 was only 94% similar to the cow and pig sequence. Likewise, human WTAP was ~96% similar to the mouse one, and ~95% similar to cow and pig WTAP. Mouse Wtap was 95% similar to cow and 94% similar to pig. Human and mouse hnRNPK are 100% identical and ~99% similar with the cow and pig protein. In contrast, Human and mouse hnRNPK share 97% similarity, with human and mouse showing ~99% and 98% similarity to cow and ~98% and ~97% similarity with pig, respectively. In summary, putative *XIST* protein partners were highly conserved: CIZ1 (~70%) < LBR (~80%), SPEN (~80%) < RBM15 (~95%), WTAP (~95%) < hnRNPK, hnRNPU (~99%).

Table 2.4. High conservation of putative protein partners of XIST in human, mouse, cow and pig.

Percent identity matrices were generated from amino acid sequences aligned with Clustal-ω (default settings) as a proxy for conservation and represent similarity between the two compared sequences. The default transition matrix is Gonnet, gap opening penalty is 6 bits, gap extension is 1 bit. The following peptide sequences were retrieved from UniProt for mouse, human, pig and cow, respectively: SPEN (Q96T58, Q62504, A0A287BPC2, F1MRK2), WTAP (Q15007, Q9ER69, F1SB61, F1MN80), RBM15 (Q96T37, Q6PHZ5, F1S619, E1BFX6) and CIZ1 (Q9ULV3, Q8VEH2, F1RRW9, F1MZB8).

		Human (against)		
		Mouse	Cow	Pig
Sequence	Gene name	Amino acid % identity		
Peptide	**SPEN**	82.7	88.4	88.3
	WTAP	96.5	95.2	94.7
	RBM15	94.5	97.8	96.8
	LBR	80.3	87.2	85.7
	CIZ1	72.5	79.9	80.3
	hnRNPK	100	98.9	98.9
	hnRNPU	97.4	98.7	98

The majority of SPEN's protein domains appear to be found within stretches of complete sequence similarity as indicated by the stars in **Figure 2.1A-E**. Two exceptions were the first 20 aa in the RNA recognition motif 1 (RRM1) and RRM2 in all species, and the lack of a Spen paralogue and orthologue, C-terminal (SPOC) domain in the pig (**Figure 2.2A&E**). Given a poor annotation score on Uniprot, it is likely this protein has not been well-characterised in pig. All domains predicted by InterPro for RBM15, WTAP and hnRNPK proteins were conserved (**Figures 2.3-2.5**). All domains of the hnRNPU protein were conserved with the exception of the pig sequence lacking a SAP domain (**Figure 2.6**). A large portion of each domain of the LBR and CIZ1 proteins was conserved, with several amino acid stretches lacking a conserved consensus sequence (**Figure 2.7** and **2.8**). Overall, in most cases the regions of high conservation seen for the amino acid sequence of the proteins tested, completely overlapped the proteins' functional domains.

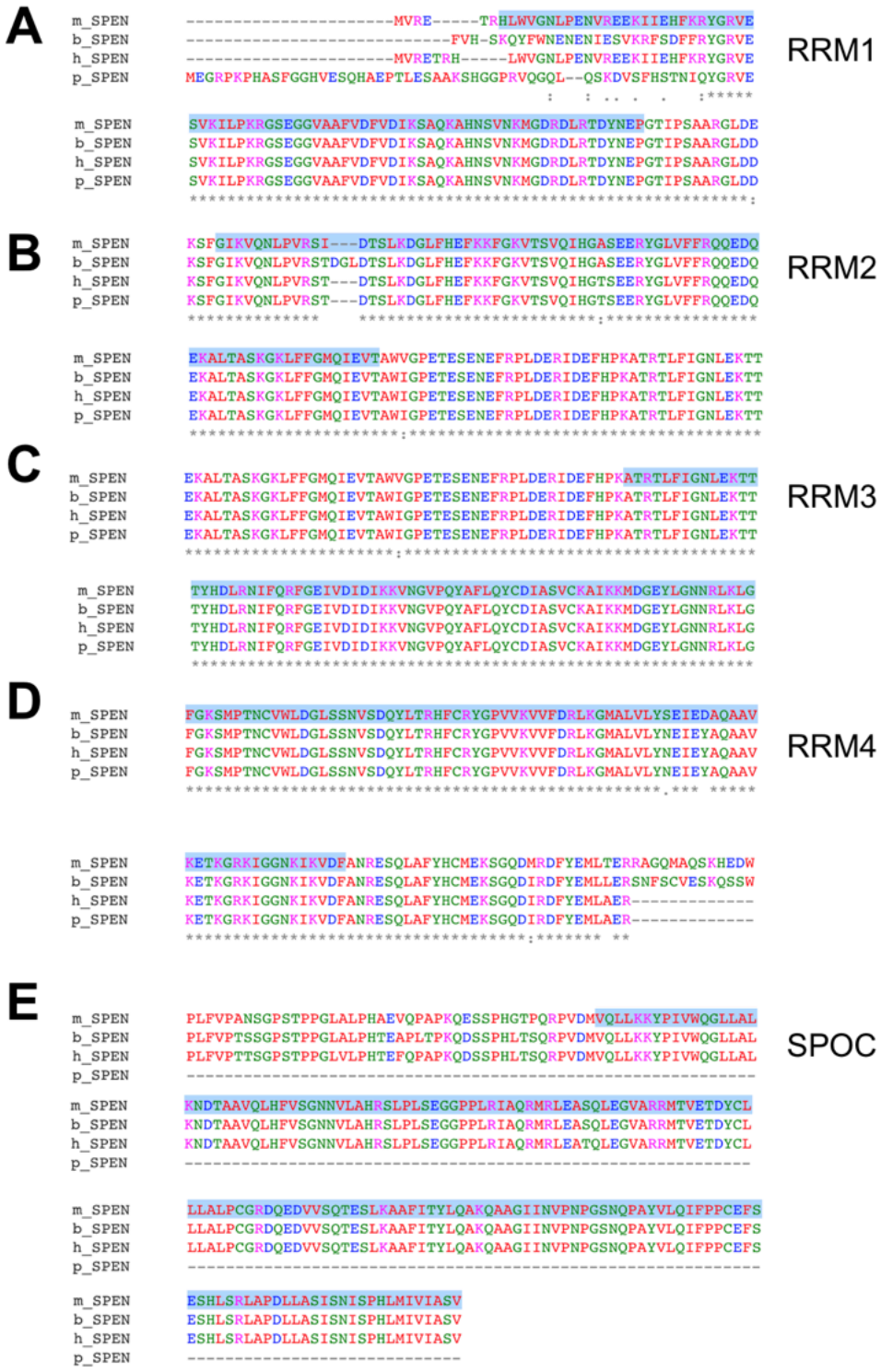


Figure 2.1. Amino acid alignment of SPEN protein domains.

Clustal-ω alignment of SPEN amino acid sequences from mouse (m_SPEN), cow (b_SPEN), human (h_SPEN) and pig (p_SPEN) (default settings). Highlighted in blue are domains (functional units) predicted by InterPro. RRM, RNA Recognition Motif; SPOC, Spen Paralogue and Orthologue C-terminal

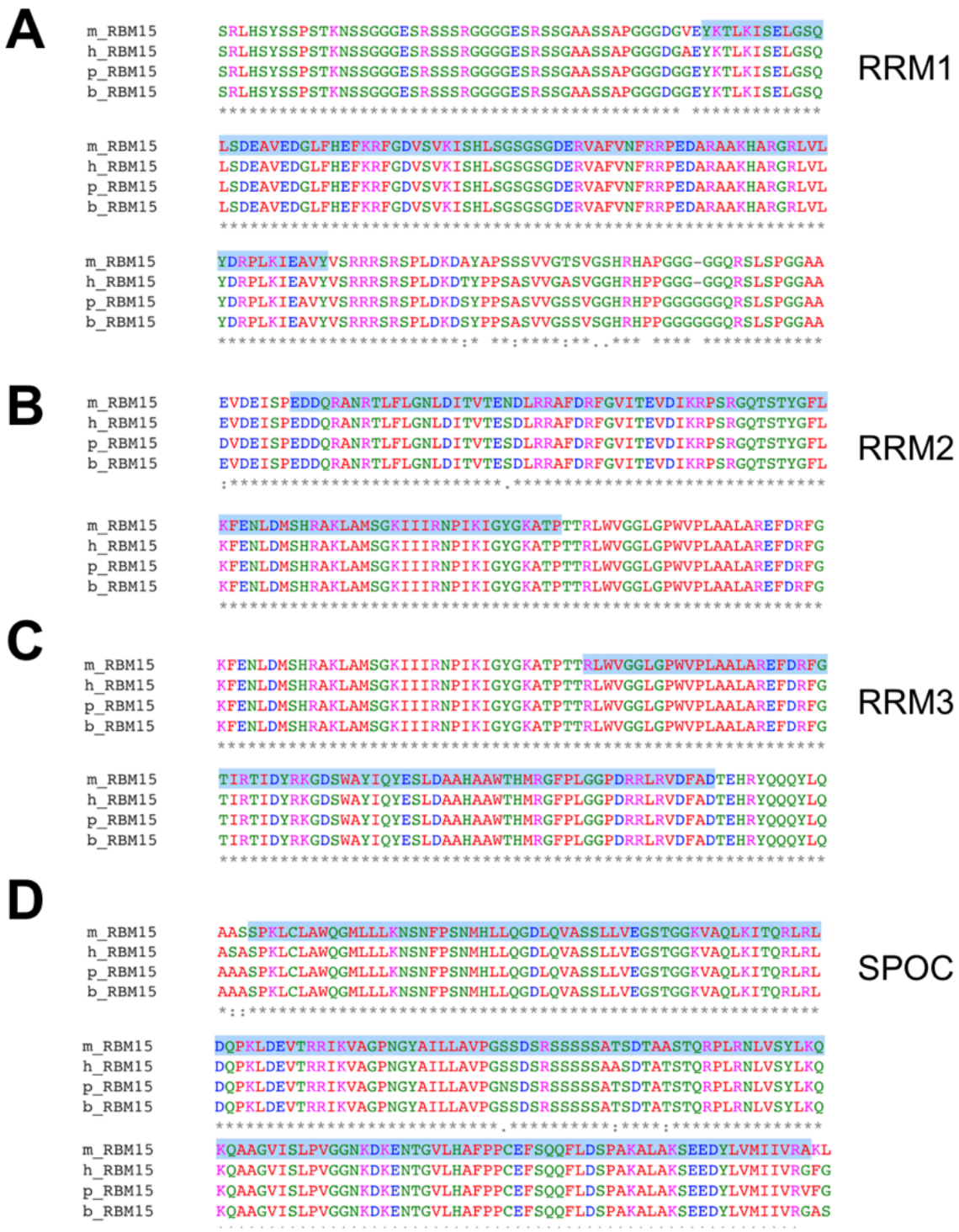


Figure 2.2. Amino acid alignment of RBM15 protein domains.

Clustal-ω alignment of RBM15 amino acid sequences from mouse (m_RBM15), cow (b_RBM15), human (h_RBM15) and pig (p_RBM15) (default settings). Highlighted in blue are domains predicted by InterPro. RRM, RNA Recognition Motif; SPOC, Spen Paralogue and Orthologue C-terminal

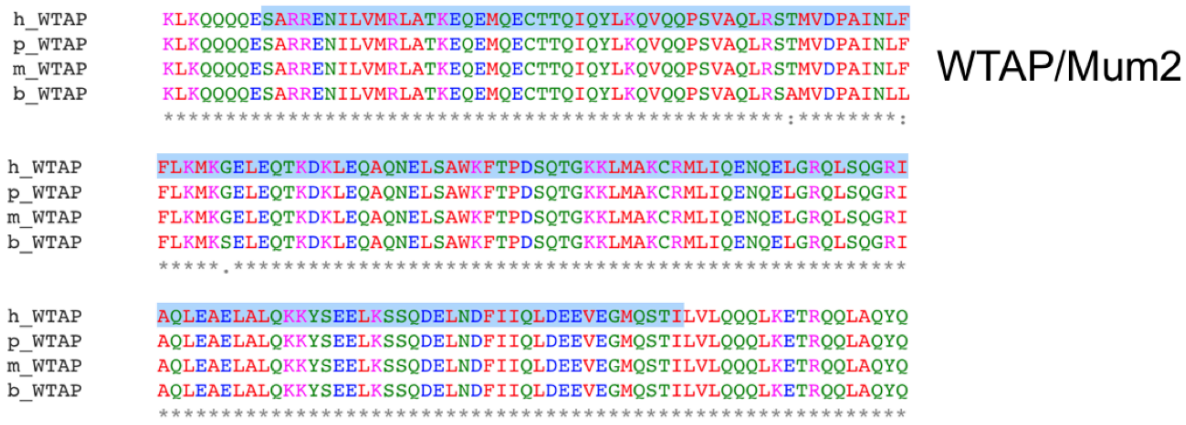


Figure 2.3. Amino acid alignment of WTAP protein domains

Clustal-ω alignment of WTAP amino acid sequences from mouse (m_WTAP), cow (b_WTAP), human (h_WTAP) and pig (p_WTAP) (default settings). Highlighted in blue are domains predicted by InterPro. Mum2, Muddled meiosis protein 2 domain

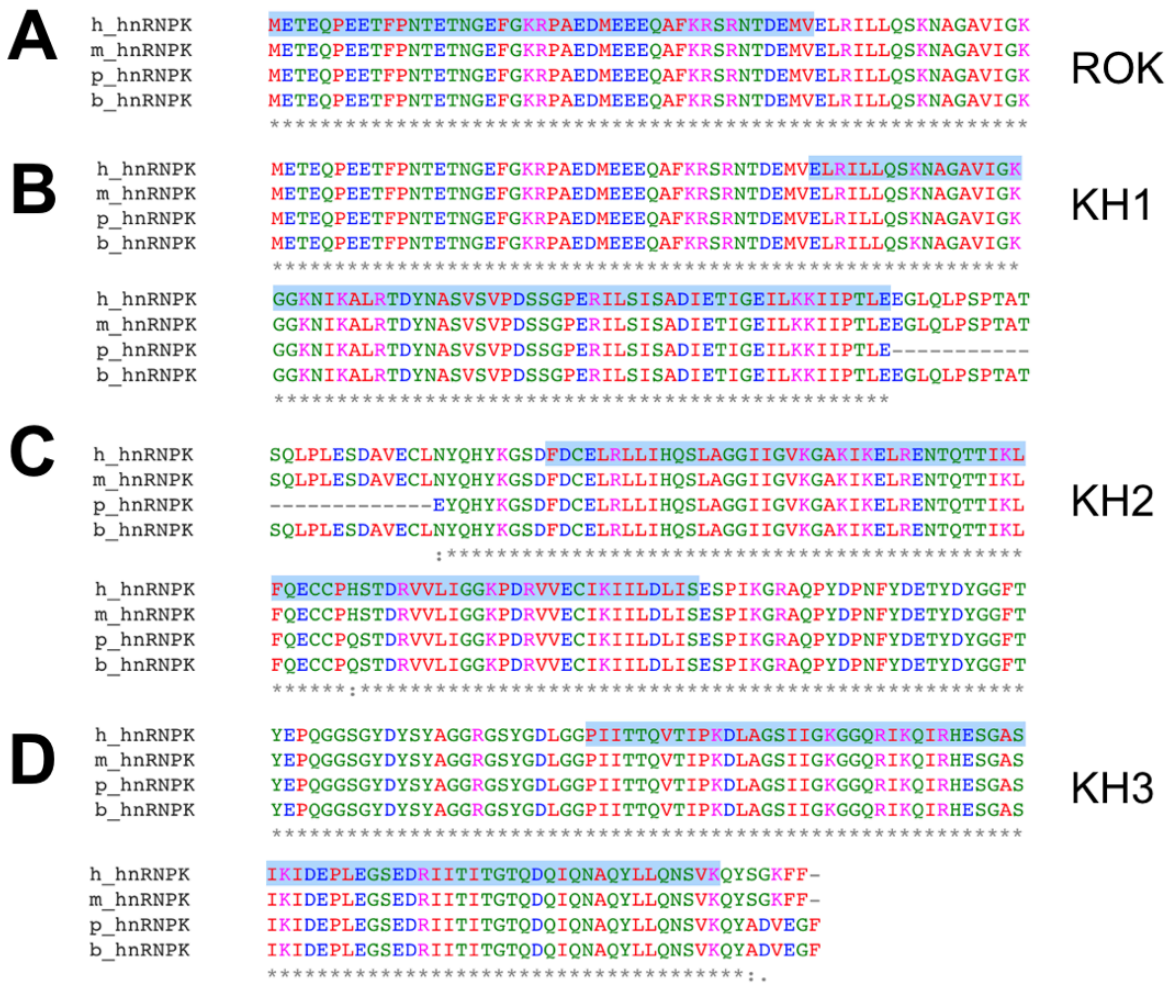


Figure 2.4. Amino acid alignment of hnRNPK protein domains.

Clustal-ω alignment of hnRNPK amino acid sequences from mouse (m_hnRNPK), cow (b_hnRNPK), human (h_hnRNPK) and pig (p_hnRNPK) (default settings). Highlighted in blue are domains predicted by InterPro. ROK, Repressor Open reading frame Kinase; KH, K Homology

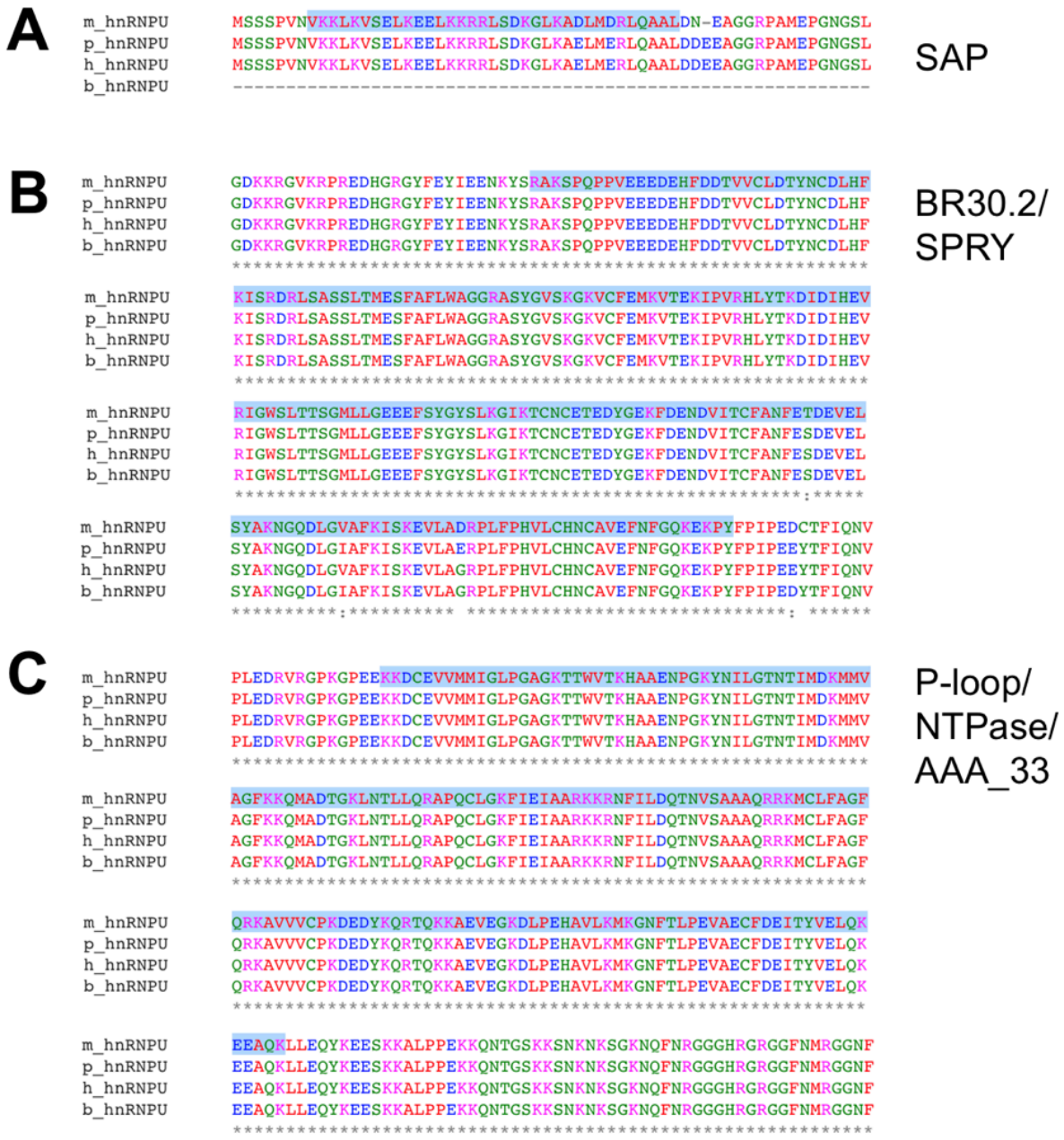


Figure 2.5. Amino acid alignment of hnRNPU protein domains.

Clustal-ω alignment of hnRNPU amino acid sequences from mouse (m_hnRNPU), cow (b_hnRNPU), human (h_hnRNPU) and pig (p_hnRNPU) (default settings). Highlighted in blue are domains predicted by InterPro. SAP, SAF-A/B, Acinus and PIAS domain; SPRY, SPla and RYanodine Receptor; P-loop, phosphate loop; NTPase, nucleoside-triphosphatase

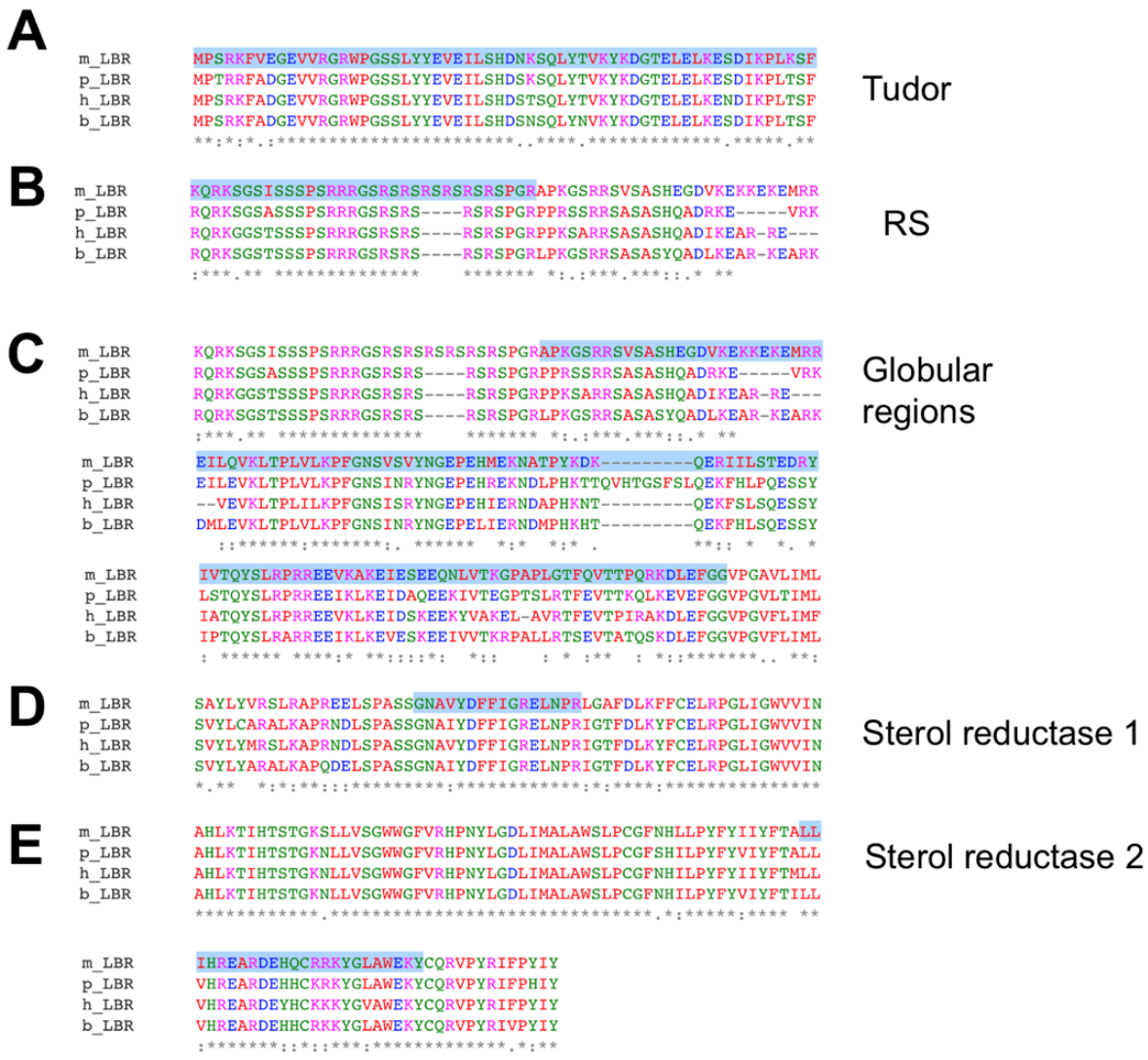


Figure 2.6. Amino acid alignment of LBR protein domains

Clustal-ω alignment of LBR amino acid sequences from mouse (m_LBR), cow (b_LBR), human (h_LBR) and pig (p_LBR) (default settings). Highlighted in blue are domains predicted by InterPro. RS, alternating arginine and serine residues

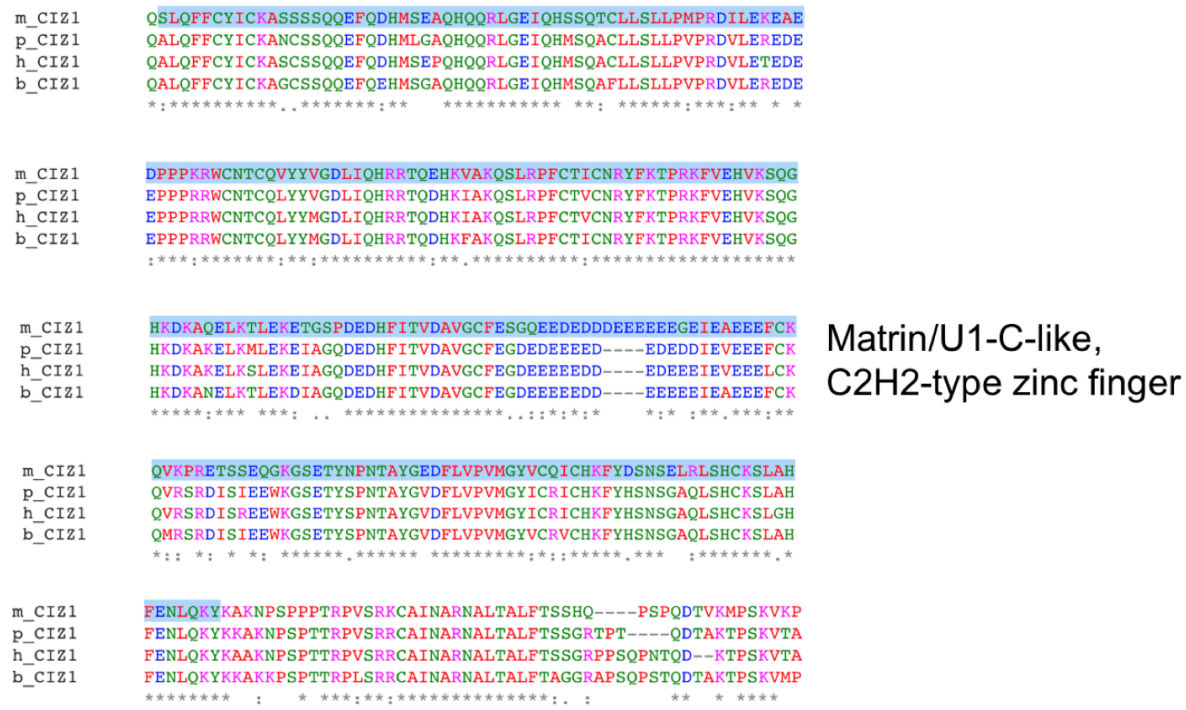


Figure 2.7. Amino acid alignment of CIZ1 protein domains

Clustal-ω alignment of CIZ1 amino acid sequences from mouse (m_CIZ1), cow (b_CIZ1), human (h_CIZ1) and pig (p_CIZ1) (default settings). Highlighted in blue are domains predicted by InterPro.

2.2. Examination of *XIST* expression levels in placental mammals of interest

The second aim of this chapter was to investigate whether *XIST* and putative protein partners are co-ordinately expressed in the same tissues/cells, a requirement before assessing for a biochemical interaction between the two. To probe for the expression levels of *XIST* in uterine tissue, data from the GTEx portal (https://www.gtexportal.org/home/), which contains tissue-specific gene expression data from 54 non-diseased tissue sites across nearly 1000 human individuals were accessed. *XIST* abundance was highest in human reproductive tissues, including the endometrium and lowly expressed in other tissues, e.g. liver or blood (**Figure 2.8**).

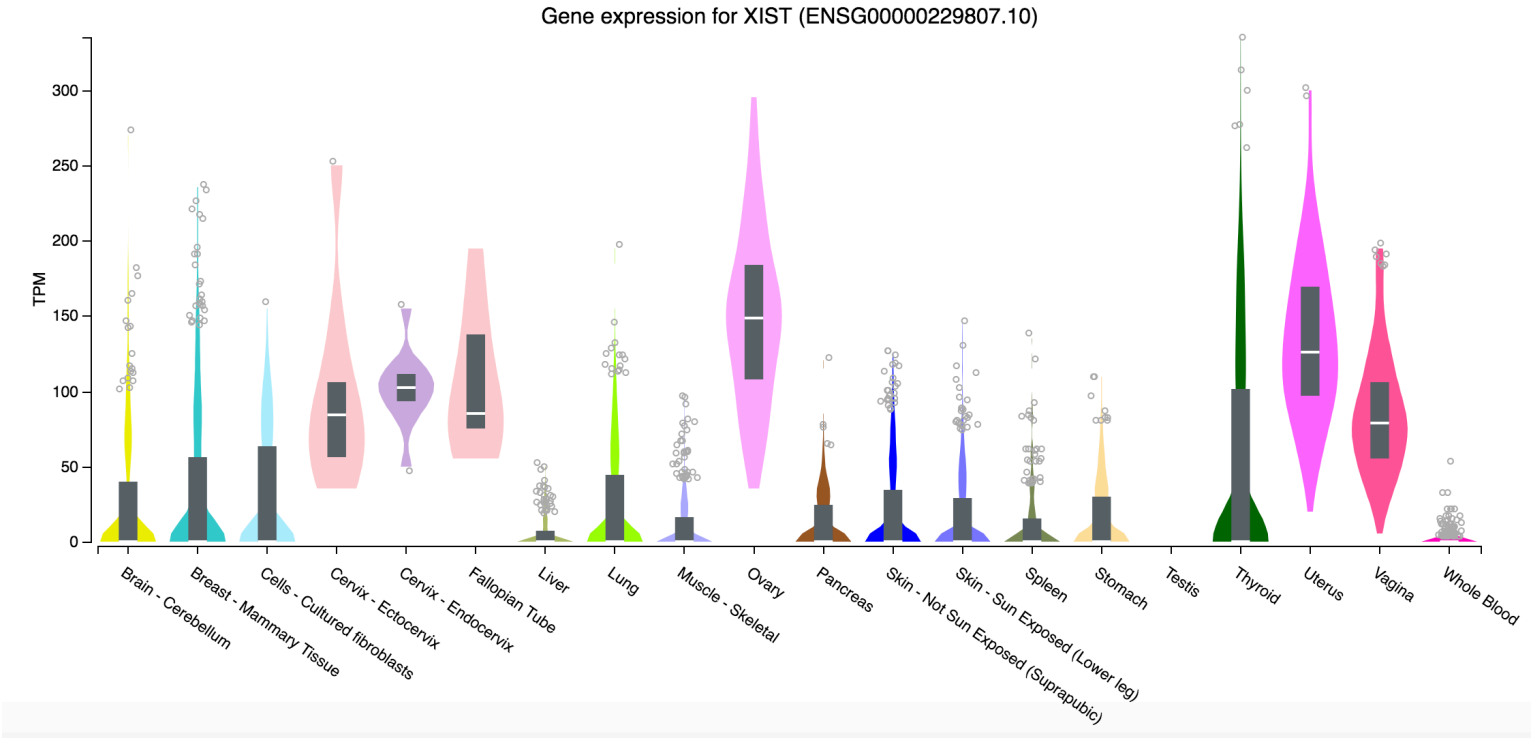


Figure 2.8. Tissue-wide global RNA sequencing highlights XIST (ENSG00000229807.10) as most enriched in human reproductive tissues

Data generated by non-strand specific polyA+ selected Illumina TruSeq library and obtained from: https://gtexportal.org/home/gene/XIST on 29/12/2020. Expression values shown as Transcripts Per Million (TPM), calculated from a gene model with isoforms collapsed to a single gene. Box plots are shown as median and 25th and 75th percentiles. Outliers are displayed as dots if they are above or below 1.5 times the interquartile range. n=~1000 human individuals. Colours indicate different tissue types and shades of same colour indicate tissue subtypes.

However, obtaining tissue samples from human endometrium is extremely difficult and getting consent from private clinics only happens in rare cases. In search of an alternative, the Expression Atlas v26 database (https://www.ebi.ac.uk/gxa/home) was queried for cell lines of endometrial origin. A human endometrial adenocarcinoma cell line (ISHIKAWA) was selected as a model for the human endometrium (**Table 2.5**) based on its endometrial origin and epithelial-like morphology (Nishida, 2002).

Table 2.5. Cell lines of reproductive origin and XIST expression levels.

Cell line name and XIST expression values were extracted from the Expression Atlas v26 database. RNA expression values shown as Transcripts Per Million (TPM), from RNA-Seq (Barretina et al., 2012). Cell lines of endometrial origin are in bold.

Origin	Cell line	XIST Expression level (TPM)
Breast Ductal Carcinoma	ZR-75-30	330
Uterine Cervix, Cervical Small Cell Carcinoma	TC-YIK	310
Uterine Cervical Squamous Cell Carcinoma	SiHa	301
Uterine Cervical Carcinoma	SKG-IIIa	271
Uterine Cervical Squamous Cell Carcinoma	SW756	261
Uterine Cervical Carcinoma	SISO	241
Ovarian Adenocarcinoma	DOV13	223
Uterine Cervical Carcinoma	C-4-I	213
Uterine Cervical Carcinoma	ME-180	194
Vulvar Squamous Cell Carcinoma	SW954	193
Ovarian Mucinous Adenocarcinoma	JHOM-2B	189
Vulvar Carcinoma	SW962	177
Breast Ductal Adenocarcinoma	HCC1187	67
Breast Ductal Adenocarcinoma	HCC38	4
Endometrium adenocarcinoma	**Ishikawa (Heraklio) 02 ER-**	**3**

To determine *XIST* expression levels, total RNA was extracted and RT-qPCR performed on ISHIKAWA cells and uterine tissues from mouse, cow and pig. *ACTB* was used as a control, and thus *XIST* levels were normalised to *ACTB. XIST* was detected in ISHIKAWA cells, albeit at a very low level (**Figure 2.9A**), consistent with a previous report in human patient endometrial cancer samples (Zhang et al., 2014). The expression of *Xist* was confirmed in the mouse uterus but varied between biological replicates (**Figure 2.9B**). *XIST* was also detected in cow and pig endometrium at similar expression levels (**Figure 2.9C&D**). Consistently, high levels of *XIST* expression in cow endometrium has been previously reported (Forde et al., 2016). Whereas expression of cow *XIST* was variable as seen for mouse uteri (**Figure 2.9**: B mouse and C cow), variation in expression of pig *XIST* was similar to ISHIKAWA cells (**Figure 2.9**: A human and D pig). Taken together, *XIST* is present and detectable in all species models tested here. *XIST*'s expression appears to be less pronounced and less variable in the ISHIKAWA cell line compared to what was seen in mouse and cow uterine tissue.

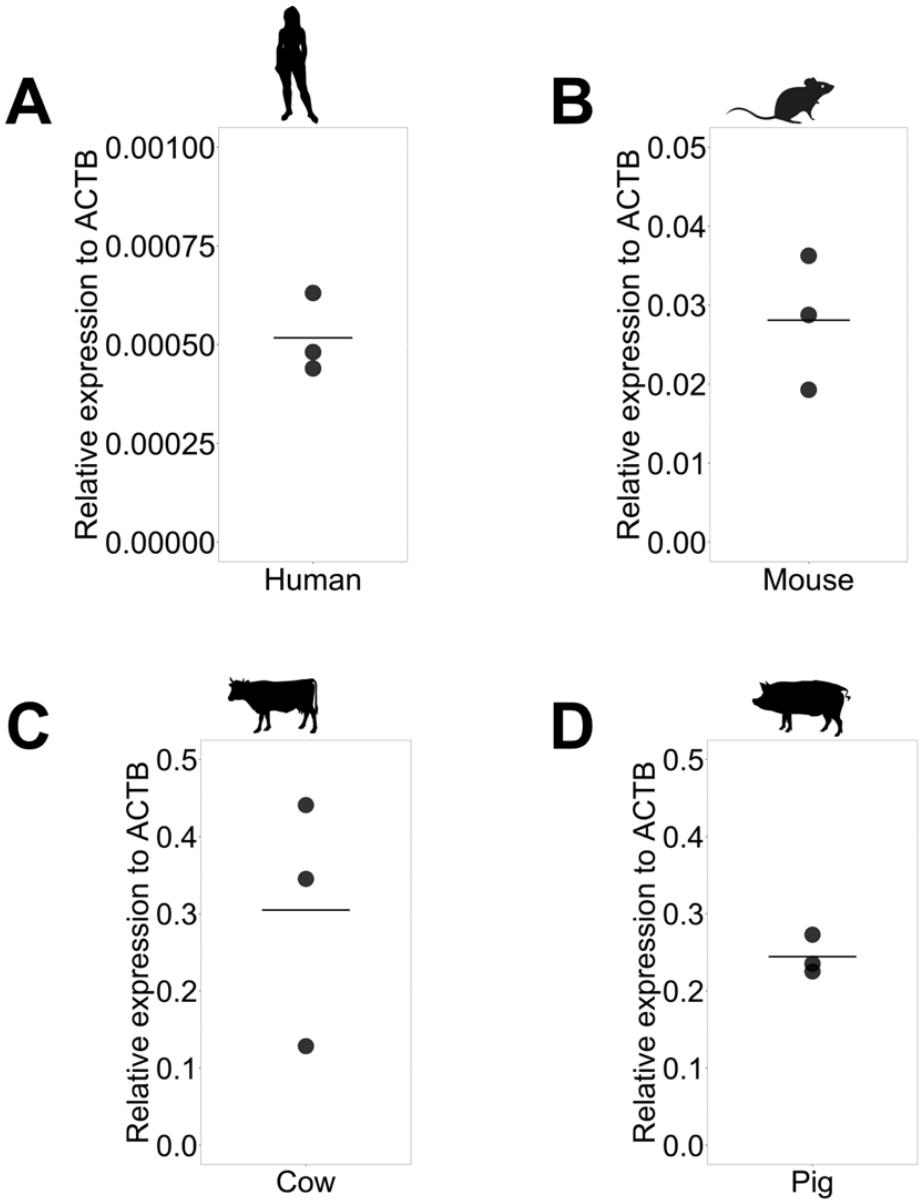


Figure 2.9. XIST lncRNA is variably expressed among different species and within individual female animals.

XIST RNA levels were measured by RT-qPCR with species-specific *ACTB* as a reference gene in **A**) human ISHIKAWA cells, **B**) mouse uterus, **C**) cow endometrium and **D**) pig endometrium. Dots denote relative expression normalised to *ACTB* and dashes indicate the mean for n=3 biological replicates. For each biological replicate, a total of three technical replicates were averaged. cDNA equivalent to 10 ng RNA were loaded per well of technical replicate. Normalisation to ACTB was performed using the $2^{\wedge\text{-(Target-ACTB)}}$ formula. Amplification efficiency of primers was checked using a standard curve of serial dilutions of a cDNA pool consisting of all samples run on the plate (1:10, 1:50, 1:250, 1:1250, 1:6250).

2.3. Examination of *XIST* putative protein partner expression levels across placental mammals of interest

To assess for the presence of putative *XIST* protein partners in the same tissues where *XIST* expression was detected, mRNA levels of *SPEN, CIZ1, hnRNPK, RBM15, WTAP, LBR and hnRNPU* were measured. RT-qPCR indicated that mRNA from all seven candidate genes was detected across human, mouse, cow and pig (**Figure 2.10A-D**). *SPEN, CIZ1, RBM15, WTAP* and *LBR* fluctuated at similar, low levels relative to *ACTB* across the four species. A pattern of high *hnRNPK* and *hnRNPU* expression relative to *ACTB* was seen consistently across all four species.

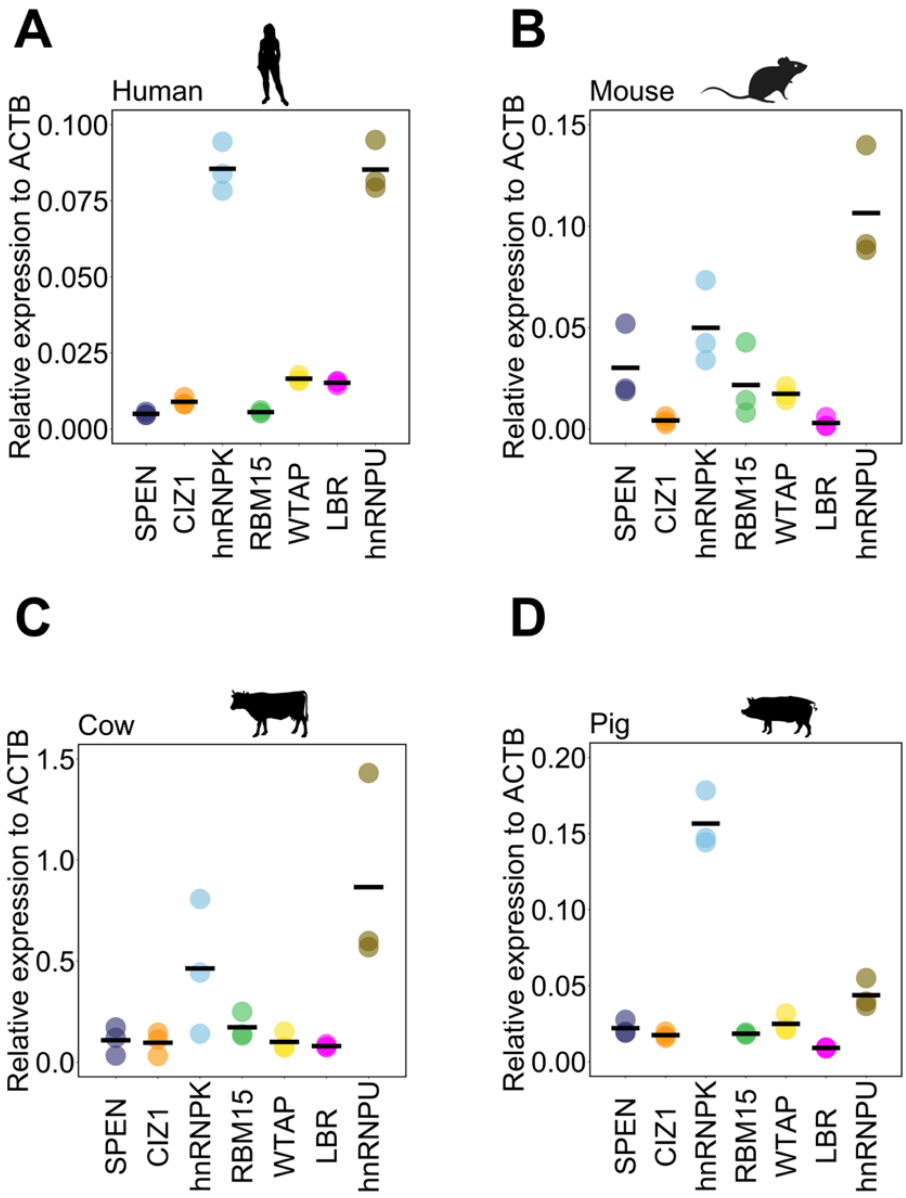


Figure 2.10. Putative XIST protein partner mRNAs are variably expressed among different species and within individual female animals.

mRNA levels were measured by RT-qPCR with species-specific ACTB as a reference gene in **A**) human ISHIKAWA cells, **B**) mouse uterus, **C**) cow endometrium and **D**) pig endometrium. Dots denote relative expression normalised to ACTB and dashes indicate the mean for n=3 female animals or ISHIKAWA cells. For each biological replicate, a total of three technical replicates were averaged. cDNA equivalent to 10 ng RNA were loaded per well of technical replicate. Normalisation to ACTB was performed using the $2^{-(Target-ACTB)}$ formula. Amplification efficiency of primers was checked using a standard curve of serial dilutions of a cDNA pool consisting of all samples run on the plate (1:10, 1:50, 1:250, 1:1250, 1:6250).

Table 2.6. Raw Ct values of *XIST* and putative protein partners as measured by RT-qPCR. Three independent replicates are shown which denote averages from three technical replicates. Human refers to the ISHIKAWA cell line, cow and pig refer to endometrium dissected whereas for mouse, whole uteri were used.

	Human		
	Replicate 1	Replicate 2	Replicate 3
ACTB	19.08	18.72	18.79
XIST	30.24	29.36	29.82
SPEN	26.86	26.45	26.27
WTAP	24.91	24.71	24.74
RBM15	26.65	26.07	26.38
CIZ1	25.8	25.68	25.24
LBR	24.87	24.85	24.63
hnRNPK	22.28	22.31	22.33
hnRNPU	22.49	22.39	22.06

	Mouse		
	Replicate 1	Replicate 2	Replicate 3
Actb	17.5	18.64	18.15
Xist	23.87	23.88	23.59
Spen	26.62	26.86	26.48
Wtap	26.79	26.73	26.21
Rbm15	25.69	25.55	24.9
Ciz1	25.45	25.41	24.89
Lbr	26.47	26.59	26.13
Hnrnpk	22.22	22.38	21.91
Hnrnpk	22.9	22.8	22.39

	Cow		
	Replicate 1	Replicate 2	Replicate 3
ACTB	24.8	25.7	22.93
XIST	25.98	28.67	24.47
SPEN	27.34	30.63	26.01
WTAP	27.52	29.52	26.6
RBM15	26.81	28.63	25.79
CIZ1	27.54	30.4	25.86
LBR	28.52	28.87	26.41
hnRNPK	25.06	28.22	23.89
hnRNPU	24.24	26.2	23.45

	Pig		
	Replicate 1	Replicate 2	Replicate 3
ACTB	20.85	20.21	19.27
XIST	22.73	22.3	21.42
SPEN	26.04	25.93	24.92
WTAP	25.84	25.78	24.78
RBM15	26.67	25.98	24.97
CIZ1	26.52	26.22	25.43
LBR	27.63	27.24	26.06
hnRNPK	23.63	23.13	21.92
hnRNPU	25.64	25.01	23.62

Having determined the mRNAs of the candidate proteins were expressed, actual protein levels were then assessed using western blotting on protein samples extracted from human ISHIKAWA cells and uterine tissues from mouse, cow and pig. Suitable antibodies for western blotting could not be identified for all proteins (e.g. LBR and SPEN). Therefore, only a subset of putative *XIST* protein partners could be assayed (i.e. hnRNPU, CIZ1, RBM15, hnRNPK and WTAP). A band at ~120 kDa was detected for human samples when probing with the anti-hnRNPU antibody and similar sized bands were found in the other species tested (**Figure 2.11**). Two bands were consistently detected for the CIZ1 protein in all species assayed, one at 95 kDa and another at 120 kDa. Similarly, two bands at 100-107 kDa were present for the RBM15 protein in all species tested, which was what was expected. When samples were probed with the anti-hnRNPK antibody, a band at roughly 60 kDa in all four species was seen. Two antibodies were used to probe for the WTAP protein. One was found to recognise the 50 kDa protein from all four species, whereas the other one was human-specific. Overall, signal was detected for all assayed proteins in all four species (**Figure 2.11**).

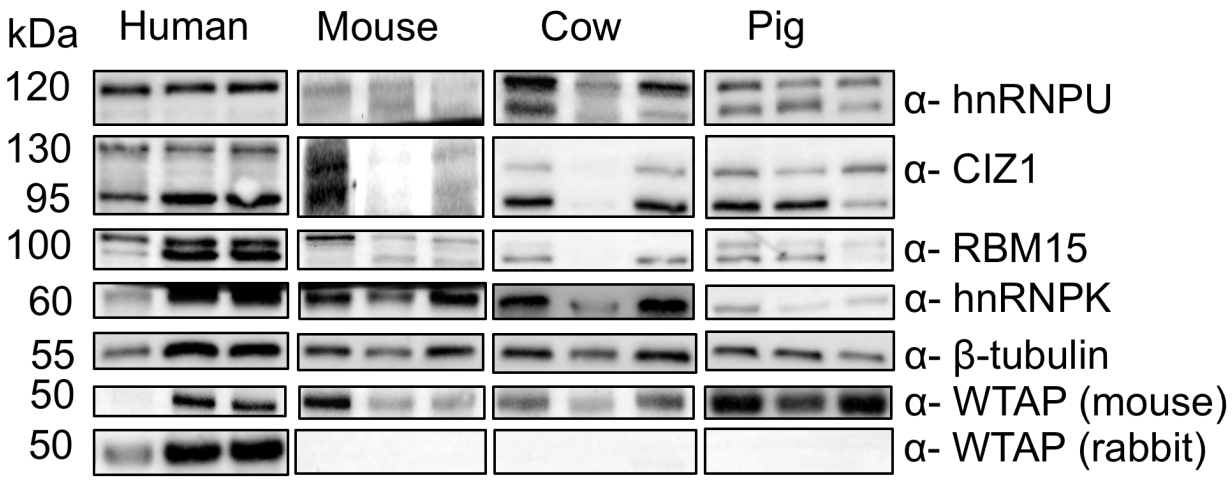


Figure 2.11. Putative XIST protein partners are present in cells and tissues of reproductive origin from human, mouse, cow and pig.

ISHIKAWA cells were used for human whereas whole uteri were collected for mouse. Endometrial tissue from the ipsilateral horn of the uterus was dissected for cow and pig. n=3 independent biological replicates shown per species. Equal amounts of protein were loaded from each animal sample per well (~30 µg) and wet-transferred into 0.45 um PVDF membranes. All antibodies were used at 1:1000 dilution in PBS with 0.5% Tween-20 (PBS-T). β-tubulin was used as a loading control. Blots containing three biological replicates from each species were probed separately for each protein per species as indicated by black boxes around blots.

2.4 Discussion

In this chapter, the aim was to examine whether protein partners previously described for mouse *Xist* are conserved and present in tissues from human, cow and pig, placental mammals with different implantation strategies. The rationale was that since eutherian mammals exhibit differences with regards to their early pregnancy events, timing and nature of XCI hallmarks as well as reproductive morphologies, the molecular effectors of XCI could have diverged as well, implicating species-specific *XIST* protein partners.

2.4.1 *XIST* and putative protein partner conservation could offer hints to cross species interactions

As a first step in understanding the conservation of *XIST*-protein complexes across placental mammals, the question that was addressed pertained to the conservation (or divergence) of a) the *XIST* lncRNA sequence and b) of its putative protein partners across mouse, human, pig and cattle. In line with previous reports (Yen et al., 2007), a moderate conservation of *XIST* (60-70% similarity) was found across diverse placental mammals (**Table 2.3**). It is worth noting that the RNA sequence of human *XIST* was found to be ~10% more similar to cow or pig than mouse *Xist*. From a genetics standpoint, this is despite humans and mice sharing an earlier common ancestor than human and cow or pig (Elsik et al., 2009) and does not fit the phylogenetic placement of the different species. From a reproductive standpoint, the placental morphology of a human is more similar to mouse than cow or pig but the timing of implantation and XCI onset are more similar between human and cow or pig than mouse (see **1.1 Main Introduction**). This could perhaps indicate a case of lineage-specific evolution whereby *XIST* in some species could have potentially diverged due to a difference in early pregnancy events. However, having too few informative sites in the multiple sequence alignment combined with a low power (only 4 species included) cannot be ruled out, thus this observation could be artefactual. It is also possible that if DNA sequences were used for the alignment, a different pattern could emerge.

However, considering previous reports citing rapid evolution of primary sequence for lncRNAs (Ulitsky et al., 2011), it is unlikely the whole sequence of a lncRNA would be crucial to eliciting its function. Instead, shorter interspersed regions could be responsible for a biochemical function. It has been hypothesized that lncRNAs can harbour repetitive elements which under the influence of evolutionary forces can obtain a function over time (Johnson and Guigo, 2014). Previously, several repetitive regions have been mapped on the *XIST* RNA and these have been shown to be of transposable element and retroviral origin (see **Main Introduction**). Recently several protein binding sites have been mapped on these repeats, associating *XIST*'s biochemical function to very small regions (see **Main Introduction**). In some respect repetitive sequences ascribe lncRNAs their 'modular' structure, which can link to function, as was the case with *UCHL1-AS* (Carrieri et al., 2012), *NEAT1* (Yamazaki et al., 2018) and *MALAT1* (Nguyen et al., 2020) to name a few. In fact, this was known to be true for *XIST* as well, which harbours several repetitive sequences along its sequence (**Figure 1.2**) (Brockdorff et al., 1992, Brockdorff, 2002, Brown et al., 1992, Nesterova et al., 2001, Yen et al., 2007). Therefore, to verify XIST repeat regions are more conserved when considered alone compared to the whole transcript, I used Clustal ω to align XIST repeat regions from human, mouse, cow and pig and generate % identity scores (an associated numeric estimate of similarity/conservation).

Comparing the sequences from each characterised repetitive region revealed hierarchical patterns of local conservation (**Table 2.3**). Delving into repetitive regions harboured across the XIST sequence, repeats A, B and D were more conserved than full-length *XIST* in human, mouse cow and pig. Repeats A and E were more similar across human, cow and pig whereas repeat B and D were more similar across mouse, cow and pig. In all comparisons, repeats from cow and pig were more similar between each other than to human or mouse. An expanded *XIST* repeat D can be seen in human and cow compared to mouse (**Figure 1.1**). The low similarity of human to cow can probably be explained due to cow repeat D having expanded in size and probably diverged in sequence compared to human and mouse. Whereas fewer repeat monomers found in mouse would have been retained in cow (and therefore these would appear more similar between mouse and cow), expansions of those in human and cow could have diverged in sequence and repeat number.

A different approach to the Clustal-ω alignment method used here that would allow for a more precise quantification of lncRNA identity scores would be the use of algorithms that evaluate similarity based on the abundance of short motifs (Kirk et al., 2018) or combine short motif assessment with synteny (Ross et al., 2021). These approaches do not rely on linear sequence conservation and can identify 'functional modules' that could be shared both across lncRNAs in a species but also across lncRNAs across organisms.

Despite a modest *XIST* RNA nucleotide sequence conservation across placental mammals, a high percentage of amino acid conservation (>70%) was found for the *XIST* putative protein partners assayed, Spen, Ciz1, Hnrnpk, Rbm15, Wtap, Lbr and Hnrnpu. Given the high % similarity between some *XIST* repeat regions (where protein binding sites have been mapped) and the high % similarity of most putative protein partners, it is likely some of these interactions could be conserved. For instance, *XIST* repeat A showing an 85% similarity between human and cow. In turn, the human and bovine SPEN proteins share almost identical RNA binding domains. Hence, one could hypothesise that an interaction between the two could be maintained in these species. Additional credence to this hypothesis could be derived from experiments demonstrating a temporally and spatially co-ordinated expression of the *XIST* RNA and of the putative protein partners.

2.4.2 *XIST* and putative protein partners are co-ordinately expressed

RT-qPCR was used to detect for the concomitant expression of *XIST* and putative protein partner mRNA in reproductive tissues from human, mouse, cow and pig. Due to the difficulty of obtaining human endometrial samples, the ISHIKAWA cell line was used as a proxy albeit at the cost of a low *XIST* expression (**Figure 2.8A**), consistent with a previous report in human patient endometrial cancer samples (Zhang et al., 2014). *XIST* was readily detectable by RT-qPCR of endometrial/uterine tissue from bovine and mouse respectively (**Figure 2.8B&C**). This difference in *XIST* abundance could reflect gene expression differences between *in vitro* immortalised cancer cell lines and *in vivo* animal tissues. Variability was seen with regards to *XIST* expression in the mouse and cow (**Figure 2.8B&C**), albeit it is known that *XIST* expression may vary from one individual to another from a study in humans (Tukiainen et al., 2017).

Tissue for bovine endometrium was sourced from females at the late luteal stage of the estrous cycle. Nonetheless, there is no evidence so far to support *XIST* expression is hormonally modulated (Tamm-Rosenstein et al., 2013). It is worth pointing out that ISHIKAWA cells and tissue from the endometrium represent terminally differentiated lineages and thus reflect the maintenance state of XCI, where the onset of XCI has passed and inactivation of X-linked genes is established. In mice, at that stage of XCI Xist has been shown to be dispensable (1990 studies).

One avenue to bypass low XIST expression associated with a cell line in future studies would be to perform expression profiling in human endometrial biopsy tissues (where available). This approach would also come with its own set of limitations though, such as restricted access and heterogeneous nature of tissue. Obtaining tissue samples from human endometrium is extremely difficult and getting consent from private clinics only happens in rare cases.

Western blotting was employed to probe for the expression of putative *XIST* protein partners at the same time in the same tissue in reproductive tissues from human, mouse, cow and pig. Identifying suitable antibodies that would recognise an antigen from all four species proved challenging and therefore the list of proteins assayed for was narrowed down to hnRNPU, CIZ1, RBM15, hnRNPK and WTAP. Proteins were detected in western blots for all of the above proteins in tissue/cells from all four species (**Figure 2.11**), indicative of the presence of the protein across human, mouse, cow and pig. This was consistent with transcriptomic evidence available from Expression Atlas release 37 (https://www.ebi.ac.uk/gxa/home). Expression of those proteins in human, mouse and pig was also confirmed from TISSUES (https://tissues.jensenlab.org/Search), a database which manually curates literature, proteomics and transcriptomics screens for tissue expression data (Palasca et al., 2018). However, none of these resources contained expression data from uterine-derived samples, making this the first report with uterine-specific expression data for these proteins across human, mouse, cow and pig.

Since parts of the same tissues were used for western blotting and RT-qPCR, a co-ordinate presence of *XIST* and putative proteins in the same tissue/cells at the same time raises the possibility that the two could interact. Given these interactions have been confirmed in mouse, the expression data from mouse *Xist* and mouse protein

partners seen here could theoretically serve as a positive control. Namely, there could be a correlation between the combined *XIST* and putative protein partner co-ordinate expression and a biochemical association inferred. Going forward, such a scenario will be assessed in human and cow, which represent placental mammals with distinctly different implantation timings and placental morphology.

Altogether, despite the aforementioned differences across placental mammals with different implantation strategies, there is a high conservation between putative *XIST* protein partners. This, in conjunction with the co-ordinate temporal expression of *XIST* and these proteins in select reproductive tissues of human, mouse, cow and pig, argues for a possibility of an interaction with *XIST*, which will be experimentally tested in the next chapter.

3. Chapter 3: Detection of a biochemical interaction between *XIST* and putative protein partners in human and cow reproductive tissues

3.1 Introduction

LncRNAs have diverse functions depending on where the lncRNA is localised and also on what molecule they interact with. Once in the specified subcellular compartment of the cell, lncRNAs have been shown to act via RNA-RNA pairing, RNA-protein interactions or both, in order to elicit their functions. LncRNA-bound proteins typically act as effector molecules being guided by lncRNAs to the correct cellular location, e.g. with transcription factors (Li et al., 2014) or chromatin remodellers (Dixon-McDougall and Brown, 2021). LncRNAs can also act as scaffolds for protein-protein interactions and assembly of multi-subunit protein complexes with transcriptional co-activating, i.e. p300-CBP (Postepska-Igielska et al., 2015) or repressing activities, i.e. PRC2 (Pintacuda et al., 2017a, Pandya-Jones et al., 2020). Another function that has been ascribed to this class of molecules is acting as 'decoys', sequestrating proteins from the available cellular pool (Kino et al., 2010) or inducing RNA-binding mediated inactivation of a protein's function (Long et al., 2017). Identifying protein partners of lncRNAs can therefore offer insight into the potential function of a lncRNA and the cellular pathways it participates in.

The XIST lncRNA functions primarily via RNA-protein interactions (**Section 1.7.2**). In order to understand what proteins *XIST* binds with, there are several methods that have been developed over the years that could be used to address which proteins interact with specific RNA sequences, (reviewed by (McHugh et al., 2014, Ramanathan et al., 2019), some with a specific focus on their suitability for lncRNAs (Cao et al., 2018). A widely adopted method for studying partners of lncRNAs includes *in vitro* transcription of a specific lncRNA sequence of interest with a tag, which can interact with magnetic beads, as a way to pull-down RNA-protein interactions formed from cell lysates (Rinn et al., 2007b). While this approach is protein-agnostic, not requiring the knowledge of proteins to include in the assay, the RNA will not be in its native state (both in terms of structure and chemical modifications), which might miss some protein partners due to a lack of modifications or hidden/altered structural motifs. The aforementioned method is biased for the

identification of proteins that are more abundant, implying that weak or infrequent interactions might be missed. Finally, proteins identified might not represent true '*in vivo*' interactions taking place within a subcellular compartment.

RNA-antisense purification with mass-spectrometry (RAP-MS) couples RNA pull-down with high-throughput proteomics to elucidate the endogenous interactomes of specific RNA targets (**Figure 3.1**). This method requires the sequence of the transcript of interest to be known so that complementary biotinylated DNA probes are designed. Multiple probes are employed to tile across the whole sequence of the RNA, enabling capture of all transcript isoforms and the identification of the complete set of protein partners, given the full-length RNA is pulled down. Additionally, the use of multiple probes significantly enhances enrichment and coverage of the transcript of interest over other targets. RAP is highly specific utilising covalent UV crosslinking and a denaturing environment allowing only the direct and specific protein interactors which bind the RNA endogenously to be captured, avoiding *in vitro* artifacts. Due to its numerous strengths, RAP has been employed to elucidate protein partners of several lncRNAs such as mouse *Xist* (McHugh et al., 2015), Survival Associated Mitochondrial Melanoma Specific Oncogenic Non-Coding RNA (*SAMMSON)* (Leucci et al., 2016), Non-Coding RNA Activated By DNA Damage (*NORAD)* (Munschauer et al., 2018) and *NEAT1* (Barra et al., 2020), to name a few.

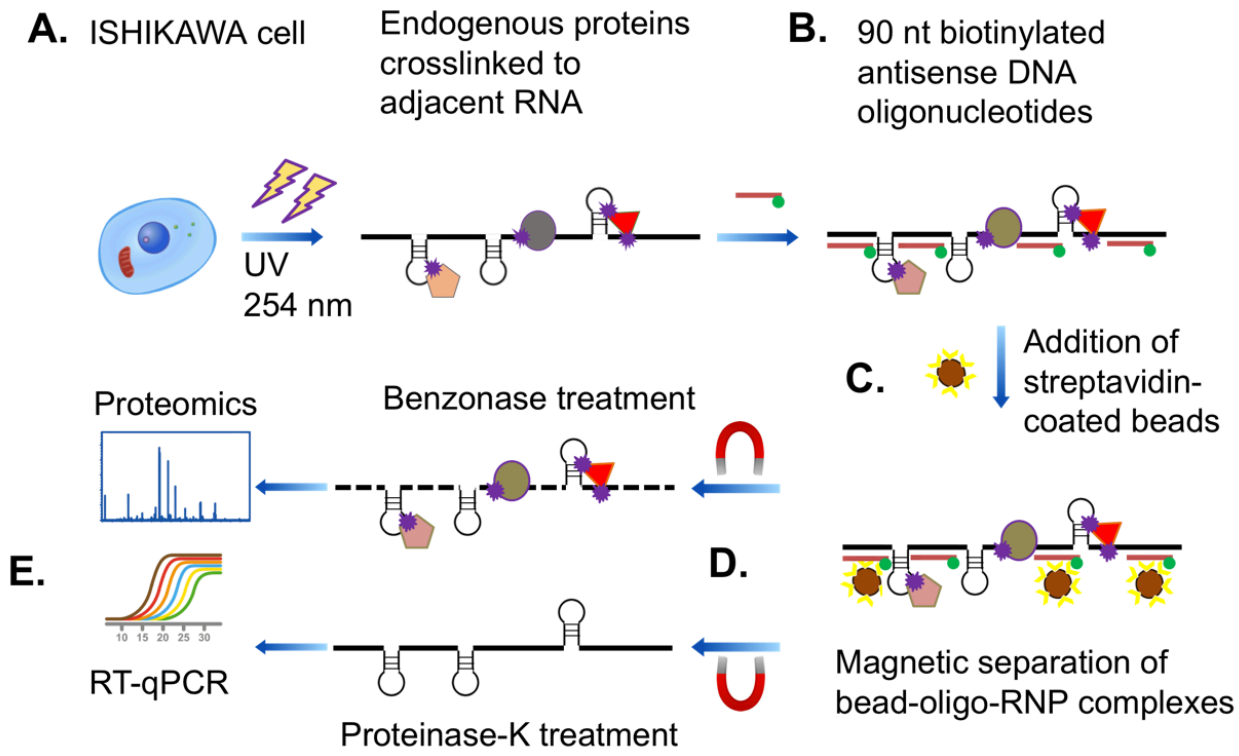


Figure 3.1. Overview of RNA antisense purification.

A) Cross-linking of RNA to adjacent proteins endogenously in intact cells by ultraviolet irradiation at 254 nm (lightning bolt). Covalent bond formation is promoted between aromatic rings from a single transcript's nitrogenous bases and candidate RNA-binding proteins' amino acids with aromatic ring structures (purple explosions). **B**) 5' biotinylated antisense DNA probes (orange lines with green circle) hybridise to the specific RNA of interest. **C**) Streptavidin-coated magnetic beads (brown circles with yellow chevrons) are used to recover the target RNA via strong interactions with the biotin group of hybridised antisense biotinylated oligos. **D**) Following magnetic separation of bead complexes, elution samples are split in two for RNA and protein analyses. **E**) Benzonase endonuclease treatment degrades nucleic acids releasing proteins enriched in elution samples which can be identified by mass-spectrometry, or western blotting if there is prior knowledge of specific partners. In parallel, RT-qPCR, or RNA-seq, can estimate the proportion of the RNA target levels enriched in the elution over the starting material.

It is important to treat protein partners identified with these approaches as candidates in a screening process, all of which would warrant further validation by other methods. Orthogonal validation of RNA-centric methods has frequently been employed via the use of RNA immunoprecipitation (RIP; **Figure 3.2**), where a protein of interest is used to perform a reciprocal pulldown experiment in order to detect the target RNA expected by RT-qPCR. The use of RIP requires the knowledge of a protein partner of an RNA and therefore is unsuitable for identifying new protein partners of an RNA. An advantage of RIP is that the protein of interest is at physiological levels when bound by the introduced antibody and that precipitated transcripts usually reflect physiological levels of the RNA present in the lysates. Hence, protein interactions with low abundance transcripts might be missed, depending on input cell number and signal-to-noise ratio (Darnell, 2010). Implementing this two-method approach for candidate screening provides additional proof of an interaction taking place rather than candidates being an artifact of a single technique. In fact, in all the aforementioned RAP-MS studies, RIP was also performed for candidate validation.

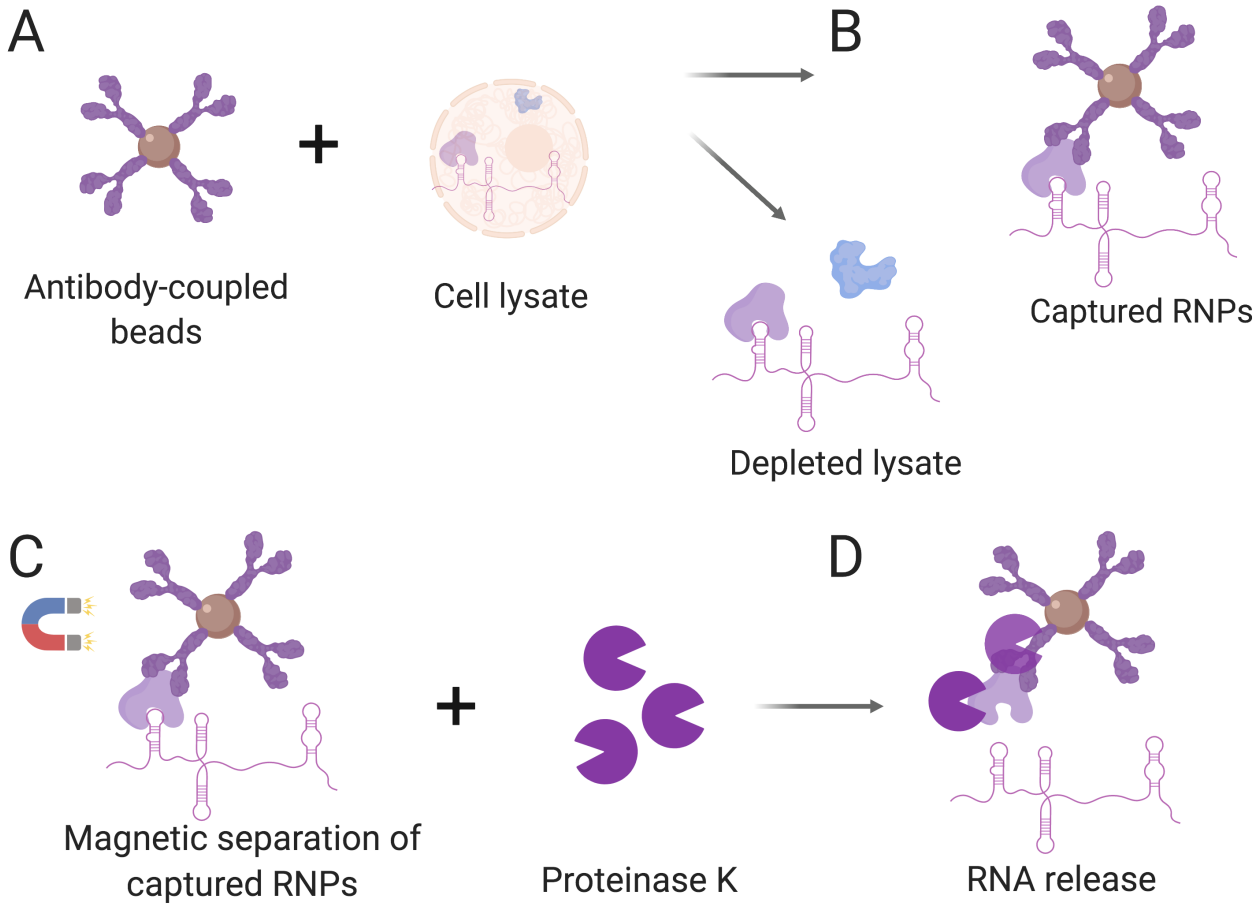


Figure 3.2. Overview of RNA immunoprecipitation.

A) The surface of protein A/B-coated magnetic beads is coupled with an antibody specific to a protein or an IgG non-specific control. Then, antibody-coupled magnetic beads are incubated with a cell lysate **B**) RNA-protein (RNP) complexes are formed and are captured by the beads during an incubation step. Unbound RNPs or non-RNA-binding proteins will remain in the lysate following capture. **C**) RNP complexes captured by beads can be specifically isolated from the cell lysate. **D**) Proteinase K treatment of bead-captured RNP complexes releases transcripts from their associated protein(s) and these proteins from the specific antibody or IgG control. Illustration created with Biorender.com.

In the previous chapter, the conservation of putative *XIST* protein partners was found to be high via multiple sequence alignment and all the candidate proteins studied were found to be co-ordinately expressed in cells/tissues from human, mouse, cow and pig via RT-qPCR and western blotting. Given the rapid divergence of lncRNA sequence over evolutionary time, this chapter aims to dissect whether the set of proteins required for the function of a lncRNA are shared across species. Alternatively, novel species-specific proteins could have arisen to compensate for the loss of specific interactions, if these were crucial to a cell's physiology. To distinguish between the two scenarios, the protein partners of the *XIST* lncRNA will be determined in human and cow. These two species were selected because they represent placental mammals with a different evolutionary trajectory, especially with regards to implantation strategies and reproductive morphologies. Despite these differences, human and cow are more similar compared to mouse or pig, since they are not litter-bearing species. Additionally, the *XIST* sequence between the two species is modestly conserved (~72.5% similar; **Table 2.3A**) and females of both species achieve dosage compensation via a single round of XCI which is random (each of the two X chromosomes has an equal chance of being inactivated).

3.2 Materials and Methods

3.2.1. Adapted RAP coupled to RT-qPCR for specific pulldown of human *XIST*

This method was previously described for mouse *Xist* (McHugh et al., 2015) and here it was adapted for human *XIST* with the following changes:

Probe design. Ten antisense DNA probes were designed to span human XIST. BLAST (megablast) was used in the Homo sapiens (taxid:9606) database to exclude probes that had a complete 30 base pair match or an incomplete (90%) identity 60 base pair match with another transcript or genomic region (McHugh and Guttman, 2018). Probes were also excluded if they contained more than 30 bases in common with a repeat annotation from RepeatMasker. The minimum E-value of probes accepted was 2e-37 whilst ensuring no other gene other than XIST would show up on the list of potential targets. Probes were HPLC-purified 90-nt DNA oligonucleotides (QIAGEN) with 5' biotin (**Table 3.1**).

UV crosslinking and lysate preparation. Given a high cell number was required for this method, the ISHIKAWA cell line was used for lysate preparation (cultured as previously described in Section 2.2.2). Cells were grown in 500 mm^2 dishes which were treated with 200 uM of 4-thiouridine (4sU) per dish overnight (~120 million cells per dish). The next day, cells were washed twice with 1x PBS, before irradiating with 0.8 Joules/cm^2 at 365 nm on ice inside a UV crosslinker (CL-1000, UVP). Cells were scraped in 1x PBS using a rubber cell scraper and pelleted by centrifugation at 1,000 xg for 5 minutes at 4^{o}C. Aliquots of 40 million cells were prepared in 2 ml tubes by centrifuging cell suspensions and discarding the supernatant. To ensure equal lysis efficiency, pellets were resuspended and pooled in Cell lysis buffer 1 [as described before (McHugh et al., 2015); 10 mM HEPES pH7.2, 20 mM KCl, 1.5 mM $MgCl_2$, 0.5 mM EDTA, 1 mM Tris(2-carboxyethyl)phosphine (TCEP), 0.5 mM phenylmethylsulfonyl fluoride (PMSF) and EDTA-free protease inhibitor cocktail] and incubated on ice for 20 minutes, inverting every 5 minutes. Cell suspensions were centrifuged at 3,300 xg for 10 minutes at 4^{o}C and pellets were resuspended and pooled in Cell lysis 1 with 0.1% dodecyl maltoside (DDM) and incubated on ice for 20 minutes. In 1 ml aliquots, lysates were homogenised using a Dounce homogeniser with a tight pestle B using 25 strokes. After centrifugation at 3,300 xg for 10 minutes

at 4°C, pellets were resuspended in 600 µl of lysis buffer 2 [20 mM Tris pH 7.5, 50 mM KCl, 1.5 mM MgCl2, 2 mM TCEP, 0.5 mM PMSF, 0.4% sodium deoxycholate, 1% DDM, and 0.1% N-lauroylsarcosine (NLS)] and pooled in a 15 ml falcon tube. To clarify the lysate, it was passed 4 times through a 23' gauge needle (0.64 mm diameter) and 5-7 times through a 27' gauge needle (0.4 mm diameter) and syringe. To remove genomic DNA, the lysate was mixed with 330 U of Turbo DNase per 200 million cells (McHugh et al., 2015) in a DNase buffer (the final 1× Buffer contains 2.5 mM $MgCl_2$ and 0.5 mM $CaCl_2$) and incubated at 37°C for 1 hour while shaking at 850 RPM. An equal volume of 2x HB buffer (the final 1× Buffer contains 10 mM Tris pH 7.5, 5 mM EDTA, 500 mM LiCl, 0.5% DDM, 0.2% SDS, 0.1% deoxycholate, 4M urea, 2.5 mM TCEP) was mixed with the lysate and after aliquoting lysates were snap-frozen in LN_2 and stored at -80°C.

Capture of RNA target. Lysates were diluted with 1 volume of 1x Hybridisation buffer (10 mM Tris–HCl pH 7.5, 5 mM EDTA, 500 mM LiCl, 0.5% DDM, 0.2% SDS, 0.1% sodium deoxycholate, 4 M urea, and 2.5 mM TCEP; Thermofisher). Per 200 million cells, 0.6 ml of Dynabeads™ MyOne™ Streptavidin C1 magnetic beads (ThermoFisher), were washed four times with 10 mM Tris-HCl pH 7.5 and twice with 1x Hybridisation buffer. To pre-clear lysates, beads were incubated with each lysate at 37°C for 30 minutes while shaking at 1000 RPM to remove endogenously biotinylated proteins. RAP probes were denatured at 85°C for 3 minutes, mixed with the pre-cleared lysate and incubated at 66°C for 2 hours while shaking at 1100 RPM. Unless otherwise stated, 1.2 ml of magnetic beads per 200 million cells (previously pre-washed as described above) were added for capture and incubated at 66°C for 30 minutes while shaking at 1100 RPM.

Elution and analysis of RNA samples. To elute captured RNA and proteins, beads were washed four times with two bead volumes of 1x Hybridisation buffer at 66°C for 5 minutes. To each reaction, 500 µl of 1x Hybridisation buffer was added to beads and 30% of this (150 µl) was used for RNA analysis. This aliquot was resuspended in 100 µl of NLS Elution Buffer (20 mM Tris–HCl pH 8.0, 10 mM EDTA, 2% NLS, and 2.5 mM TCEP) and heated at 95°C for 2 minutes to elute RNA from beads. Next, it was treated with 60 U of Turbo DNase for 1 hour at 37°C and subsequently with 600 µg of Proteinase K (20 mg/ml, ThermoFisher) for 1 hour at 55°C.

To extract RNA, the Quick-RNA Miniprep kit (Zymogen) was used according to the manufacturer's instructions. Briefly, 600 µl of RNA lysis buffer were mixed with each sample and these were cleared by passing through a Spin-Away Filter column by centrifugation at 16,000 xg for 1 minute at room temperature. The flow-through was mixed with 1 volume of 100% ethanol and passed through a Zymo-Spin column by centrifugation at 16,000 xg for 30 seconds at room temperature. To each column, 400 µl of RNA Prep Buffer were added and centrifuged at 16,000 xg for 30 seconds at room temperature. To each column, 700 µl of RNA Wash Buffer were added and centrifuged at 16,000 xg for 30 seconds at room temperature. To each column, 400 µl of RNA Wash Buffer were added and centrifuged at 16,000 xg for 120 seconds at room temperature. Columns were transferred to microfuge tubes and 50 µl of DNase/RNase-Free Water were added directly to the column and incubated at room temperature for 5 minutes. Columns were centrifuged at 16,000 xg for 30 seconds at room temperature. RNA samples were subjected to quantification using the NanoDrop1000 instrument.

RNA enrichment analysis via RT-qPCR. For the cDNA synthesis reaction, the qSCRIPT kit (Quanta/VWR) was used according to the manufacturer's instructions with 1 µg of RNA, or less of each sample, equally used across all samples. For RT-qPCR, the PowerUp SYBR Green Master Mix kit was used (ThermoFisher). All samples were diluted accordingly so that 12.5 ng of cDNA were loaded per 20 µl RT-qPCR reaction and primers were used at 0.3 µM final concentration (for both forward and reverse primer) per reaction (**Table 3.2**). All primers were designed to anneal at 60^{o}C. Standard curves were run on each primer set using a cDNA pool from all samples and primer efficiencies were calculated from the standard curves. All reactions were performed in triplicate in 96-well white plates. No-template controls were run for each primer pair and no-reverse-transcription reactions were run with each RNA sample used to make the cDNA. The thermocycler used was the CFX Connect Real-Time PCR Detection System (BioRad, UK) according to the PowerUp SYBR Green Master Mix thermocycling conditions (**Table 3.3**). To calculate fold-enrichment of a specific transcript over the input, the following formula was used: $2^{-(Ct_{elution} - Ct_{input})}$.

Elution and analysis of protein samples. Western blotting of protein samples and

antibodies used were listed in chapter 2 (**Table 2.2**). Membranes were developed using a ChemiDoc XRS+ imaging system (BioRad, UK). For silver staining, samples run under denaturing conditions on SDS-PAGE gels were silver stained with the ProteoSilver Plus Silver Stain kit, according to the manufacturer's instructions (Sigma-Aldrich). Briefly, gels following electrophoresis were immersed in 100 ml of fixing solution (50% ethanol, 40% acetic acid in ddH2O) and incubated overnight. All incubation steps were performed whilst shaking. The next day, gels were washed with 100 ml of 30% ethanol for 10 minutes. Then, gels were washed in 200 ml of ddH2O for 10 minutes before incubating with 100 ml of 1% v/v sensitizer solution for 10 minutes. Next, gels were washed twice with 100 ml of ddH2O for 10 minutes each. Subsequently, gels were equilibrated with 100 ml of 1% silver solution for 10 minutes and then gels were washed with 200 ml of ddH2O for 1.5 minutes. Finally, gels were developed using 100 ml of developing solution (5% v/v Proteosilver developer 1, 0.1% Proteosilver developer 2 in ddH2O) for a minimum of 3 minutes up to 12 minutes or until bands of interest become visible. To stop the reaction, 5 ml of stop solution were added and after 5 minutes, the gel was washed with 100 ml of ddH2O and stored at room temperature in ddH2O.

Table 3.1. List of antisense probes used for RAP hybridisation against human XIST (ENSG00000229807.10).

Probe number	Target region	Tm (°C)	Sequence (5'-3')
1	Exon 1	72.1	GCTTTGGCAGAGAATGACTCTGCAGTTAAGCTAAGGGCGTGTTCAGATTGTGGAGGAAAAGT GGCCGCCATTTTAGACTTGCCGCATAAC
2	Exon1	73.8	TGGTGTACCGCCCACTGGGAGACATACACGTGGCCCCTCCACTTCTTTCTCCTGACTGGCTA AAGACAGCTGCGAAGTGCCATGCTAATT
3	Exon1	71.2	CAGTAGGGTGCCTTTGTTTAATAGGCAAAGCTATGACAATTGGGACTGAGCATTTTAACTGTC CAACAAAAGACGGGTTGTCTGCGACCC
4	Exon1	70.2	CCATGCAATAAAGCAAAGAGGGTGTGATAGGTCAGAAACCCAAGTCTAATTGAAGGACCATT GACAACTGCAATTACATGCCATCTACAG
5	Exon1	71.3	ATCACCACATGGTTCATCCTAATTAAACAAAGTTCTACCTTCTCACCCTCCATTTGCAGTATAC CAGGGTTGCTGACCCCCTAAGTCCCC
6	Exon1	71.6	CCATCAGTCCAAGATCTCCCTACCACTTTGGTGTGTTGGTGCAGTGTTGACTATGAAAAGCAG GCCTGAACTAGGTGGATAAGCCTTCAC
7	Exons 3-4	70.6	AAAGATCTTCCTCAGAAGAATAGGCTTGTTGTTTTACAGTGTTAGTGATCCATTCCCTTTGACG ATCCCTAGGTGGAGATGGGGCATGAG
8	Exon6	71.7	AGTGGCCAAATAATTTGGTGGACTGTGCCAACGCTACTCCTGGGTTTAATACCCATCTCTAGG CTTAAAGATGAGAGAACCTGGGACTGT
9	Exon6	68.1	GAGAGATGAGGGCATTAGATCACTGACAGCTGAAGATAGAAGAACATCTTTGGTTTGATTGTT TAAATAATATTTCAATGCCTATTCTCT
10	Exon6	68.3	GTAACTAAAAGCCTTCCTTTCACAGTTTCTGGCATCACTACCACTACTGATTAAACAAGAATAA GAGAACATTTTATCATCATCTGCTTT

Table 3.2. List of primers used for RT-qPCR assessment of transcript enrichment in pull-down RAP assays in human.

	Primer		
Target transcript	Orientation	Location	Sequence (5'-3')
ACTB	Forward	Exon-exon junction	GCACGGCATCGTCACCAAC
	Reverse	Exon	GTCCAGACGCAGGATGGC
U2	Forward	Exon	GGAGCAGGGAGATGGAATAGG
	Reverse	Exon	GCACCGTTCCTGGAGGTA
XIST-ex1	Forward	Exon	GCATAACAGCAGTGGGACTGAC
	Reverse	Exon	AGGTAGTTCACACTATCTAGGAGC
XIST-ex2-3	Forward	Exon-exon junction	GGCTCCTCTTGGACATTCTGAG
	Reverse	Exon	AGCTTGGCCAGATTCTCAAAG
XIST-ex4-5	Forward	Exon-exon junction	CTCCAGGGGAAAAGCTCACTAC
	Reverse	Exon	GAAGAGCTTGACGTGTGGTG
XIST-ex6	Forward	Exon	GCTCGGAACTACATGCCC
	Reverse	Exon	ACAGGACTTTATCTCTCTACTCAGC

Table 3.3. RT-qPCR thermocycling conditions.

Step	Temperature (°C)	Time	Cycles
Uracil-DNA	50	2 min	
Dual-Lock DNA polymerase	95	2 min	
Denaturation	95	3 secs	40
Annealing/extension	60	30 secs	
Melt curve	65-95, increasing by 0.5	5 secs	

3.2.2. Primary endometrial bovine stromal cell isolation and culture

Unless otherwise stated, all reagents were purchased from Sigma (UK). Some of the material described here was also generated by the following people: Haidee Tinning, Stefania Moutevelidou and Irene Malo Estepa.

Bovine stromal cell isolation was performed as previously described (Tinning et al., 2020). Uteri from non-pregnant cows (*Bos taurus*) were selected on the basis of corpus luteum morphology as previously described (Ireland et al., 1980). Endometrial tissue from the ipsilateral uterine horn (the horn attached to the ovary with the corpus luteum) was dissected in strips from the underlying myometrium and washed with 25 ml of endometrial wash solution (DPBS, 1% Antibiotic-Antimycotic (ABAM); ThermoFisher, UK). Next, tissue strips were washed twice with 25 ml HBSS (-$CaCl_2$, -$MgCl_2$), 1% ABAM). Subsequently, tissue was cut into 3-5 mm pieces and washed with 25 ml of HBSS before incubating in 50 ml of filter-sterilized digestion solution (50 mL HBSS (-$CaCl_2$, -$MgCl_2$), 25 mg collagenase II, 50 mg BSA, 125 µL 4% DNase I, 500 µL 0.0175% trypsin in HBSS) in a rocking hot box for 1 hour at 37°C. The solution with the tissue pieces was filtered through a 100 µm mesh cell strainer above a 40 µm strainer into a 50 ml sterile falcon containing 5 ml of stop solution (HBSS with 10% charcoal-stripped FBS). The flow-through was centrifuged at 500 xg for 5 minutes and the pellet was vortexed with 5 ml of sterile H_2O at 37°C to lyse erythrocytes. Immediately after, 45 ml of stop solution was added and the solution was centrifuged at 500 xg for 5 minutes. The pellet was resuspended with 10 ml of complete medium (Gibco RPMI 1640, 10% charcoal-stripped Fetal Bovine Serum, 1% ABAM) and transferred into a T-75 flask in a 37°C/5% CO_2 incubator. Two days later, medium was aspirated, washed once with PBS and 10 ml of fresh medium was added. Once the stromal cells had reached ~90% confluency, they were harvested by trypsinization (5 ml 0.025% trypsin in PBS), counted, snap-frozen and stored at -80°C.

3.2.3. RIP coupled to RT-qPCR for pulldown of putative *XIST* protein partners in human and cow

Lysate preparation. Cell pellets of 40 million ISHIKAWA or bovine stromal cells were snap-frozen and stored at -80°C. Cells were lysed with 100 µl with RIP lysis buffer (Magna RIP kit, Merck) supplemented with 10 U RNaseIn and 0.5 µl protease-inhibitor cocktail; per 20 million cells while incubating on ice for 15 minutes, with frequent inversion of samples every three minutes to mix. Lysates were then stored at -80°C. Lysates were passed through a 27' gauge needle and syringe 3-5 times to homogenise. Pooled lysates were centrifuged at 17,000 xg for 1 minute to remove cell debris and the supernatant was kept. From that supernatant, 10 µl were kept for RNA and 10 µl for protein downstream analyses ("input").

Coupling antibody to beads. In each RIP reaction, 50 µl of magnetic bead suspension (provided in the kit) were mixed with 500 µl of RIP Wash Buffer by vortexing for 3 seconds and the tubes were placed on a magnet for 1 minute after which the supernatant was discarded. This wash step was repeated once more. The beads were resuspended in 100 µl of RIP Wash Buffer and mixed with ~5 µg of specific antibody (~1.2 µg for hnRNPK) or 5 µg of the IgG negative control. Antibody-bead mixtures were incubated for 30 minutes at room temperature while rotating. The tubes were centrifuged briefly and placed on a magnet after which they were washed with 500 µl of RIP Wash Buffer. A total of three washes were performed.

Capture of RNA-protein complexes. For each RIP reaction, 900 µL of RIP Immunoprecipitation Buffer (Magna RIP kit, Merck) supplemented with 200 U RNaseIn was mixed with 100 µl of the lysate and this mixture was used to resuspend the antibody-coupled beads. Due to tissue availability and kit cost, replicates for IgG reactions were not always run in parallel with every replicate for a specific antibody, even though they were used with the same lysates. Reactions were incubated overnight at 4°C while rotating. The next day, reactions were briefly centrifuged and the supernatant was kept for RNA and protein downstream analyses ("depleted lysate"). RNP complexes on beads were washed with 500 µl of RIP Wash Buffer for a total of six times. From the sixth wash, 50 µl (10% of the elution) were kept for RNA and protein downstream analyses ("elution"). For mass spectrometry analyses, 500 µl (100% of the elution) were kept.

Elution of RNP complexes after RIP for RNA analysis. To elute RNP complexes, the input and elution samples were treated with 150 µl of Proteinase K buffer and samples were incubated at 55ºC for 30 minutes while shaking at 1100 RPM. Tubes were briefly centrifuged and the supernatant was kept for downstream RNA analyses.

For each sample, 250 µl of ddH2O was added and mixed with 400 µl of phenol:chloroform:isoamyl alcohol while vortexing for 15 seconds. Samples were centrifuged at 17,000 xg for 10 minutes at room temperature. The aqueous layer was retrieved and mixed with 400 µl of chloroform. Samples were vortexed for 15 seconds and centrifuged at 17,000 xg for 10 minutes at room temperature, after which the aqueous layer was retrieved and mixed with 920 µl of Ethanol Precipitation buffer (provided by the kit and supplemented with 2 µl of GlycoBlue; ThermoFisher). Samples were kept at -80ºC for 3 hours to overnight. Samples were centrifuged at 17,000 xg for 30 minutes at 4ºC and the supernatant was discarded. The pellets were washed with 1 ml of 75% ethanol and centrifuged at 17,000 xg for 15 minutes at 4ºC and the supernatant was discarded. The pellet was resuspended in 15 µl of RNase-free water and mixed before measuring the RNA concentration with a NanoDrop1000 instrument.

RNA enrichment analysis via RT-qPCR. For cDNA synthesis, 500 ng of RNA from ISHIKAWA or bovine stromal cells were used per sample to make cDNA (for SPEN, 1000 ng were used per sample). For ISHIKAWA samples, cDNA synthesis and RT-qPCR were performed as described above (Section 3.2.1). Primers used for human *XIST* and negative control transcripts were the same as described above for RAP (**Table 3.2**). For bovine samples, cDNA synthesis and RT-qPCR were performed as described in a previous chapter (**3.2.4 Methods**). Primers used for bovine *XIST* and negative control transcripts are listed (**Table 3.4**). To calculate fold-enrichment of a specific transcript over the input, the same formula was used as described above for RAP (Section 3.2.1).

Table 3.4. List of primers used for RT-qPCR assessment of transcript enrichment in RIP from cow.

	Primer		
Target transcript	Orientation	Location	Sequence (5'-3')
ACTB	Forward	Exon-exon junction	CGCCATGGATGATGATATTGC
	Reverse	Exon	AAGCCGGCCTTGCACAT
RPL19	Forward	Exon	GGGTATAGGTAAGCGAAAGGG
	Reverse	Exon	TCACGGTATCGTCTAAGCAGC
XIST-exon 1	Forward	Exon	CTGCTCTTCTGCGTTGTGG
	Reverse	Exon	CAGGGATTCCTCTTCTGCC
XIST-exon 5	Forward	Exon-exon junction	CCAATCATCATTCTGGACCCTC
	Reverse	Exon	CTGTCAATTAGCAGGCAGAGC

Elution of RNP complexes after RIP for protein analysis. Samples kept from the last wash of the captured RNPs were placed on a magnet, beads were resuspended with 1x of Laemmli buffer (described in Section 2.2.6) and heated at 95^{o}C for 5 minutes.

For Western blotting, the equivalent of 500,000 cells for input and depleted lysate as well as the whole elution samples (20 µl) were mixed with 1x Laemmli buffer and heated at 95^{o}C for 5 minutes before loading in SDS-PAGE. Samples were run under denaturing conditions on SDS-PAGE and were blotted as described above (Section 3.2.1).

For proteomics analysis, 100% of RIP elutions (20 µL samples) were sent to the Biomolecular Mass Spectrometry Facility at Leeds for protein identification. Samples were injected onto an in house-packed 20cm capillary column (inner diameter 75µm, 3.5µm Kromasil C18 media). An EasyLC nano liquid chromatography system was used to apply a gradient of 4–40% ACN in 0.1% formic acid over 30 min at a flow rate of 250 nl/min. Total acquisition time was 60 minutes including column wash and re-equilibration. Separated peptides were eluted directly from the column and sprayed into an Orbitrap Velos Mass Spectrometer (ThermoFisher Scientific, Hemel Hempstead) using an electrospray capillary voltage of 2.7 kV. Precursor ion scans were acquired in the Orbitrap with resolution of 60000. Up to 20 ions per precursor scan were selected for fragmentation in the ion-trap. Dynamic exclusion of 30 s was used. Peptide MS/MS data were processed with PEAKS Studio X+ (Bioinformatic Solutions Inc, Waterloo, Ontario, Canada) and searched against the Uniprot databases (release 2020_01). Carbamiodomethylation was selected as a fixed modification, variable modifications were set for oxidation of methionine and deamidation of glutamine and asparagine. MS mass tolerance was 5 ppm, and fragment ion mass tolerance was 0.3 Da. The peptide false discovery rate was set to 1% for human and 0.1% for bovine samples.

3.3 Results

The aim of this chapter was to identify, compare and contrast protein partners of *XIST* in species with different pregnancy morphologies and XCI timing (human and cow).

3.3.1.1. Assessing the steady state abundance of different *XIST* regions

To establish an RNA-protein co-precipitation method that will allow for the experimental characterisation the protein interactome of human and cow *XIST*, RAP was chosen as a suitable biochemical pulldown method. RAP is based on the purification of proteins (and/or RNAs) interacting with a specific target RNA pulled-down using biotinylated antisense complementary DNA probes tilled across its whole sequence and streptavidin-coated beads from *in vivo* crosslinked cells (see **Figure 3.1**).

Human *XIST* is a lncRNA comprised of 6 exons, across a transcript length of 19.2 kbp. Its genomic structure is unusual in that exons 1 (~11 kbp) and 6 (~7 kbp) are the longest and intervening exons vary in length from 60-200 bp. In Chapter 2, the steady-state abundance of *XIST* was determined in the ISHIKAWA cell line using a single primer set targeting *XIST's* first exon. To get a sense of whether the detected expression level of *XIST* varied depending on which of its regions was amplified, primers were designed across different exons and RT-qPCR performed in ISHIKAWA cells (**Table 3.2** and **Figure 3.3**). Primers were designed to span exon-exon junctions to avoid the amplification of potential genomic DNA available in the sample as well as they were designed to target regions far away from repetitive regions harboured in the first and last exons of human *XIST*. The exons selected were not of particular note, rather the ones where appropriate primers could be designed.

In this experiment, RNA from ISHIKAWA cell lysates was isolated and following cDNA synthesis, RT-qPCR was performed where the new primers were tested. The aim was to characterise expression levels of human *XIST* available in the ISHIKAWA

cell line. Signal was detected for all primers designed, revealing the steady state abundance of human XIST in the ISHIKAWA cell line (**Table 3.5** and **Figure 3.3**).
In summary, this results section exhibits that primers designed for different regions of human *XIST* are efficient and the abundance of *XIST* can be quantified by RT-qPCR across its length.

Table 3.5. **Amplification efficiencies for primers targeting various human *XIST* regions in RT-qPCR**. Total RNA was extracted from ISHIKAWA cells. Primer efficiencies were calculated as percentages using the formula $[1-10^{(-1/slope)}]*100$ where the slope was derived from a plot of the Ct values for each primer across serial dilutions on the y axis and the log of 1 over the dilution factor on the x axis.

Primer set	Target	Amplification efficiency (%)
XIST-exon 1	Exon	96.1
XIST-exons 2-3	Exon-exon junction	94.1
XIST-exons 4-5	Exon-exon junction	127.5
XIST-exon 6	Exon	85.5

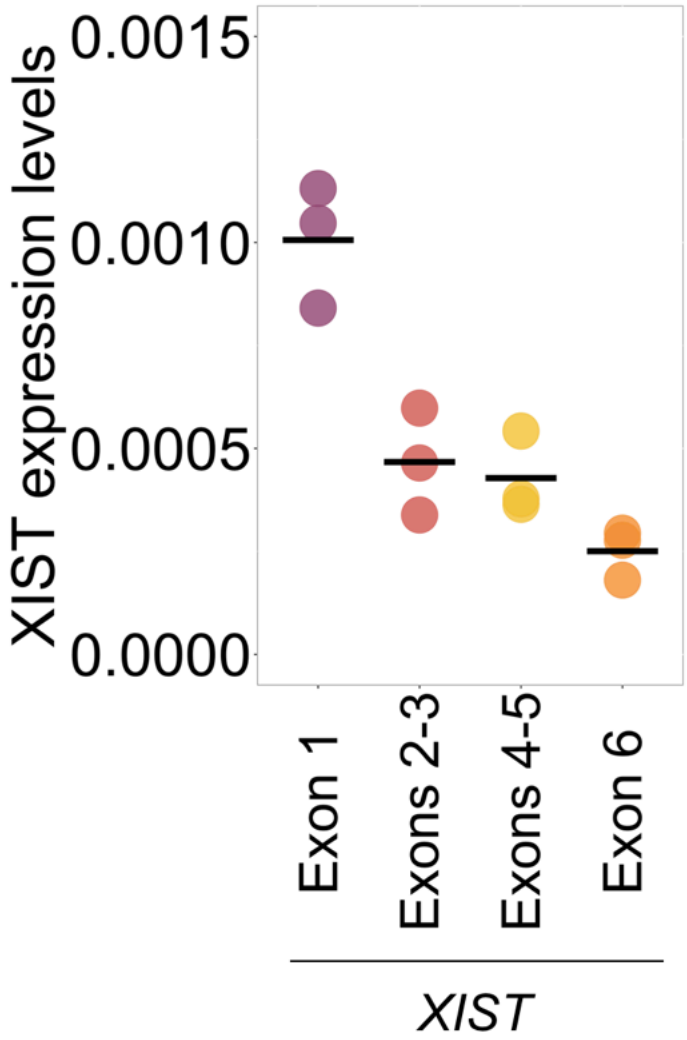


Figure 3.3. Abundance of human XIST across its length in the ISHIKAWA cell line.

Baseline human *XIST* lncRNA expression was established using RT-qPCR with a standard curve and starting quantity was estimated from a pool of 12.5 ng cDNA-equivalent prepared from whole cell ISHIKAWA lysates. Primers were designed to amplify different parts of *XIST*. XIST expression levels shown were normalised to the expression of *ACTB* mRNA and were $\log_{10}$ transformed. Three technical replicates were performed for each of three independent biological repeats.

3.3.1.2. Examination of the extent of RNA damage caused by UV crosslinking

In the previous section (Section 3.3.1.1), the steady state abundance of *XIST* was measured in intact, untreated ISHIKAWA cells. The end goal here was to perform the RAP protocol for the pulldown of as much of the available *XIST* RNA as possible to enable the identification of its protein partners. UV crosslinking is a crucial part of the RAP protocol which enables direct RNA-protein partner capture. UV crosslinking has traditionally been used as a “zero-distance” crosslinking method to covalently link RNAs with proteins in close proximity (Moller et al., 1978, Harrison et al., 1982, Hockensmith et al., 1986, Brimacombe et al., 1988, Pashev et al., 1991), ensuring only direct protein partners of a given RNA are identified. However, UV treatment of cells has been described to have a harmful effect on RNA (Wurtmann and Wolin, 2009, Urdaneta et al., 2019).

In this next experiment, the aim was to assess the RNA damage sustained by the RNA of interest, *XIST* and a positive control of high abundance, *ACTB*. Cells were irradiated on a dish with 0.8 Joules/cm^2 of UV at 254 nm and lysed. Following RNA isolation, RT-qPCR was performed to quantify the amount of *XIST* and ACTB pre- and post-UV-C irradiation at 254 nm. Approximately, ~1000 times less *XIST* RNA was observed in samples crosslinked with UV at 254 nm versus non-crosslinked samples based on RT-qPCR starting quantity values (**Figure 3.4**). *ACTB* abundance was more prominently affected being >1,000 times less abundant in samples crosslinked with UV at 254 nm versus non-crosslinked samples (**Figure 3.4**)., consistent with a detrimental side effect of UV crosslinking on RNA integrity. A compromised RNA integrity due to UV treatment was confirmed by agarose gel electrophoresis of total RNA (**Figure 3.5**).

UV irradiation at 365 nm in combination with the incorporation of a uracil analogue, 4-thiouridine (4sU), has previously been shown to be less damaging for RNA (Strein et al., 2014). To circumvent further loss of an already lowly expressed *XIST* lncRNA in the ISHIKAWA cell line upon UV treatment, 4sU treatment of ISHIKAWA cells was coupled to short-wavelength UV-A crosslinking (at 365 nm) to mitigate UV damage and attain an improved *XIST* abundance.

In this experiment, cells were treated with 200 μM of 4sU overnight and irradiated with 0.8 Joules/cm^2 of UV at 365 nm. Following lysis and RNA isolation, RT-qPCR was performed to quantify the amount of *XIST* and ACTB pre- and post-UV-A irradiation at 365 nm. The aim of this experiment was to examine for an improved *XIST* abundance with the combination of 4sU treatment and irradiation at a longer wavelength.

qRT-PCR of RNA from these samples revealed that UV treatment at 365 nm in the presence of 4sU still resulted in a ~100-fold reduction of *XIST* compared to non-crosslinked samples. There was also a 10-fold improvement of XIST abundance when samples were UV crosslinked at 365 nm in the presence of 4sU, compared to samples treated with UV at 254 nm (**Figure 3.4**). The mitigation on RNA damage at UV crosslinking at 365 compared to 254 was also detected in ACTB levels. Compared to non-crosslinked samples, UV treatment at 365 nm in the presence of 4sU reduced the loss of ACTB to ~10-fold (~2.6 Ct shift), versus >1000-fold reduction in samples treated with UV at 254 nm. The integrity of RNA was also assessed by agarose gel electrophoresis. An improvement in RNA integrity was also reflected of total RNA from these samples (**Figure 3.5**). In summary, comparing and contrasting the two UV crosslinking conditions, treatment with 4sU coupled to UV irradiation at 365 nm mitigated UV damage and improved *XIST* abundance.

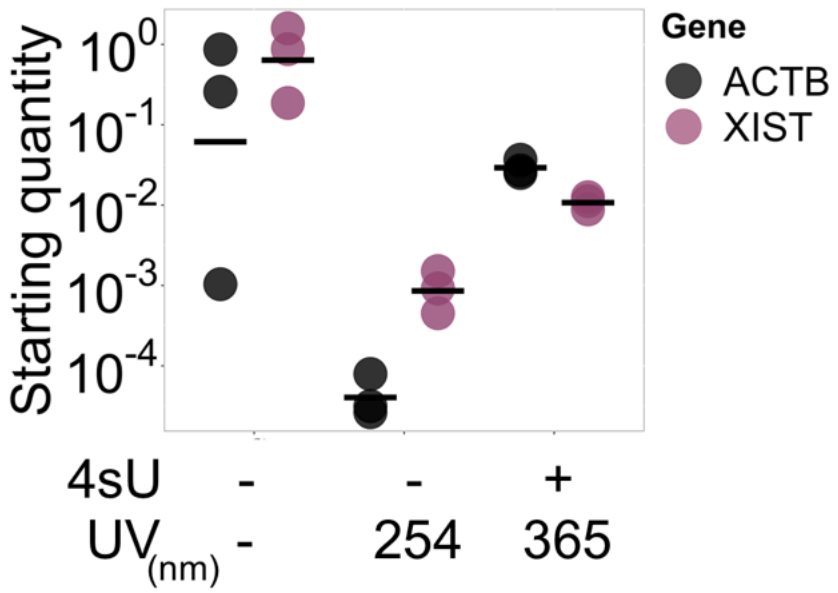


Figure 3.4. XIST abundance shifts following UV treatment.

RT-qPCR measurements of starting quantity for *XIST* and *ACTB* pre- and post- UV treatment to determine detrimental effect of UV on RNA abundance. Results reflect two technical replicates for each of three biological replicates. Two UV crosslinking treatments are shown: at 254 nm and at 365 nm. UV crosslinking at 365 nm was coupled to treating cells with 200 μM of 4sU overnight prior to irradiation. All irradiation steps delivered a 0.8 J/cm^2 dose. UV nm refers to wavelength. 4sU, 4-thiouridine

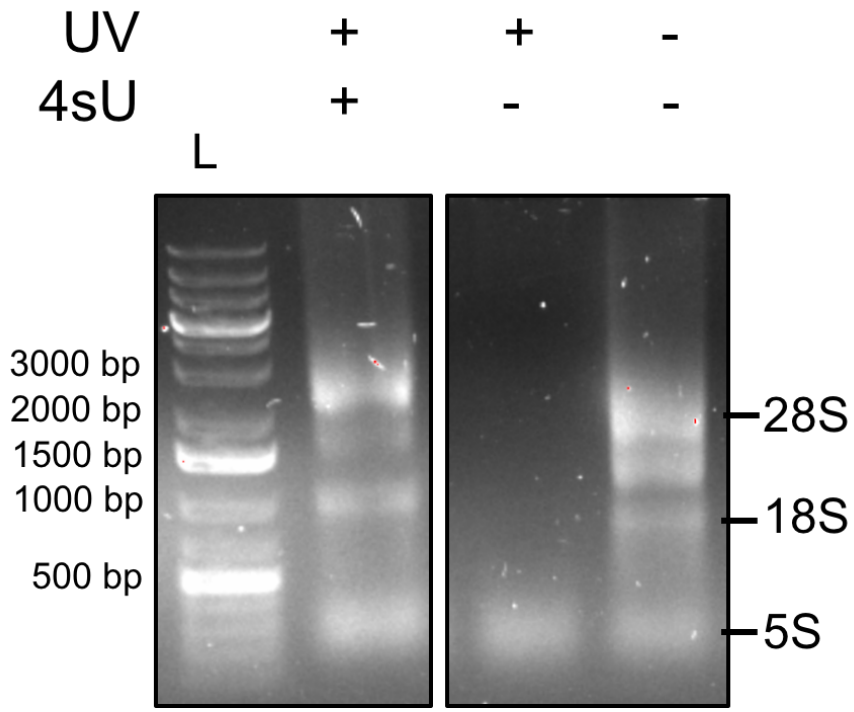


Figure 3.5. Agarose gel electrophoresis of total RNA from crosslinked and non-crosslinked ISHIKAWA cells.

Total RNA from ISHIKAWA cells across various UV treatments was extracted and subjected to agarose gel electrophoresis to examine RNA integrity. 700 ng of each sample was run on the same gel together and relevant samples were cropped and shown next to one another. Treatment with 4sU was combined with UV treatment at 365 nm. Cells were not treated with 4sU when UV was used at 254 nm. The detection of two bands corresponding to the 18S and 28S rRNAs is indicative of good quality RNA. A double-stranded DNA ladder was used here. 4sU, 4-thiouridine

3.3.1.3. Evaluating efficiency of *XIST* enrichment from RAP in non-crosslinked lysates

Tiling antisense DNA probes against human XIST confers high specificity and sequence coverage. Previously, 1 probe per ~120 nt of sequence had been successfully used with mouse *Xist* (McHugh et al., 2015). Given the 19.2 kb length of human *XIST*, the 160 probes required to tile across the entire length combined with a high cost for biotinylated 90-nt antisense DNA probes were limiting factors in implementing the method as originally described. A total of 10 probes were designed here (**Table 3.1** and **Figure 3.6**) and tested for their capacity to pull down a sufficient amount of *XIST* in order to ensure protein identification.

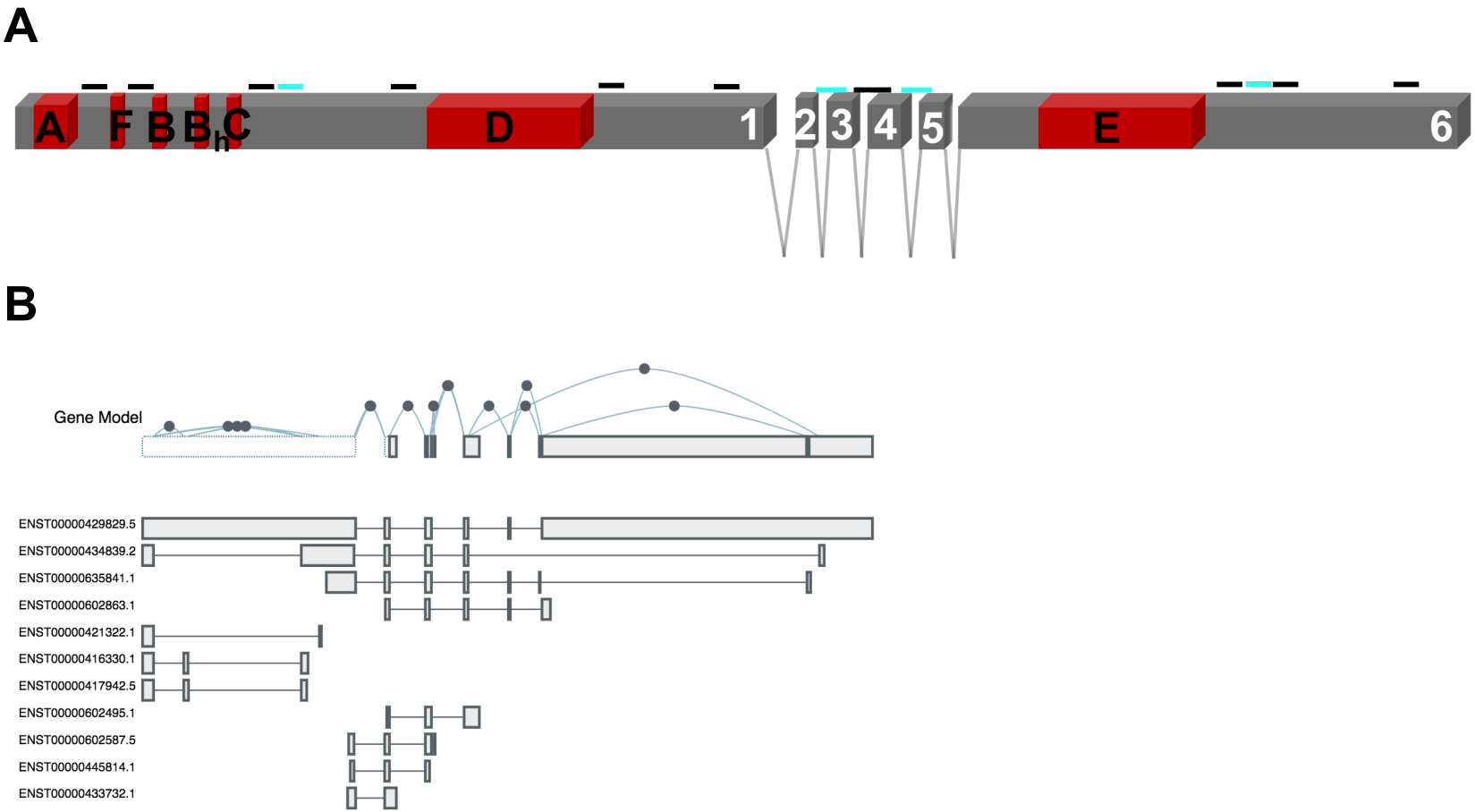


Figure 3.6. Schematic of RAP probe and XIST primer location on human XIST.

A) The location of probes and primers used for RAP capture of human *XIST* are shown here (5'-3'). Straight lines denote probes (black) or primers (cyan) used. Boxes in grey denote exons, with repetitive regions highlighted in red. Numbers on grey boxes denote exon numbers. **B**) Isoforms of human XIST (3'-5'). Available from https://www.gtexportal.org/home/gene/XIST [Accessed 03 March 2022]

To determine the baseline level of *XIST* RNA enrichment achievable with this setup prior to the use of crosslinking, I performed RAP using 135 million ISHIKAWA cells, ~6 mg of beads for pre-clearing and ~4 mg of beads for capture. Following RAP pulldown, the level of *XIST* enrichment was determined from elution samples by RT-qPCR. In addition, the *ACTB* and *U2* transcripts were used as negative controls to a) detect non-specific probe binding and b) establish a baseline for background levels of non-specific transcript enrichment. *XIST* was enriched in the elution over the input, albeit at varying levels, depending on which region was targeted for amplification. On average, a 1-fold enrichment was seen for XIST exon 1, ~24-fold enrichment for *XIST* exons 2-3 and ~12-fold for *XIST* exon 6 over the input (**Figure 3.7**) whereas the amount of *ACTB* seen over the input was negligible. The levels of the negative control *U2* observed were comparable to those for *XIST* exon 1, indicating a low *XIST* pulldown efficiency for that particular region or *XIST* fragmentation. A pattern of differential amplification depending on XIST region previously seen in steady-state RNA from ISHIKAWA cells was also evident here, however the pattern was not the same (**Figure 3.3**). Exons 2-3 were more highly enriched here (**Figure 3.7**), followed by exon 6 and then exon 1 whereas exon 1 was more readily detectable in steady-state RNA, followed by exons 2-3 and exon 6 (**Figure 3.3**). Therefore, this denotes a lack of full-length *XIST* pulldown. Whilst a ~24-fold enrichment for *XIST* exons 2-3 was higher compared to negative control transcript levels, conditions were sought to be optimised for a higher *XIST* enrichment in order to ensure a sufficient amount of proteins would be pulled down for protein identification via mass spectrometry.

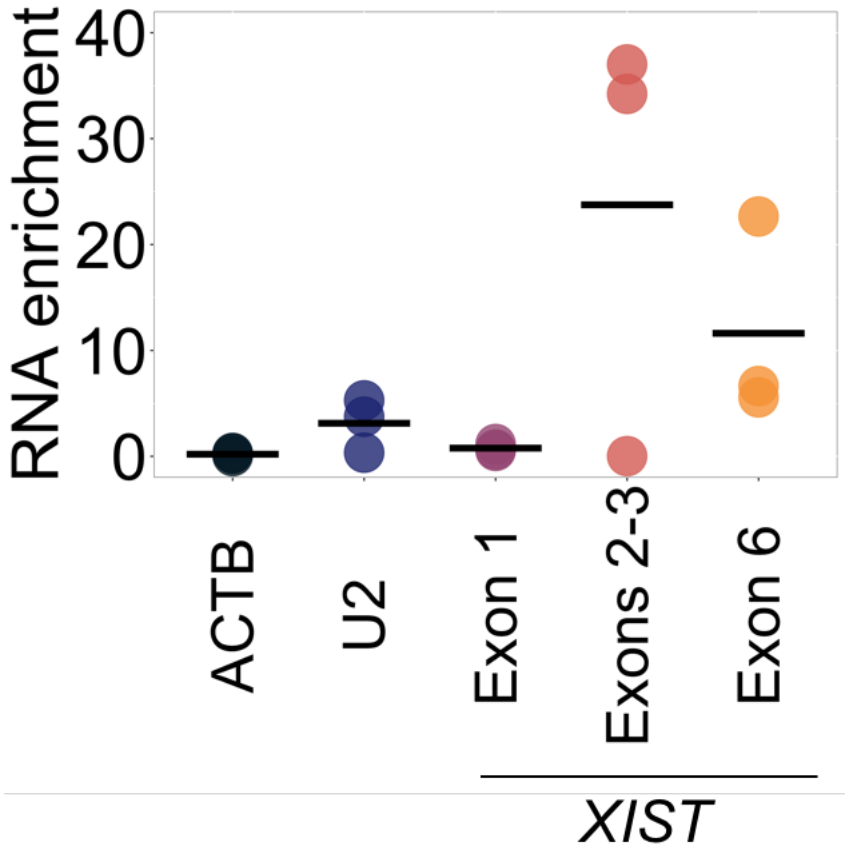


Figure 3.7. Adapted RAP displayed a lack of full-length XIST enrichment in ISHIKAWA cells.

RT-qPCR in elution of samples following RAP with *XIST* probes in 135 million whole cell ISHIKAWA non-crosslinked lysates with ~6 mg of beads for pre-clearing and ~4 mg of beads for capture. Fold enrichment of each transcript's abundance was normalised to input using the $2^{\wedge-(Ct_{elution}- Ct_{input})}$ formula and reported as RNA enrichment. Three technical replicates were performed. RAP, RNA affinity purification; RT-qPCR, reverse transcription quantitative PCR

The end goal was to isolate a sufficient amount of XIST to enable identification of protein partners via mass spectrometry. Assuming the number of cells and amount of bead suspension for capture used previously were not conditions in which a sufficient amount of *XIST* could be pulled down, the number of cells used in RAP were next raised to 160 million together with an increase in the use of ~3 mg of beads for pre-clearing and 9.6 mg for capture (2.4-fold more compared to the previous attempt).

In the next experiment, RAP was performed for XIST using 160 million cells and following pulldown, RT-qPCR was performed in elution samples to measure the enrichment levels of *XIST* and negative control transcripts. *XIST* enrichment levels varied according to primer sets targeting different regions, mirroring the pattern seen in the previous experiment, hinting again at a lack of full-length *XIST* pulled down. On average, a 6.5-fold enrichment was seen for *XIST* exon 1, 307-fold enrichment for XIST exons 2-3 and 135-fold enrichment for *XIST* exon 6 (**Figure 3.8**). The observed enrichment for *XIST* exons 2-3 and exon 6 was not only much higher compared to *ACTB* and *U2* levels, but also compared to the XIST enrichment previously achieved here (**Figure 3.7**). Overall, a two-fold increase in the bead amount used and using 25 million cells more, resulted in a 12.5-fold increase in RNA enrichment compared to the previous RAP attempt.

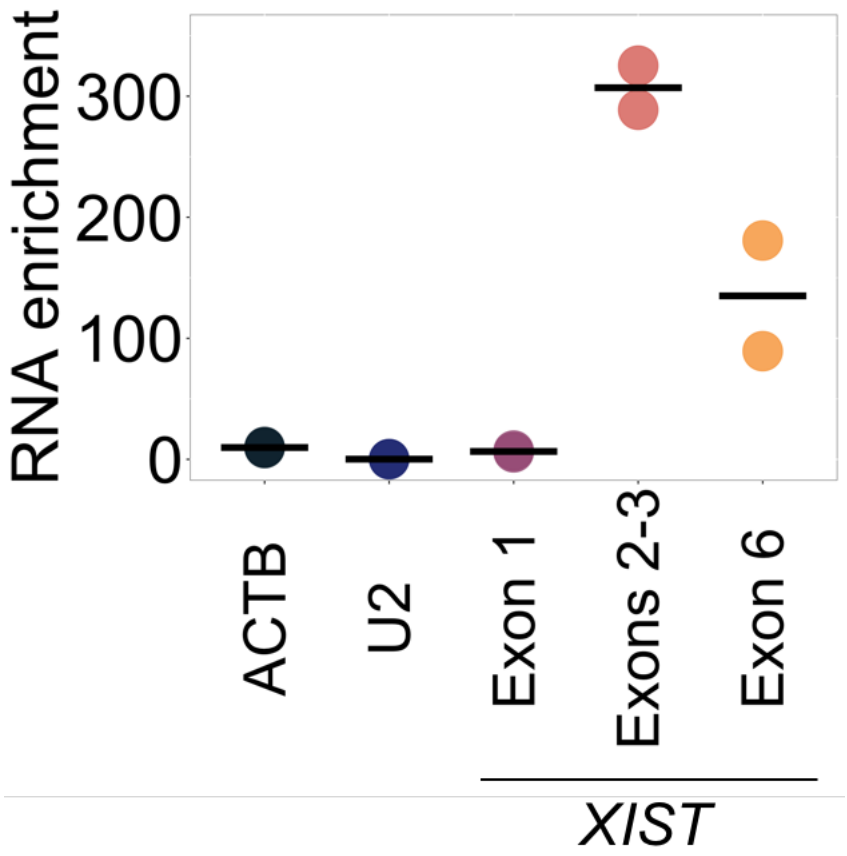


Figure 3.8. Adapted RAP specifically enriches for human XIST at a high level in ISHIKAWA cells.

RT-qPCR in elution of samples following RAP with *XIST* probes in 160 million whole cell ISHIKAWA non-crosslinked lysates with ~3 mg of beads for pre-clearing and 9.6 mg for capture. Fold enrichment of each transcript's abundance was normalised to input using the $2^{\wedge-(Ct_{elution}- Ct_{input})}$ formula and reported as RNA enrichment. Two technical replicates were performed. RAP, RNA affinity purification; RT-qPCR, reverse transcription quantitative PCR

3.3.1.4. Assessing *XIST* protein partner enrichment following RAP in crosslinked lysates

The *XIST* probe pool used here contained ~7,096 pmoles in 20.3 µl (equivalent to 20 µg). According to the manufacturer of the beads, 1 mg of the streptavidin-coated magnetic beads is capable of binding at most ~500 pmoles of single-stranded oligonucleotides (ThermoFisher Scientific, 2016). Therefore, ~1,419 µl of 10 mg/ml (~14.2 mg) streptavidin-coated beads would be required to capture ~7,096 pmoles of biotinylated single-stranded oligos.

To increase the number of *XIST* molecules pulled down and to determine whether the current RNA enrichment was sufficient to enable protein partner identification, 258 million ISHIKAWA cells were treated with 4sU and crosslinked with UV at 365 nm. In this experiment, crosslinked nuclear-enriched cell lysates were pre-cleared with ~3.8 mg of beads. To fully saturate the beads used with available *XIST* probes from the pool, ~15.5 mg beads were used for capture. Following pulldown, the elution was split 70:30 for protein and RNA analysis. Consistent to the previous RAP attempts with non-crosslinked lysates, *XIST* enrichment levels varied depending on the region amplified. On average, a ~9-fold enrichment was seen for XIST exon 1, 253-fold enrichment for XIST exons 2-3 and ~66-fold for XIST exon 6 (**Figure 3.9A**). The level of enrichment seen here was lower than what was seen in the previous RAP attempt (**Figure 3.9A vs Figure 3.8**), which was unexpected given the higher cell number and bead amount used.

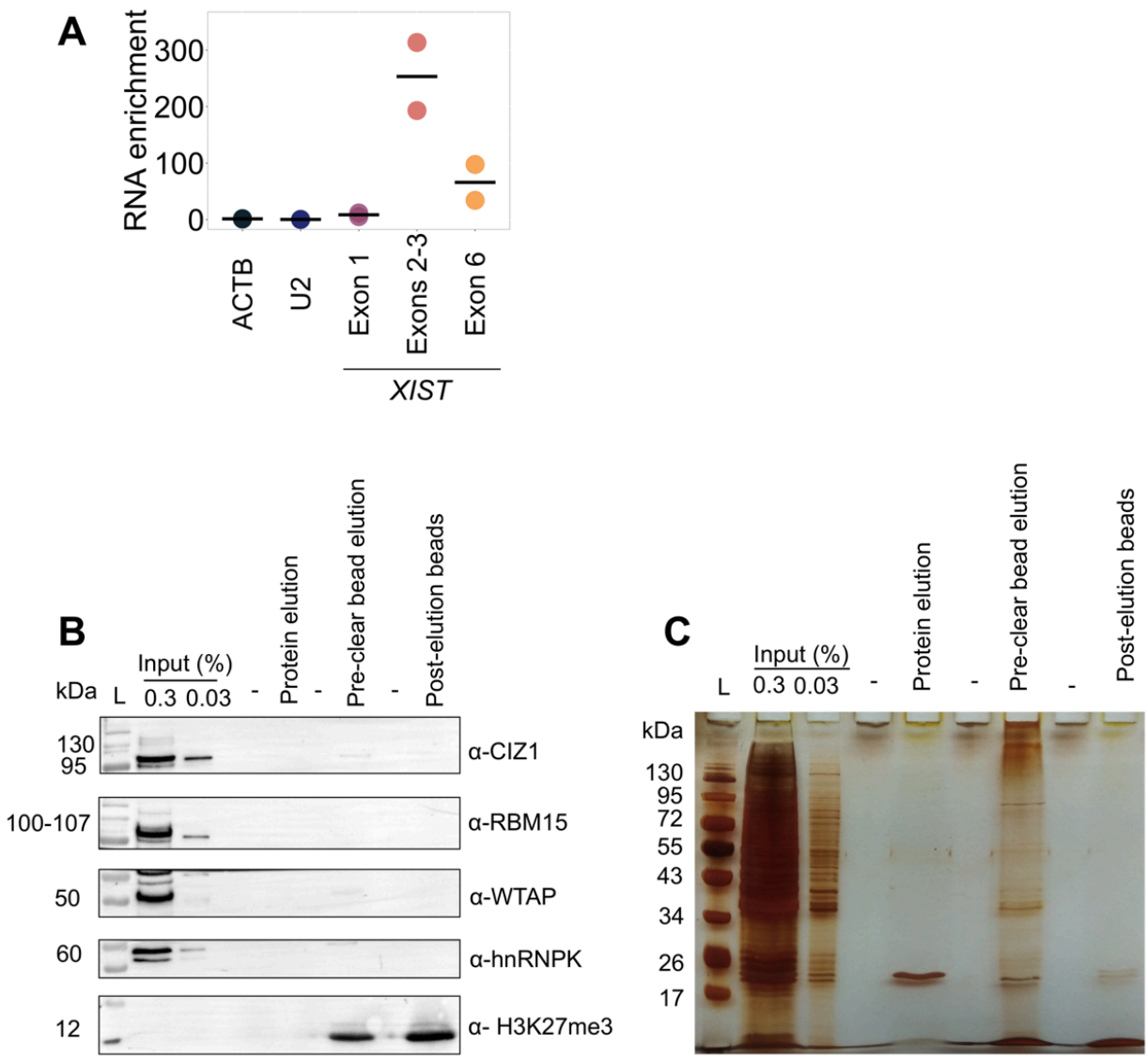


Figure 3.9. Adapted RAP enriches for human XIST but not highly enough to detect known protein partners.

A) RT-qPCR in elution of samples following RAP with *XIST* probes in 258 million whole cell ISHIKAWA crosslinked lysates with ~3.8 mg of beads for pre-clearing and ~15.5 mg for capture. RNA enrichment indicates fold enrichment normalised to input. Two technical replicates were performed. **B**) Western blot of putative protein partners of *XIST* in elution samples following RAP with XIST probes in ISHIKAWA cells. None of the putative protein partners were detected in the elution samples following RAP. H3K27me3 served as a negative control. All antibodies used at 1:1000 in PBS-T. **C**) Silver staining of elution sample shows a general lack of protein enrichment following RAP of XIST in ISHIKAWA cells. RAP, RNA affinity purification; RT-qPCR, reverse transcription quantitative PCR

To determine whether protein partners of *XIST* were co-precipitated with the *XIST* RNA, western blotting was performed on elutions from the pull-down, alongside input. Following the elution from the beads used for capture, another elution step was performed from the same beads using 1x Laemmli buffer and boiling samples (this served as a 'harsher' elution; referred to as 'post-elution'). This was done as a check for the capacity of the elution solution to detach proteins from the beads. Additionally, in this experiment the beads that had been used for pre-clearing had been saved, proteins eluted and sample run on the same gel (referred to as 'pre-clear elution'). This was to test whether the amount of pre-clearing beads was high enough to remove the assayed proteins. Overall, none of the putative protein partners probed for were found bound by the beads when eluted either with the standard elution approach (Elution lane, **Figure 3.9B**) or with a harsher treatment aimed to release all bound proteins (Post-elution lane, **Figure 3.9B**). Looking at proteins bound to pre-clearing beads revealed negligible amounts of the proteins blotted for were bound and lost in this instance (Pre-clear bead elution lane, **Figure 3.9B**).

A lack of putative protein partner enrichment could either result from a general lack of protein enrichment indicative of weak RNA enrichment, or from a lack of an interaction between these proteins and human *XIST*. To distinguish between these scenarios, elution samples were run on a denaturing polyacrylamide gel and silver stained. Silver staining of the elution sample did not highlight any proteins in the elution samples, besides benzonase at ~26 kDa (**Figure 3.7C**). Benzonase was used in the elution sample to digest the nucleic acid of the probes linking the proteins to the beads as well as any captured nucleic acid that might alter protein migration following protein electrophoresis.

Taken together, these data provide evidence that performing RAP with 10 anti-sense DNA probes in ~258 million ISHIKAWA cells with a 1-fold excess of streptavidin-coated beads to the oligo probe pool can enrich for human *XIST* up to 300-fold over input. However, these conditions did not result in an RNA enrichment level sufficient to enable the pull-down of human *XIST* protein partners to the lower detection limit of silver staining.

3.3.2. Employing RIP coupled to RT-qPCR to characterise putative protein partners of XIST in human

Given the RNA-centric approach trialled did not efficiently pulldown a sufficient amount of protein for the identification of *XIST* protein partners, a protein-centric approach was taken instead. Protein partners have previously been characterised for *Xist* in mouse and the hypothesis tested here was that that these proteins would also interact with *XIST* in humans and cow, based on similarities of both XIST nucleic acid sequence and protein amino acid sequence. Therefore, RIP coupled to RT-qPCR was a suitable alternative in which the association of candidate proteins with *XIST* based on those identified in mouse could be assessed. In selecting candidate proteins to assay for an association with *XIST*, the following parameters were considered: 1) amino acid conservation of putative protein partner across species of interest, 2) efficacy of antibody at detecting the protein in western blotting and 3) functional studies in the literature linking the role of the protein to XCI. Therefore CIZ1, WTAP, hnRNPK, SPEN and RBM15 were selected.

Given *XIST* is predominantly nuclear-enriched (Clemson et al., 1996), initial RIP experiments were performed using both whole cell and nuclear-enriched extracts of ISHIKAWA cells to find out whether increased specificity could be achieved by reducing the complexity of the lysates, i.e. using a nuclear-enriched lysate.

In the first instance, ISHIKAWA cell lysates were prepared from 20 million cells each and RIP was performed using 5 µg of the CIZ1 antibody. The RIP approach involves incubating the lysate with magnetic beads coupled to the CIZ1 antibody. This allows the CIZ1 protein to be isolated from the rest of the lysate. Eluting material captured by the beads can be split in two samples to a) analyse proteins captured via western blot, verifying RIP specificity and b) amplify candidate CIZ1-interacting RNAs via PCR. In the western blots performed here, β-tubulin was also probed for as a negative control, meaning that an interaction between specific antibodies such as CIZ1 here and β-tubulin was not expected. Equally, in the RT-PCR assessment of elutions, *ACTB*, *GAPDH* and *U2* RNAs were used as negative controls, indicating that an interaction between CIZ1 and these RNAs was not expected.

In this experiment, RIP was performed and following that western blot was performed. Looking in the depleted lysate fraction where whole cell lysates were used, the CIZ1 protein was depleted regardless of which antibody was used for RIP (**Figure 3.10A**). This demonstrating most of the available CIZ1 was captured by the beads in whole cell extracts. The same pattern was observed for the non-specific protein control, β-tubulin. Importantly, there was no notable difference in the depletion of β-tubulin across the anti-CIZ1 or IgG antibodies (**Figure 3.10A&B**), indicative of non-specific capture. Looking in the elution samples, CIZ1 protein was pulled down, demonstrated by the presence of two distinct bands (at 95 and 130 kDa, as expected) when the anti-CIZ1 antibody was used. However, some CIZ1 protein was also pulled down with the use of the IgG control in whole cell extracts (**Figure 3.10A**). More importantly, despite the presence of β-tubulin in both elution samples, there was no difference in enrichment when using the anti-CIZ1 or the IgG antibody, indicative of non-specific β-tubulin pulldown.

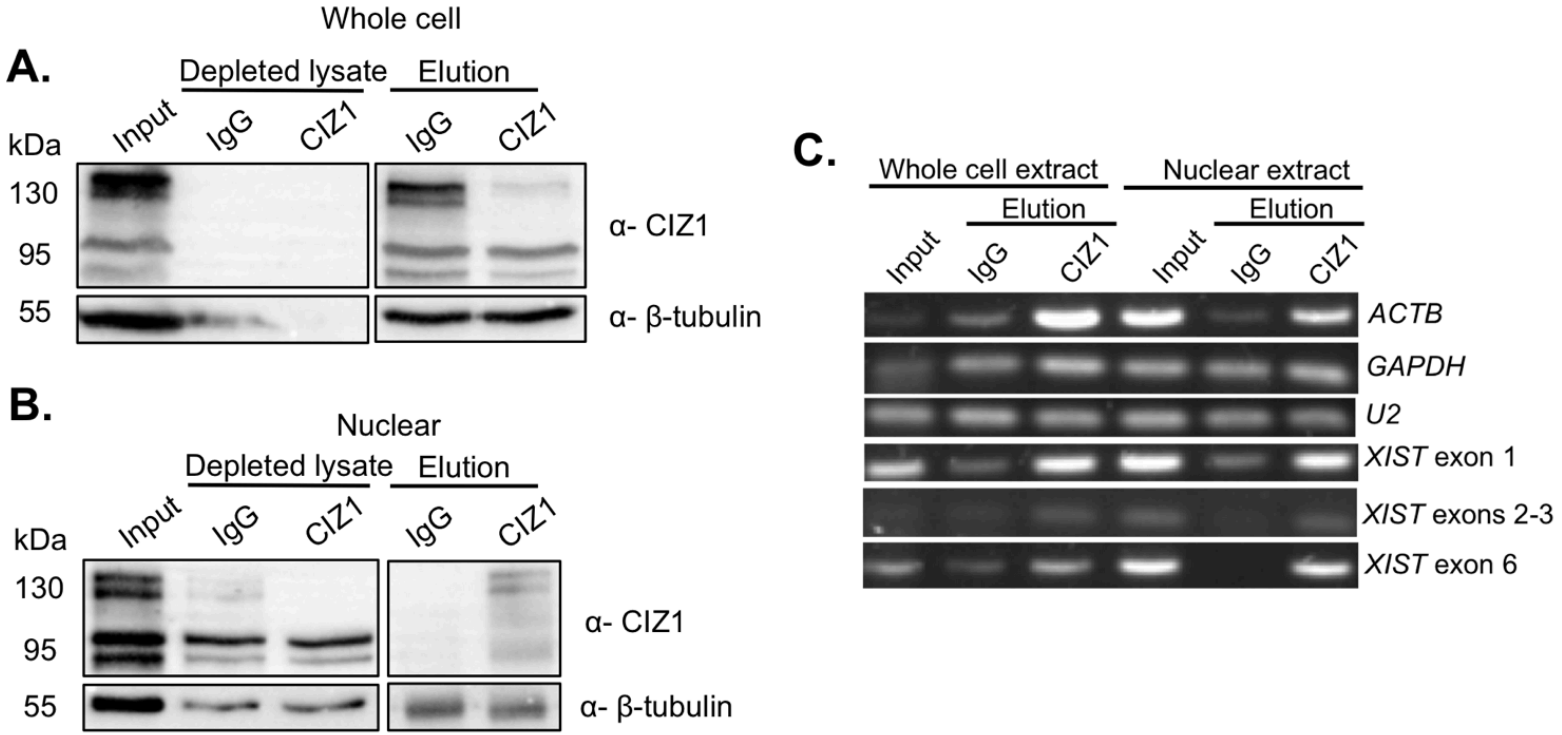


Figure 3.10. Comparison of CIZ1-XIST enrichment from RIP in whole cell and nuclear-enriched extracts.

A) Western blot of CIZ1 RIP samples using whole cell extract of ISHIKAWA cells. Two bands at 95 and 130 kDa expected (four isoforms). **B**) Western blot of CIZ1 RIP samples using nuclear-enriched extract of ISHIKAWA cells. All antibodies used at 1:1000 in PBS-T. **C**) RT-PCR of CIZ1 RIP from whole cell and nuclear extract of ISHIKAWA cells. IgG is a non-specific control in RIP experiments. β-tubulin serves as a non-specific interacting protein negative control in western blot. *ACTB* serves as a specific interacting transcript positive control whereas *U2* serves as non-specific transcript negative control. N=1 biological replicate. RIP, RNA immunoprecipitation; RT-qPCR, reverse transcription quantitative PCR

In nuclear-enriched extracts, more of the heavier isoform of CIZ1 (at 130 kDa) was preferentially captured by the CIZ1 antibody compared to IgG, in depleted lysates (**Figure 3.10B**). Conversely, there was no notable difference in the depletion of β-tubulin across the anti-CIZ1 or IgG antibodies. Examining the elution samples, the heavier isoform of CIZ1 was found enriched when the anti-CIZ1 antibody was used in nuclear-enriched extracts whereas no CIZ1 protein was found eluted with the IgG antibody (**Figure 3.10B**). Even though some β-tubulin signal was seen for both elution samples, there was no enrichment for it, regardless of antibody used, indicative of non-specific capture.

In summary, the CIZ1 protein was completely depleted regardless of antibody used in in whole cell extracts whereas only the heavier CIZ1 isoform was completely depleted in nuclear-eneriched lysates (**Figure 3.10A&B**). The same depletion pattern was seen for β-tubulin. In terms of the elution samples, CIZ1 was found more enriched with the IgG than the anti-CIZ1 antibody in whole cell extracts. Conversely, hardly any CIZ1 was seen in the elution of nuclear-enriched lysates, with most of it showing in the elution with the anti-CIZ1 antibody.

Following RIP, RT-PCR was performed to identify transcripts associated with the assayed CIZ1 protein with *ACTB, GAPDH* and *U2* were used as non-specific controls meaning an interaction between these transcripts with the CIZ1 protein was not expected. Indeed, there was no enrichment of *GAPDH* and *U2* RNA in elution samples of IgG or CIZ1, for either whole cell or nuclear-enriched extracts was observed (**Figure 3.10A&B**), indicative of high specificity. An enrichment of *ACTB* was observed in the elutions from CIZ1 pulldown compared to input and IgG control, both in whole cell and nuclear-enriched extracts (**Figure 3.10C** and **Appendix I**), suggesting that *ACTB* may be specific partner of CIZ1. This has not been previously reported in the literature. Therefore, *ACTB* could serve as an additional positive control of RIP performance and efficiency in future experiments with the anti-CIZ1 antibody.

To ensure full-length *XIST* was recovered, all *XIST* primers described previously were used (cyan bars in **Figure 3.6A**). *XIST* RNA was found enriched in the anti-CIZ1 elution over the IgG elution in both whole cell and nuclear-enriched extracts,

with all primers used (**Figure 3.10C** and **Appendix I**), demonstrating a specific association of CIZ1 with human *XIST* in ISHIKAWA cells. Enrichment levels varied according to primer set used, with *XIST* exon 1 and *XIST* exon 6 showing the highest enrichment in the elution samples of anti-CIZ1 versus the IgG (**Figure 3.10C** and **Appendix I**).

Given no gross differences were seen in the levels of *XIST* RNA enrichment across whole cell and nuclear-enriched extracts, whole cell extracts were used for the rest of the RIP experiments (Merck Millipore, 2009).

In the next experiment, RIP was performed using whole cell lysates this time in three biological replicates with 10% of elution used for western blots and the remaining was analysed by RT-qPCR. A reduction in the signal of CIZ1 protein was seen from western blotting of the depleted lysate fraction with the anti-CIZ1 antibody, but not with the IgG (**Figure 3.11A**). Additionally, an enrichment of the CIZ1 protein was seen in the elution with the anti-CIZ1 antibody but not with the IgG. In contrast, the negative control protein, β-tubulin, was not present in elutions and therefore not pulled-down (**Figure 3.11A**). RT-qPCR was then used to assess the binding of XIST RNA to CIZ1 protein. Consistent with RT-PCR results from the previous experiment, *XIST* was found to be more enriched in elutions of anti-CIZ1 versus IgG elutions in three biological replicates (**Figure 3.11B**). More specifically, *XIST* exon 1, exons 2-3 and exon 6 regions were 4-, ~14-, and 7-fold enriched, respectively, over the input (**Figure 3.11B**). *ACTB* was found to be more enriched in in CIZ1 pulldowns when compared to the non-specific IgG control (4-fold enrichment over input; **Figure 3.11B**), consistent with what was previously seen with RT-PCR (**Figure 3.11C**). *U2* was not enriched in either CIZ1 or IgG pulldown, indicating that pulldowns were specific. *ACTB* has not been previously recognised as a CIZ1 interactor in the literature. Conversely, *XIST* has been shown to bind CIZ1 in HEK293FT cells previously (Sunwoo et al., 2017). Overall, these data demonstrate an association between human CIZ1 and human XIST in the human endometrial ISHIKAWA cell culture model.

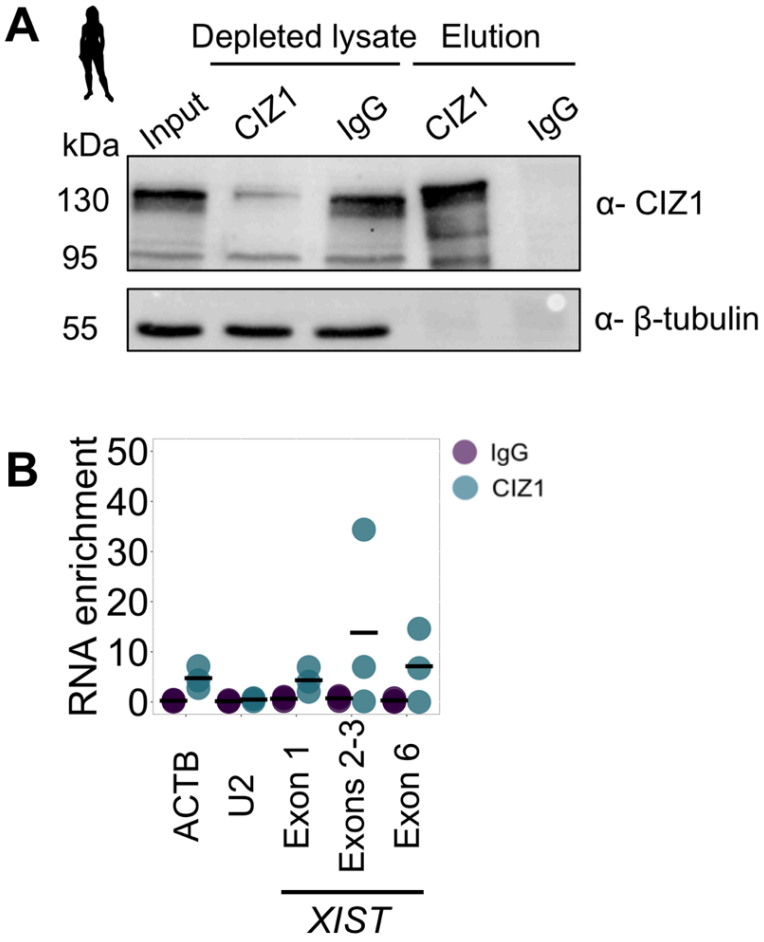


Figure 3.11. The CIZ1 protein associates with human XIST.

A) Representative image from western blot of CIZ1 RIP samples in whole cell extract of ISHIKAWA cells. Two bands at 95 and 130 kDa expected (four isoforms). The amount of protein loaded across input and depleted lysate samples was consistent and equivalent to the same number of cells (~500,000) whereas 10% of the elution was used for western blotting. n=3 biological replicates. **B**) RT-qPCR of CIZ1 RIP in whole cell extracts of ISHIKAWA cells. Fold enrichment of each transcript's abundance in RIP elutions was normalised to input using the $2^{\wedge-(Ct_{elution} - Ct_{input})}$ formula and reported as RNA enrichment. Three technical replicates were performed for each of three biological replicates. Antibodies were used at 1:1000 in PBS-T. IgG serves as a non-specific negative control in RIP experiments. β-tubulin serves as a non-specific negative control in western blot. *ACTB* serves as a specific interacting transcript positive control whereas *U2* serves as non-specific transcript negative control. RIP, RNA immunoprecipitation; RT-qPCR, reverse transcription quantitative PCR

Next, RIP was performed using the RBM15 antibody as previously with whole cell ISHIKAWA lysates in two independent biological replicates. Western blotting of RIP elution samples indicated that RBM15 is specifically pulled down with the anti-RBM15 but not the IgG antibody, given a higher abundance of the RBM15 protein in the former compared to the latter (**Figure 3.12A**). This was also reflected by the depletion of RMB15 levels in the lysate following pulldown. To determine whether *XIST* RNA was bound to RBM15, RT-qPCR was performed on RNA eluted from the pulldown. Neither *ACTB* or *U2* negative control transcripts were specifically bound to the RBM15 protein from elution samples (**Figure 3.12B**). Likewise, *XIST* was not enriched in the RBM15 pulldown when compared to the IgG control, regardless of which primer set was used for its detection (**Figure 3.12B**). This result suggests *XIST* is not bound by RMB15 in human ISHIKAWA cells. This is in contrast to results from previous studies, which detect a specific RBM15-*XIST* interaction (Patil et al., 2016, Graindorge et al., 2019). Taken together, RBM15 could not be robustly validated as a specific protein partner of human *XIST* here, given a weak capture of the RBM15 protein from RIP.

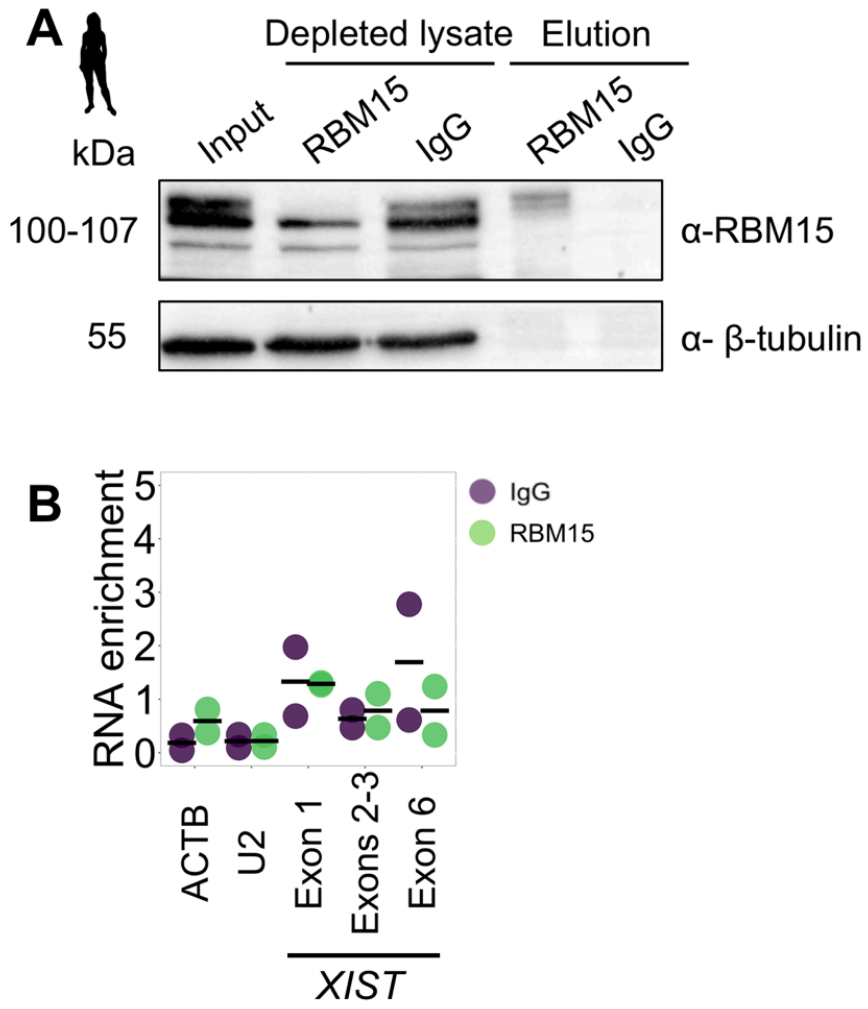


Figure 3.12. An association between the RBM15 protein and human XIST could not be inferred

A) Representative western blot of RBM15 RIP samples in whole cell extract of ISHIKAWA cells. Bands expected between 100-107 kDa (four isoforms). The amount of protein loaded across input and depleted lysate samples was consistent and equivalent to the same number of cells (~500,000) whereas 10% of the elution was used for western blotting. n=3 biological replicates. **B**) RT-qPCR of RBM15 RIP in whole cell extracts of ISHIKAWA cells. Fold enrichment of each transcript's abundance in RIP elutions was normalised to input using the $2^{\wedge-(Ct_{elution} - Ct_{input})}$ formula and reported as RNA enrichment. Three technical replicates were performed for each of two biological replicates. Antibodies were used at 1:1000 in PBS-T. IgG serves as a non-specific negative control in RIP experiments. β-tubulin serves as a non-specific negative control in western blot. *ACTB* and *U2* serve as non-specific transcript negative controls. RIP, RNA immunoprecipitation; RT-qPCR, reverse transcription quantitative PCR

An association between the WTAP protein and *XIST* was sought to be examined. Two antibodies were trialled for the detection of the WTAP protein in chapter 2 (**Figure 2.11**), one recognised the protein from all species (mouse anti-WTAP) and the other was human-specific (rabbit anti-WTAP). Here, RIP was performed using both WTAP antibodies in a single biological replicate of whole cell ISHIKAWA lysates to determine which one was more suitable for pulldowns. RIP elution samples were assessed for protein pulldown efficiency via western blot. Western blotting of depleted lysates indicated that both antibodies could capture most of the available WTAP protein (**Figure 3.13A**). Additionally, probing for WTAP in elution samples indicated that both antibodies could pulldown the WTAP protein, albeit the rabbit anti-WTAP antibody showed a stronger signal (**Figure 3.13A**). The size difference observed between the two is owed to the antibodies. The mouse one was raised against endogenous WTAP and therefore has the capacity to recognise phosphorylated WTAP whereas the rabbit one was raised against recombinant WTAP. Given WTAP had a similar size to β-tubulin (50 kDa vs 55 kDa, respectively), RpS5 was used as a non-specific negative control in western blotting. The presence of the RBM15 protein was also probed due to previous studies showing its co-immunoprecipitation with WTAP in human and mouse (Horiuchi et al., 2013, Patil et al., 2016, Coker et al., 2020).

Having established the rabbit anti-WTAP antibody was a more suitable candidate for RIP, RIP was performed in two more independent biological replicates. Here, the WTAP protein was found greatly enriched in the elution of the anti-WTAP antibody over the IgG, although at a slightly lower molecular weight compared to other lanes (**Figure 3.13A & B**). Whereas in the first WTAP RIP replicate, there was no detectible RBM15 protein in any of the elutions with the different anti-WTAP antibodies (**Figure 3.13A**), a marked enrichment of RBM15 was seen in the elution of rabbit anti-WTAP in the second replicate (**Figure 3.13B**). Conversely, the abundance of RpS5 was low (if any) in the elution samples, as expected from a negative control.

RT-qPCR was then used on eluted RNA to determine the binding of XIST RNA to the WTAP protein. Two biological repeats were performed for this experiment, each of which was supported with three technical replicates for RT-qPCR. Analysis from

both biological replicates showed that neither of the non-specifc RNA controls, *ACTB* and *U2,* were enriched in the WTAP pulldown comapred to the non-specific IgG control (**Figure 3.13C**). Conversely, *XIST* was enriched in the WTAP pulldown. On average, *XIST* exons 2-3 and *XIST* exon 6 were found to be 2- and 1.2-fold enriched in the the anti-WTAP elution over the input respectively, whereas there was no enrichment of *XIST* in the IgG elution (**Figure 3.13C**). Together these data indicate that WTAP is bound to *XIST* in human endometrial cells. Despite a robust WTAP immunoprecipitation achieved, the presence of the RBM15 protein was not consistent in elution samples here. Thus, an association between WTAP and RBM15 could not be reliably inferred under these conditions described here.

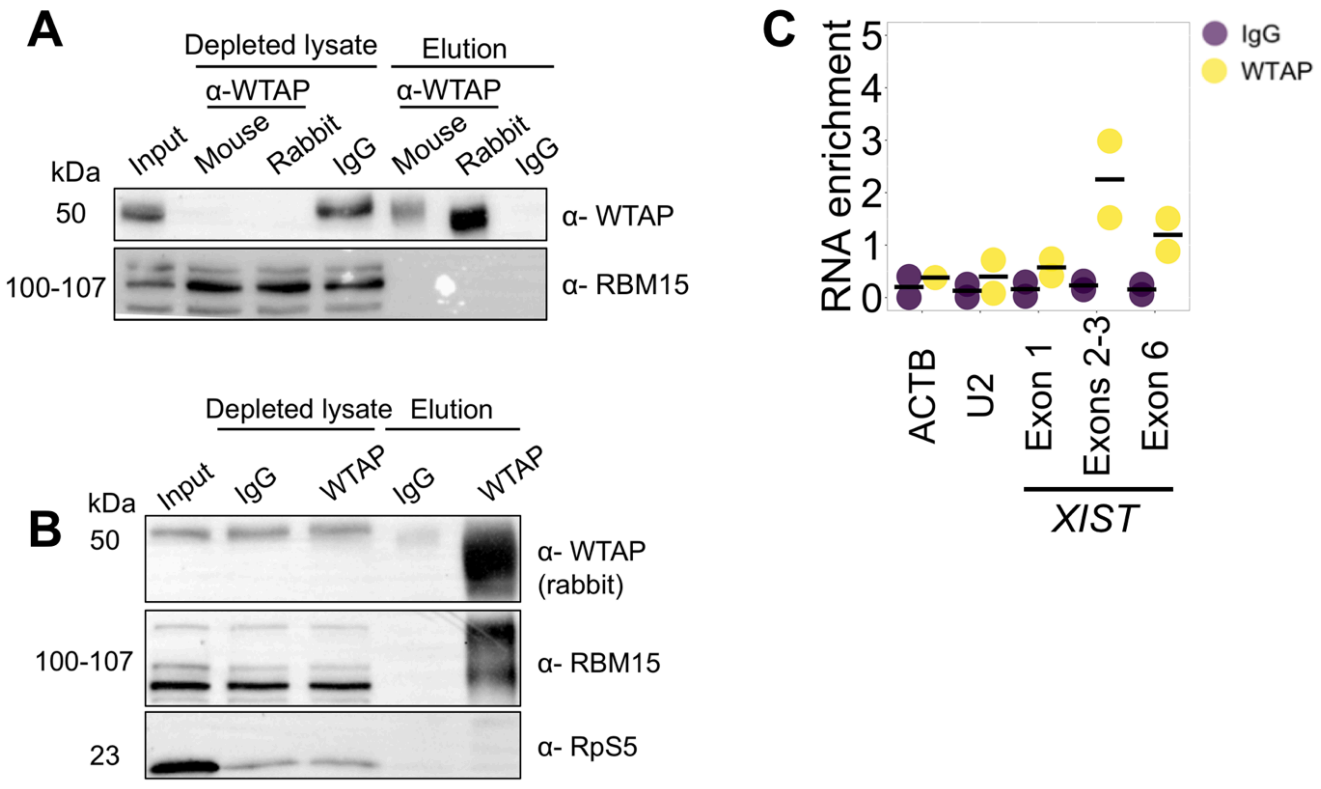


Figure 3.13. The WTAP protein associates with human XIST.

A) Representative image from western blot of WTAP RIP samples in whole cell extract of ISHIKAWA cells. One isoform expected at 44 kDa. The human-specific rabbit anti-WTAP antibody was used for western blotting. The amount of protein loaded across input and depleted lysate samples was consistent and equivalent to the same number of cells (~500,000) whereas 10% of the elution was used for western blotting. n=2 biological replicates. **C**) RT-qPCR of WTAP RIP in whole cell extracts of ISHIKAWA cells. Fold enrichment of each transcript's abundance in RIP elutions was normalised to input using the $2^{\wedge-(Ct_{elution}- Ct_{input})}$ formula and reported as RNA enrichment. Results here are only from the use of the rabbit anti-WTAP antibody. Three technical replicates were performed for each of two biological replicates. Antibodies were used at 1:1000 in PBS-T. IgG serves as a non-specific interacting protein negative control in RIP experiments. RpS5 serves as a non-specific negative control in western blot. *ACTB* serves as a specific interacting transcript positive control whereas *U2* serves as non-specific transcript negative control. RIP, RNA immunoprecipitation; RT-qPCR, reverse transcription quantitative PCR

An interaction between mouse *Xist* and hnRNPK has been extensively characterised. To assess whether human XIST could associate with human hnRNPK, RIP was performed in three independent biological replicates using whole cell ISHIKAWA lysates. Western blotting of samples where RIP was performed with the anti-hnRNPK antibody, exhibited a modest, albeit higher depletion of the available hnRNPK protein in the input, compared to the IgG (**Figure 3.14A**). hnRNPK protein was detected in the hnRNPK pulldown elution but not in the IgG control. Given an interaction was not seen previously between *XIST* and RBM15, it was used as a negative control. However, a modest depletion of RBM15 was observed when performing RIP with the anti-hnRNPK antibody but no RBM15 depletion was seen with the IgG (**Figure 3.14A**). Nonetheless, no RBM15 protein was present in the elution of either anti-hnRNPK or IgG, perhaps indicating that RBM15 could be depleted due to being bound to the same transcript(s) as hnRNPK.

Similar to what was seen with CIZ1 RIP, *ACTB* was more highly enriched in the hnRNPK pulldown compared to the IgG (**Figure 3.14B**). More specifically, *ACTB* was enriched ~6-fold over the input. In contrast, *U2* levels were no different between the two elutions (**Figure 3.14B**). This observation demonstrated a specific association between hnRNPK and ACTB. Compared to the IgG, *XIST* was also found to be more enriched when the anti-hnRNPK antibody was used versus the IgG (**Figure 3.14B**). In fact, enrichment of *XIST* over the input for exon 1, exons 2-3 and exon 6 was on average 8-fold, ~3-fold and 1.5-fold higher respectively (**Figure 3.14B**). This was in agreement with previous evidence of a hnRNPK-*XIST* interaction in human HepG2 (liver), K562 (myeloid) and HEK293T (kidney) cells (Van Nostrand et al., 2016, Graindorge et al., 2019, Lu et al., 2020a). Overall, these data demonstrate that *XIST* specifically associates with human hnRNPK in ISHIKAWA cells.

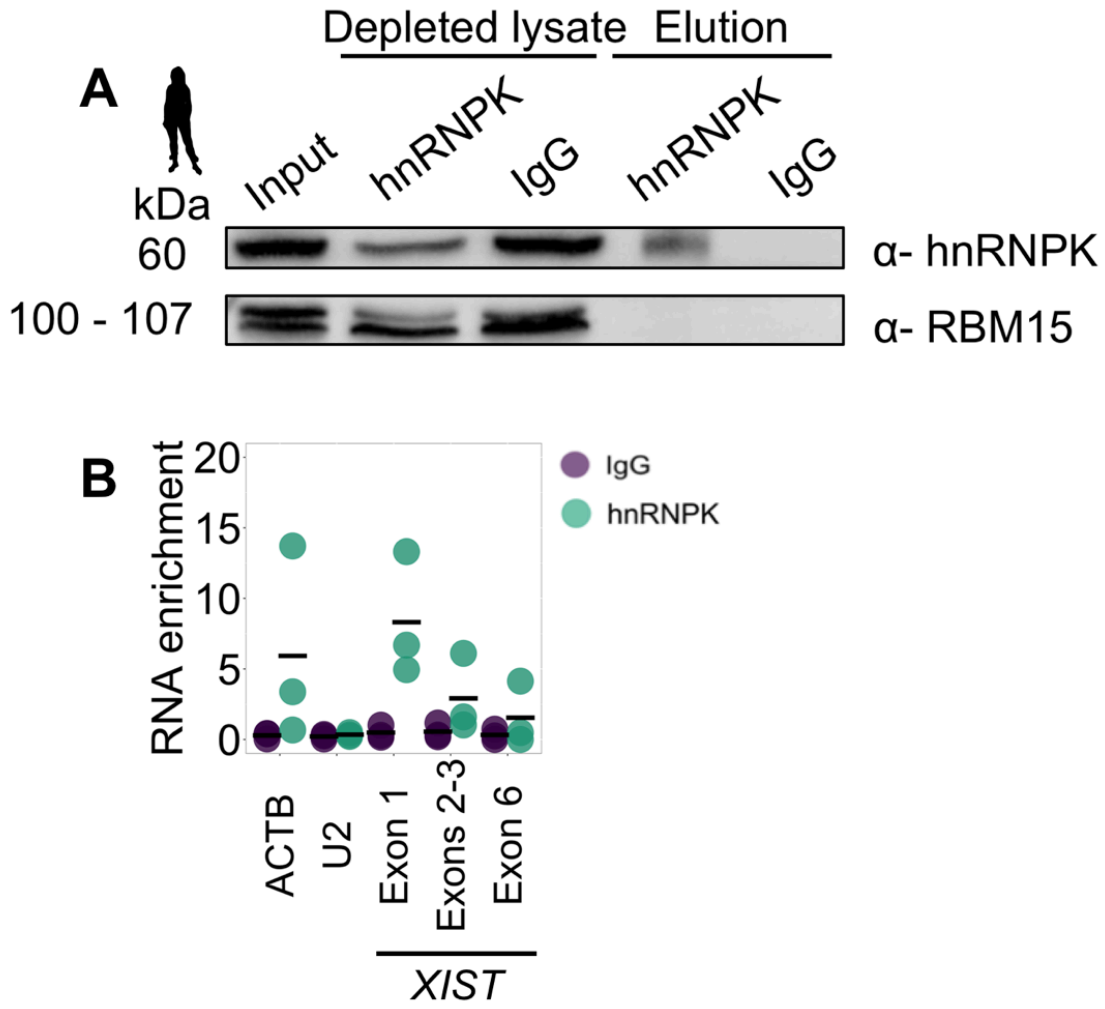


Figure 3.14. The hnRNPK protein associates with human XIST.

A) Representative western blot of hnRNPK RIP samples in whole cell extract of ISHIKAWA cells. One isoform expected at 60 kDa (according to datasheet; 51 kDa according to Uniprot). The amount of protein loaded across input and depleted lysate samples was consistent and equivalent to the same number of cells (~500,000) whereas 10% of the elution was used for western blotting. n=3 biological replicates. **B**) RT-qPCR of hnRNPK RIP in whole cell extracts of ISHIKAWA cells. Fold enrichment of each transcript's abundance in RIP elutions was normalised to input using the $2^{\wedge-(Ct_{elution} - Ct_{input})}$ formula and reported as RNA enrichment. Three technical replicates were performed for each of three biological replicates. Antibodies were used at 1:1000 in PBS-T. IgG serves as a non-specific negative control in RIP experiments. β-tubulin serves as a non-specific negative control in western blot. *ACTB* serves as a specific interacting transcript positive control whereas *U2* serves as non-specific transcript negative control. RIP, RNA immunoprecipitation; RT-qPCR, reverse transcription quantitative PCR

The mouse *Xist*-Spen interaction has been found to be at the forefront of XCI gene silencing onset. Therefore, here the assessment of whether this interaction also holds true in human was carried out. An anti-SPEN antibody previously failed to consistently detect the SPEN protein in western blotting of human or other placental mammal tissue. Here, the anti-SPEN antibody was tested for its suitability in RIP . Thus, RIP was performed in four independent biological replicates using the anti-SPEN antibody in whole cell ISHIKAWA cell extracts. In three of those replicates, the elution was split 90:10 for RNA:protein analyses. Since no western blot could be performed to biochemically establish whether the pulldown had efficiently worked, tThe elution of SPEN and the elution of the IgG negative control were analysed via mass-spectrometry for protein identification instead. SPEN was found in the anti-SPEN elution sample with a spectral count of 30 and 29 unique peptides, whereas none was detected in the IgG elution (**Table 3.6**) confirming SPEN protein is pulled down and the anti-SPEN antibody is suitable for RIP.

Then, RT-qPCR was performed on RIP samples for the examination of transcripts pulled down by SPEN. RT-qPCR revealed no enrichment of negative controls ACTB or U2 in the elution where the anti-SPEN antibody was used over the IgG (**Figure 3.15**). *XIST* was found enriched in the elution of the SPEN pulldown compared to the IgG pulldown (**Figure 3.15**). Primers targeting *XIST* exon 1 and exons 2-3 were the two most highly enriched regions with a ~12- and ~20-fold enrichment over IgG, respectively whereas exon 6 showed a ~3-fold enrichment. This is consistent with previous studies showing SPEN interacts with XIST in human cells (Graindorge et al., 2019, Lu et al., 2020a). Taken together, it was demonstrated that SPEN associates with human *XIST* in human endometrial cells.

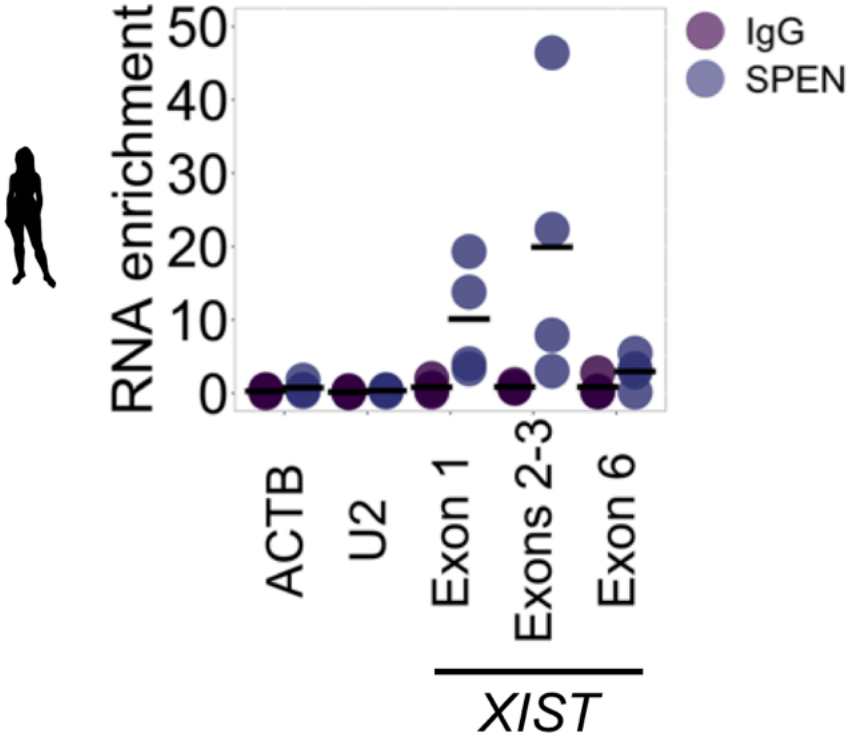


Figure 3.15. RT-qPCR from RIP of SPEN protein in human.

RT-qPCR of SPEN RIP in whole cell extracts of ISHIKAWA cells. Fold enrichment of each transcript's abundance in RIP elutions was normalised to input using the $2^{-(Ct_{elution} - Ct_{input})}$ formula and reported as RNA enrichment. Three technical replicates were performed for each of four biological replicates. Antibodies were used at 1:1000 in PBS-T. IgG serves as a non-specific negative control in RIP experiments. β-tubulin serves as a non-specific negative control in western blot. *ACTB* and *U2* serve as non-specific transcript negative controls. RIP, RNA immunoprecipitation; RT-qPCR, reverse transcription quantitative PCR

Table 3.5. List of proteins identified as specific to the SPEN elution following RIP of SPEN in ISHIKAWA cells.

Protein	Spectral count ratio (SPEN/IgG)	#Unique Peptides
GRIPAP1	113	81
SPEN	30	29
PABPC1	24	16
MKI67	16	14
IMPDH	14	12
IMPDH2	14	12
MCM7	14	14
HNRNPC	13	11
RBMX	12	10
DDX48	10	9
EIF4A3	10	9
DKFZp686K23100	9	9
MATR3	9	9
SRSF1	7	6
PNN	7	6
SAP18	6	6
SRSF7	5	5
KTN1	5	5
KIF1C	5	4
SEMG1	5	4
SRSF3	4	4
SFRS3	4	4
SRSF6	4	4
HEL-S-91	4	4
DKFZp667N107	4	4
ACIN1	4	4
FLG2	4	4
DCD	4	3
HNRNPM	4	4
PCMT1	3	3
GAPDH	3	3
DKFZp686F18120	3	3
SRSF10	3	3
PABPN1	3	3
TRA2B	3	3
HNRNPH1	3	3
UBA52	3	3

GJA1	2.923076923	24
HEL-S-89n	2	14
HSPA5	2	14
H4C1	2	3
RPL38	2	2
hCG_21173	2	2
FLJ10292	2	2
RPL12	2	2
cICK0721Q.2	2	2
TRAF3	2	2
PRSS1	2	2
PRSS3	2	2
E2k	2	2
DLST	2	2
PRPF19	2	2
TRIOBP	2	2
RPL23	2	3
RPLP0	2	2
PGAM5	2	2
RPLP0P6	2	2
RALY	2	2
HIST1H2BC	1.75	6
HIST1H2BK	1.75	6
H2BC12	1.75	6
H2BC4	1.75	6
HIST2H2BF	1.75	6
HIST1H2BH	1.75	6
HIST1H2BD	1.75	6
H2BC15	1.75	6
H2BC14	1.75	6
GRN	1.5	10
HSPA8	1.5	7
HEL-S-72p	1.5	7
HRNR	1.35	24
TUBB4B	1.277777778	2
DSP	1.166666667	10
TUBB	1.15	2

An additional set of proteins were found to be co-immunoprecipitated with SPEN from the mass-spectrometry results (**Table 3.6 and Figure 3.16**). Some of these proteins have been previously identified in a biochemical pulldown using SPEN's SPOC domain in mESCs (Dossin et al., 2020)(see **3.4 Discussion**). Among them were Sin3A Associated Protein 18 (SAP18), Serine/arginine-rich splicing factor 1 (SRSF1), RNA Binding Motif Protein X-Linked (RBMX), Polyadenylate-binding nuclear protein 1 (PABN1) and pinin (PNN).

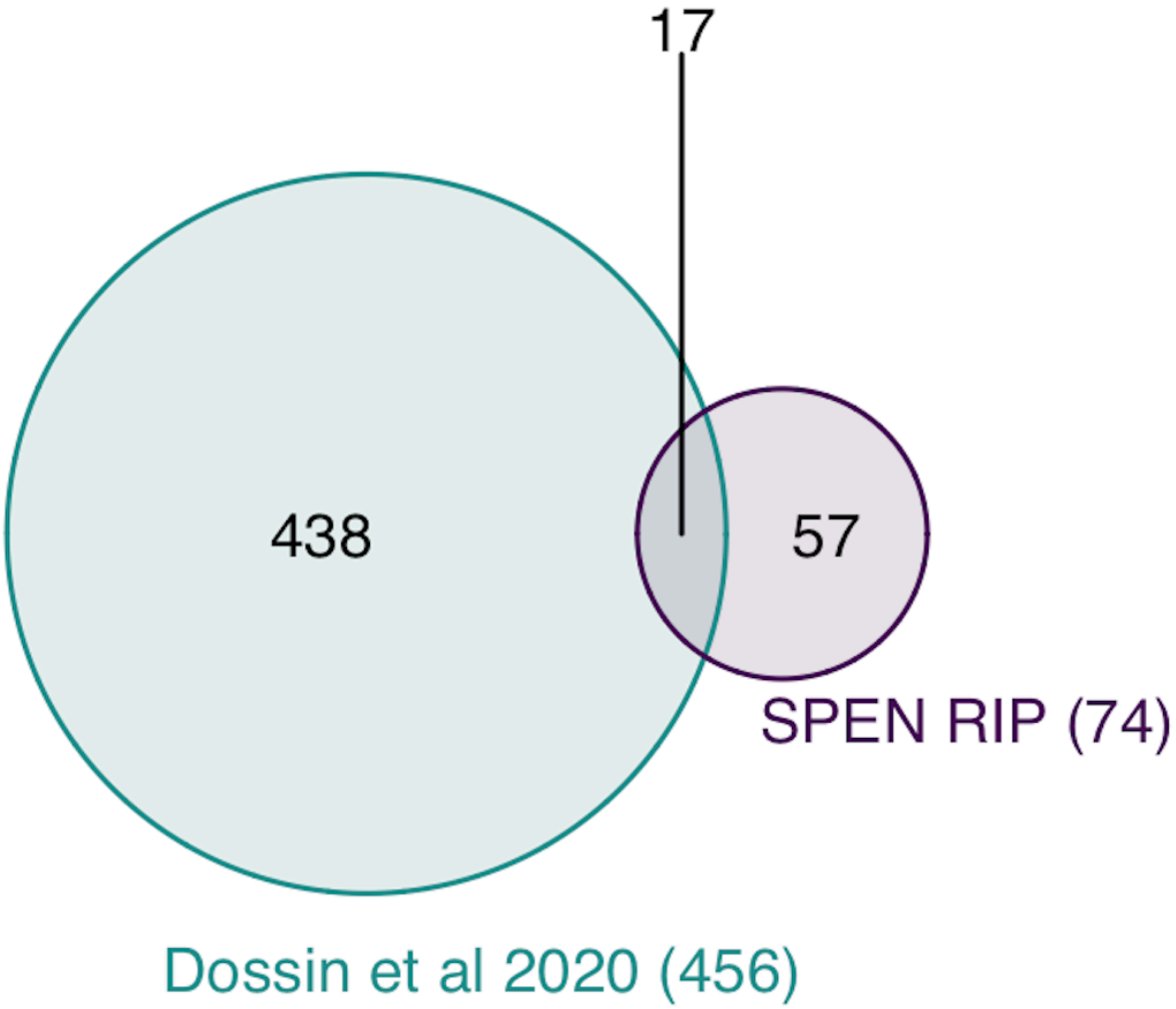


Figure 3.16. MS of SPEN pulldown identifies SPEN and previously described co-immunoprecipitated proteins.

Venn diagram showing 17/74 (~23%) of the identified proteins specific to SPEN from RIP-MS of SPEN here (purple) had also been found in mESCs (Dossin et al., 2020)(green). Trypsin, immunoglobulin- and keratin-related proteins, uncharacterised, had less than two unique peptides and a SPEN/Control ratio below one, were filtered out from both datasets. A single biological replicate was performed with 100% of the RIP elution sample from each antibody. RIP, RNA immunoprecipitation; MS, mass-spectrometry.

3.3.3. Employing RIP coupled to RT-qPCR to characterise putative protein partners of *XIST* in cow

Performing pulldowns in ISHIKAWA cells revealed some of the human XIST protein partners are shared with mouse *Xist*. To establish whether the same proteins partners of human *XIST* are shared across cow, RIP was carried out in primary bovine endometrial stromal cells using the same antibodies.

More specifically, the aim of this experiment was to assessed whether cow CIZ1 associates with cow XIST. Hence, RIP was performed here for CIZ1 in whole cell extracts from bovine stromal cells in three independent biological replicates. The antibody used for RIP was the same as the one successfully used for pulldowns in ISHIKAWA cells. Following RIP, samples were assessed for the presence of CIZ1 by western blotting using the same anti-CIZ1 antibody. Western blotting revealed the anti-CIZ1 antibody detected bovine CIZ1, but also depleted it (with a preference for the heavier isoform at 130 kDa) in bovine stromal cells more so than the IgG (**Figure 3.17A**). CIZ1 protein was found at a higher level in the elution of the anti-CIZ1 antibody compared to the IgG. A similar pattern of enrichment was observed for β-tubulin, albeit at much lower levels.

Subsequently, RT-qPCR was used to characterise whether cow *XIST* interacts with cow CIZ1. Using *ACTB* and *RPL19* were used as non-specific interactors of the CIZ1 protein in cow (negative controls) levels of both transcripts in the elution samples did not differ between CIZ1 and IgG in two of the three biological replicates but in the third replicate background was detected in CIZ1 pulldown (**Figure 3.17B**). On average, *XIST* exons 1 and 5 failed to enrich higher than 2-fold over the input and were not specific to CIZ1 compared to IgG (**Figure 3.17B**). This also occurred for the XIST exon 1 primer set, perhaps indicating a difference related to the particular replicate, where there was high background non-specificity. Overall, the weak enrichment of *XIST* over input was similar to the enrichment seen for negative control transcripts, *ACTB* and *RPL19*. Hence, an association between bovine CIZ1 and *XIST* could not be reliably shown under these conditions.

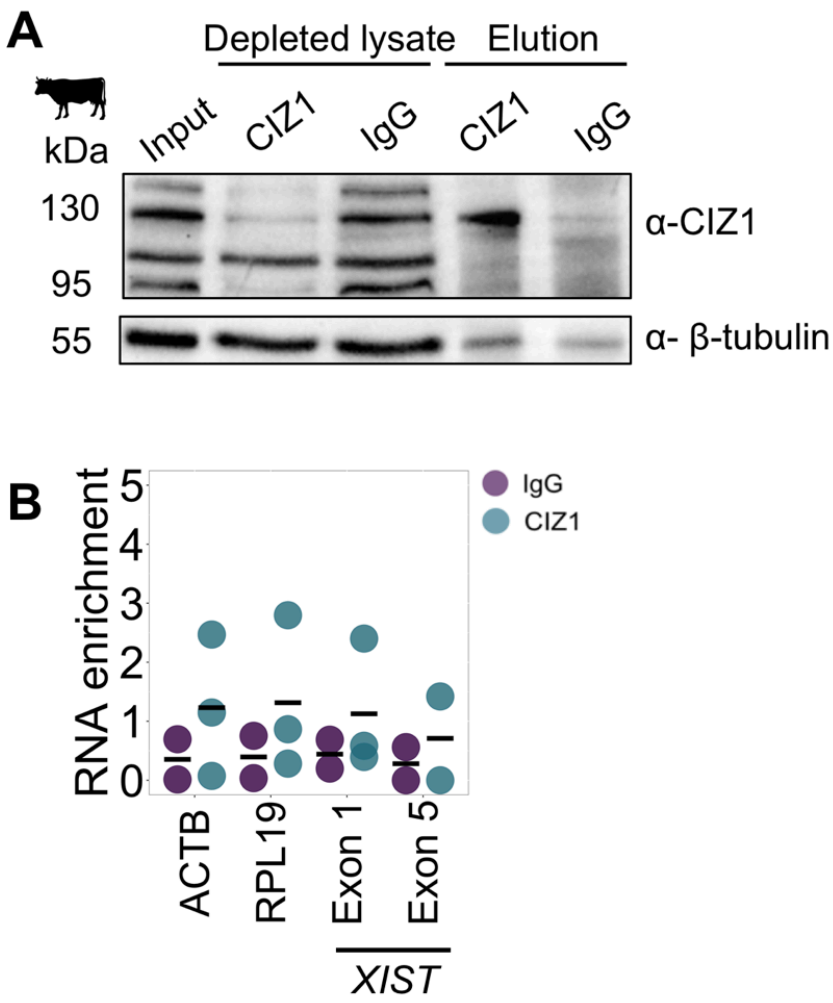


Figure 3.17. RIP of CIZ1 protein in cow.

A) Representative image from western blot of CIZ1 RIP samples in whole cell extract of ISHIKAWA cells. One isoform at 91 kDa predicted by Uniprot. The amount of protein loaded across input and depleted lysate samples was consistent and equivalent to the same number of cells (~500,000) whereas 10% of the elution was used for western blotting. n=3 biological replicates. **B**) RT-qPCR of CIZ1 RIP in whole cell extracts of ISHIKAWA cells. Fold enrichment of each transcript's abundance in RIP elutions was normalised to input using the $2^{\wedge-(Ct_{elution} - Ct_{input})}$ formula and reported as RNA enrichment. Three technical replicates were performed for each of three biological replicates. Antibodies were used at 1:1000 in PBS-T. IgG serves as a non-specific negative control in RIP experiments. β-tubulin serves as a non-specific negative control in western blot. *ACTB* serves as a specific interacting transcript positive control whereas *U2* serves as non-specific transcript negative control. RIP, RNA immunoprecipitation; RT-qPCR, reverse transcription quantitative PCR

RBM15 has been shown to be a bona fide interactor of mouse and human XIST. To examine whether cow RBM15 associates with cow *XIST*, RIP was performed here in whole cell bovine stromal extracts with the same antibody used for ISHIKAWA pulldowns. Western blotting was used to assess whether pulldown of RBM15 was successful. The anti-RBM15 antibody was no better than the IgG at efficiently depleting RBM15 from bovine stromal cells, since no discernible difference was seen between the RBM15 protein levels across the input and depleted lysate samples (**Figure 3.18A**). Nevertheless, a slight enrichment of the RBM15 protein (although of slightly higher molecular weight than in other lanes) was seen in the elution sample of the anti-RBM15 antibody compared to the IgG. β-tubulin was found to be enriched in the elution sample from the IgG compared to the anti-RBM15 antibody. To determine whether cow RBM15 could bind cow *XIST* RNA, RT-qPCR was performed on RNA eluted from the pulldown. Similar to the levels of negative controls *ACTB* and *RPL19*, *XIST* was not found to be enriched in the elution sample of the anti-RBM15 antibody compared to the IgG (**Figure 3.18B**). This was consistent with the relatively low presence of the RBM15 protein in the same elution sample as detected by western blotting. Hence, under these conditions an association between cow RBM15 and cow XIST could not be assessed, given RBM15 could not be robustly pulled down.

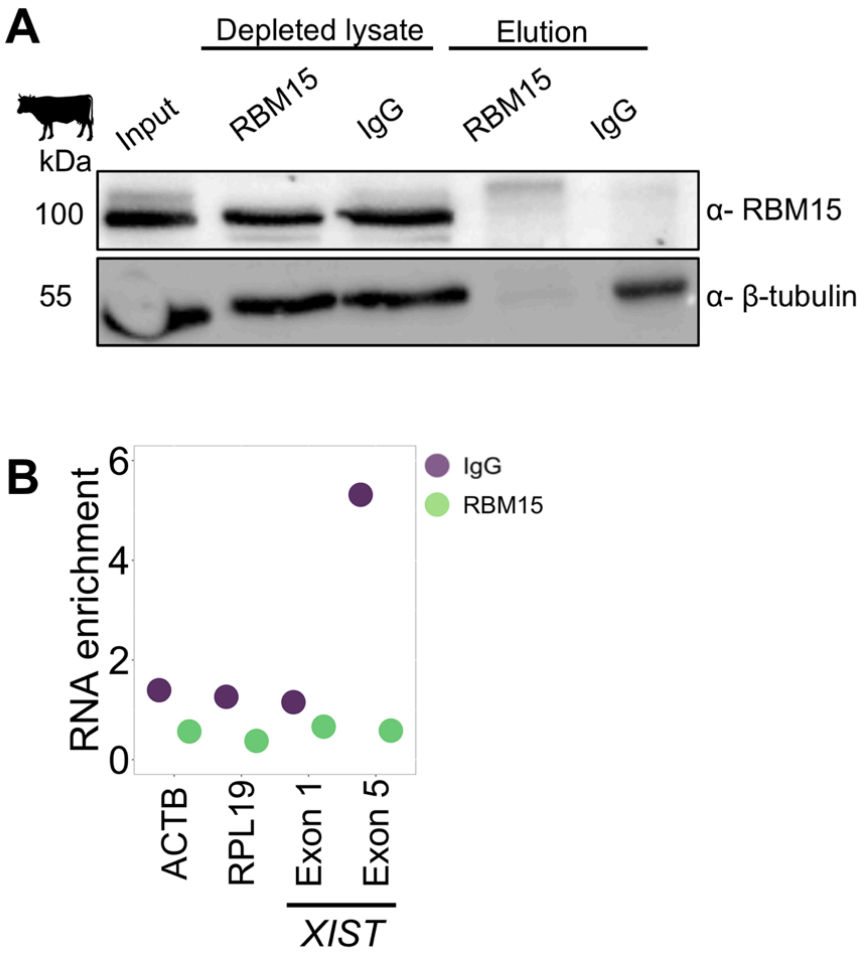


Figure 3.18. RIP of RBM15 protein in cow.

A) Representative image from western blot of RBM15 RIP samples in whole cell extract of ISHIKAWA cells. One isoform at 105 kDa predicted from Uniprot. The amount of protein loaded across input and depleted lysate samples was consistent and equivalent to the same number of cells (~500,000) whereas 10% of the elution was used for western blotting. n=1 biological replicate. **B**) RT-qPCR of RBM15 RIP in whole cell extracts of ISHIKAWA cells. Fold enrichment of each transcript's abundance in RIP elutions was normalised to input using the $2^{\wedge-(Ct_{elution} - Ct_{input})}$ formula and reported as RNA enrichment. Three technical replicates were performed for the single biological replicate. Antibodies were used at 1:1000 in PBS-T. IgG serves as a non-specific negative control in RIP experiments. β-tubulin serves as a non-specific negative control in western blot. *ACTB* serves as a specific interacting transcript positive control whereas *U2* serves as non-specific transcript negative control. RIP, RNA immunoprecipitation; RT-qPCR, reverse transcription quantitative PCR

Next, the potential of a cow WTAP-*XIST* interaction was examined. Thus, RIP was performed in whole cell extracts from bovine stromal cells in a single biological replicate. For this, the mouse anti-WTAP antibody was employed, given it was the only WTAP available that could recognise cow WTAP (**Figure 2.11**). Western blotting of RIP samples with the same antibody was used to assess for pulldown efficiency. Cow WTAP is predicted to be 44 kDa, however another band was evident at ~50 kDa (which could be a phosphorylated fraction of the protein). There was no notable difference in the depletion of the bovine WTAP 44 kDa protein from the input using either anti-WTAP or IgG antibodies. There was a difference for the 50 kDa band though (**Figure 3.19A**). This band was detected at a higher amount in the elution sample when the anti-WTAP antibody was used compared to the IgG (**Figure 3.19A**). A slight enrichment of the RBM15 protein could also be discerned in the same anti-WTAP elution sample compared to the IgG. To characterise whether cow XIST was bound to the pulled down WTAP protein, RT-qPCR was performed. Results showed no *ACTB* or *RPL19* enrichment was found in any elution samples (**Figure 3.19B**), indicating these transcripts do not preferentially associate with either WTAP or IgG. Although *XIST* exon 1 was not more highly enriched in the elution from anti-WTAP compared to the IgG (**Figure 3.19B**), *XIST* exon 5 was (1-fold vs 0.13-fold over input for WTAP vs IgG antibodies used). Thus, in the absence of at least a 2-fold *XIST* enrichment over input and more biological replicates, an association of bovine WTAP with *XIST*, could not be confidently examined.

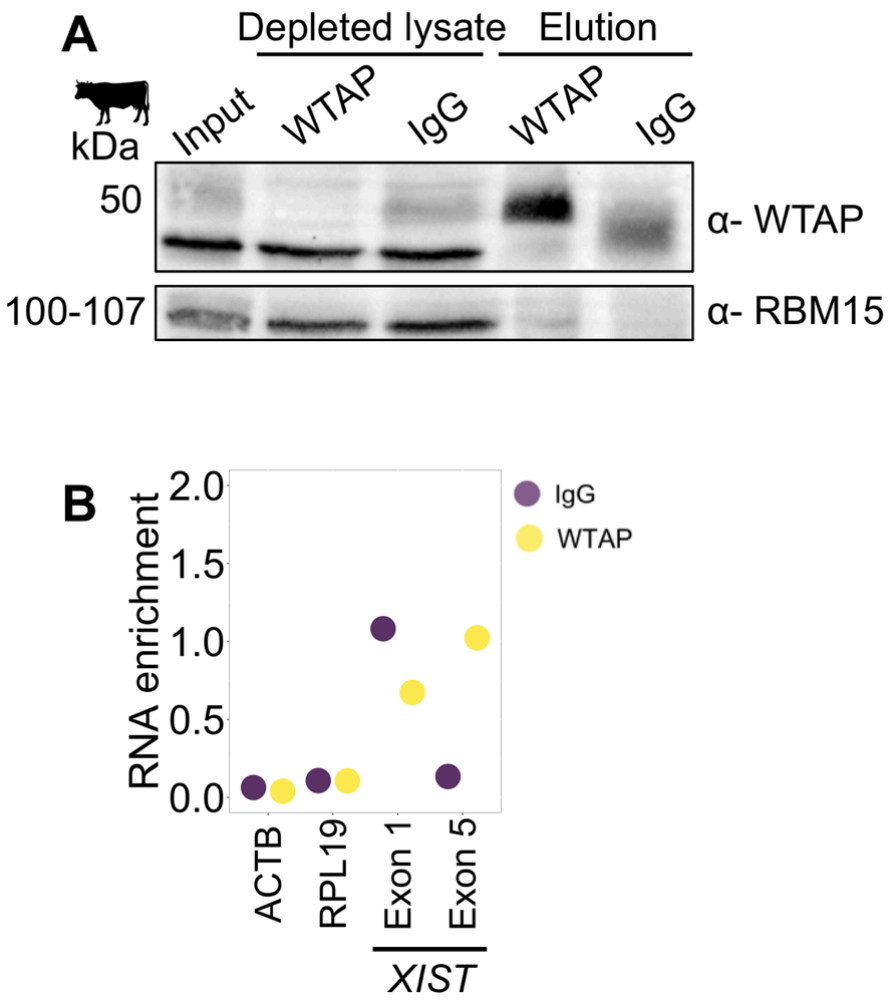


Figure 3.19. RIP of WTAP protein in cow.

A) Representative image from western blot of WTAP RIP samples in whole cell extract of ISHIKAWA cells. One isoform at 44 kDa predicted from Uniprot. The amount of protein loaded across input and depleted lysate samples was consistent and equivalent to the same number of cells (~500,000) whereas 10% of the elution was used for western blotting. n=1 biological replicate. **B**) RT-qPCR of WTAP RIP in whole cell extracts of ISHIKAWA cells. Fold enrichment of each transcript's abundance in RIP elutions was normalised to input using the $2^{\wedge-(Ct_{elution} - Ct_{input})}$ formula and reported as RNA enrichment. Three technical replicates were performed for the single biological replicate. Antibodies were used at 1:1000 in PBS-T. IgG serves as a non-specific negative control in RIP experiments. β-tubulin serves as a non-specific negative control in western blot. *ACTB* serves as a specific interacting transcript positive control whereas *U2* serves as non-specific transcript negative control. RIP, RNA immunoprecipitation; RT-qPCR, reverse transcription quantitative PCR

An hnRNPK-XIST interaction is closely linked to the establishment of a repressive chromatin environment for chromosomal-wide gene silencing of the X chromosome. In this experiment, an association between cow hnRNPK and cow XIST was tested. For this, RIP of hnRNPK was performed in a single biological replicate of whole cell bovine stromal extracts using the same antibody as for the ISHIKAWA pulldowns. Again, western blotting was carried out to address the suitability of this antibody for cow hnRNPK pulldowns. A noteworthy depletion of the hnRNPK protein was seen with the use of the anti-hnRNPK antibody, compared to the IgG (**Figure 3.20A**). Consistently, a clear enrichment of the hnRNPK protein was witnessed in the anti-hnRNPK elution sample versus the IgG. The presence of cow *XIST* and non-specific transcript controls bound to the pulled down hnRNPK protein was assessed via RT-qPCR. No enrichment of negative controls *ACTB* or *RPL19* was seen by RT-qPCR in any of the elution samples. In spite of a clear enrichment of bovine hnRNPK protein in the elution sample with the anti-hnRNPK antibody, a related concomitant *XIST* enrichment was not detected regardless of *XIST* primers used, with *XIST* abundance being similar across both anti-hnRNPK and IgG elution samples (**Figure 3.20B**). Therefore, no association could be reliably tested between bovine hnRNPK and *XIST* with a single replicate under these conditions.

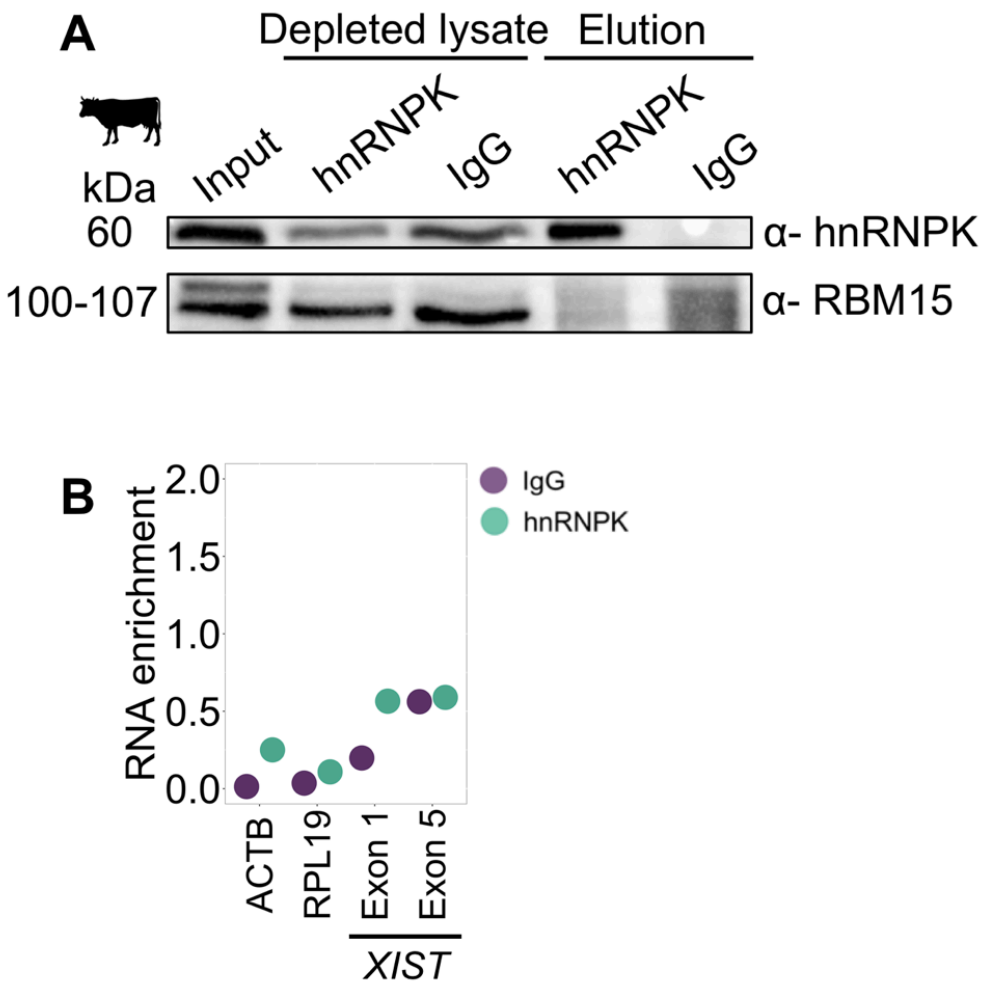


Figure 3.20. RIP of hnRNPK protein in cow.

A) Representative image from western blot of hnRNPK RIP samples in whole cell extract of ISHIKAWA cells. One isoform at 60 kDa from datasheet (51 kDa from Uniprot). The amount of protein loaded across input and depleted lysate samples was consistent and equivalent to the same number of cells (~500,000) whereas 10% of the elution was used for western blotting. n=1 biological replicate. **B**) RT-qPCR of hnRNPK RIP in whole cell extracts of ISHIKAWA cells. Fold enrichment of each transcript's abundance in RIP elutions was normalised to input using the $2^{\wedge-(Ct_{elution}-Ct_{input})}$ formula and reported as RNA enrichment. Three technical replicates were performed for the single biological replicate. Antibodies were used at 1:1000 in PBS-T. IgG serves as a non-specific negative control in RIP experiments. β-tubulin serves as a non-specific negative control in western blot. *ACTB* serves as a specific interacting transcript positive control whereas *U2* serves as non-specific transcript negative control. RIP, RNA immunoprecipitation; RT-qPCR, reverse transcription quantitative PCR

Given the central role of the SPEN-XIST interaction in human and mouse, the aim of the next experiment was focused on examining whether such an interaction would also exist in cow. Since the anti-SPEN antibody had previously worked successfully in ISHIKAWA pulldowns, RIP was performed here in whole cell bovine stromal cell extracts using the same antibody. Despite no information was provided in the SPEN antibody manufacturer's datasheet with regards to the peptide used to raise the antibody, an ~88% amino acid similarity between human and cow SPEN was shown previously in Chapter 2 (**Figure 2.11**). Following RIP, a single biological replicate of the elution with the SPEN and the IgG antibodies were analysed with mass spectrometry to confirm the successful pulldown of cow SPEN in bovine stromal cells, as performed for ISHIKAWA cells. Cow SPEN protein was not found in either elution sample from anti-SPEN or IgG from the mass-spec report (data not shown). Therefore, it is likely the anti-SPEN antibody used is not suitable for RIP of bovine SPEN. Therefore, an association between cow SPEN and cow *XIST* could not be assessed here since the anti-SPEN antibody used probably did not recognise the bovine SPEN protein.

3.4. Discussion

In this chapter, work was focused on experimentally demonstrating whether putative protein partners of *XIST* biochemically interact with the *XIST* RNA in cells from placental mammals with different implantation strategies, given their co-ordinate expression seen in reproductive tissues/cells in the previous chapter.

3.4.1. Adapted RAP coupled to RT-qPCR enriches for human XIST but not its interactome

RAP had previously been used to identify protein partners of a specific lncRNA candidate, therefore this technique was selected to be adapted for the detection of the human *XIST* protein interactome in ISHIKAWA cells. RAP employs UV crosslinking to identify specific proteins bound to an RNA target, however, UV treatment is known to lead to RNA damage. A comparison of XIST abundance in UV crosslinked versus non-crosslinked cells revealed a decrease in *XIST* abundance upon UV treatment (**Figure 3.4** and **Figure 3.5**), further compounded by an already low steady-state *XIST* expression in the ISHIKAWA cell line (**Figure 2.9 & 3.3**). This has implications for the use of RAP given the low proportion of a single RNA in the transcriptomic pool of a cell and the need to attain high enrichment levels in order to robustly and specifically pulldown its protein partners.

Despite these implications, an *XIST* enrichment of up to ~300-fold was achieved in RAP elutions (**Figure 3.8**). Following pulldown, *XIST* enrichment was not consistent across the whole sequence and varied depending on regions amplified. More specifically, the highest level of enrichment was seen across *XIST* exons 2-3, followed by exon 6 whereas exon 1 was never found enriched at a level higher than 9-fold over input (**Figure 3.8A**). The differences in enrichment detected by different primer pairs might be the result of primer efficiency biases, amplification of different XIST isoforms, a lack of accessibility from some antisense probes on *XIST* or a lack of full-length *XIST* isolation. Differential enrichment across *XIST* regions following RAP is unlikely to be the result of primer efficiencies as the pattern observed in elution samples was different to steady-state *XIST* abundance (compare **Figure 3.3**

and **Figure 3.7**). Several *XIST* isoforms are present in humans (**Figure 3.6B**). It is evident that primers targeting exons 2-3 and 4-5 would amplify more *XIST* isoforms than primers targeting exon 1 or 6 (**Figure 3.6B**), which is consistent with the amplification pattern seen (**Figure 3.3**). The pattern of *XIST* enrichment could also be related to the location of the probes and the presence of local secondary structure. However, care was taken during the design stage to avoid probes overlapping known repetitive regions on XIST and also incubation of ISHIKAWA lysates with the XIST probe pool occurred at 66^{o}C, which would melt low to medium strength local structures. One of the reasons why full-length *XIST* would not be pulled down could be due to fragmentation, perhaps due to shearing during lysate preparation, the use of too few probes for the pulldown of such a large transcript or RNase-mediate degradation owed to lysate contamination. RNA fragmentation during lysate preparation is very likely given the passing of the lysate through a needle and syringe, although none of the above were explicitly addressed here. Overall, the pattern of *XIST* enrichment seen is likely a combination of preferential isoform capture and amplification by the exon 2-3 targeting primers and likely degradation of the *XIST* RNA.

The highest *XIST* enrichment (~307-fold; **Figure 3.8**) was achieved with the use of 160 million ISHIKAWA cells and 9.6 mg of beads for capture. Nevertheless, when 258 million cells were used with ~15.5 mg of beads, more beads than just scaling for cell number, a lower *XIST* enrichment was noted (**Figure 3.9A**). Whilst a higher number of beads may have been used, it is likely that due to the addition of almost 100 million cells relative to the previous experiment, the extra copies of *XIST* present in the starting lysate (input) would decrease the observed fold enrichment as it is calculated relative to input. Despite the increase in cell number used as input, no proteins were detected bound to *XIST*, which indicated that the achieved level of enrichment was not sufficiently high (**Figure 3.9B&C**). Nevertheless, the rationale behind increasing the number of cells used was two-fold. A low starting quantity of *XIST* would mean a higher difficulty in enriching it at sufficient levels to detect its protein partners, therefore having more available *XIST* would allow for more probe binding. Moreover, a higher cell number would help in overcoming the inherently low UV crosslinking inefficiency ranging from 1-5% (Fecko et al., 2007), which would reduce the abundance of lncRNA-protein complexes. Going forward, to further

augment *XIST* RNA enrichment with RAP, one avenue would be to design more antisense probes against the XIST RNA given its transcript length. Another experiment one could do even before that would be to test the levels of *XIST* enrichment with fewer than 10 probes. An alternative route to boost RNA enrichment would involve the use of a higher amount of bead to probe (maximum 2-2.2-fold excess)(ThermoFisher Scientific, 2016) or the use of a different cell line where *XIST*'s abundance would be higher.

3.4.2. RIP coupled to RT-qPCR demonstrates an association between human *XIST* and several putative protein partners in endometrial cells

RIP coupled to RT-qPCR has often been employed to reciprocally validate lncRNA-protein interactions following identification of candidate protein partners of a specific lncRNA via an RNA-centric approach. Thus, using a protein-centric approach here the ability of putative protein partners of *XIST* to enrich for the *XIST* RNA was assessed in human and cow. CIZ1, WTAP, hnRNPK, SPEN and RBM15 proteins were the ones selected to be assayed for an interaction with XIST. These were selected primarily on the availability of functional studies linking the role of the protein to XCI and secondarily, on amino acid conservation across species of interest and availability and efficacy of antibody at detecting the protein in western blotting.

Employing RIP of CIZ1 in whole cell ISHIKAWA extracts depleted all of the CIZ1 protein from the input as seen in depleted lysates (**Figure 3.10A**), indicative of efficient capture. Given the strict nuclear localisation of CIZ1 however (Warder and Keherly, 2003), it could also be indicative of a 'diluted' initial abundance of CIZ1 in a complex pool of proteins from both the cytoplasm and the nucleus. Thus, CIZ1 would be less represented in the amount of whole cell lysate loaded on the gel for western blotting compared to a more 'concentrated' presence in nuclear-enriched lysates (**Figure 3.10,** compare A & B). Hence, it could give the impression that depletion of CIZ1 in whole cell extracts is more efficient than in nuclear-enriched lysates. This would also explain the differences in the presence of β-tubulin in depleted lysate and elution samples across whole cell and nuclear-enriched lysates.

Notably, CIZ1 was found to associate with human *XIST* in both whole cell and nuclear-enriched lysates (**Figure 3.10 & Figure 3.11**), in agreement with other pulldown studies in human HEK293FT and K562 cells (Sunwoo et al., 2017, Lu et al., 2020a). Another noteworthy finding reported here was the *ACTB* enrichment in elutions from the CIZ1 antibody over the IgG, revealing a previously unappreciated association between human CIZ1 and *ACTB*, confirmed both by RT-PCR (**Figure 3.10C**) and RT-qPCR (**Figure 3.11B**) across several biological replicates. Similar to CIZ1, hnRNPK was also shown here to associate with *ACTB* and *XIST* in human endometrial cells (**Figure 3.14**). This was in line with previous studies in human demonstrating a hnRNPK-*XIST* interaction from pulldown assays (Van Nostrand et al., 2016, Graindorge et al., 2019). Although a direct hnRNPK-*ACTB* interaction could not be traced in the literature, hnRNPK binding sites can be found on *ACTB* (Van Nostrand et al., 2016, Dominguez et al., 2018).

An association between RBM15, WTAP and human *XIST* could not be reliably validated under the conditions in which RIP was used in here (**Figure 3.12 & Figure 3.13**). This was in contrast to previous studies RBM15 using CLIP data, which showed a robust interaction of XIST with RBM15 (Patil et al., 2016, Van Nostrand et al., 2016, Lu et al., 2020a). Equally, an interaction cannot not be ruled out because of the weak RBM15 protein enrichment following western blotting of RBM15 RIP elution samples. (**Figure 3.12**), perhaps due to a small fraction of available RBM15 being immunoprecipitated or owed to the anti-RBM15 antibody not being suitable for RIP. The low abundance of *XIST* in the ISHIKAWA model system combined with a potential low proportion of RBM15 or WTAP bound to *XIST* at any one time could comprise another reason as to why these interactions were not observed here. The detection of *XIST* in elution samples could be further compounded depending on whether proteins such as RBM15, WTAP, SPEN and LBR or others not described here, bind to the *XIST* repeat A competitively with one another. Competitive binding of RBM15 and SPEN has been speculated in the literature previously based on mouse eCLIP data (Cirillo et al., 2016, Pintacuda et al., 2017b).

A previous study aiming to identify protein partners of *XIST* used a CRISPR-dead Cas9 fusion with biotin ligase and guide RNAs navigating the biotin ligase to *XIST* locations (Yi et al., 2020). After feeding free biotin to cells, biotin ligase labels

proteins nearby *XIST* allowing for their identification. Although RBM15 was detected, it was not above the enrichment cut-off as a specific *XIST* partner in their dataset (Yi et al., 2020). In another study performing RIP-seq with RBM15 in a human bone marrow MEG-01 cell line, *XIST* was not identified in their dataset (Zhang et al., 2015). These two studies and work presented here are in contrast to several lines of evidence that support a direct interaction between RBM15, WTAP and human *XIST* (see **2.1 Introduction**). Technical shortcomings aside, the results presented here could argue for a cell/tissue-type specific *XIST* interactome whereby RBM15 is not bound by XIST either at all or at a high enough level in endometrial cells whereas this interaction could occur in kidney cells. Furthermore, these observations could reflect differences in the dynamic *XIST* interactome across XCI stages, with different interactions preserved following the establishment of XCI. In support of this logic, a recent study reported a 71.3% overlap between human *XIST* protein partners across B cells (GM12878 cell line) and myeloid cells (K562 cell line)(Yu et al., 2021). Moreover, this study described a 57.8% overlap between protein partners of Xist across B cells and ESCs with inducible expression of *XIST*. Finally, observations made here could also be taken to mean that the RBM15-*XIST* interaction could be dispensable for endometrial cells and a different protein has been co-opted to undertake RBM15's function in this particular cell/tissue type. Redundancy in the m6A methylation pathway, in which RBM15 has been described to participate in, has been previously documented with mouse Ythdf m6A reader proteins (Lasman et al., 2020) and human RBM15 with its paralog RBM15B (Patil et al., 2016). RIP with the RBM15B protein was not performed here, however.

In line with previous studies in human HEK293T and K562 cells (Graindorge et al., 2019, Lu et al., 2020a), an association between SPEN and human *XIST* was shown in human endometrial cells (**Figure 3.15**). Due to a lack of a suitable antibody for western blotting of the SPEN protein, elutions from RIP with the SPEN antibody were subjected to mass spectrometry for protein identification. Only 114 of those had more than two unique peptides and 74 were specific to SPEN compared to the IgG, as measured by a spectral count ratio >1 (**Table 3.6**). These could comprise protein partners of SPEN (i.e. part of the same protein complex as SPEN), protein partners of SPEN targets (i.e. bound on the same transcript as SPEN, e.g. *XIST*) or both. Of

that set, 17/74 (~23%) had been previously identified in a previous study using pulldowns with SPEN's SPOC domain in mESCs (Dossin et al., 2020).

Some of those proteins identified to be specifically enriched in the SPEN elution have been shown to bind to mouse (Chu et al., 2015b, McHugh et al., 2015, Minajigi et al., 2015) and human (Graindorge et al., 2019, Yi et al., 2020, Yu et al., 2021) *XIST* in previous studies such as SAP18, RBMX, PNN, EIF4A3, PABPN1, TRA2B and SRSF1/6/10. GRIPAP1, the most highly enriched protein in this dataset (81 unique peptides, 113 spectral count ratio), has not been seen in any *XIST* interactome studies or genetic screens. Hence it is unknown whether it interacts with SPEN or plays a role in XCI via *XIST* association. A role for GRIPAP1 that has been described, unrelated to XCI, involves GRIPAP1 acting as a scaffold protein, binding JNK and an upstream kinase MEKK1 facilitating JNK signaling in cultured cortical neurons. Additionally, it has been reported to interact with the GRIP1 protein and AMPA receptors in the rat brain mediating plasticity and synaptic transmission, since its overexpression specifically disrupted AMPA receptor synaptic targeting (Ye et al., 2000).

SPEN has previously been shown to interact and recruit proteins from the NCOR2/SMRT and NuRD complexes (Shi et al., 2001, McHugh et al., 2015), however, no proteins from these complexes were detected here. Moreover, besides RBMX, other proteins from the m6A methylation machinery such as RBM15, RBM15B or WTAP were not seen here, despite their co-immunoprecipitation with SPEN in mESCs in other studies (Malovannaya et al., 2011, Horiuchi et al., 2013, McHugh et al., 2015, Coker et al., 2020, Dossin et al., 2020). This was unexpected given a central role of those complexes in aiding SPEN with establishing X-linked gene silencing (McHugh et al., 2015, Dossin et al., 2020). However, it is known that SPEN binding to chromatin (where it engages other protein partners) decreases, after X-linked gene transcription diminishes as XCI establishment ensues (Dossin et al., 2020). Given the differentiated status of ISHIKAWA cells used here which have established XCI, it is likely that a smaller fraction of SPEN would be found associated with these complexes as less SPEN would be expected to be bound on chromatin, perhaps making it more difficult to pick those associations up. Hence, it's possible that SPEN's interactome is dynamic and dependent on XCI status.

In pulldown experiments it is reasonable to expect regions where proteins bind on *XIST* to be more enriched over other *XIST* regions where protein binding is not observed (or predicted). The hnRNPK protein has been shown to interact with repeat B (part of exon 1) on *XIST*. Following RIP of hnRNPK here, RT-qPCR results from the elution with the hnRNPK antibody revealed that *XIST* exon 1 was more highly enriched compared to exons 2-3 or exon 6. CIZ1 has been shown to interact with repeat E (part of the 5' end of exon 6) on *XIST*. Here, the primer for *XIST* exon 2-3 was in closer proximity to repeat E than the primer for exon 6 (684 bp vs 2.3 kbp away from repeat E, respectively), justifying the result observed. Nevertheless, human SPEN and WTAP have been shown to interact with repeat A (part of exon 1) on *XIST*, whilst here the enrichment of *XIST* exon 2-3 was higher than that of exon 1 in the elutions for both proteins. Perhaps this pattern could be justified by RNA fragmentation/shearing, due to the size of full-length XIST, eliminating exon 1 primer binding sites following capture of the protein and pulldown of XIST. However, RNA extracted from RIP experiments was not examined via agarose gel electrophoresis due to limiting amounts of RNA. Therefore, RNA degradation as a driving force behind this observation cannot be ruled out.

3.4.3. RIP coupled to RT-qPCR in bovine stromal cells highlights potentially shared *XIST* protein partners across human and cow

The cow is a placental mammal with different early pregnancy events and implantation strategies compared to human (Chapter 2.1.). Here it was used as a model to enable the identification of shared and novel putative protein partners bound across human and cow.

Mass spectrometry of elution samples from RIP with the SPEN antibody in bovine stromal cells showed a lack of cow SPEN enrichment. Thus, RIP pulldowns were not successful for cow SPEN given the antibody likely did not recognise the protein. Western blots used to evaluate success of RIP pulldowns showed that cow CIZ1, WTAP, RBM15 and hnRNPK proteins could be found enriched in elution samples with specific antibodies compared to IgG controls, indicative that pulldowns worked. However, *XIST* binding was not consistently observed for any of the above proteins.

A previous study used selective pressure variation analyses across 43 vertebrate species and did not report any sites on the WTAP gene as being under positive selection, implying purifying selection is acting on this protein-coding gene. Given the amino acid sequence of WTAP was found to be >95% across mouse, human and cow and the functions of this protein are conserved (based on the study above), a lack of *XIST* binding was unexpected here. As previously state above, it is possible these observations relate to a cell-type specific *XIST* interactome that is dynamically different across tissues and XCI stage.

A technical shortcoming that could account for both the inconsistency in bovine *XIST* enrichment from CIZ1 and WTAP RIP as well as the lack of an inferred association with bovine *XIST* from RBM15 and hnRNPK RIP experiments would be related to lysate preparation. More specifically, a weak lysis buffer unable to achieve complete cell lysis of the specific cell number required to perform RIP could translate into three biological replicates each with varying cell numbers below the threshold for efficient pulldown. In turn this could underlie suboptimal protein depletion from input samples and insufficient enrichment in elution samples, consequently limiting the available *XIST* RNA for RT-qPCR detection.

In future experiments, lysis buffers should be tested and proven they achieve complete lysis consistently despite kit recommendations since various cell lines might require buffers with different ionic strength. Additionally, future experiments should focus on generating additional biological replicates of RIP with RBM15, WTAP and hnRNPK antibodies in human and cow to ensure one-off effects are not interfering with data interpretation and robustly establish whether these proteins associate with *XIST*. Given the anti-SPEN antibody used here was not able to pulldown the bovine SPEN protein, future approaches could employ RIP with antibodies raised against a specific RRM domain of bovine SPEN or using an antibody against a tag to bind a constitutively expressed tagged version of the protein.

Due to RIP being sensitive to detecting artefactual interactions (see **3.1 Introduction**), future attempts to establish a direct and specific interaction between XIST and its putative protein partners could employ the use of lysates prepared from

UV-treated cells. This way interactions that are taking place endogenously in a cell are fixed prior to cell lysis and harsh washing buffers can be used to remove non-specific associations. This approach is also better at determining whether a protein directly interacts with an RNA, or whether the interaction is mediated by another protein which is part of the same complex.

In summary, complementing findings from the previous chapter, RIP in human and cow revealed an association between *XIST* and a subset of the earlier described mouse *Xist* protein partners. Notably, while there was an overlap in the subset of putative protein partners between human and cow, protein partners of human *XIST* were different to those of bovine *XIST*. Human *XIST* was found to associate with CIZ1, WTAP, hnRNPK and SPEN, whereas an association with RBM15 could not be robustly confirmed here. In contrast, none of the proteins tested could be reliably verified as associating with cow *XIST* under the described RIP conditions. These results will be orthogonally validated, and protein partners of cow *XIST* further explored in the next chapter.

4. Chapter 4: Identification of human and cow *XIST* repeat region protein partners

4.1. Introduction

Characterising the protein interactome of a lncRNA can offer insight into the various pathways it can participate in, as well as provide hints to the mechanism behind achieving that. Understanding where proteins bind on a lncRNA, can highlight parts of the lncRNA with functional importance. Modern high-throughput approaches aimed at identifying protein partners of an endogenous transcript (e.g. RAP described in **Section 3.1**) provide no information about the part of the RNA that each protein is associated with. Early methods to identify protein partners of putative lncRNAs were based on *in vitro* transcription pulldowns. This approach relies on the exogenous *in vitro* transcription of a lncRNA sequence from a plasmid or PCR amplicon (**Figure 4.1A**) coupled to a reaction appending biotin groups to the transcription product(s) (**Figure 4.1B**). The biotinylated transcription products would be introduced into a cell lysate (**Figure 4.1C**) and lncRNA-protein complexes allowed to form (**Figure 4.1D**). The introduction of streptavidin-coated magnetic beads (**Figure 4.1E**) enables the isolation of biotinylated lncRNA-protein complexes (**Figure 4.1F**). The identification of proteins from RNA pulldowns is then performed by mass spectrometry (MS). Such an approach has been used to characterise the protein interactome of several lncRNAs, including *HOTAIR* (Rinn et al., 2007a), *NORAD* (Tichon et al., 2016) and mouse *Xist* (Pintacuda et al., 2017a). It has also been used to explore the protein partners of specific elements harboured within lncRNAs that ascribe them a function, such as the short interspersed nuclear elements (SINE), which upregulate target mRNA translation (SINEUP class of lncRNAs)(Toki et al., 2020).

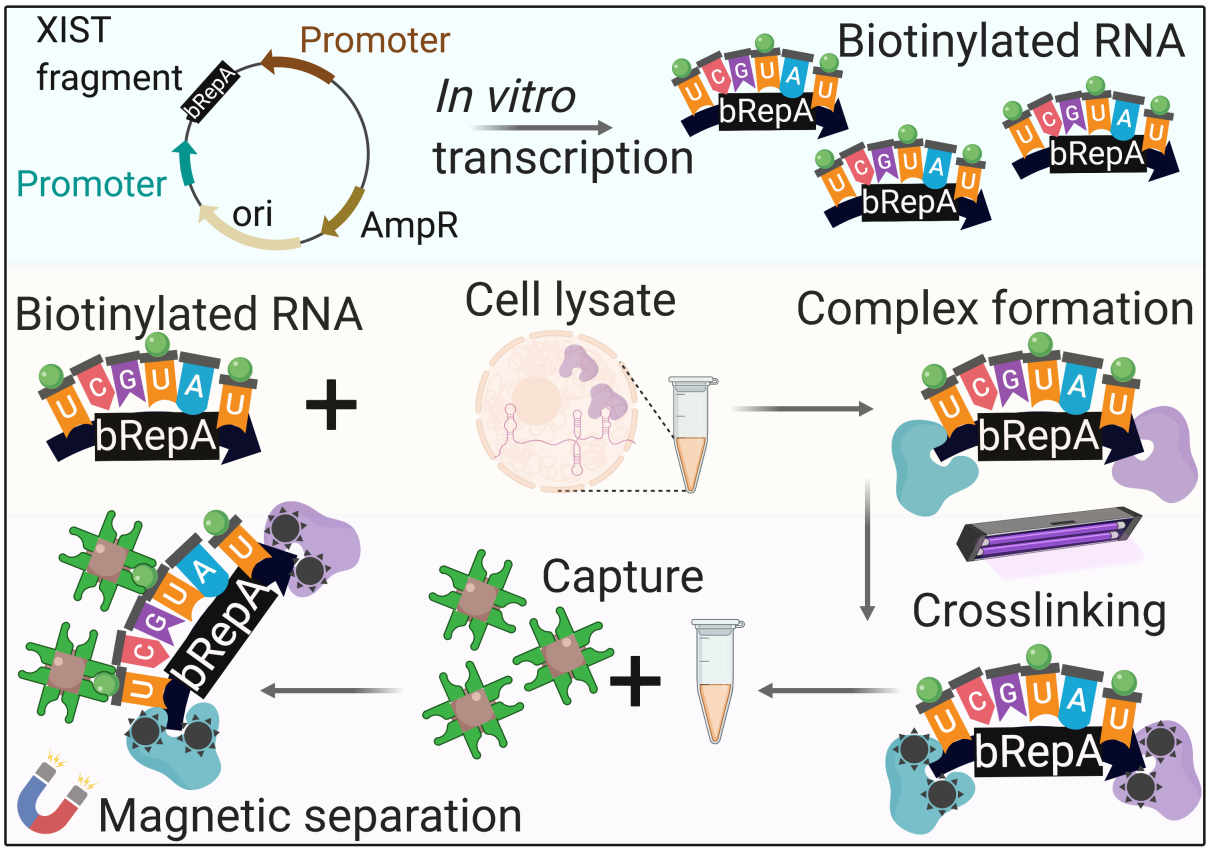


Figure 4.1. Schematic of XIST in vitro transcription and pulldown approach.

A) The template for *in vitro* transcription (black 'Repeat' box) is cloned into a plasmid containing an upstream and downstream promoter (for sense or antisense transcripts, respectively). **B**) For run-off *in vitro* transcription, the plasmid is linearised shortly downstream of the 3' end of Xist sequence cloned and transcribed with either T7 or SP6 RNA polymerase, a pool of dNTPs (grey 'C', purple 'G' and blue 'A') and biotin-modified dUTPs (orange 'U' with green spheres), which generates biotinylated RNA. **C&D**) Biotinylated RNA is introduced in a protein lysate and RNA-protein complexes are allowed to form. **E**) Streptavidin-coated magnetic beads (brown spheres with green chevrons) are introduced to recover the target RNA by interacting with the biotin group of the biotinylated *in vitro* transcribed RNA. **F**) Magnetic separation is employed to isolate bead RNA-protein complexes from the lysate, which are washed and then RNA-protein complexes eluted. Created with BioRender.com.

The specific type of mass spectrometry used has varied across the studies of lncRNA protein partners through the years. Initial studies utilised label-free proteomics, which involves trypsin digestion of proteins in a control and a treated sample prepared and analysed separately by liquid chromatography and MS. Relative quantification is achieved by measuring signal intensity from each peptide in the control sample and comparing to the intensity of the same peptides from the treated sample (Bantscheff et al., 2007). With a label-free approach, samples are analysed one by one successively, which can result in weak comparisons across different samples. The median protein coefficient of variation (variability compared to the mean of the sample size) has been reported to be 20% across replicates (Pappireddi et al., 2019). Additionally, label-free approaches cannot distinguish between a peptide missing due to a) being misidentified, b) having an abundance below the lower detection limit or c) being absent from the sample (Karpievitch et al., 2012).

More recently, quantitative proteomics approaches using isobaric mass tag labelling, such as tandem mass tags (TMT), are frequently employed to enable multiplexing of samples run in parallel, mitigating some of the shortfalls of label-free mass spectrometry. TMT-MS has some obvious advantages over stable isotope labelling by amino acids in cell culture (SILAC), a gold standard in proteomic analyses due to its accuracy. In SILAC, cells in culture are fed amino acid reagents that have been labelled with stable isotopes, therefore generating nascent isotope-labelled proteomes by live, metabolically active cells (Ong and Mann, 2006). SILAC was one of the first technologies that enabled multiplexing up to three different conditions. TMT-MS, however, does not depend on growing a sufficient number of cells and is not limited by three conditions compared in parallel (comparisons up to 11 conditions are possible). In TMT-MS, each isobaric tag contains an amine group and a reporter group, bridged by a balancer group (Thompson et al., 2003)(**Figure 4.2A**). Isotopes present on the reporter and balancer groups are distributed so that all tags have the same mass. Characteristic isotope patterns remain following reporter group cleavage during the mass spectrometry run, since each tag will have a signature of light/heavy isotope distribution (Dayon et al., 2008)**(Figure 4.2B**). In 6-plex TMT each of the six tags has a specific reporter ion that appears at a mass-to-charge ratio of 126.1, 127.1, 128.1, 129.1, 130.1, and 131.1, facilitating the tracking of which sample the

labelled peptide was present in (Dayon et al., 2008). TMT mass spectrometry has been preferred in general because this technology allows relative quantitation of protein abundance across samples which are analysed in parallel due to multiplexing, resulting in a reduced replicate-to-replicate variation with a median coefficient of variation of <10% (Sonnett et al., 2018, Pappireddi et al., 2019). Furthermore, given only tagged peptides are analysed, it mitigates the issue of comparing signal intensity from low abundance proteins in the sample since even low abundance proteins would be tagged.

In the previous chapter, an association between human *XIST* and CIZ1, WTAP, hnRNPK and SPEN proteins, but not RBM15, was evident from RIP in ISHIKAWA cells. On the contrary, none of these interactions were found to occur with bovine *XIST*. The aim of this chapter is to map the location of protein partner binding on human and bovine *XIST*, to explore previously uncharacterised protein partners of bovine *XIST* via mass spectrometry as well as orthogonally validate interactions found in the previous chapter. Studies in mouse have showed an interaction of Spen, Wtap and Rbm15 with *Xist* repeat A, Ciz1 with Repeat E and hnRNPK with repeats B and C (**Section 1.6.2**). In chapter 2, the similarity of *XIST* repeat A and repeat E between human and cow were estimated to be the most and least conserved, at ~85% and ~54.6%, respectively (**Table 2.3**). The hypothesis here was that a highly conserved *XIST* repeat A would be more likely to have maintained protein partners compared to a modestly conserved *XIST* repeat E. Therefore, interactions observed for human *XIST* repeat A are expected to be seen for *XIST* repeat A in cow, whereas interactions observed for human *XIST* repeat E are not expected to be seen maintained in cow.

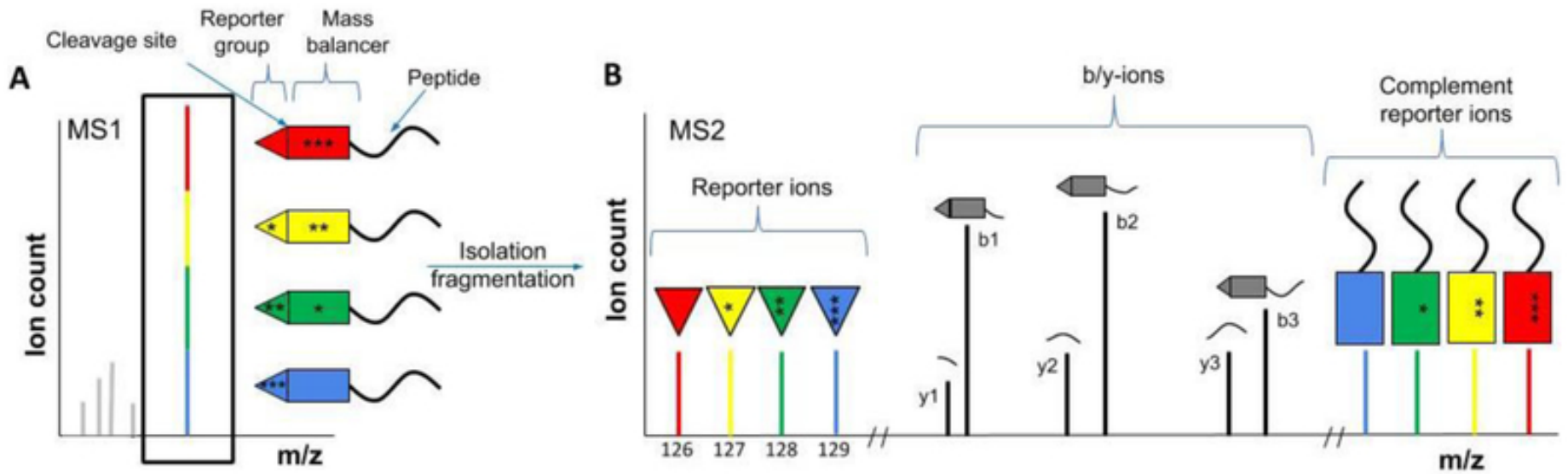


Figure 4.2. Quantitative proteomics with isobaric tags overview.

A) Four peptides tagged with different isobaric tags of equal total mass but disproportionate distribution of heavy isotopes across the reporter group and mass balancer group. Isobaric tags are coloured coded. In the first phase of liquid chromatography-mass spectrometry (LC-MS), tagged peptides are detected as a single peak in the first analytical stage of mass spec. **B**) Following collision-induced dissociation, peptides fragment and either the reporter ions or the peptide backbone is cleaved off. In the former case, reporter ions will show different masses in the second analytical stage of mass spec which can be used or relative quantification. In the latter case, the balancer groups will also have a different mass-to-charge ratio and can be used for quantification. Following fragmentation of peptides, if the charge is retained near the N-terminus, the product is called a b-ion whereas if the charge is at the C-terminus, it is a y-ion. These can be used to reconstruct the identity of each protein. Adapted from (Pappireddi et al., 2019).

4.2. Materials and Methods

Unless otherwise stated, all reagents were purchased from ThermoFisher, UK.

4.2.1 Cloning of human and bovine XIST repeats

Human and bovine *XIST* repeats A and E were PCR amplified from cDNA generated from either total RNA of ISHIKAWA (human) or bovine stromal cells (see **Table 4.1** for primers used) with the Q5 High-Fidelity DNA polymerase (NEB, UK) or the non-proofreading Taq DNA Polymerase (EP040; ThermoFisher, UK). Thermocycling was followed as shown in **Table 4.2**.

Amplicons were run on 1-1.5% agarose gels, bands corresponding to expected sizes were excised and DNA was purified using the QIAquick Gel Extraction Kit (QIAGEN, UK). Human XIST repeat A was inserted into T3 plasmid vector (pEASY-T3 Cloning kit; TransGen Biotech, UK) downstream of a T7 promoter and upstream of a SP6 promoter. All other XIST fragment sequences were inserted into a TOPO Blunt vector (Zero Blunt™ TOPO™ PCR Cloning Kit; ThermoFisher, UK), downstream of an SP6 promoter and upstream of a T7 promoter. Plasmids were introduced into One Shot™ TOP10 Chemically Competent *E. coli* by mixing the ligation reaction with 15 µl of competent cells and incubating on ice for 30 minutes. Next, cells were heat-shocked for 30 seconds at 42°C, followed by briefly incubating on ice for 2 minutes. After the addition of 150 µl of room temperature S.O.C. medium, cells were incubated at 37°C for 1 hour while shaking at 1000 RPM in a thermoshaker block. In LB agar plates (with 50 µg/mL kanamycin for the TOPO Blunt vector or 100 µg/mL ampicillin/carbenicillin for the T3 vector), 150 µl from each transformation reactions were evenly spread using a glass spreader and incubated overnight at 37°C.

Bacterial cultures were grown from single colonies inoculated into 3-8 ml of LB media (50 µg/mL kanamycin for the TOPO Blunt vector or 100 µg/mL ampicillin/carbenicillin for the T3 vector). Cultures were incubated at 37ºC for 14-16 hours whilst shaking at 220 RPM. Plasmid DNA was isolated from bacterial cultures using either the E.Z.N.A.® Plasmid DNA Mini Kit I (Omega Bio-tek through VWR,

UK) and the QIAGEN® Plasmid Plus Midi Kit (QIAGEN, UK). Plasmid DNA was isolated following culture centrifugation at 10,000 xg for 10 minutes at room temperature according to manufacturer’s instructions. Plasmid DNA concentration and purity were examined by Nanodrop measurements. The correct sequence of plasmid constructs was confirmed using Sanger sequencing with M13F and M13R primers (GENEWIZ, UK).

Table 4.1. List of primers used for cloning of XIST Repeat A and E.

All primers target exons and contain restriction enzyme recognition sites (bold) separated with a 3-4 bp 'stuffer' sequence (lowercase). Forward primers contain a 5' recognition site for NotI whereas reverse primers a 3' recognition site for SalI. Primer melting temperatures were estimated before the addition of the restriction sites or the stuffer sequence using the IDT OligoAnalyzer tool (https://eu.idtdna.com/calc/analyzer). Amplicon sizes were predicted by the *in silico* PCR software MFEprimer (https://mfeprimer3.igenetech.com/spec).

Species	Region	Amplicon size (bp)	Primer	
			Orientation	Sequence (5'-3')
Cow	Repeat A	1156	Forward	**GCGGCCGC**attaGGATTTCTTTGCCTGTGTGGT
			Reverse	**GTCGAC**atgtTCTTCTCCCGCTCATTTTCC
	Repeat E	923	Forward	**GCGGCCGC**taacTGCTCATCACTGTAGTTTGTCTCT
			Reverse	**GTCGAC**tcaTGAGTCTTCTATCCAACTCCAGTC
Human	Repeat A	931	Forward	**GCGGCCGC**tacCCCCCAACACCCTTTATGG
			Reverse	**GTCGAC**tcagGACTTCCTCTGCCTGACCTG
	Repeat E	1881	Forward	**GCGGCCGC**atcatGCACTCTAGCACTTGAGGATAGC
			Reverse	**GTCGAC**cttgctGAGTAGCGTTGGCACAGTCCA

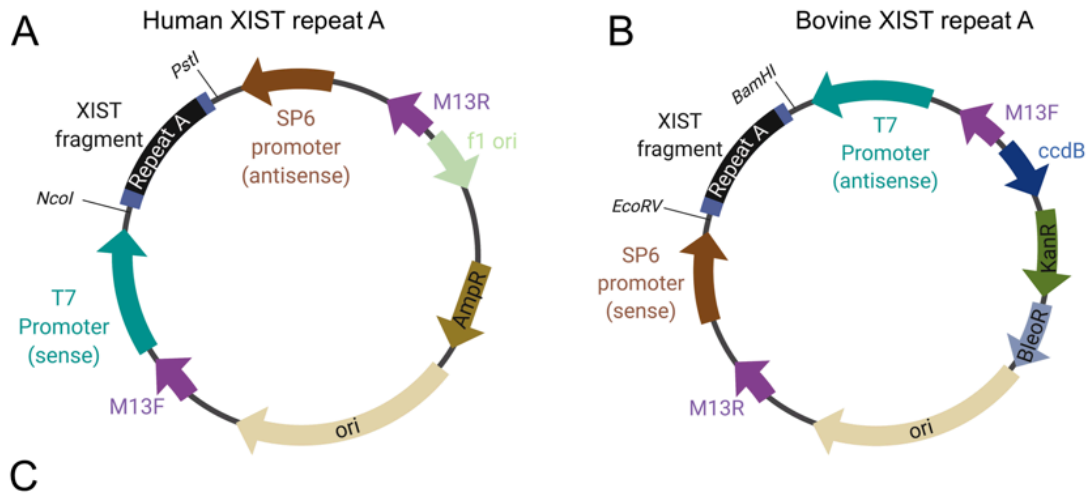
Table 4.2. Thermocycling conditions for PCR amplification of XIST fragments.

STEP	TEMPERATURE	TIME
Initial Denaturation	98°C	30 seconds
	98°C	10 seconds
	*__57.5__°C	30 seconds
35 Cycles	72°C	40 seconds/kb
Final Extension	72°C	2 minutes
Hold	4°C	

*Annealing temperature was tailored according to primer set

4.2.2. Run-off *in vitro* transcription of human and bovine XIST repeats using biotinylated nucleotides

Plasmids were linearised with excess restriction enzyme (see below for enzymes; NEB) for 3 hours at 37°C. Typical reactions included 5 µg of plasmid DNA, 1x CutSmart Bufer (NEB, UK), 130 U of restriction enzyme (NEB, UK) in ddH2O in a 100 µl total volume. The use of restriction enzyme was dependent on the plasmid vector, XIST repeat construct and promoter sequence to be utilised for the transcription of the sense (assayed treatment) or antisense (negative control) construct needed (summarised in **Figure 4.3**).



C

Species	XIST region	Restriction enzyme for linearisation	Promoter used	Orientation	RNA length (nt)
Human	Repeat A	PstI	T7	Sense	1062
		NcoI	SP6	Antisense	1010
	Repeat E	EcoRV	SP6	Sense	1979
		BamHI	T7	Antisense	1994
Cow	Repeat A	EcoRV	SP6	Sense	1254
		BamHI	T7	Antisense	1268
	Repeat E	EcoRV	SP6	Sense	1036
		BamHI	T7	Antisense	1021

Figure 4.3. XIST fragment transcription and RNA polymerase use per construct.

A) Plasmid map for human XIST fragment containing repeat A. B) Plasmid map for bovine XIST fragments A and E as well as human XIST fragment containing repeat E. C) Orientation of *XIST* fragment transcription depends on restriction enzyme and polymerase used. Expected sizes were estimated by counting all nucleotides downstream from the last 'G' or 'T' of the T7 or the SP6 promoter sequence, respectively.

4.2.2.1. Plasmid linearization and purification

Linearised plasmid DNA was purified using acidified phenol:chloroform (ThermoFisher) and precipitated with ethanol. Briefly, samples were made up to 300 µl with ddH2O, mixed with 300 µl of acidified phenol/chloroform (pH 4.5 ± 0.2) and vortexed for 10 seconds. Then, samples were centrifuged at 17,000 xg for 10 minutes at 4°C and the aqueous layer was transferred into a fresh tube. Next, 2.5x volumes of 100% ethanol, 1 µl of GlycoBlue and a final concentration of 0.3 M NaCl were mixed with the aqueous layer and vortexed for 10 seconds. Samples were precipitated at -80°C overnight and the next day, they were centrifuged at 17,000 xg for 30 minutes at 4°C after thawing. Pellets were washed with 300 µl of ice-cold 70% ethanol and flicked to dislodge the pellet. Subsequently, the pellet was centrifuged at 17,000 xg for 15 minutes at 4°C, after which the supernatant was aspirated and the pellet left to air dry for 10 minutes. The pellet was resuspended in 30 µl of ddH2O and the purity and concentration of plasmid DNA were determined using the Nanodrop 8000 instrument.

4.2.2.2. *In vitro* transcription

RNA was transcribed using 1 µg of linearised plasmid DNA, 1x of Biotin RNA labelling mix (1x mixture: 1 mM ATP, 1 mM CTP, 1 mM GTP, 0.65 mM UTP, 0.35 mM Biotin-16-UTP, pH 7.5; Roche/Sigma) and ~40 U of either T7 (Roche/Sigma, UK) or SP6 (Roche/Sigma, UK) polymerase in 20 µl total volume at 37°C for 2 hours. For the SP6 polymerase, the HiScribe SP6 RNA Synthesis (NEB, UK) kit was also used. After transcription, reactions were DNase-treated using 6 U of Turbo DNase (ThermoFisher, UK) at 37°C for 15 minutes. RNA was purified using chloroform (ThermoFisher, UK) and precipitated with isopropanol. In brief, samples were made up to 300 µl with ddH2O, mixed with 300 µl of chloroform isoamyl alcohol (24:1) and vortexed for 10 seconds. Then, samples were left to incubate at room temperature for 3 minutes before a centrifugation step at 17,000 xg for 10 minutes at 4°C. The aqueous layer was next transferred into a fresh tube. Next, 300 µl of 100% isopropanol, 1 µl of GlycoBlue and 38.4 µl of 5 M NaCl were mixed with the aqueous layer and vortexed for 10 seconds. Samples were precipitated at -80°C overnight

then centrifuged at 17,000 xg for 30 minutes at 4°C after thawing. The pellet was washed with 300 µl of ice-cold 70% ethanol and flicked to dislodge the pellet. Subsequently, the pellet was centrifuged at 17,000 xg for 15 minutes at 4°C, after which the supernatant was aspirated and the pellet left to air dry for 10 minutes. The pellet was resuspended in 30 µl of ddH2O and the purity and concentration of plasmid DNA were determined by NanoDrop.

4.2.2.3. Assessment of biotinylated RNA products

Transcript sizes were checked on 1% denaturing formaldehyde agarose gels in 1x MOPS buffer (20 mM MOPS pH 7.0, 12 mM sodium acetate, 0.5 mM EDTA pH 8.0). To make a 60 ml denaturing gel, 0.6 g of agarose was dissolved in 43.2 ml of water, immediately after which 6 ml of 10x MOPS (200 mM MOPS pH 7.0, 120 mM sodium acetate, 5 mM EDTA pH 8.0) were mixed in and solution was cooled down to 50-60^{o}C. Subsequently, 10.8 ml of 37% formaldehyde (final concentration 6.7%) and 6 µl of x10,000 SYBR Safe staining dye were mixed in and the gel was poured in a casting tray. Prior to loading RNA samples, 1 µg of RNA was mixed with formamide dyes (95% v/v deionised formamide, 20 mM EDTA, 0.05% w/v xylene cyanol, 0.05% w/v bromophenol blue) at a final concentration of 63% formamide and heated at 68^{o}C for 15 minutes. Samples were electrophoresed at 90 V for 45 minutes. Sizes of RNAs were estimated based on an RNA ladder (RiboRuler High Range; NEB, UK).

The presence of biotin was confirmed via RNA slot blot. Briefly, 500 ng of RNA were diluted in Dot buffer (10mM NaOH and 1mM EDTA in ddH2O) and applied until dry onto an Amersham Hybond-XL nylon membrane (GE Healthcare, UK), pre-washed in ddH2O for 10 minutes. The membrane was blocked for 1 hour with 1% BSA in PBS + 0.5% SDS and then blotted with 1:1000 NeutrAvidin-HRP (ThermoFisher, UK) in PBS +10% SDS for 30 minutes at room temperature. The membrane was washed for 10 minutes in each of 10% SDS in PBS, 1% SDS in PBS and 0.1% SDS in PBS and then developed using ECL (Biological Industries, UK). Chemiluminescent signal was detected using a ChemiDoc XRS+ imaging system (BioRad, UK).

4.2.3. Generation of nuclear cell extracts

Pellets of 10 million ISHIKAWA or bovine stromal cells (described in **Section 2.2.2**) were fractionated as previously with a few modifications (Werner and Ruthenburg, 2015). Briefly, cell pellets were resuspended in 250 µl of Buffer A (10 mM HEPES pH 7.5, 10 mM KCl, 10% glycerol, 340 mM sucrose, 4 mM $MgCl_2$, 1 mM DTT in ddH2O supplemented with 1x cOmplete™, Mini, EDTA-free protease inhibitor cocktail; Sigma, UK) and 250 µl of Buffer A+Triton (2% Triton-X100, 10 mM HEPES pH 7.5, 10 mM KCl, 10% glycerol, 340 mM sucrose, 4 mM $MgCl_2$, 1 mM DTT in ddH2O supplemented with 1x cOmplete™, Mini, EDTA-free protease inhibitor cocktail). The cell suspension was incubated on ice for 12 minutes with occasional inverting to mix followed by centrifugation at 1,200 xg for 5 minutes at 4°C. Supernatants were kept as the cytoplasmic fraction. Nuclear pellets were resuspended in 250 µl of Buffer A and 250 µl of Buffer A+Triton and centrifuged at 1,200 xg for 5 minutes at 4°C. Nuclear pellets were resuspended with 250 µl of RNP lysis buffer (20 mM Tris-HCl, pH 7.5, 500 mM LiCl, 0.5% LiDS, 1 mM EDTA, 5 mM DTT in ddH2O, supplemented with protease inhibitor cocktail and 1U/µl RNase inhibitor) and incubated on ice for 20 minutes. Nuclear extracts were then homogenised by passing through a 27’ gauge needle and syringe 5-7 times and then centrifuged at 17,000 xg for 1 minute. The supernatants were kept as nuclear extracts. Cytoplasmic fractions were subjected to another centrifugation at 1,200 xg for 5 minutes at 4°C to remove any traces of nuclei. Subsequently, the concentration of protein in the samples was determined via the Protein Qubit assay according to the manufacturer’s instructions (ThermoFisher, UK) and samples were stored at -80°C.

To examine the purity of fractions, samples were prepared for immunoblotting of cytoplasmic and nuclear markers by heating 10 µg of protein mixed with Laemmli Sample Buffer (at 1x: 31.5 mM Tris-HCl, pH 6.8, 10% glycerol, 1% SDS, 0.005% Bromophenol Blue; BioRad, UK; supplemented with 2-mercaptoethanol at a final concentration of 355 mM; Sigma, UK) at 95^{o}C for 5 minutes. Proteins were separated on a denaturing 10% SDS-polyacrylamide gel at 90 V for 30 minutes and then at 150 V for 80 minutes (total run time 110 minutes) using 1x running buffer (1x running buffer, 0.025 M Tris, 0.25 M Glycine and 0.1% SDS in ddH2O). Subsequently, proteins were onto a 0.22 µm PVDF membrane at 200 mA for 1:30

hours using 1x transfer buffer (25 mM Tris and 192 mM Glycine with 20% methanol in ddH2O) and wet-blotter (Mini-PROTEAN Tetra Vertical electrophoresis cell). Membranes were blocked with 5% Marvel milk powder in PBS-T (0.5% Tween) for 1 hour at room temperature while rolling. Blots were incubated with primary antibodies (**Section 2.2**) overnight at 4°C while rolling, washed three times in PBS-T (0.5% Tween) at room temperature for 10 minutes each while rolling and incubated with secondary HRP-conjugated antibodies for 1 hour at room temperature while rolling. Membranes were washed three times in PBS-T (0.5% Tween) at room temperature for 10 minutes each while rolling and developed using ECL (Biological Industries, UK). Chemiluminescent signal was detected using a ChemiDoc XRS+ imaging system (BioRad, UK).

4.2.4. *In vitro* transcribed RNA pulldown

RNA-protein complexes were isolated as previously described using an *in vitro* transcribed RNA pulldown approach (Tichon et al., 2018). Nuclear-enriched extracts from ISHIKAWA or whole-cell extracts from bovine stromal cells were generated. Nuclear-enriched extracts could not be generated here from bovine cells (see **Figure 4.6**), so whole-cell lysates were used instead. Each lysate containing 0.5-1 mg protein was mixed with the incubation buffer (20 mM Tris-HCl pH 7.5, 150 mM NaCl, 1.5 mM $MgCl_2$, 2 mM DTT, 0.5% sodium deoxycholate, 0.5% NP-40 in ddH2O supplemented with protease inhibitors and 1 U/µl RNase Inhibitor). The ratio of lysis buffer to incubation buffer was ~1:4. Lysates were pre-cleared by incubating with 50 µl of Dynabeads™ MyOne™ Streptavidin C1 magnetic beads (10 mg/ml; ThermoFisher, UK) for 1 hour at 4ºC while rotating, discarding the beads at the end. From each of sense or antisense *in vitro* transcribed biotinylated transcripts, 10 µg were denatured at 85ºC for 3 minutes and snap-chilled on ice for 2 minutes before incubating with pre-cleared lysates for 2 hours while rotating at 4°C. 50 µl of beads were then added and incubated for further 1 hour at 4°C while rotating to recover biotinylated XIST-protein complexes. Beads were washed six times with washing buffer (20 mM Tris-HCl pH 7.5, 150 mM NaCl, 1.5 mM $MgCl_2$, 2 mM DTT, 0.5% sodium deoxycholate, 0.5% NP-40 in ddH2O). The standard elution approach was performed in elution buffer (15 mM Tris HCl pH 7.5, 0.02% SDS) for 15 min at 65°C. Following that and for later downstream analyses via western blot, beads were

resuspended in 1x Laemmli buffer and heated at 95°C for 5 minutes while shaking at 1100 RPM (referred to as post-elution in Fig. 4.8). Pulled down proteins were examined via western blotting as previously described (**Section 4.2.3**). For downstream analyses via TMT-MS, beads were resuspended in 10 mM Tris, pH 7.5, 2 mM $MgCl_2$ in ddH2O, snap-frozen and stored at -80°C. Samples were analysed via TMT-MS at the Proteomics Facility of the University of Bristol.

4.2.5.1. Proteomic analyses: TMT labelling and high pH reversed-phase chromatography

The following proteomic analysis was performed by Dr Kate Heesom, Proteomics Facility at the University of Bristol. For each pulldown replicate, an equal amount of starting material (lysate) was used to perform pulldowns and 100% of the captured material by streptavidin-coated beads was labelled by TMT-MS 6-plex reagents according to the manufacturer's protocol (Thermo Fisher Scientific) and the labelled samples were pooled. Pooled samples of 100 µg were evaporated until dry, resuspended in 5% formic acid and then desalted using a SepPak cartridge according to the manufacturer's instructions (Waters). The resulting eluates from the cartridge were evaporated dry and resuspended in buffer C (20 mM ammonium hydroxide, pH 10) before fractionating by high pH reversed-phase chromatography with the Ultimate 3000 liquid chromatography system (Thermo Scientific). Briefly, samples were loaded onto an XBridge BEH C18 Column (130 Å, 3.5 µm, 2.1 mm × 150 mm, Waters) in buffer C and peptides eluted with increasing gradient of buffer D (20 mM ammonium hydroxide in acetonitrile, pH 10) from 0 to 95% over 60 min. The resulting fractions were until dry and resuspended in 1% formic acid prior to analysis by nano-LC–MS/MS using an Orbitrap Fusion Tribrid mass spectrometer (Thermo Scientific).

4.2.5.2. Nano-LC mass spectrometry

Fractions collected were further fractionated using an Ultimate 3000 nano-LC system in line with an Orbitrap Fusion Tribrid mass spectrometer (Thermo Scientific). Briefly, peptides in 1% (vol/vol) formic acid were injected onto an Acclaim PepMap C18 nano-trap column (Thermo Scientific). After washing with 0.5% (vol/vol) acetonitrile 0.1% (vol/vol), formic acid peptides were resolved on a 250 mm × 75 µm Acclaim

PepMap C18 reverse phase analytical column (Thermo Scientific) over a 150 minute organic gradient, using seven gradient segments (1–6% solvent B over 1 min, 6–15% B over 58 min, 15–32% solvent B over 58 min, 32–40% solvent B over 5 min, 40–90% solvent B over 1 min, held at 90% solvent B for 6 min and then reduced to 1% solvent B over 1 min) with a flow rate of 300 nl/min. Solvent B was aqueous 80% acetonitrile in 0.1% formic acid. Peptides were ionized by nano-electrospray ionization at 2.0 kV using a stainless-steel emitter with an internal diameter of 30 μm (Thermo Scientific) and a capillary temperature of 275°C.

Spectra were acquired from an Orbitrap Fusion Tribrid mass spectrometer using the Xcalibur 2.1 software (Thermo Scientific) in data-dependent acquisition mode under the SPS-MS3 workflow. Fourier transform mass analyzer 1 (FTMS1) spectra were collected at a resolution of 120,000 with an automatic gain control (AGC) target of 200,000 and a max injection time of 50 ms. Precursors were filtered with an intensity threshold of 5,000, according to charge state (to include charge states 2–7) and with monoisotopic peak determination set to peptide. Previously interrogated precursors were excluded using a dynamic window (60 s ± 10 ppm). The MS2 precursor ions were isolated with a quadrupole isolation window of 1.2 *m*/*z*. FTMS2 spectra were collected with an AGC target of 10,000, max injection time of 70 ms and CID collision energy of 35%. For FTMS3 analysis, the Orbitrap was operated at 50 000 resolution with an AGC target of 50,000 and a max injection time of 105 ms. Precursors were fragmented by high energy collision dissociation (HCD) at a normalised collision energy of 60% to ensure maximal TMT reporter ion yield. Synchronous Precursor Selection (SPS) was enabled to include up to 5 MS2 fragment ions in the FTMS3 scan.

4.2.5.3. TMT-MS data analysis

The described data analysis was performed by Dr Phil Lewis, Proteomics Facility at Bristol. Raw data files were processed and quantified using Proteome Discoverer software v2.1 (Thermo Scientific) and queried against the UniProt *Homo sapiens and Bos taurus* databases using the SEQUEST HT algorithm. Peptide precursor mass tolerance was set at 10 ppm, and MS/MS tolerance was set at 0.6 Da. Search criteria included oxidation of methionine (+15.995 Da) and acetylation of the protein

N-terminus (+42.011 Da) as variable modifications and carbamidomethylation of cysteine (+57.021 Da) and the addition of the TMT mass tag (+229.163 Da) to peptide N-termini and lysine as fixed modifications. UniProt searches were performed with full tryptic digestion and a maximum of two missed cleavages were allowed. The reverse database search option was enabled and all data were filtered to a false discovery rate (FDR) of 5%.

Peptide IDs not corresponding to *H. sapiens or B. taurus* proteins were removed from all TMT replicates. Using the protein grouping decided by PD2.1, master protein selection was improved using an in-house script to select the UniProt accession (database downloaded January 2021; 42,818 sequences) with the best annotation whilst maintaining confidence in protein identification and quantitation. Abundances are the sum of the signal-to-noise ratio values for the TMT reporter groups for all peptide-spectrum matches (PSMs) matched to the protein. Normalised abundances of these values were obtained by normalising the Total Peptide Amount in each sample such that the total signal from each TMT tag is the same. Normalised abundances were log2 transformed to bring them closer to a normal distribution. Pairwise comparisons were then used to calculate the $\log_2$ fold-change difference ($\log_2$FC) of proteins between sense and antisense samples. The standard Student's t-test was used to test the statistical significance of $\log_2$FC values across sense and antisense samples. Principal component and volcano plot analyses were performed by Dr Phil Lewis, Proteomics Facility at Bristol. All other analyses of TMT data (violin plots and venn diagrams) were performed in R by Ioannis Tsagakis using custom scripts.

4.2.6. Gene ontology (GO) term over-representation analysis

GO term analysis for biological process and cellular component was performed using the Gene ontology (Ashburner et al., 2000, The Gene Ontology Consortium, 2021) and PANTHER (Mi et al., 2021) platforms with a consensus list of proteins identified across all three biological replicates as enriched in the sense transcript. The background gene set used includes all protein-coding genes in the genome selected (Homo sapiens or Bos taurus). Fisher's exact test was used to infer significance (p-

value <0.05) and results were filtered for a false discovery rate (FDR) with a p-value <0.05. Platform used can be accessed here: http://geneontology.org/

4.3. Results

The binding sites of a subset of the *XIST* protein interactome have been clearly documented to be within repetitive regions of the *XIST* RNA in mouse and human. Spen has been shown to interact with repeat A of *Xist* both in mouse embryonic stem cells (Monfort et al., 2015, Chen et al., 2016, Lu et al., 2016, Lu et al., 2020) and in human embryonic kidney cells (HEK293T)(Graindorge et al., 2019). Rbm15 and Wtap have been shown to bind repeat A in mouse embryonic stem cells (Chu et al., 2015b, Moindrot et al., 2015) whereas RBM15 has also been shown to bind human *XIST* repeat A in HEK293T (Graindorge et al., 2019). WTAP was found associated with repeat C of *XIST* in HEK293T (Graindorge et al., 2019). Ciz1 has been found to associate with repeat E in both mouse and human (Sunwoo et al., 2017, Graindorge et al., 2019). To determine which repetitive regions of the XIST RNA are bound by specific proteins, *XIST* RNA fragments were *in vitro* transcribed and used in pulldown reactions.

4.3.1. *In vitro* transcription generates biotinylated human and bovine *XIST* fragments

In order to perform RNA pulldowns, the first step was to generate biotinylated *XIST* fragments. To this end, human and bovine *XIST* repeat A and E fragments were cloned into plasmids (**Section 4.2.1**), *in vitro* transcribed and RNA was purified. To assess the size and quality of these transcripts, they were electrophoresed on denaturing agarose gels. When electrophoresed, transcripts appeared almost twice in size compared to the size of the transcript predicted to be transcribed from the plasmid. This is owed to the presence of biotin in the modified nucleotides, slowing down the migration of RNA on the gel (**Figure 4.4**)(New England Biolabs, 2021).

To confirm biotin incorporation, RNA slot blot assays were performed. *In vitro* biotinylated transcribed RNA was run alongside total RNA purified from cells as a negative control. RNA slot blot demonstrated the presence of biotin in samples with 800 ng of RNA that had been *in vitro* transcribed using biotin-modified nucleotides, whereas background levels of biotin (if any) were detected in up to 800 ng of total RNA (**Figure 4.5**). Background levels of biotin were no longer detectable when using

200 ng of total RNA, whereas a clear signal was evident for RNA that had been *in vitro* transcribed using biotin-modified nucleotides. In summary, *in vitro* transcribed RNA of roughly the appropriate size could be generated and modified with biotin groups.

Species	XIST region	Orientation	RNA length (nt)
Human	Repeat A	Sense	1062
		Antisense	1010
	Repeat E	Sense	1979
		Antisense	1994
Cow	Repeat A	Sense	1254
		Antisense	1268
	Repeat E	Sense	1036
		Antisense	1021

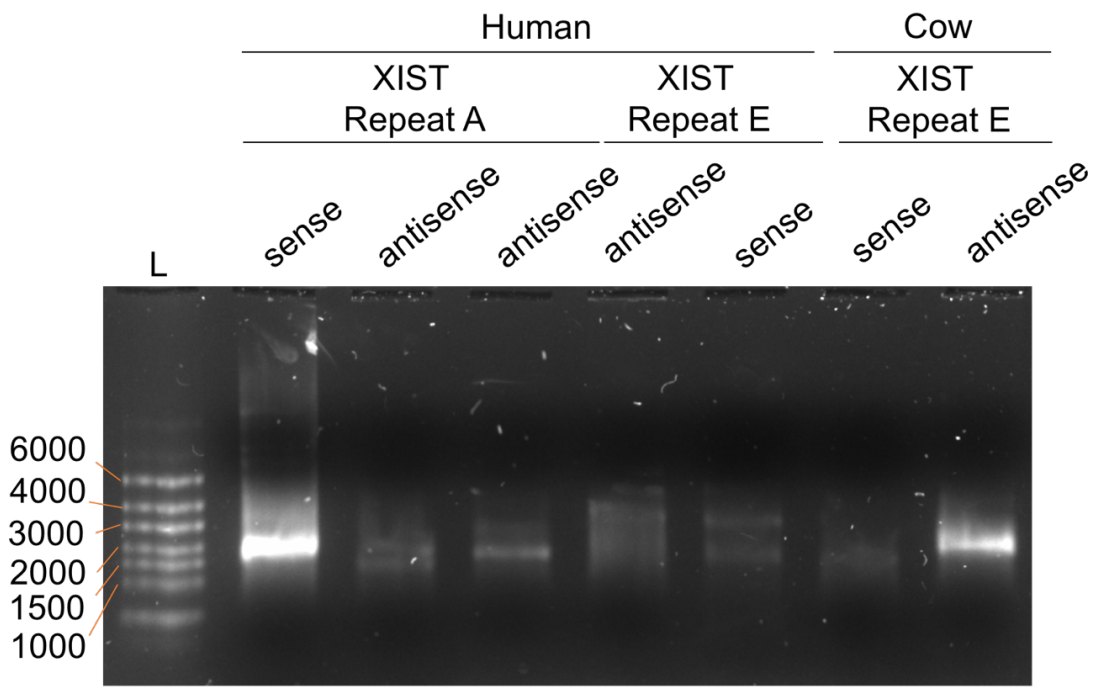


Figure 4.4. Denaturing agarose gel electrophoresis of in vitro transcribed biotinylated XIST RNA repeats.

Sizes of fragments transcribed from plasmids are given in table (left). *In vitro* transcribed *XIST* fragments containing biotin were separated on a 1% agarose gel under denaturing conditions with formaldehyde (right). To aid in band size estimation, a single-stranded RNA ladder was used (200 to 6000 bp; Riboruler High Range). One µg of each RNA was run on the gel. High molecular weight products indicate spurious *in vitro* transcription products.

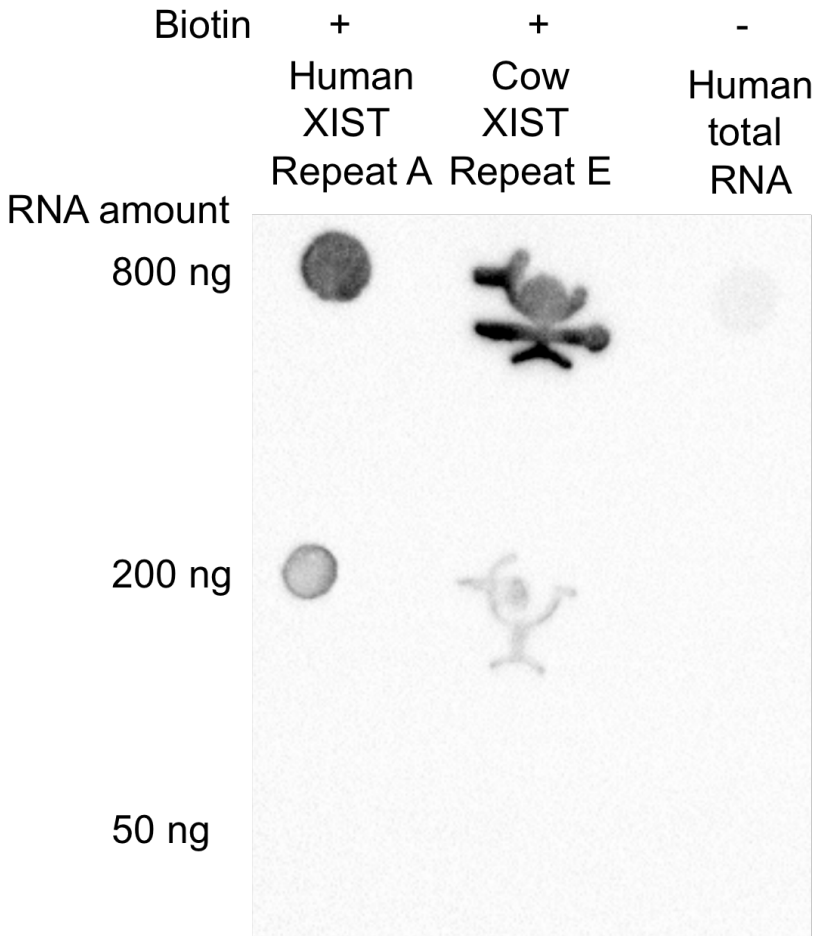


Figure 4.5. RNA slot blot of select in vitro transcribed XIST RNA repeat constructs.

In vitro transcribed RNA using biotinylated dUTP and total RNA isolated from ISHIKAWA cells were blotted with a single replicate per serial dilution. Equal amounts of RNA were used for all samples and serial dilutions of 1/4 were performed. The pattern of the second column was owed to sample leaking from the cast used. Membrane was probed with a biotin-specific NeutrAvidin antibody at 1:1000.

4.3.2. Subcellular fractionation isolates a pure nuclear compartment in ISHIKAWA but not in bovine stromal cells

To identify nuclear protein partners of *in vitro* transcribed *XIST* fragments given the exclusive nuclear localisation of endogenous *XIST*, nuclear lysates were prepared and used in pulldown assays. To this end, subcellular fractionation was performed on human ISHIKAWA and bovine stromal cells. Purity of fractions was then assessed by western blotting with a cytoplasmic marker, β-tubulin, and a nuclear marker, Lamin B. Following fractionation of ISHIKAWA cells, the cytoplasmic marker, β-tubulin, was abundant in the cytoplasmic fraction and less so in the nuclear fraction whereas the nuclear marker, Lamin B, was only found in the nuclear fraction (**Figure 4.6**). Despite a pure cytoplasmic fraction lacking Lamin B, the presence of β-tubulin in the nuclear fraction suggested the protocol used was sufficient to generate nuclear-enriched lysates from ISHIKAWA cells, but not pure nuclear lysates. However, the same fractionation protocol was not effective for bovine stromal cells, given both cytoplasmic and nuclear markers were found in the nuclear fraction, indicative of a whole cell lysate (**Figure 4.6**). Increasing the ionic strength of the plasma membrane lysis buffer resulted in a lower abundance of β-tubulin in the nuclear fraction of ISHIKAWA cells, but was not strong enough to completely deplete it (**Figure 4.7**). These conditions were not trialled in bovine stromal cells. In summary, the lysis conditions used were suitable for the generation of nuclear-enriched lysates for ISHIKAWA cells, but not for bovine stromal cells. Whole cell lysates were used for bovine stromal cells.

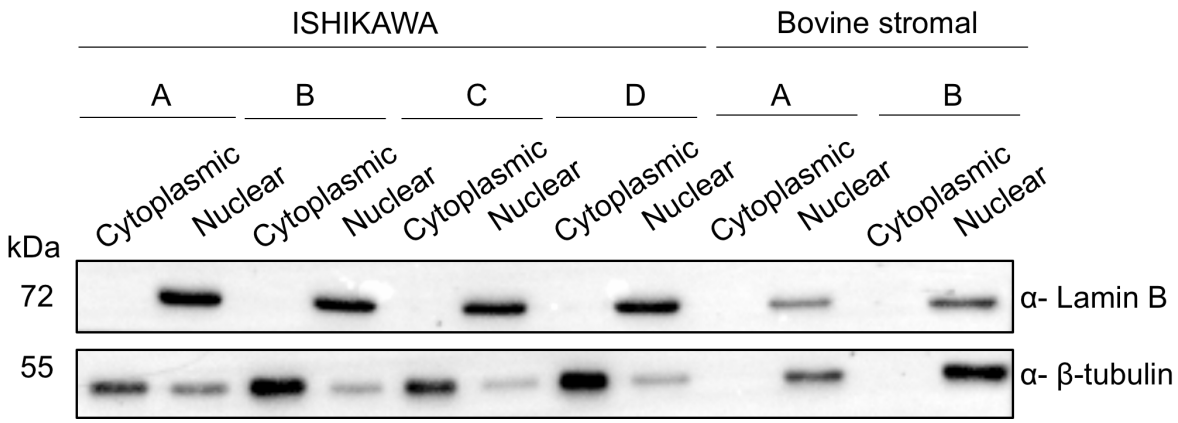


Figure 4.6. Subcellular fractionation of ISHIKAWA and bovine stromal cells.

Protein lysates prepared from 10 million ISHIKAWA or bovine stromal cells. Plasma membrane was lysed using a 2% Triton containing solution. n=4 independent biological replicate shown for ISHIKAWA and n=2 biological replicate shown for bovine stromal cells. Equal amounts of protein were loaded from each sample (~5 µg). β-tubulin was used as a cytoplasmic marker and Lamin B was used as a nuclear marker,. All antibodies were used at 1:1000 dilution in PBS with 0.5% Tween-20 (PBS-T).

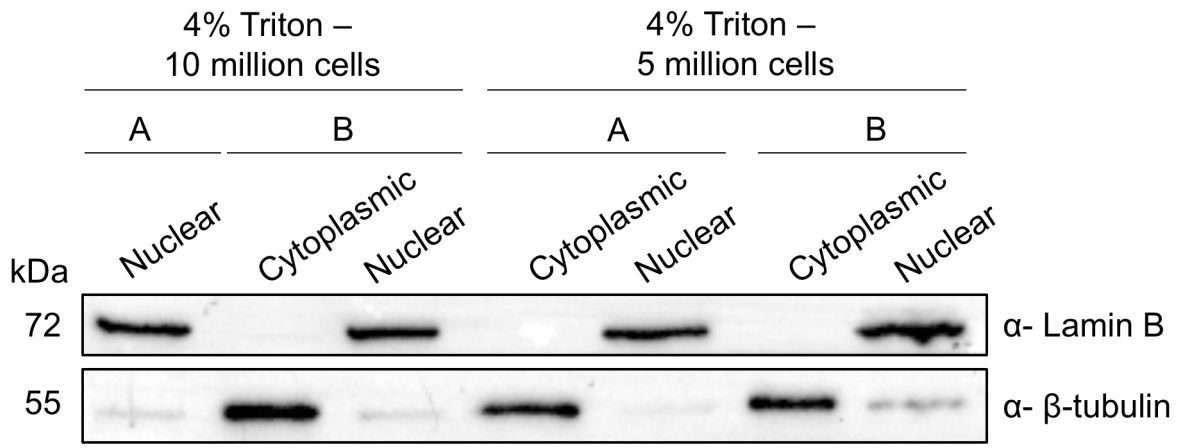


Figure 4.7. Subcellular fractionation of ISHIKAWA.

Protein lysates prepared from 5 and 10 million ISHIKAWA cells. Plasma membrane was lysed using a 4% Triton containing solution. n=4 independent biological replicate shown for ISHIKAWA and n=1 biological replicate shown for bovine stromal cells. Equal amounts of protein were loaded from each sample (~5 µg). β-tubulin was used as a cytoplasmic marker and Lamin B was used as a nuclear marker. All antibodies were used at 1:1000 dilution in PBS with 0.5% Tween-20 (PBS-T).

4.3.3. Human CIZ1 interacts with *XIST* repeat E in nuclear-enriched human endometrial cell lysates

The *in vitro* transcription pulldown approach involves incubating lysates with either sense or antisense transcript, allowing protein partners of the *XIST* region to be isolated from the rest of the lysate. The antisense transcript is used as a negative control to differentiate sequence-specific partners of the sense transcript from sequence non-specific protein partners of the antisense transcript. RNA-protein interactions captured by the beads can then be eluted and proteins captured analysed via western blot or mass spectrometry. Lamin B is not expected to associate with any *XIST* regions specifically and was thus used as a negative control in western blots to assess background levels of non-specific protein binding.

In the first instance, 0.5 mg of nuclear-enriched ISHIKAWA cell lysates were prepared and pulldowns were performed using 10 µg of biotinylated *in vitro* transcribed *XIST* repeat E sense or antisense transcripts. Samples from the pulldown were analysed by western blotting against CIZ1, known to bind human and mouse *Xist* repeat E. Probing for CIZ1 indicated that its abundance was higher in the elution from the sense transcript compared to the antisense in nuclear-enriched ISHIKAWA lysates (**Figure 4.8A**). Nonetheless, Lamin B (negative control) was not detected in either elutions, indicating there is no non-specific protein binding. In a second biological replicate of this experiment, CIZ1 was again detected at higher levels in the sense elution than in the antisense. Although the overall abundance of the CIZ1 protein was low, requiring a long exposure time to detect CIZ1 in the elutions (**Figure 4.8B**). However, Lamin B was also detected at higher levels in the sense when compared to antisense, although the signal difference between sense and antisense did not appear as great as the one for CIZ1. Overall, a greater CIZ1 enrichment was found in the elution from the sense than the antisense transcript, demonstrating a specific association of human *XIST* repeat E with CIZ1.

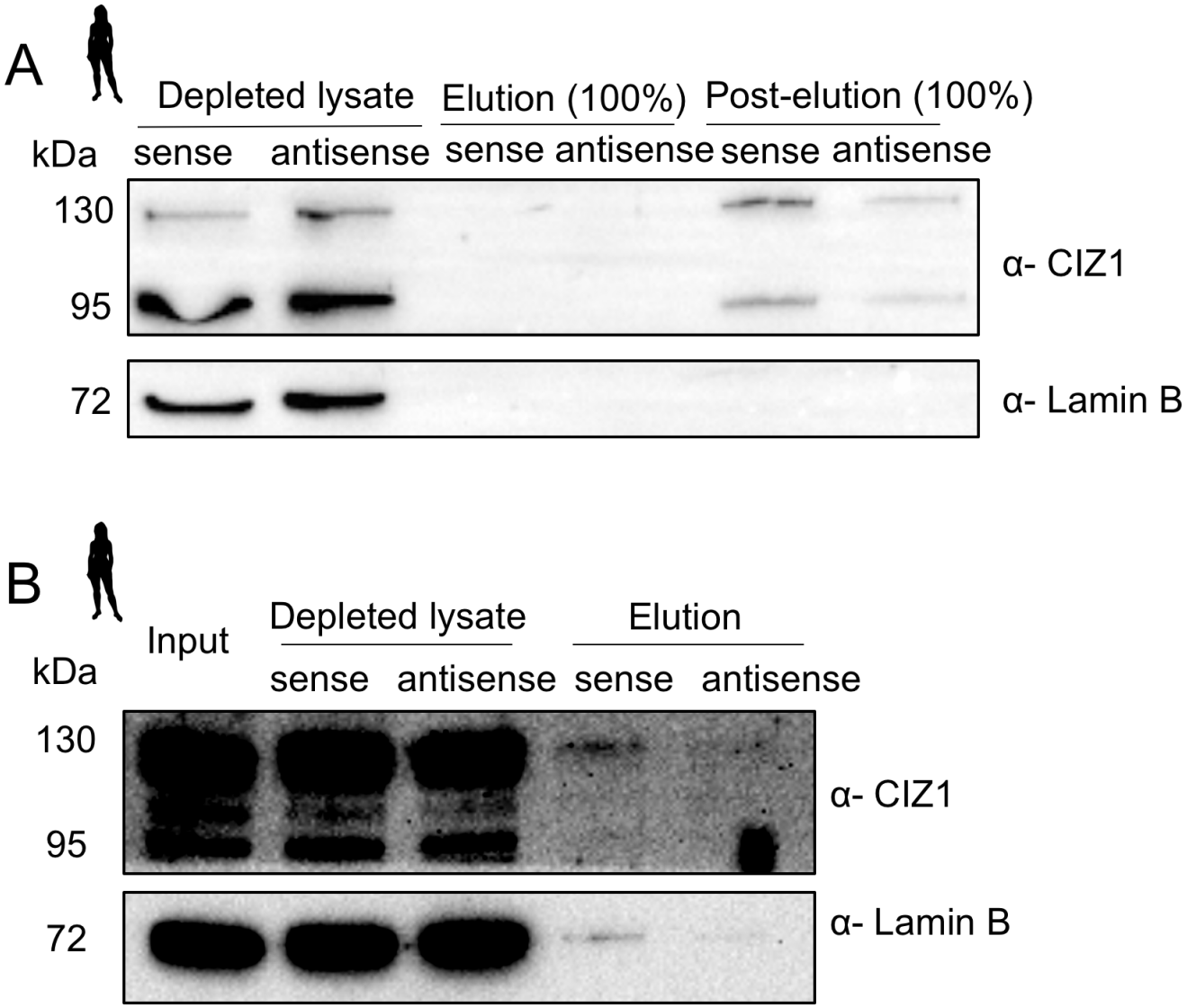


Figure 4.8. Human CIZ1 binds to human XIST repeat E.

A&**B**) Pulldowns were performed in nuclear-enriched ISHIKAWA protein lysates. Equal amounts of protein were loaded from each sample (~5 µg) and 100% of the elution sample. In B) equal amounts of input and depleted lysate were used. Input here corresponds to 0.5% of starting amount. Lamin B was used as a negative control. All antibodies were used at 1:1000 dilution in PBS with 0.5% Tween-20 (PBS-T). Two independent biological replicates shown.

4.3.4. Bovine CIZ1 interacts with *XIST* repeat E in whole-cell bovine stromal cell lysates

CIZ1 binding on mouse and human *XIST* has been documented to be specifically restricted in the repeat E region (**Section 1.6.2**). To test whether cow CIZ1 associates with the bovine XIST repeat E region, a biotinylated RNA pulldown was performed. To this end, 10 µg of bovine XIST repeat E was *in vitro* transcribed and used in pulldown assays with 1 mg of bovine stromal whole cell lysates. Following pulldown with bovine XIST repeat E sense and antisense transcripts, western blotting was performed on input, depleted lysates and elutions to probe for a differential enrichment (and hence specificity) of the CIZ1 protein for the sense or antisense transcript. In the first replicate, CIZ1 protein binding was higher in the elution from the sense compared to the antisense transcript (**Figure 4.9A**). Lamin B was absent from both elution samples, consistent with a lack of non-specific protein binding to the biotinylated transcripts. In the second replicate, again CIZ1 depletion was no different across the sense and antisense transcripts whilst a higher enrichment of CIZ1 was seen in the elution of the sense versus the antisense transcript (**Figure 4.9B**). As with the previous replicate, there was no difference in the depletion of Lamin B between sense and antisense transcripts, and despite the presence of Lamin B in both elution samples, again there was no great difference in enrichment in the sense over the antisense transcripts (**Figure 4.9B**). A low CIZ1 abundance was observed in both elution samples of the third replicate. In order to determine whether CIZ1 was more abundant in the sense over the antisense transcript, the exposure was increased. This revealed a slightly higher CIZ1 abundance in the sense over the antisense transcript (**Figure 4.9C**). Although Lamin B was present in both elution samples, there was no indication of an enrichment in either, suggestive of a lack of non-specific protein enrichment across all three replicates (**Figure 4.9**). Taken together, these observations suggested a weak although discernible preferential presence of cow CIZ1 in the elution of the sense transcript, indicative of a specific association with bovine XIST repeat E.

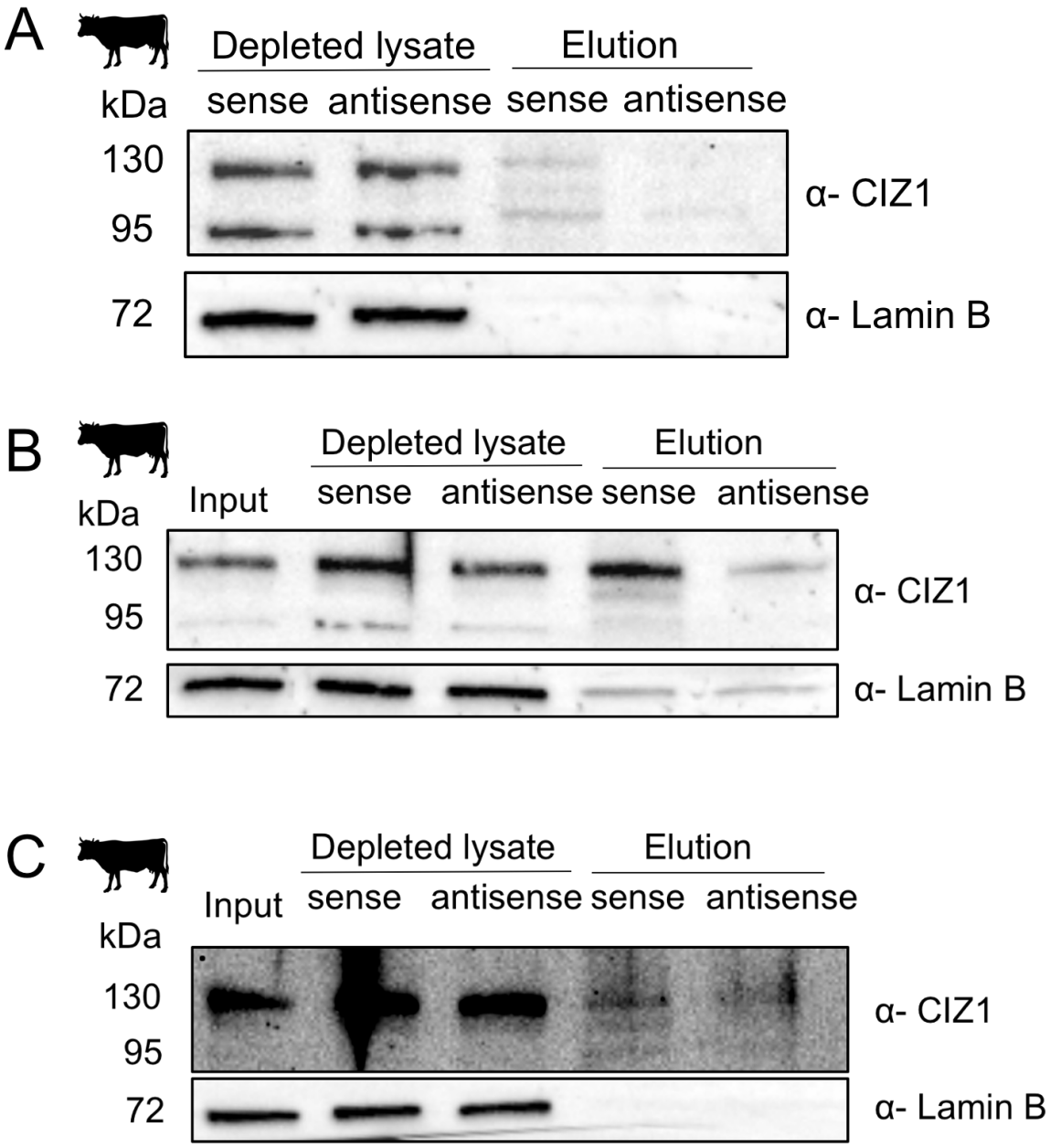


Figure 4.9. Cow CIZ1 associates with bovine XIST repeat E.

A-C) Pulldowns were performed with whole cell bovine stromal protein lysates. Equal amounts of protein were loaded from each sample (~5 µg) and 100% of the elution sample. Input here corresponds to 0.5% of starting amount. Lamin B was used as a negative control. All antibodies were used at 1:1000 dilution in PBS with 0.5% Tween-20 (PBS-T). Three independent biological replicates shown. Material from different animals was used for each biological replicate here. Occasionally, due to efficiency of cell purification from animals, material from two animals was pooled and used once as an individual replicate.

4.3.5. Human CIZ1 from nuclear-enriched human endometrial cell lysates does not bind bovine XIST repeat E

In this thesis, the CIZ1 protein has been shown to interact with *XIST*, binding repeat E in both human and cow. Next, given that *XIST* repeat E was estimated to be ~54.6% similar between human and cow and the CIZ1 protein showed ~80% sequence similarity between human and cow, it was sought to be determined if human CIZ1 could bind bovine XIST repeat E. To address whether a sufficiently high conservation has been maintained to enable a biochemical interaction, bovine XIST repeat E was assayed for an association with human CIZ1 in ISHIKAWA lysates.

To this end, bovine XIST repeat E was *in vitro* transcribed and used in pulldowns with nuclear-enriched human endometrial cell lysates (ISHIKAWA). Subsequently, the presence of the human CIZ1 protein was probed by western blotting in elution samples of the sense and antisense bovine XIST repeat E transcripts. Only a small amount of human CIZ1 could be detected in the elutions and there was no difference between sense or antisense, suggesting non-specific binding (**Figure 4.18A**). Lamin B was slightly more enriched in the sense elution (**Figure 4.18A**), which could be indicative of non-specific protein associations occurring with the sense transcript compared to the antisense. In the second replicate a small amount of human CIZ1 was detected in elutions but more in antisense compared to sense, again indicating non-specific binding of human CIZ1 to bovine XIST repeat E (**Figure 4.18B**). Lamin B binding was also higher in antisense compared to sense suggesting higher background protein binding to the antisense. In summary, under these pulldown conditions, human CIZ1 does not bind bovine XIST.

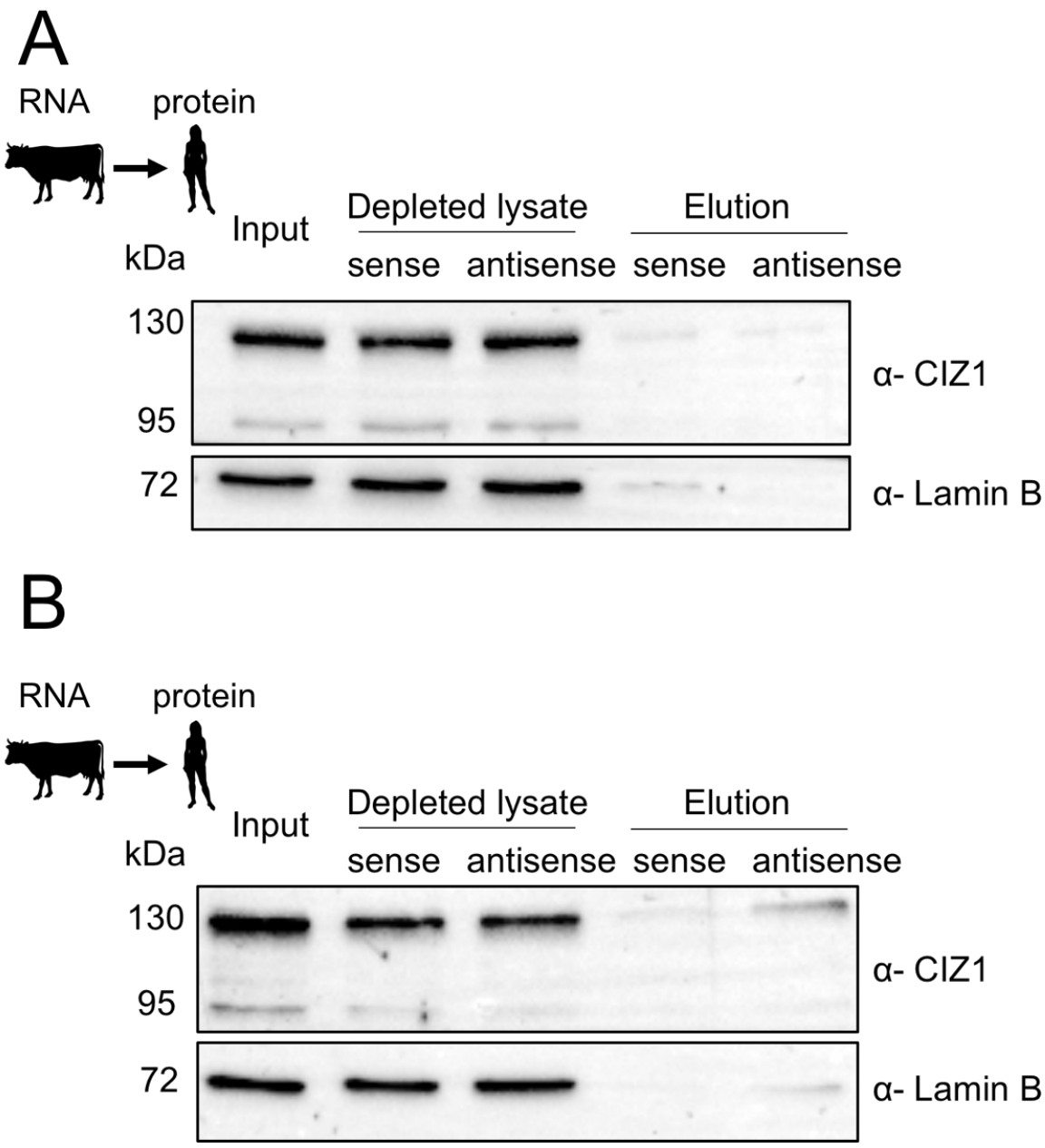


Figure 4.10. Human CIZ1 does not associate with bovine XIST repeat E.

A-B) Pulldowns were performed in nuclear-enriched ISHIKAWA protein lysates. Equal amounts of protein were loaded from each sample (~5 µg) and 100% of the elution sample. Input here corresponds to 0.5% of starting amount. Lamin B was used as a negative control. All antibodies were used at 1:1000 dilution in PBS with 0.5% Tween-20 (PBS-T). Two independent biological replicates shown.

4.3.6. Human RBM15 but not WTAP associates with *XIST* repeat A in nuclear-enriched human endometrial cell lysates

In the previous chapter, the WTAP and RBM15 proteins were used to pulldown *XIST* in RIP where WTAP, but not RBM15, demonstrated a specific association with human *XIST* in human endometrial cells. To determine if WTAP associates with *XIST* via the repeat A region and verify a lack of binding with RBM15, the reverse approach was employed here, where *in vitro* transcribed and biotinylated *XIST* repeat A was used in pulldowns. In this experiment, 10 µg of biotinylated *in vitro* transcribed *XIST* repeat A were used in a pulldown assay in ~500 µg of nuclear-enriched lysate from ISHIKAWA cells. Western blotting was performed in the elution samples from sense and antisense transcripts to assess whether human *XIST* repeat A binds RBM15 and/or WTAP. There was a greater enrichment of two RBM15 protein isoforms in the elution of the sense than in the antisense sample, indicating RBM15 specifically binds the human *XIST* A repeat. UniProt lists four RBM15 isoforms for human (100-107 kDa in size), three of which were seen with RIP in ISHIKAWA cells (**Figure 3.12**), but it is unknown whether all of them are expected to associate with *XIST*. Samples were also probed for WTAP, previously seen to associate with *XIST* in RIP-RT-qPCR (**Figure 3.13**). WTAP was detected in the sense elution but at low levels and required a long exposure time. WTAP was not detected in the antisense elution (**Figure 4.8A**). However, levels of Lamin B were also higher in the sense compared to the antisense, indicating a higher non-specific protein enrichment. Therefore, the preferential, albeit weak, presence of the WTAP protein in the elution from the sense transcript could be non-specific. In contrast, the difference between the enrichment of the RBM15 protein in the sense over antisense elution, was too great to be explained by Lamin B levels seen.

In the second replicate, RBM15 was still found enriched in the elution from the sense over the antisense transcript, albeit at a lower abundance compared to the first replicate (**Figure 4.8B**). Again, WTAP was found present in both samples. Lamin B levels were almost non-existent in any elution sample in this replicate, suggestive of no background, non-specific protein enrichment (**Figure 4.8B**). Taken together, RBM15 was shown to be greatly enriched in the elution of the sense compared to the antisense *XIST* repeat A transcript, suggesting a specific interaction. The

difference in sense over antisense enrichment was not as clear for WTAP here, preventing its definitive assignment as a specific interactor of *XIST* repeat A.

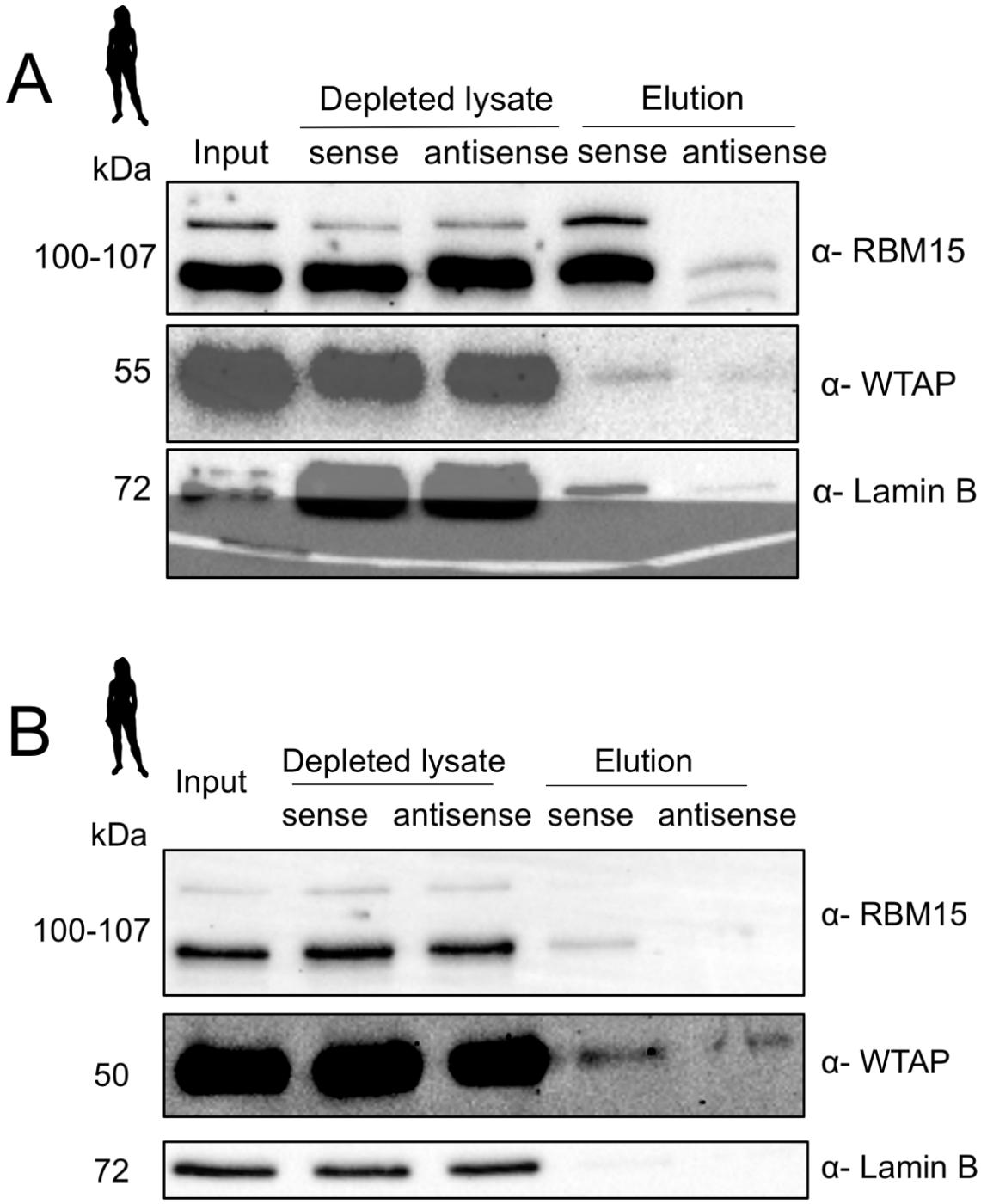


Figure 4.11. Human RBM15 associates with XIST repeat A.

A&**B**) Pulldowns were performed in nuclear-enriched ISHIKAWA protein lysates. Equal amounts of protein were loaded from each sample (~5 µg) and 100% of the elution sample. Input here corresponds to 0.5% of starting amount. Lamin B was used as a negative control. All antibodies were used at 1:1000 dilution in PBS with 0.5% Tween-20 (PBS-T). Two independent biological replicates shown.

4.3.7. Bovine hnRNPU but not SPEN, RBM15 or WTAP associate with *bovine XIST* repeat A in whole-cell bovine stromal lysates

Previously, the association of cow SPEN, RBM15 and WTAP with bovine *XIST* was assayed using RIP (Chapter 3), with inconsistent results and a number of replicates insufficient for a statistically powered analysis. Given SPEN, RBM15 and WTAP interact with repeat A of *XIST* in both mouse and human, the aim of this experiment was to determine whether cow SPEN, RBM15 and WTAP interact with the bovine *XIST* repeat A. Thus, using 10 µg of bovine *XIST* repeat A *in vitro* transcribed in the sense or antisense orientation (negative control) were mixed with 1 mg of bovine stromal whole cell lysate in pulldown assays. Since no suitable antibody was available for western blotting of the cow SPEN protein, the RNA-protein complexes bound to the streptavidin-coated magnetic beads were subjected to quantitative TMT mass spectrometry. Three biological replicates for each pulldown using sense and antisense *XIST* repeat A biotinylated fragments were performed and all 6 samples were run together on a 6-plex TMT-MS run (**Section 4.2.5**).

However, after data analyses it became evident that some samples had been mislabelled and thus switched. Namely, replicates 2 and 3 exhibited a pattern which was the reverse of what was expected (these analyses are shown in Supplementary Information). To account for this, an adjustment was implemented and the analyses using the adjusted data are shown below.

Given the hypothesis that sense samples had been misidentified as antisense (and vice versa) in replicates 2 and 3, samples were switched and reanalysed to account for potential human error. In order to determine how similar the three biological replicates were to one another, principal component analysis (PCA) was performed (**Figure 4.12;** by Dr Phil Lewis). Here, replicates 2 and 3 for sense cluster better together than replicate 1 and the same is true for replicates 2 and 3 for antisense. The difference seen here following data reanalysis, is that replicates 2 and 3 for sense cluster at the top and right part of the plot, whereas replicates 2 and 3 for antisense occupy the bottom and left part of the plot, indicating they have smaller values than the sense replicates. Despite the data adjustment implemented here,

there was still agreement between two of the three replicates for pulldowns performed with either the sense or antisense transcripts.

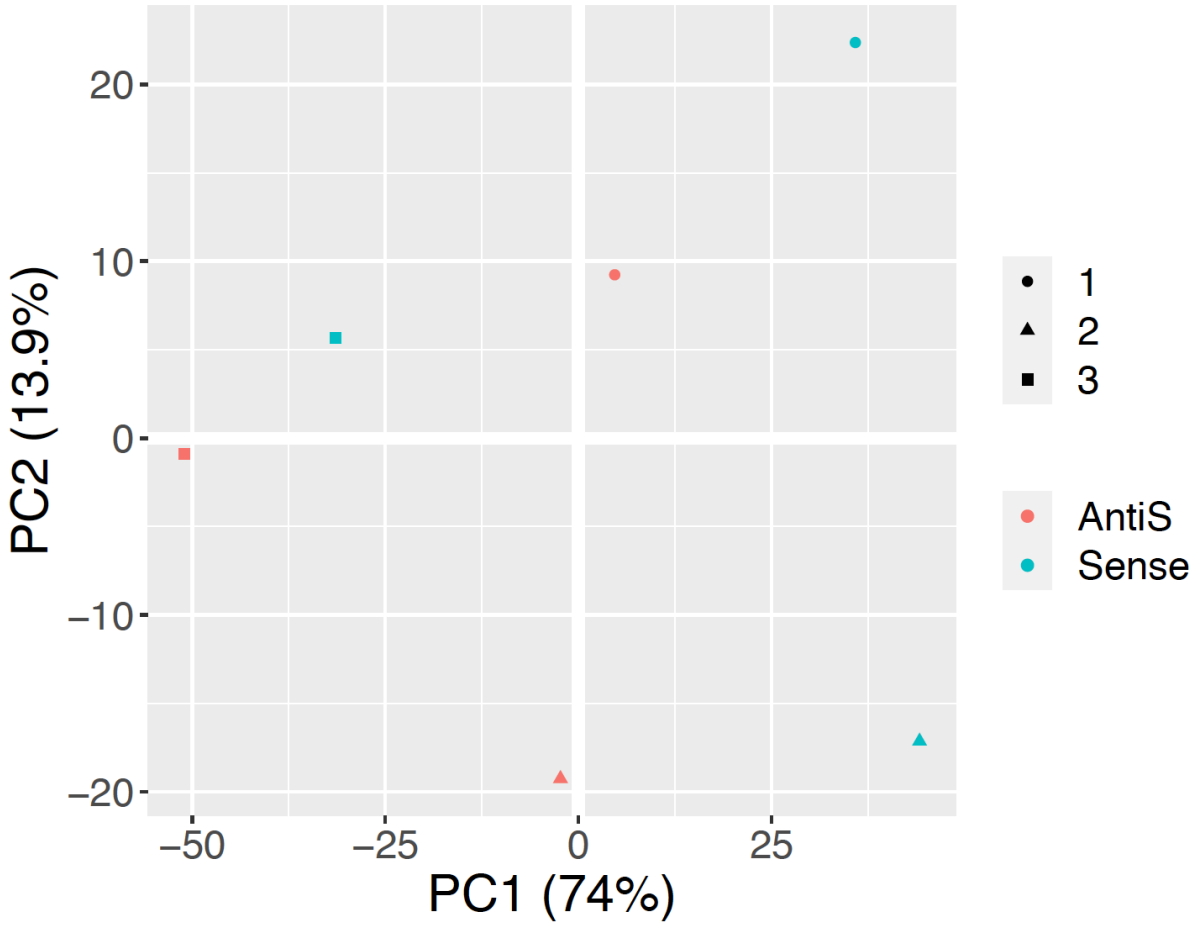


Figure 4.12. Bovine XIST repeat A sense replicates in cow cells are more similar to each other than bovine XIST repeat A antisense replicates following log2FC adjustment.

PC1 on the x-axis accounts for 74% of the variation between sense and antisense bovine XIST repeat A samples and PC2 on the y-axis accounts for 13.9% of the variation between samples. Two out of three pulldown replicates with the sense transcript (green) clustered above 0, and two out of three pulldown replicates with the antisense transcript (red) were below 0 for PC1 and PC2. Replicates shown are independent biological replicates. Generated by Dr Phil Lewis (Proteomics Facility at University of Bristol). PC, principle component

A violin plot of the distribution of log_2FC protein abundance differences between sense and antisense in replicates 2 and 3 were now mostly found in the same range of 0 to 2.5 as replicate 1 (**Figure 4.13**). This suggested that most proteins were enriched in the sense, i.e. supporting specificity of protein binding. This was in line with pulldown experiments, whereby proteins that recognise a specific sequence are expected to be found enriched in the sense compared to the antisense (negative control) transcript (hence log2FC changes for most proteins would show as positive). Therefore, whilst there was still a difference observed between replicate 1 and the adjusted replicates 2 and 3, the directionality of log_2FC change was in agreement across all replicates and the difference across replicates seen is more likely to represent differences in lysate preparations or pulldown efficiencies.

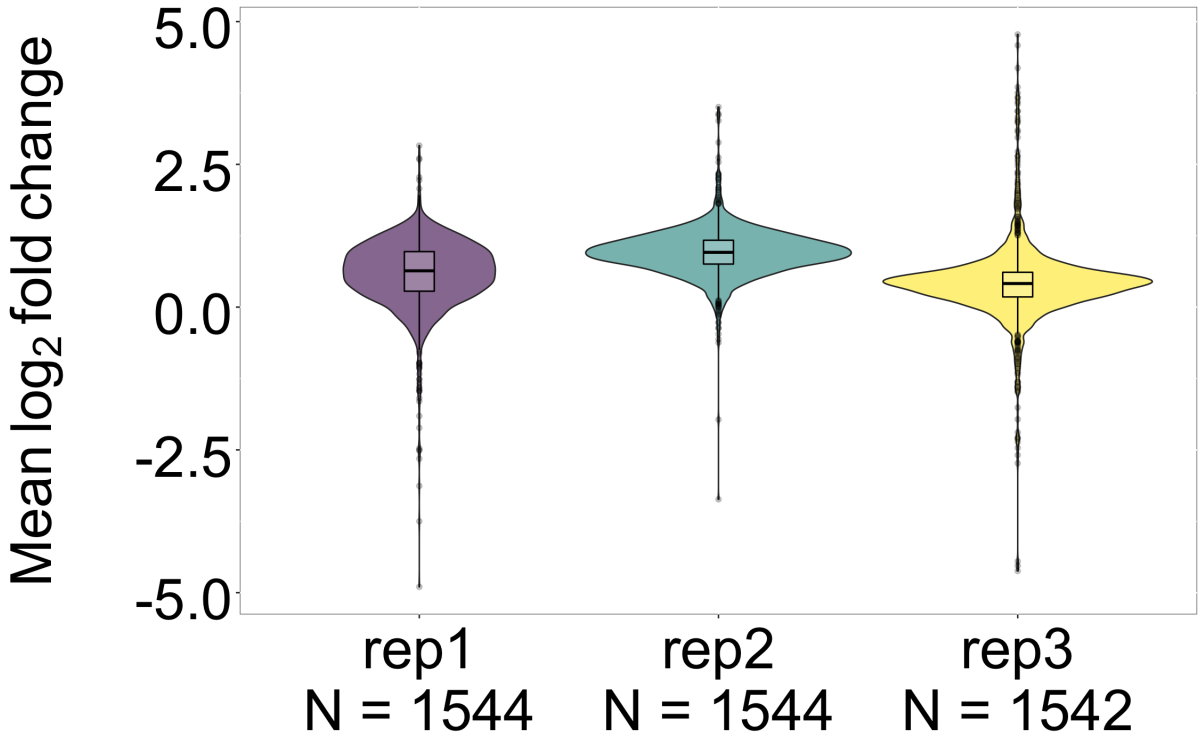


Figure 4.13. Distribution of differential protein abundance across replicates in the cow dataset following log2FC adjustment between sense and antisense.

Violin plot illustrating the distribution of differential protein abundance in replicate 1 is similar to that in replicates 2 and 3 following adjustment. Differential protein abundance (log2 fold change) was estimated by taking the mean of the log2 normalised protein abundance difference of antisense from sense: $\Sigma \Delta log_2^{(sense-antisense)}$. Violin width reflects the frequency of data points in each region. Cross bars denote the median, box limits represent the 25th and 75th percentiles, whiskers show the 1.5x interquartile range and dots indicate individual data points. Number of proteins detected in total per replicate are shown under each violin plot. three replicates shown are independent biological replicates.

Following log_2FC adjustment for replicates 2 and 3, the overlap of proteins found enriched or depleted across replicates of the bovine XIST repeat A sense transcript was examined by plotting Venn diagrams. Proteins with a log_2FC cut-off of higher than 1 were plotted as enriched and those with a log_2FC cut-off lower than -1 were plotted as depleted. The number of proteins enriched on the sense transcript were found to be 342 for replicate 1, 666 proteins for replicate 2 and 130 for replicate 3 (**Figure 4.14**). There were three proteins that were found to be enriched across all three replicates (**Figure 4.14**). Additionally, there were 105 proteins shared between replicates 1 and 2, 80 proteins shared between replicates 2 and 3, and 11 proteins shared between replicates 1 and 3. Examining proteins enriched in the antisense (i.e. depleted from the sense) could highlight proteins that display promiscuous binding, irrespective of sequence specificity. When looking at the proteins showing a higher abundance in the antisense compared the sense elution, no overlap was found across all three replicates, suggesting different proteins were binding to the antisense with every pulldown replicate. In fact, there was no overlap even between replicates 1 and 2 or 1 and 3. A single protein, MICALL1, was shared between replicates 2 and 3. Taken together, not only were several proteins found to be enriched in each replicate, but also there was an overlap in the number of enriched proteins shared.

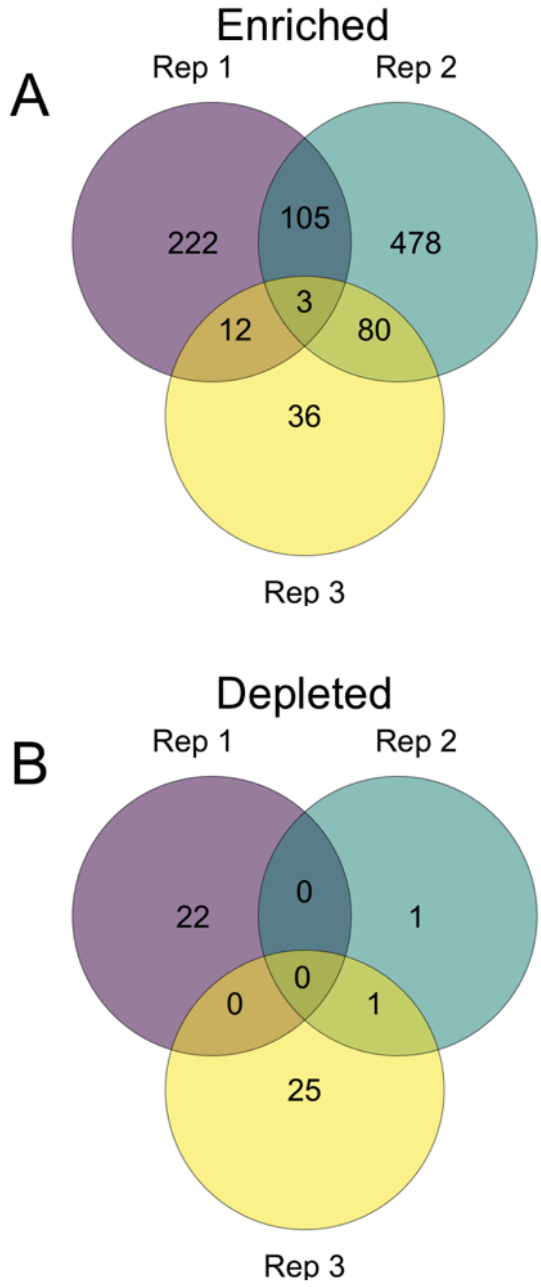


Figure 4.14. Higher overlap in enriched compared to depleted proteins pulled down by bovine XIST repeat A sense transcripts across cow replicates following log2FC adjustment.

Venn diagram showing overlap between proteins **A**) enriched or **B**) depleted in sense over antisense bovine XIST repeat A transcripts. Three proteins were common across all three replicates from those found to be enriched whereas no proteins were common across all three replicates from proteins found to be depleted. Proteins were called 'enriched' or 'depleted' based on log_2FC cut-offs of >1 or <-1, respectively. Proteins with log_2FC values between -1 and 1 were not plotted here. Replicate 1 is shown in purple, replicate 2 in green and replicate 3 in yellow.

One of the aims of this chapter was to characterise the protein partners of bovine *XIST*. To identify proteins that were differentially bound between sense and antisense elution samples from all three biological replicates, log_2FC scores and associated p-values were plotted, following the inversion of log_2FC scores for replicates 2 and 3. A total of 1544 proteins were found for the first and second replicates whereas 1542 proteins for the third replicate. Across the three replicates, a total of 376 proteins were statistically significant ($p<0.05$), all of which were enriched in the sense over the antisense transcript, but only 40 were above the log_2FC >1 cut-off (**Figure 4.15;** by Dr Phil Lewis and **Table 4.3**). This constitutes the high-confidence list of bovine *XIST* repeat A protein partners.

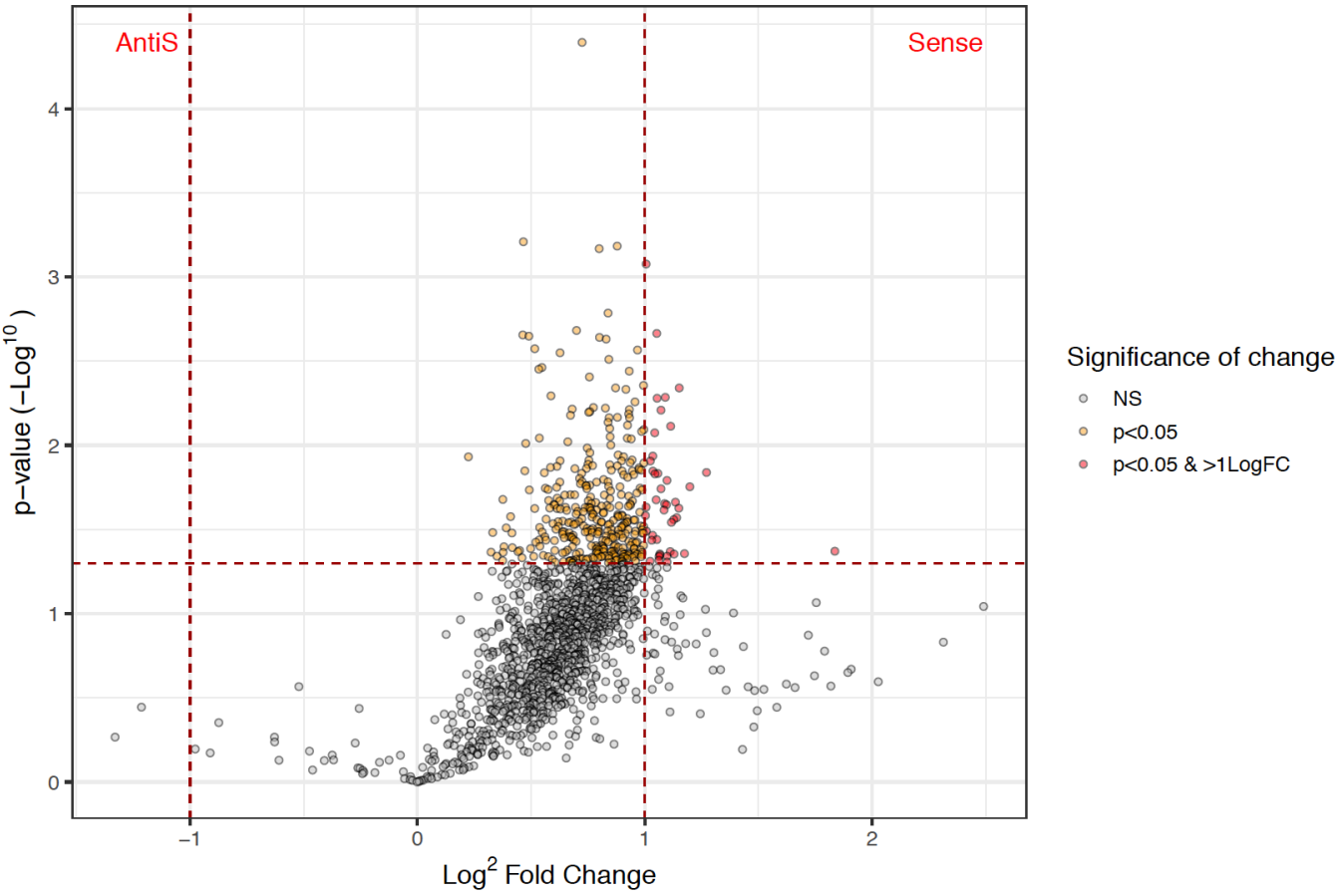


Figure 4.15. Volcano plot of proteins differentially bound between sense and antisense bovine XIST repeat A.

In total, 40 proteins (in red) were found to be significantly enriched ($log_2FC>1$ and p-value <0.05) in the bovine XIST repeat A sense transcript over the antisense transcript. Proteins that were significantly enriched at a $log_2FC<1$ (p-value <0.05) are shown in yellow. Proteins that were not enriched or did not reach statistical significance are shown in grey. Data generated from three biological replicates. Statistical analysis was performed with a Student's t-test (significance was inferred at p-value <0.05). Generated by Dr Phil Lewis (Proteomics Facility at University of Bristol).

Table 4.3. High-confidence list of proteins identified by TMT-MS from bovine XIST repeat A pulldowns in cow lysates.

Candidates fulfilled the criteria of having a $log_2FC>1$ and being statistically significant (Student's t-test, p-value<0.05). TMT-MS, tandem mass tag mass spectrometry

Protein symbol	Protein name	Unique Peptides	Average log_2FC	T-test (p-value)	Previously described
TAF3	TATA-box binding protein associated factor 3	15	1.84	0.04259	
MIA3	Transport and Golgi organization protein 1 homolog (TANGO1)	7	1.27	0.014506	
SEPTIN11	Septin-11	2	1.2	0.017633	
SEC23B	Protein transport protein Sec23B (SEC23-related protein B)	1	1.17	0.044025	
ATP5MG	ATP synthase subunit g, mitochondrial (ATPase subunit g) (ATP synthase membrane subunit g)	1	1.15	0.004566	
SGPL1	Sphingosine-1-phosphate lyase 1	2	1.15	0.023653	
KCTD12	Potassium channel tetramerization domain containing 12	2	1.14	0.026913	
RPL9	60S ribosomal protein L9	9	1.14	0.021768	
RAB2A	RAB2A, member RAS oncogene family	1	1.13	0.044279	
SLC25A5	ADP/ATP translocase 2 (ADP, ATP carrier protein 2) (Adenine nucleotide translocator 2) (ANT 2) (Solute carrier family 25 member 5) [Cleaved into: ADP/ATP translocase 2, N-terminally processed]	1	1.13	0.027615	
URB1	URB1 ribosome biogenesis homolog	1	1.12	0.028702	
GLYR1	Glyoxylate reductase 1 homolog (Nuclear protein NP60) (Putative oxidoreductase GLYR1)	2	1.11	0.00771	
RPL35A	60S ribosomal protein L35a	7	1.11	0.042789	

OSTC	Oligosaccharyltransferase complex subunit OSTC	1	1.1	0.049372	
NDUFS7	NADH dehydrogenase [ubiquinone] iron-sulfur protein 7, mitochondrial (EC 7.1.1.2) (Complex I-20kD) (CI-20kD) (NADH-ubiquinone oxidoreductase 20 kDa subunit) (PSST subunit)	3	1.1	0.045566	
HNRNPA0	Heterogeneous nuclear ribonucleoprotein A0	7	1.1	0.01617	Mouse (Chu et al., 2015b), Human (Graindorge et al., 2019, Yu et al., 2021)
RCL1	RNA 3'-terminal phosphate cyclase-like protein	2	1.1	0.022557	
FLOT1	Flotillin-1	2	1.09	0.005191	
RPL31	60S ribosomal protein L31	5	1.09	0.022215	
RPL22	60S ribosomal protein L22	2	1.09	0.024261	
HNRNPA2B1	Heterogeneous nuclear ribonucleoproteins A2/B1	13	1.07	0.018127	Mouse (Nguyen et al., 2018), Human (Nguyen et al., 2018)
NUDT21	Cleavage and polyadenylation specificity factor subunit 5 (Nucleoside diphosphate-linked moiety X motif 21) (Nudix motif 21)	8	1.07	0.006186	
RPL23	60S ribosomal protein L23	5	1.07	0.044941	
SLC25A6	ADP/ATP translocase 3 (ADP, ATP carrier protein 3) (ADP, ATP carrier protein, isoform T2) (ANT 2) (Adenine nucleotide translocator 3) (ANT 3) (Solute carrier family 25 member 6) [Cleaved into:	6	1.07	0.0442	

	ADP/ATP translocase 3, N-terminally processed]				
MLEC	MLEC protein (Malectin)	4	1.06	0.045755	
HNRNPU	Heterogeneous nuclear ribonucleoprotein U (Heterogeneous nuclear ribonucleoprotein U (Scaffold attachment factor A))	3	1.06	0.048149	Mouse (Chu et al., 2015b, McHugh et al., 2015), Human (Kolpa et al., 2016)
RPL30	60S ribosomal protein L30	8	1.06	0.0147	
SKIV2L2	SKIV2L2 protein	9	1.05	0.005256	
TOP1	TOP1 protein (Fragment)	2	1.05	0.036306	Mouse (Minajigi et al., 2015)
SYPL1	Synaptophysin like 1	1	1.05	0.002165	
IARS1	Isoleucyl-tRNA synthetase (EC 6.1.1.5)	6	1.05	0.014789	
FYN	Tyrosine-protein kinase Fyn (EC 2.7.10.2) (Proto-oncogene c-Fyn) (p59-Fyn)	2	1.04	0.008434	
RPL14	60S ribosomal protein L14	1	1.04	0.014238	
PRPF3	U4/U6 small nuclear ribonucleoprotein Prp3 (Pre-mRNA-splicing factor 3)	4	1.04	0.011587	Human (Graindorge et al., 2019).
ARPC1B	Actin-related protein 2/3 complex subunit 1B (Arp2/3 complex 41 kDa subunit) (p41-ARC)	5	1.03	0.034236	
SLC25A1	Tricarboxylate transport protein, mitochondrial (Citrate transport protein) (CTP) (Solute carrier family 25 member 1) (Tricarboxylate carrier protein)	4	1.03	0.036637	
KIAA1217	KIAA1217	4	1.02	0.012406	
SLC25A3	Phosphate carrier protein, mitochondrial (Phosphate transport protein) (PTP) (Solute carrier family 25 member 3)	9	1.01	0.032351	

SENP3	ULP_PROTEASE domain-containing protein	4	1.01	0.000839	
RPL11	60S ribosomal protein L11	5	1.01	0.023303	

Bovine XIST is predominantly localised in the nucleus (Yu et al., 2020). The bovine stromal cell lysates used for pulldown assays of bovine XIST repeat A were whole cell lysates, based on cytoplasmic and nuclear markers (shown in **Figure 4.6**). Pulldown assays lacking UV crosslinking prior to cell lysis are inherently biased to identify artefactual associations (Mili and Steitz, 2004), which can be further exacerbated by exogenously introducing a transcript at a supraphysiological concentration. The aim of the next analysis was to distinguish proteins identified as enriched in the sense transcript, due to an artefactual spatial co-localisation of bovine *XIST* repeat A with cytosolic components. To this end, Gene Ontology (GO) term over-representation analysis for biological processes and cellular components was performed on the 40 proteins classified as specifically enriched in the bovine *XIST* repeat A sense transcript.

The predominant biological processes (most highly enriched with p-value <0.05) computed were ‘positive regulation of signal transduction by p53 regulator’ and ‘negative regulation of ubiquitin-dependent protein catabolic process’, both of which processes were represented by the presence of two genes, RPL11 and RPL23 (**Figure 4.16**). It’s worth noting that ribosomal proteins are known contaminants in pulldown experiments (Chen and Gingras, 2007, Rees and Lilley, 2011, Pardo and Choudhary, 2012, Mellacheruvu et al., 2013). Among the most statistically significant processes estimated were ‘translation’ and ‘cellular nitrogen compound metabolic process’, the former of which was represented by 10 genes and the latter by 20 genes, again featuring a high proportion of ribosomal proteins. This suggested that some of the proteins that were pulled down in bovine whole cell lysates with the bovine XIST repeat A sense transcript, were proteins that would carry out their function in the cytosol and cytoplasm. Following this up with GO term cellular component analysis revealed that up to 36/40 (~90%) proteins were classified under a category that indicates cytoplasmic localisation (**Figure 4.17**). Two of the proteins in this category were hnRNPU and TOP1, both of which were previously found to interact with *Xist* in mouse (Chu et al., 2015b, McHugh et al., 2015, Minajigi et al., 2015) and human (Yu et al., 2021). PRPF3, a known mRNA splicing factor, was also in the ‘cytoplasm’ category and has previously been described as a human *XIST* partner (Graindorge et al., 2019). The ‘ribonucleoprotein complex’ category was comprised of 11 genes, eight of which were ribosomal proteins and three splicing

factors (hnRNPU, hnRNPA2B1 and PRPF3). Taken together, these findings highlight a large proportion of the proteins found to interact with the bovine XIST repeat A sense transcript by TMT-MS was predominantly cytosolic proteins, or proteins which could shuttle between the cytosol/cytoplasm and the nucleus.

The protein with the highest log_2FC was TATA-binding protein-associated factor 3 / Transcription initiation factor TFIID subunit 3 (TAF3; log_2FC ~1.84). TAF3 is a component of the core promoter-recognition complex TFIID, which has been shown to recognise H3K4me3 histone marks (associated with active transcription and promoters) and is involved in lineage commitment regulation in ESCs (Lauberth et al., 2013, Ong and Corces, 2014). Another protein found here that was previously described to interact with human *XIST* was HNRNPA0 (Graindorge et al., 2019, Yu et al., 2021), which has known roles in RNA metabolism and translation. Equally, HNRNPA2B1 has been reported to bind human and mouse *Xist* (Nguyen et al., 2018). SPEN, RBM15 and WTAP proteins (previously shown to interact with human *XIST*; **Figures 3.15, 4.8** and **3.13**, respectively) were not found to be significantly enriched (p-value <0.05) with a log_2FC>1 in the sense bovine *XIST* repeat A over the antisense transcript. In fact, RBM15 was found to be enriched at 0.8-fold on average in the sense over the antisense transcript (with 11 unique peptides) whereas SPEN and WTAP proteins were not detected at all in this dataset, in either sense or antisense elutions. This analysis has identified potentially novel protein interactors of *XIST* which have not been previously found in cow, human or mouse, including TAF3 and PRPF3. In summary, *in vitro* transcription of bovine *XIST* coupled to TMT-MS revealed several protein partners that interact with bovine *XIST* repeat A, expanding the list of conserved protein partners and introducing novel players with a potential role in bovine *XIST* processing or cow XCI. It would be necessary to validate these interactions and test for a role in XCI before characterising them as *bona fide* functional *XIST* interactors.

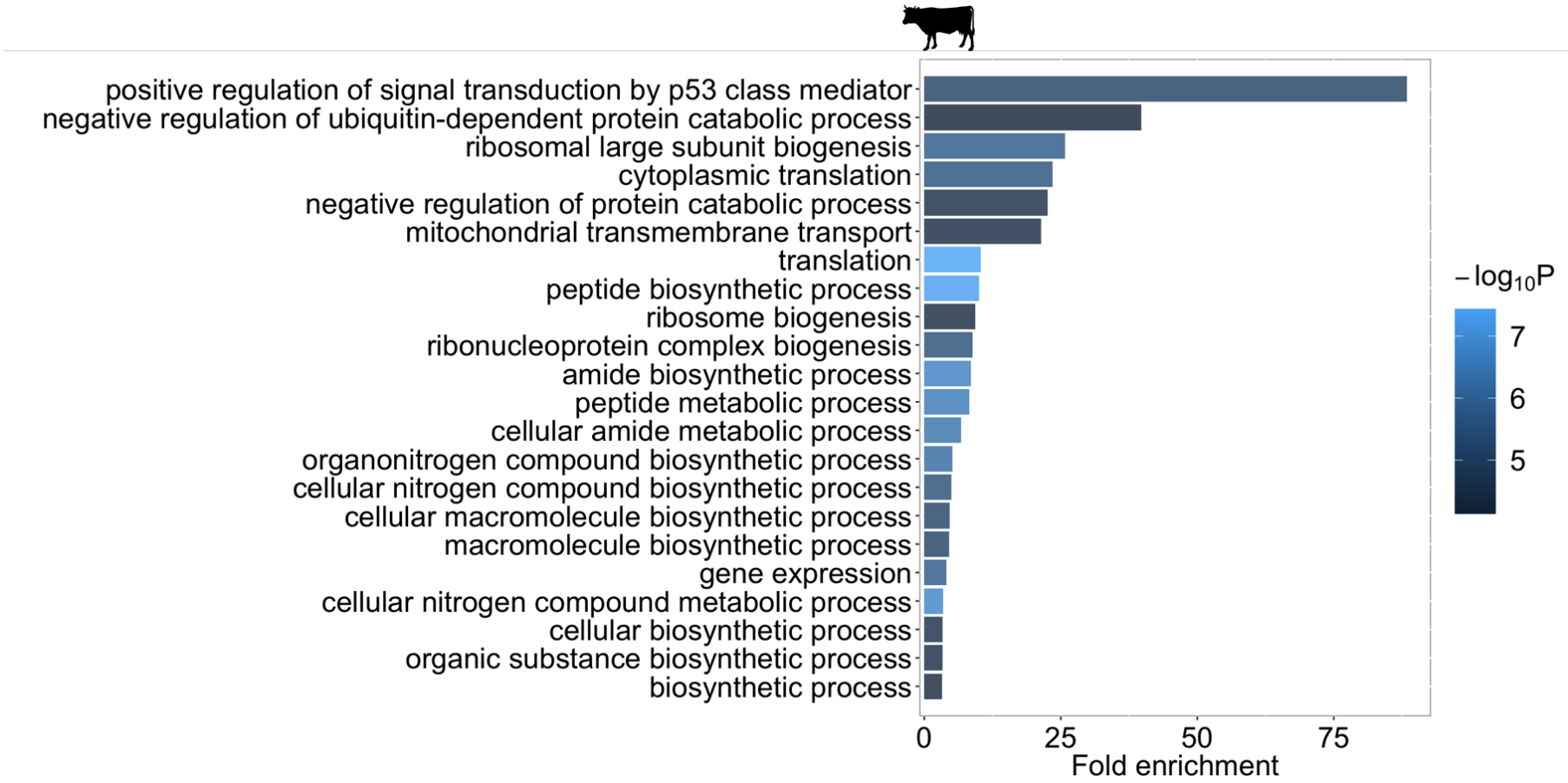


Figure 4.16. Most proteins bound to bovine XIST repeat A sense in bovine cells are involved in biological processes in the cytoplasm.

GO term analysis for biological process was performed using the Gene ontology and PANTHER platforms with the consensus list of 40 proteins identified across all three biological replicates as enriched in the bovine *XIST* repeat A sense transcript mixed in bovine cells. Fisher's exact test was used to infer significance (p-value <0.05). $-\log_{10}P$, $\log_{10}$-transformed p-value.

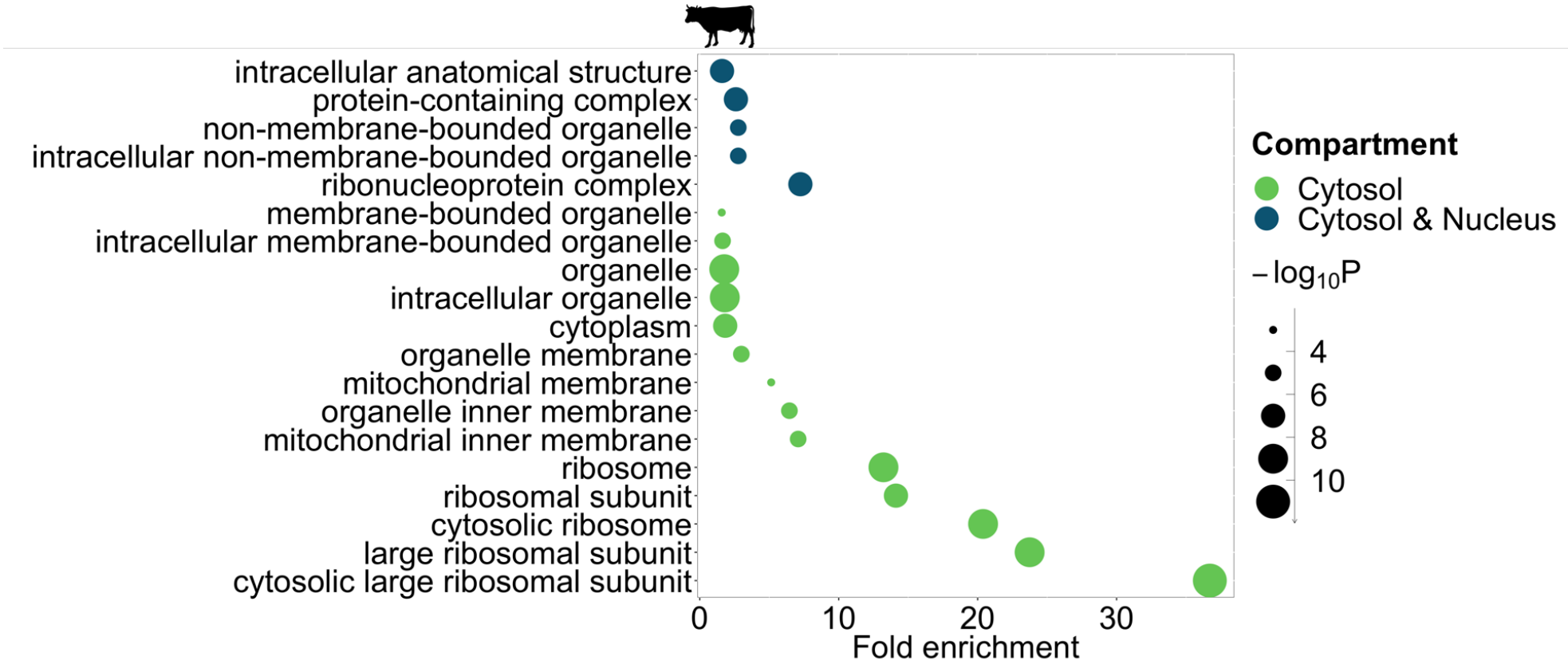


Figure 4.17. A high proportion of proteins enriched in bovine XIST repeat A sense in bovine cells have a role in the cytosol.

GO term analysis for cellular process was performed using the Gene ontology and PANTHER platforms with the consensus list of 40 proteins identified across all three biological replicates as enriched in the bovine *XIST* repeat A sense transcript mixed in bovine cells. Fisher's exact test was used to infer significance (p-value <0.05). $-\log_{10}P$, $\log_{10}$-transformed p-value.

4.3.8. Pulldown of bovine *XIST* repeat A in human lysates elucidates protein partners shared with *XIST* from other placental mammals but not with bovine *XIST*

Repeat A is the region on *XIST* with the highest conservation across placental mammals (>76% across human, mouse, cow and pig; **Table 2.3**) compared to other repetitive regions or full-length *XIST*. Protein partners of mouse and human *XIST* repeat A were shown to be very similar in amino acid sequence comparisons (LBR >80%, SPEN >80%, RBM15 >94% and WTAP >95%; **Table 2.4**). Therefore, it is possible that *XIST* repeat A from one placental mammal could bind proteins from another placental mammal, which would be more likely if the function of the complex would be conserved. The aim of this next experiment was to test whether the sequence of bovine *XIST* repeat A could interact with proteins from human lysates, which had previously been shown to interact with human *XIST*. To this end, bovine *XIST* repeat A was *in vitro* transcribed and used in pulldowns with nuclear-enriched ISHIKAWA lysates. Subsequently, streptavidin-coated magnetic beads used for the capture of RNA-protein complexes from three independent biological replicates were subjected to a 6-plex quantitative TMT mass spectrometry run to identify all protein partners of bovine *XIST* repeat A in human cells. Mass spectrometry was employed here since an extensive list of validated protein partners of bovine *XIST* is lacking and several new partners were sought to be identified, instead of testing predicted ones via western blotting.

After data analyses, it was evident that all sense samples had been misidentified to antisense and vice versa. This was based on the expectation that more proteins would be enriched on the sense transcript (containing a specific sequence) rather than the non-biologically relevant antisense negative control (analyses leading to this conclusion are shown in Supplementary Information).

Given the hypothesis that sense samples had been misidentified as antisense (and vice versa), data were reanalysed to account for potential human error. To examine the variation across biological replicates after applying the proposed adjustment of inverting values for all replicates, PCA was performed. There was agreement

between two of the three replicates for pulldowns performed with either the sense or antisense transcripts (**Figure 4.18;** by Dr Phil Lewis).

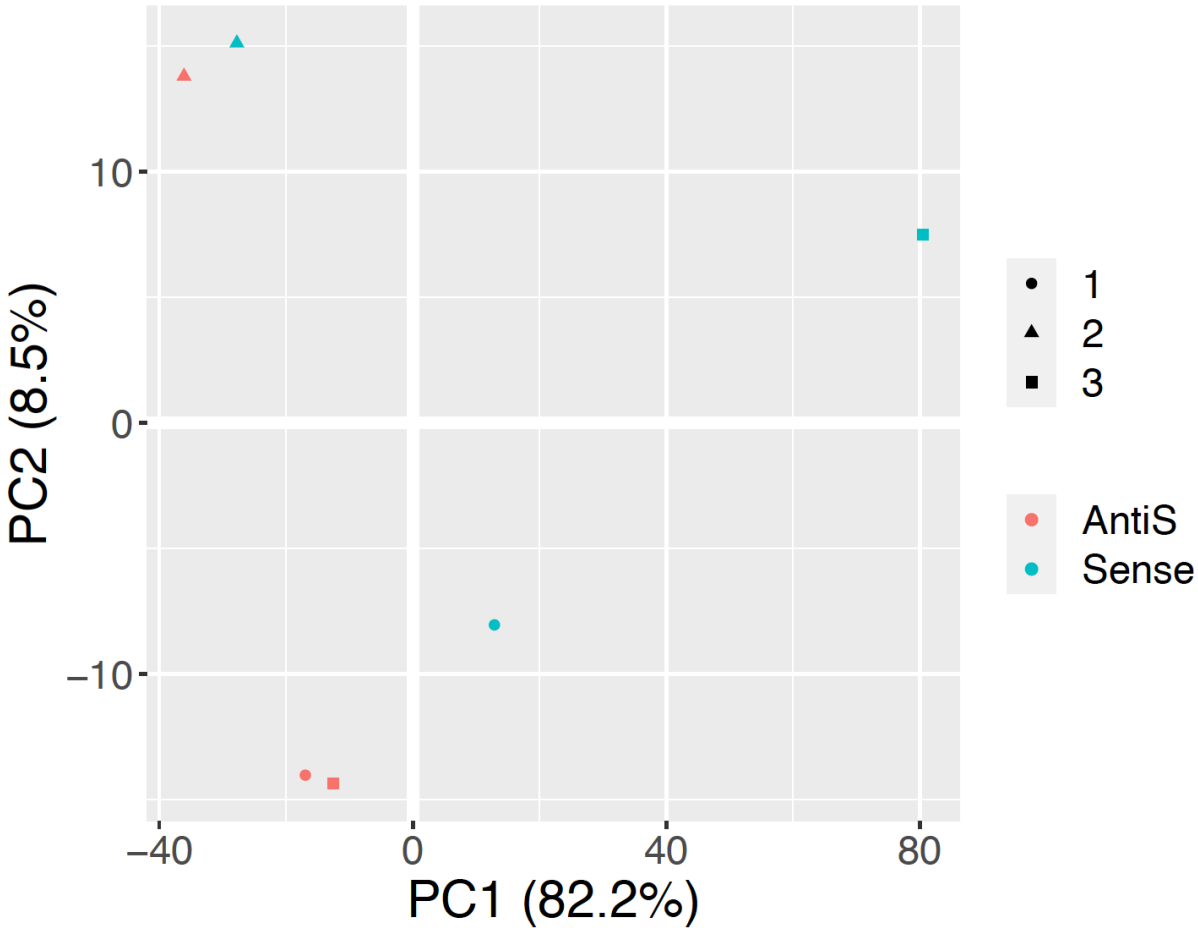


Figure 4.18. Bovine XIST repeat A antisense replicates in human cells are more similar to each other than bovine XIST repeat A sense replicates following log2FC adjustment.

PC1 on the x-axis accounts for 82.2% of the variation between sense and antisense bovine *XIST* repeat A samples and PC2 on the y-axis accounts for 8.5% of the variation between samples. Two out of three pulldown replicates with the sense transcript (green) clustered above 0 for PC1, and two of these were above 0 for PC2 as well. All three pulldown replicates with the antisense transcript (red) were below 0 for PC1 and two of these were below 0 for PC2. All replicates shown are independent biological replicates. Generated by Dr Phil Lewis (Proteomics Facility at University of Bristol). PC, principle component.

The distribution of protein abundance differences between sense and antisense across replicates was examined using violin plots of the distribution of log_2FC across all proteins per replicate. Protein abundance fold-changes of replicate 1 varied from 1.5 to -0.5 range as expected, revealing that a large proportion of proteins demonstrated a slight preference for the sense over the antisense transcript (**Figure 4.19**). For replicate 2, the range was from 1 to -1, still indicating most proteins were not showing a clear preference for either the sense or antisense transcript. Conversely, protein abundance fold-changes in replicate 3 ranged from 3 to 0.5, suggestive of a clear preference of most proteins for the sense over the antisense transcript. Replicate 3 was different to the other two replicates, given the higher number of proteins that displayed preferential enrichment towards the sense transcript. Altogether, all three replicates agreed on a positive directionality of differential protein abundance, implying most proteins in the dataset were preferentially binding the sense over the antisense transcript, as expected.

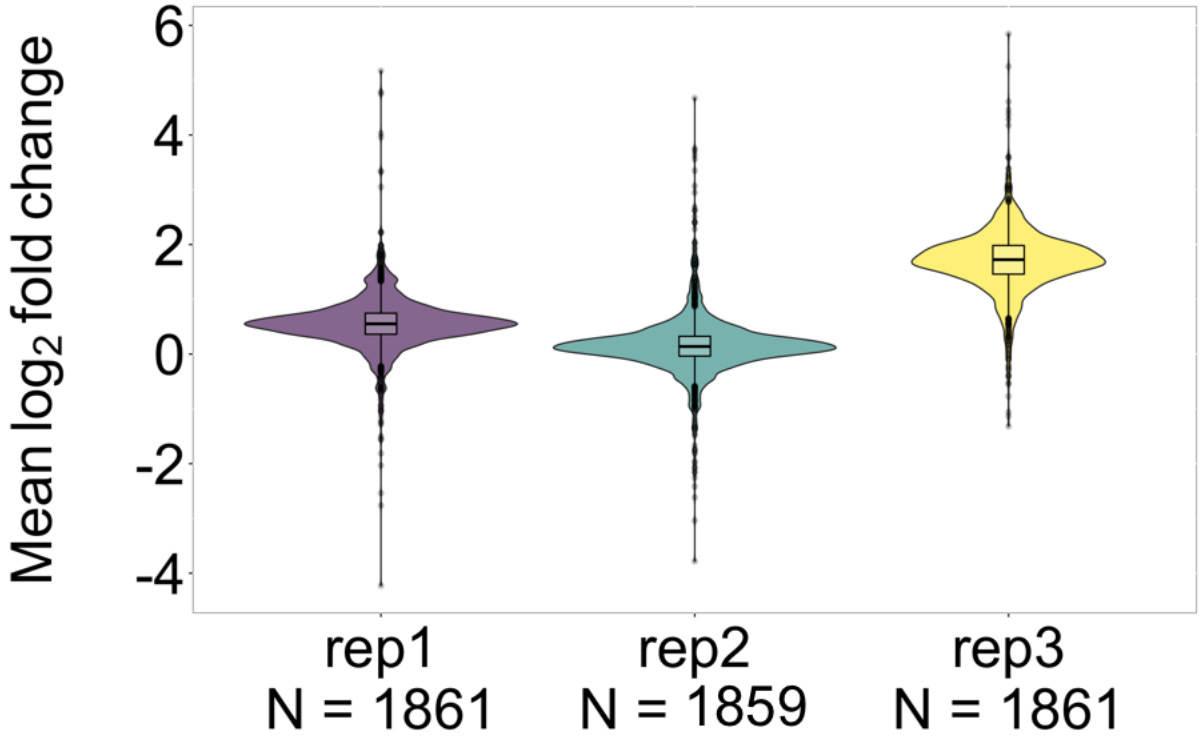


Figure 4.19. Distribution of differential protein abundance across replicates in the cow dataset following data reanalysis.

Violin plot illustrating the distribution of differential protein abundance in replicate 1 is similar to replicate 2 whereas replicate 3 is slightly different. Following adjustment, changes appear predominantly positive, indicative of protein enrichment in the bovine *XIST* repeat A sense transcript instead of depletion. Differential protein abundance (log2 fold change) was estimated by taking the mean of the log2 normalised protein abundance difference of antisense from sense: $\Sigma\ \Delta log_2^{(sense-antisense)}$. Violin width reflects the rough frequency of data points in each region. Cross bars denote the median, box limits represent the 25th and 75th percentiles, whiskers show the 1.5x interquartile range and dots indicate individual data points. Number of proteins per replicate are shown under each violin plot. Replicates performed were independent biological replicates.

Next, the overlap of proteins found enriched or depleted across replicates of the bovine *XIST* repeat A sense transcript was examined. Proteins with a log_2FC cut-off of higher than 1 were classified as enriched and those with a log_2FC cut-off lower than -1 were classified as depleted. More proteins were found to be enriched than depleted in each replicate (compare panels in **Figure 4.20**), and the overlap between replicates in the sense-enriched proteins was much larger: 40 proteins were enriched in the sense compared to antisense in all 3 replicates. Overall, the pattern observed fitted with the expectation that there would be a consensus list of proteins bound to the sense transcript upon recognition of a specific sequence. In agreement, variation in proteins binding to the antisense would be higher, consistent with a lack of great overlap seen in proteins found depleted from the sense.

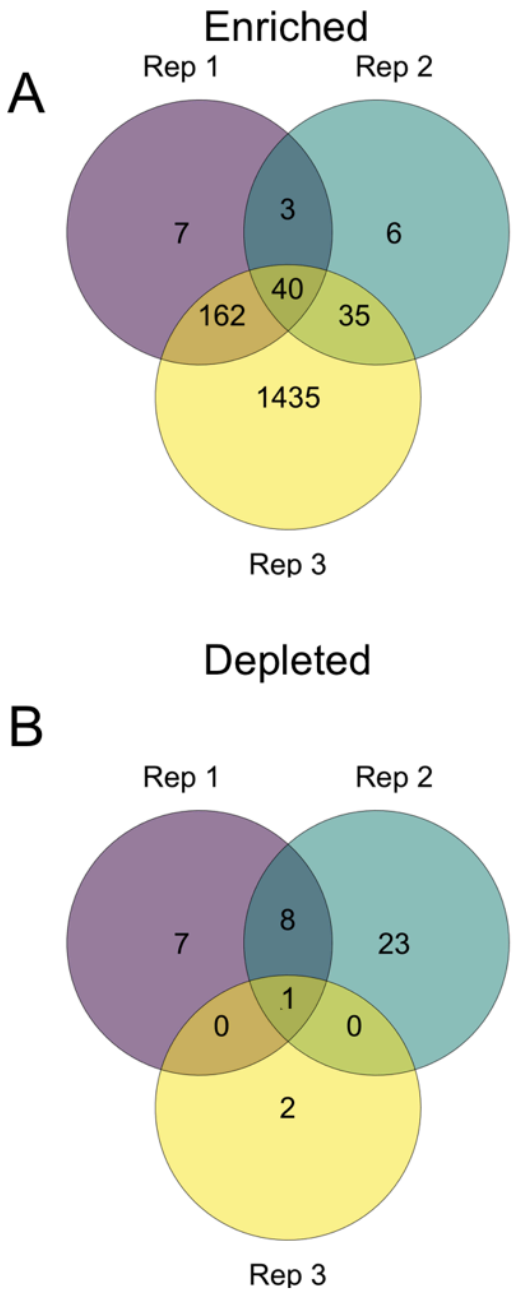


Figure 4.20. Higher overlap in enriched than depleted proteins pulled down by bovine XIST repeat A sense transcripts in human cells across replicates following reanalysis.

Venn diagram showing overlap between proteins **A**) enriched or **B**) depleted in sense over antisense bovine *XIST* repeat A transcripts. 40 proteins were common across all three replicates from those found to be enriched whereas a single protein was common across all three replicates from proteins found to be depleted. Proteins were called 'enriched' or 'depleted' based on log_2FC cut-offs of >1 or <-1, respectively. Proteins with log_2FC values between -1 and 1 were not plotted here. Replicate 1 is shown in purple, replicate 2 in green and replicate 3 in yellow.

The aim of using bovine *XIST* repeat A for pulldowns in human cells was to examine the potential of a highly conserved region of *XIST* (repeat A; 85% similarity between cow and human) from one placental mammal species to associate with proteins from another placental mammal species. To identify proteins that were statistically differentially bound between sense and antisense elution samples, log_2FC values and associated p-values were calculated, using all 3 replicates and then plotted as a volcano plot. A total of 1861 proteins were seen for the first and third replicates whereas 1859 proteins were found in the second replicate. Across all three replicates, 55 proteins were found to be statistically significant ($p<0.05$), out of which only 33 were enriched in the sense compared to antisense (above the minimum enrichment cut-off of log_2FC >1; **Figure 4.21;** by Dr Phil Lewis and **Table 4.4**). This constituted the high-confidence list of human proteins specifically binding to the bovine *XIST* repeat A sense transcript.

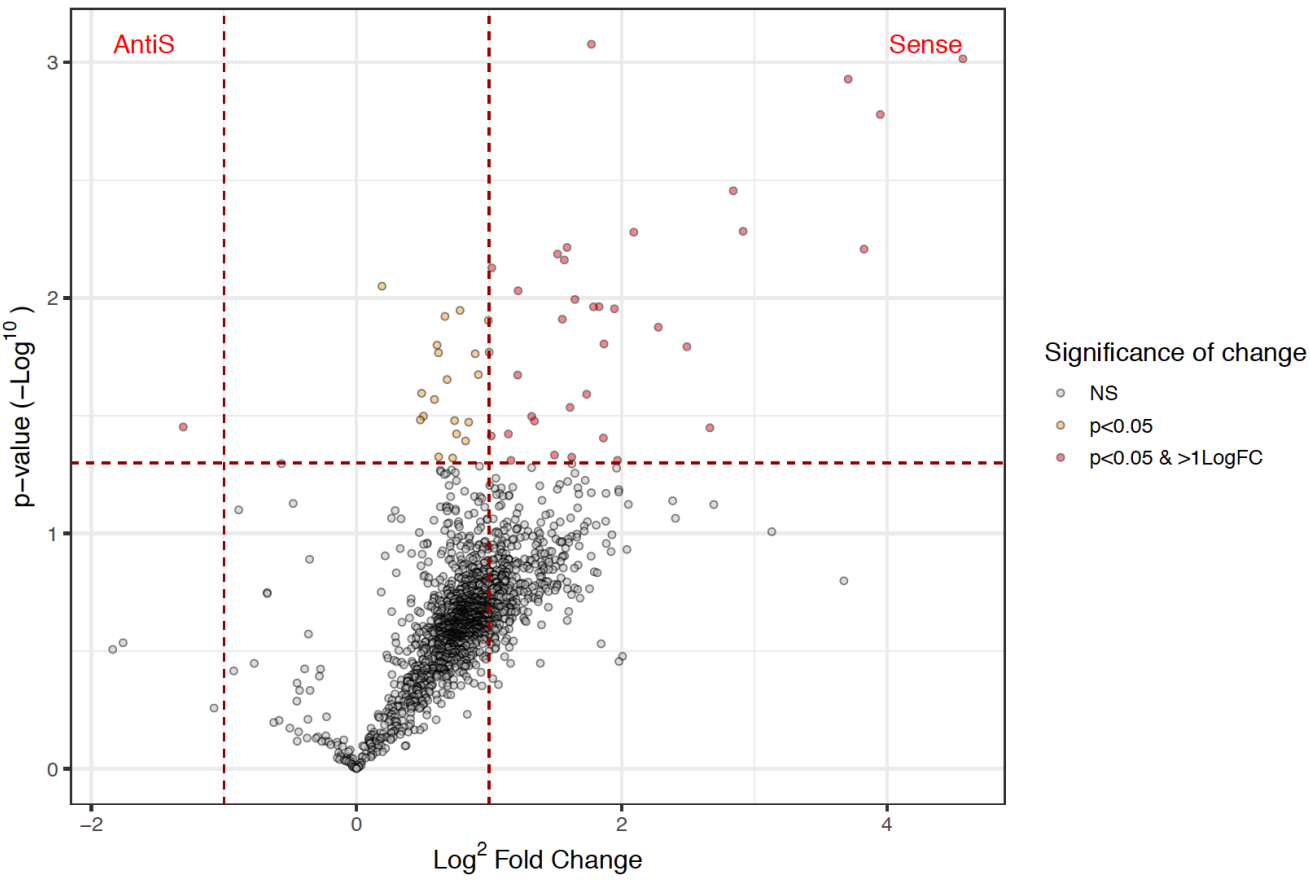


Figure 4.21. Volcano plot of proteins found enriched in sense or antisense bovine XIST repeat A transcripts from human cells following log2FC adjustment.

In total, 33 proteins (in red) were found to be significantly enriched ($log_2FC>1$ and p-value <0.05) in the bovine *XIST* repeat A sense transcript over the antisense transcript. Proteins that were significantly enriched at a $log_2FC<1$ (p-value <0.05) are shown in yellow. Proteins that were not enriched or did not reach statistical significance are shown in grey. N=3 independent biological replicates performed. Statistical analysis was performed with a Student's t-test (significance was inferred at p-value <0.05). Generated by Dr Phil Lewis (Proteomics Facility at University of Bristol).

Table 4.4. High-confidence list of proteins identified by TMT-MS from bovine XIST repeat A pulldowns in human lysates.

Candidates fulfilled the criteria of having a $log_2FC>1$ and being statistically significant (Student's t-test, p-value<0.05). TMT-MS, tandem mass tag mass spectrometry

Protein symbol	Protein name	Unique Peptides	Average log_2FC	T-test (p-value)
HNRNPC	Heterogeneous nuclear ribonucleoproteins C1/C2 (hnRNP C1/C2)	2	4.57	0.000964
RALYL	RNA-binding Raly-like protein (hRALYL) (Heterogeneous nuclear ribonucleoprotein C-like 3) (hnRNP core protein C-like 3)	1	3.95	0.001663
PCCA	Propionyl-CoA carboxylase alpha chain, mitochondrial (PCCase subunit alpha) (EC 6.4.1.3) (Propanoyl-CoA:carbon dioxide ligase subunit alpha)	1	3.83	0.006199
HNRNPCL1	Heterogeneous nuclear ribonucleoprotein C-like 1 (hnRNP C-like-1) (hnRNP core protein C-like 1)	1	2.91	0.00521
GRHL2	Grainyhead-like protein 2 homolog (Brother of mammalian grainyhead) (Transcription factor CP2-like 3)	3	2.84	0.003505
ZCCHC10	Zinc finger CCHC domain-containing protein 10	2	2.66	0.035613
SRSF1	Serine/arginine-rich splicing factor 1 (Alternative-splicing factor 1) (ASF-1) (Splicing factor, arginine/serine-rich 1) (pre-mRNA-splicing factor SF2, P33 subunit)	11	2.49	0.01611
WDR89	WD repeat-containing protein 89	1	2.28	0.013311
STEEP1	STING ER exit protein	3	2.09	0.005256
SRSF8	Serine/arginine-rich splicing factor 8 (Pre-mRNA-splicing factor SRP46) (Splicing factor SRp46) (Splicing factor, arginine/serine-rich 2B)	3	1.97	0.048946
MCCC1	Methylcrotonoyl-CoA carboxylase subunit alpha, mitochondrial (MCCase subunit alpha) (EC 6.4.1.4) (3-	20	1.94	0.011116

	methylcrotonyl-CoA carboxylase 1) (3-methylcrotonyl-CoA carboxylase biotin-containing subunit) (3-methylcrotonyl-CoA:carbon dioxide ligase subunit alpha)			
YY1	Transcriptional repressor protein YY1 (Delta transcription factor) (INO80 complex subunit S) (NF-E1) (Yin and yang 1) (YY-1)	3	1.87	0.015682
MATR3	Matrin-3	1	1.86	0.039355
SNRPE	Small nuclear ribonucleoprotein E (snRNP-E) (Sm protein E) (Sm-E) (SmE)	4	1.83	0.010893
PC	Pyruvate carboxylase, mitochondrial (EC 6.4.1.1) (Pyruvic carboxylase) (PCB)	18	1.79	0.010906
C19orf53	Leydig cell tumor 10 kDa protein homolog	5	1.77	0.000836
MLAA-44	Antigen MLAA-44	1	1.73	0.025646
SNRPD3	Small nuclear ribonucleoprotein Sm D3 (Sm-D3) (snRNP core protein D3)	5	1.65	0.010147
THRAP3	Thyroid hormone receptor-associated protein 3 (BCLAF1 and THRAP3 family member 2) (Thyroid hormone receptor-associated protein complex 150 kDa component) (Trap150)	14	1.62	0.047444
SNRPF	Small nuclear ribonucleoprotein F (snRNP-F) (Sm protein F) (Sm-F) (SmF)	2	1.61	0.029166
NSF	Vesicle-fusing ATPase (EC 3.6.4.6) (N-ethylmaleimide-sensitive fusion protein) (NEM-sensitive fusion protein) (Vesicular-fusion protein NSF)	2	1.59	0.006102
POLR1B	DNA-directed RNA polymerase I subunit RPA2 (RNA polymerase I subunit 2) (EC 2.7.7.6) (DNA-directed RNA polymerase I 135 kDa polypeptide) (RPA135)	3	1.57	0.006896
ATP5PO	ATP synthase subunit O, mitochondrial (ATP synthase peripheral stalk subunit OSCP)	2	1.55	0.012304

	(Oligomycin sensitivity conferral protein) (OSCP)			
SNRPD2	Small nuclear ribonucleoprotein Sm D2 (Sm-D2) (snRNP core protein D2)	9	1.52	0.006511
ZNF43	Zinc finger protein 43 (Zinc finger protein 39) (Zinc finger protein HTF6) (Zinc finger protein KOX27)	1	1.49	0.046503
SPCS2	Signal peptidase complex subunit 2 (EC 3.4.-.-) (Microsomal signal peptidase 25 kDa subunit) (SPase 25 kDa subunit)	2	1.34	0.033359
SNRPN	Small nuclear ribonucleoprotein-associated protein N (snRNP-N) (Sm protein D) (Sm-D) (Sm protein N) (Sm-N) (SmN) (Tissue-specific-splicing protein)	7	1.32	0.03185
RBM5	RNA-binding protein 5 (Protein G15) (Putative tumor suppressor LUCA15) (RNA-binding motif protein 5) (Renal carcinoma antigen NY-REN-9)	2	1.22	0.00932
SNRPD1	Small nuclear ribonucleoprotein Sm D1 (Sm-D1) (Sm-D autoantigen) (snRNP core protein D1)	3	1.22	0.021253
ITGB1	Integrin beta-1 (Fibronectin receptor subunit beta) (Glycoprotein IIa) (GPIIA) (VLA-4 subunit beta) (CD antigen CD29)	1	1.16	0.048991
OSBPL8	Oxysterol-binding protein-related protein 8 (ORP-8) (OSBP-related protein 8)	2	1.14	0.037828
SNRPB2	U2 small nuclear ribonucleoprotein B" (U2 snRNP B")	2	1.02	0.00745
Krt20	Keratin, type I cytoskeletal 20 (Cytokeratin-20) (CK-20) (Keratin-20) (K20)	1	1.02	0.03857

To explore the biological processes that proteins identified might participate in as well as the cellular compartments they localise to, GO term over-representation analysis for biological processes and cellular components was performed using this list of 33 proteins. The most enriched biological processes were related to the spliceosomal complex encompassing several splicing factors such as SNRPD1,2,3, SNRNPE and SNRNPEF (**Figure 4.22**). Among the most statistically significant processes was 'RNA splicing', which included a lot of the same splicing factor family proteins, with the addition of HNRNPC and RBM5. This suggested that a large proportion of the proteins pulled down in bovine whole cell lysates using the bovine *XIST* repeat A sense transcript, were proteins with a clear role in the nucleus. In agreement, GO term cellular component analysis revealed that 22/33 (~66.7%) proteins were classified under the 'nucleus' category (**Figure 4.23**). Both the 'ribonucleoprotein complex' and 'nuclear-protein containing complex' were comprised of splicing factor proteins. However, the latter category also included YY1, a known transcription factor, which has been implicated in the initiation of XCI in mouse (Jeon and Lee, 2011). All together, most proteins identified from the bovine *XIST* repeat A pulldown in human endometrial cells exhibited a nuclear localisation, with defined roles in splicing and RNA processing.

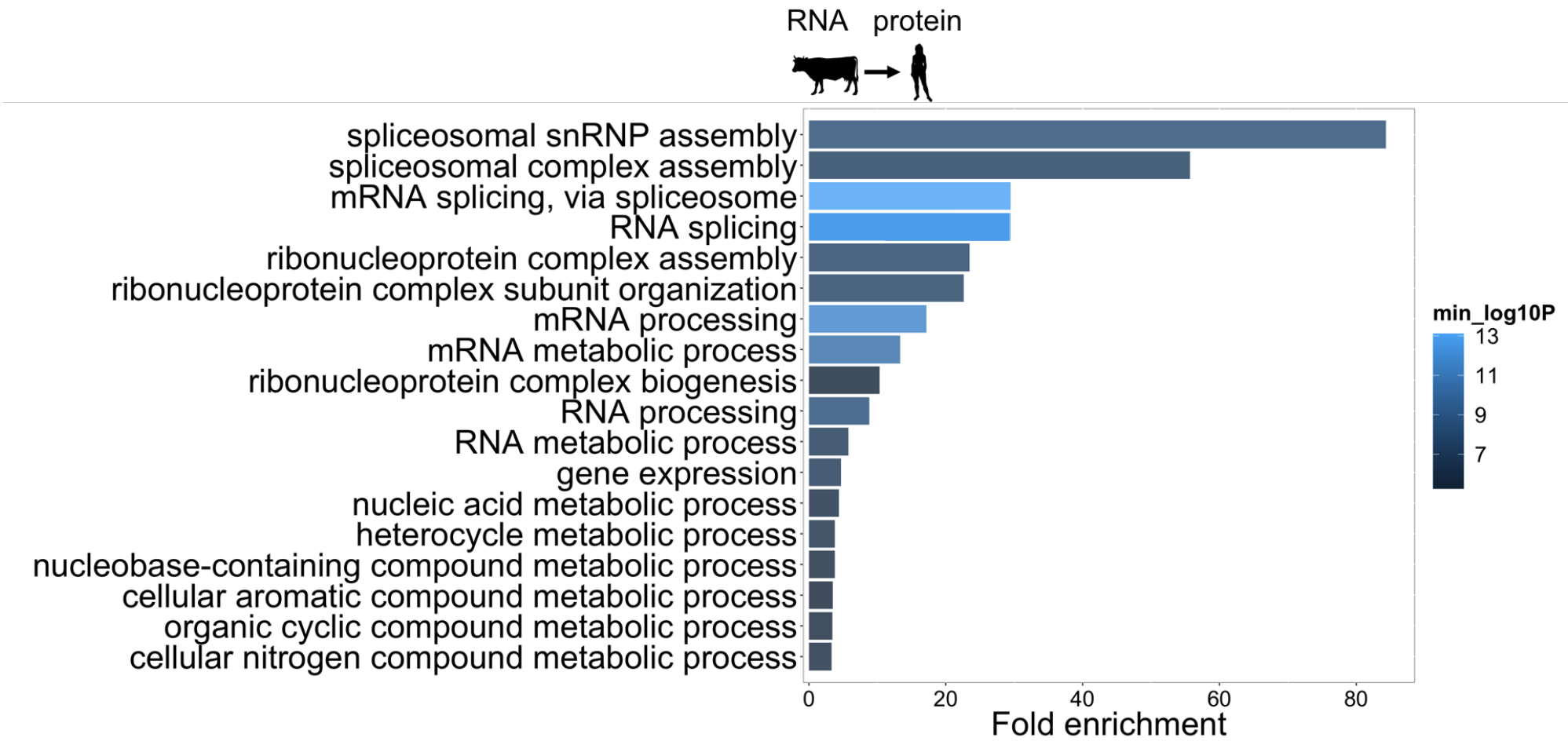


Figure 4.22. Most proteins bound to bovine XIST repeat A in human cells are involved in nuclear processes.

GO term analysis for biological process was performed using the Gene ontology and PANTHER platforms with the consensus list of 33 proteins identified across all three biological replicates as enriched in the bovine *XIST* repeat A sense transcript mixed in ISHIKAWA cells. Fisher's exact test was used to infer significance (p-value <0.05). $-\log_{10}P$, $\log_{10}$-transformed p-value.

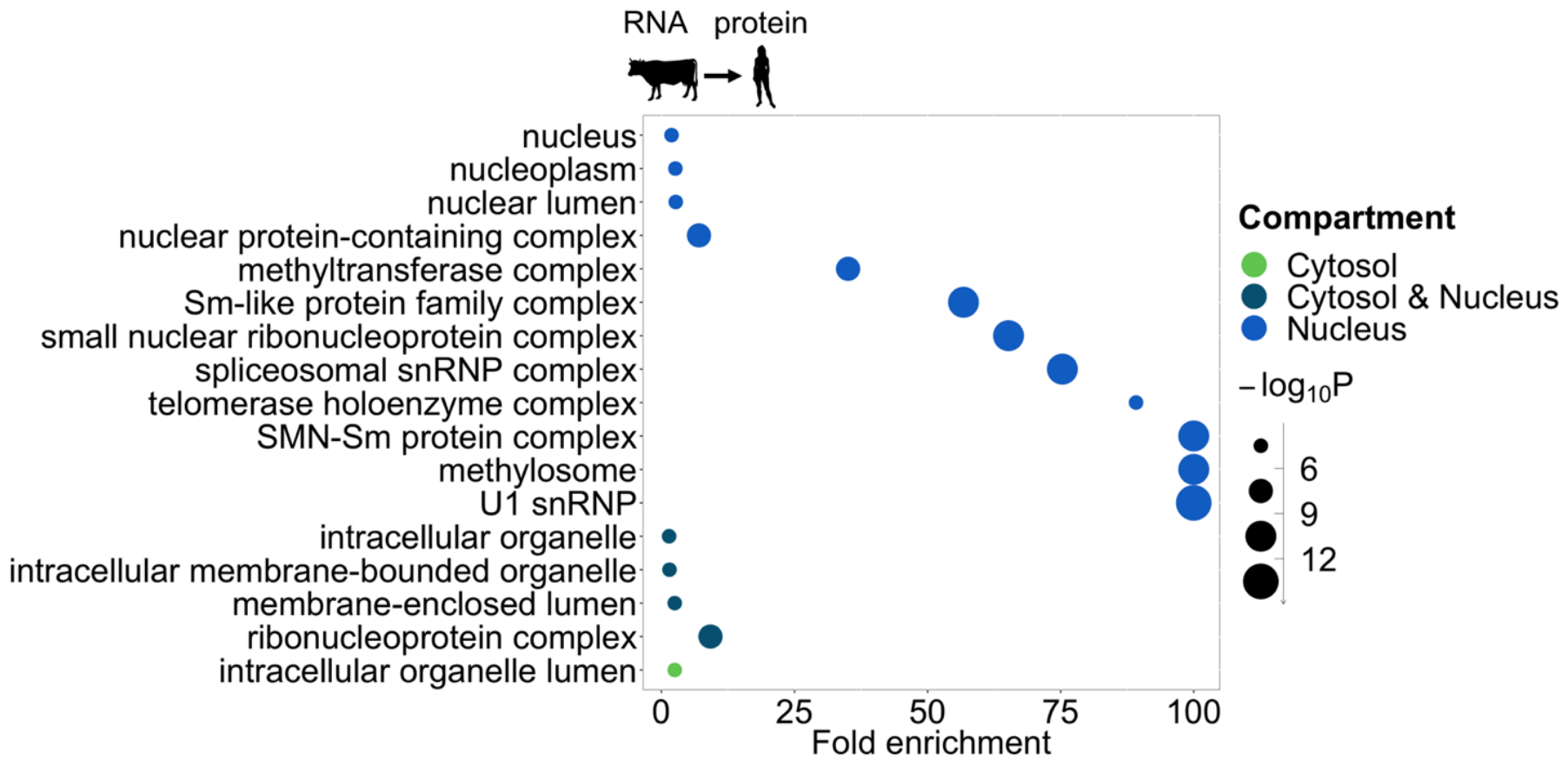


Figure 4.23. Proteins enriched in bovine XIST repeat A sense in human cells have a role in the nucleus.

GO term analysis for cellular process was performed using the Gene ontology and PANTHER platforms with the consensus list of 33 proteins identified across all three biological replicates as enriched in the bovine *XIST* repeat A sense transcript mixed in ISHIKAWA cells. Fisher's exact test was used to infer significance (p-value <0.05). $-\log_{10}P$, $\log_{10}$-transformed p-value.

Among proteins with the highest log_2FC values were hnRNPC (4.57), RALYL (a paralog of RALY; 3.95), HNRNPCL1 (log_2FC 2.91), SRSF1 (2.49) and MATR3 (1.86). All of which were previously found to interact with *XIST* both in mouse (Chu et al., 2015b, Pintacuda et al., 2017a, Dossin et al., 2020) and human (Brown and Baldry, 1996, Graindorge et al., 2019, Yu et al., 2021). YY1 was also shown to bind mouse *Xist* (Jeon and Lee, 2011b). To examine whether bovine *XIST* repeat A in bovine cells would associate with the same set of proteins when placed in a different context, such as in human cells, lists of statistically significant proteins found enriched in the bovine *XIST* repeat A sense transcript from cow (**Table 4.3**) and human lysates (**Table 4.4**) were compared. When considering proteins that passed the statistical significance (p-value < 0.05) and log_2FC>1 cut-offs across all three replicates, there was no overlap detected between the cow and human datasets. When the p-value cut-off was removed, overlap between the two lists was observed for 41 proteins (**Figure 4.24** and **Table 4.5**).

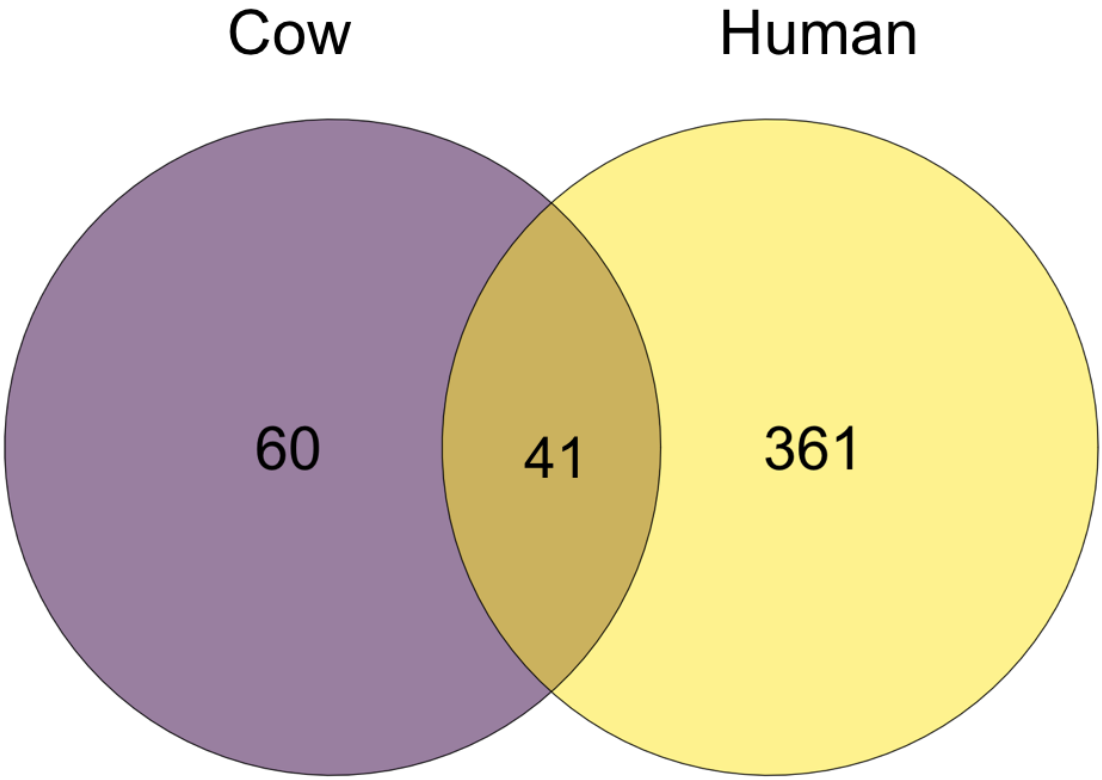


Figure 4.24. Overlap in proteins specifically binding bovine *XIST* repeat A sense in cow and human lysates.

Venn diagram displaying all proteins that exhibit $log_2FC>1$, but are not statistically significant for the cow (purple) and human (yellow) datasets across all three replicates. Almost half of the proteins (~40%) detected with a $log_2FC>1$ in cow lysates and ~10.2% of the proteins in human were common among the two datasets.

Table 4.5. List of common proteins found across three replicates of bovine XIST repeat A pulldowns in cow and human lysates.

Candidates fulfilled the criterion of having a $log_2FC>1$ but were not statistically significant (Student's t-test, p-value<0.05). Most of the common proteins display similar average log_2FC values across species.

	Average log_2FC		Overlap with other datasets	
Protein	Cow	Human	Human	Mouse
ACIN1	1.04	1.59	(Graindorge et al., 2019, Yi et al., 2020)	(Pintacuda et al., 2017)
RNPS1	1.89	3.13		
NKAP	1.27	1.65		
SRSF1	1.58	2.49	(Graindorge et al., 2019, Yi et al., 2020, Yu et al., 2021)	(Pintacuda et al., 2017)
SSRP1	1.3	1.21		
THRAP3	1.16	1.62	(Graindorge et al., 2019, Yi et al., 2020)	
PPIG	1.79	2.38		
HNRNPA0	1.1	1.45		(Chu et al., 2015)
ZC3H18	1.43	1.38		
CWC22	1.14	1.09		
HNRNPU	1.06	1.46	(Yu et al., 2021)	
ARL6IP4	1.04	1.98	(Graindorge et al., 2019)	(Chu et al., 2015)
NOLC1	1.75	1.57	(Yi et al., 2020)	(Pintacuda et al., 2017)
RPL8	1.05	1.12		
ZRANB2	1.23	1.93		
GPATCH1	1.66	1.81		
PNN	1.91	2.4		
PNISR	1.15	1.12		
RPL21	1.03	1.04		
UBTF	1.52	1.11		
ZCCHC10	1.82	2.66		
EIF4A2	1.36	1.21		(Chu et al., 2015)
SERBP1	1.01	1.39		
HNRNPA2B1	1.07	1.45		
BRD3	1.45	2.04		
RPL22	1.09	1.09		
IARS1	1.05	1.18		

BANF1	2.03	1.79		
TBC1D10B	2.31	2.69		
SUB1	1.62	1.15		
NUDT21	1.07	1.11		
TAF3	1.84	1.17		
MCCC1	1.3	1.94		
BRD2	1.11	1.68		
CSRP2	1.06	1.07		
HDGFL2	1.48	1.67		
RPL30	1.06	1.1		
PCCA	1.48	3.83		
PC	1.24	1.79		
RCL1	1.1	1.39		
SKIV2L2 / MTREX	1.05	1.04		

SPEN, RBM15 or WTAP proteins were previously shown to interact with human *XIST* (**Figures 3.15, 4.8** and **3.13**, respectively). These proteins were not found to be significantly enriched in bovine *XIST* repeat A sense in the human lysates. However, SPEN and RBM15 exhibited some level of binding, which did not pass the $log_2FC > 1$ or the statistical significance cut-off. SPEN was enriched at 0.92-fold on average in the sense over the antisense transcript (with 6 unique peptides) and RBM15 0.99-fold (with 11 unique peptides). Although not statistically significant, RBM15B, a paralog of RBM15, both of which have been shown to interact with human *XIST* and be required for efficient X-linked gene silencing (Patil et al., 2016), was enriched at 1.12-fold on average (with 9 unique peptides). WTAP was not detected in this dataset in any samples.

Analysis of TMT-MS of bovine *XIST* repeat A in human lysates also revealed a number of novel proteins (**Table 4.4**), previously not described as interactors of *XIST* in cow, human or mouse, including ZCCHC10, WDR89, POLR1B, RBM5 and splicing factors (SNRPB2, SNRPE, SNRPF, SNRPD1, SNRPD2, SNRPD3, SNRPN). In summary, mixing *in vitro* transcribed bovine *XIST* repeat A with ISHIKAWA cells and using pulldown assays coupled to TMT-MS revealed several proteins that could recognise a conserved repeat A region across human and cow.

4.4 Discussion

The aim of this chapter was both to orthogonally validate interactions that could not be robustly inferred in the previous chapter as well as dissect the region on *XIST* where protein partners bound. Additionally, work here was also directed towards identifying previously uncharacterised protein partners of bovine *XIST*.

4.4.1. *In vitro* transcription pulldowns identify RBM15 as a protein partner of human *XIST* repeat A

In vitro transcription coupled to pulldown is a suitable assay not only to identify protein partners of a specific lncRNA but also to determine the location of the interaction on the RNA and/or the elements contributing to the interaction. Exploiting such an approach here, human *XIST* repeat A was shown to associate with RBM15 (**Figure 4.8**), an observation that could not be reliably inferred from RIP assays (**Figure 3.12**). On the contrary, when the presence of WTAP, another protein known to bind human *XIST* repeat A, was also probed in the same *XIST* repeat A pulldown experiment, a WTAP-*XIST* repeat A interaction could not be reliably inferred. This was in contrast to results obtained with RIP, whereby a WTAP-*XIST* interaction was evident (**Figure 3.13**). Equally, CIZ1 was seen to associate with *XIST* when RIP was used (**Figures 3.10** and **3.11**) whereas when pulldowns were performed using human *XIST* repeat E, the CIZ1 protein could not be reproducibly recovered at a greater degree in the sense transcript compared to the antisense (**Figure 4.7**). Variation in replicates seen could reflect different amounts of input material, despite best efforts made to quantify amount of protein per lysate and ensure equal amount in each pulldown replicate. Another factor that could influence variation in pulldown efficiencies is the variable volume of the lysate (depending on its concentration) mixed with the incubation buffer, which could alter pulldown stringency. Concentrated cell lysate preparations could also affect pulldown efficiency by increasing viscosity of incubation reactions, which would make magnetic bead separation challenging.

It is worth noting that there is only one other study that showed WTAP binding to human *XIST* in a repeat-specific manner, albeit surprisingly binding at the C repeat

was more specific (Graindorge et al., 2019). This was in contrast to studies of mouse *Xist* reporting *Wtap* binding in the A repeat (Chu et al., 2015b). It would be surprising if binding preferences for this protein differ among species given a 95% conservation of the protein between human and cow (**Table 2.4**) and an identical WTAP/Mum2 functional domain between human, cow and mouse (**Figure 2.3**). Instead, it's possible the WTAP-*XIST* interaction a) is not high affinity, b) the strength of the interaction might also depend on co-binding RBM15, or c) it could be context-specific. The original study reporting *Wtap* binding in the A repeat used male mESCs with inducible *Xist* expression (Chu et al., 2015b). Conversely, when others used hybrid mESCs from two different mouse strains (C57BL/6JJcl x 129/SvJcl), an interaction of mouse *Xist* with Wtap was not reported, despite Rbm15 being bound to *Xist* in the same dataset.

It will be important for future *in vitro* transcription coupled to pulldown experiments to address inconclusive results for WTAP and CIZ1 binding human *XIST* in ISHIKAWA lysates by increasing the number of replicates. Due to the trend of low protein abundance for all proteins probed in western blots here, future experiments should consider increasing the amount of starting material to 2 mg, the upper limit suggested by the protocol followed (Tichon et al., 2018). Additionally, the stringency of the incubation buffer could be further optimised as an approach to tackle non-specificity from the presence of Lamin B in elution samples, for which a plethora of protocols/buffers are available (Liu et al., 2012, Wang et al., 2014, Lee et al., 2016, Pintacuda et al., 2017a). Despite the lack of an antibody for western blotting of SPEN, elution samples from pulldowns of human *XIST* repeat A (the region where SPEN is known to bind) could be analysed via mass spectrometry to enable identification of the most highly abundant protein interactors.

4.4.2. Bovine *XIST* protein interactome includes known and previously uncharacterised partners of *XIST*

The protein interactome of bovine *XIST* is previously uncharacterised with no reports in the literature for any protein partners and their potential contributions to X-linked gene silencing and downstream XCI. Here, the aim was to uncover protein partners of bovine *XIST* in primary bovine endometrial stromal cells. These could include

proteins that are specific to cow or shared with other placental mammals. Under a model where *XIST* is shared across placental mammals and has been co-evolving with its protein partners across species, it is likely that protein partners of bovine *XIST* would share protein binding sites on the *XIST* RNA. Based on protein partners previously described for human and mouse *XIST*, the bovine *XIST* repeat E was chosen to be studied for an interaction with CIZ1 and the bovine *XIST* repeat A region for an interaction with SPEN, RBM15 and WTAP. To test whether cow CIZ1 associates with bovine *XIST* via its repeat E element as seen for mouse and human, bovine *XIST* repeat E was *in vitro* transcribed and tested in pulldowns with bovine stromal whole cell lysates.

Pulldown of bovine *XIST* repeat E in bovine stromal whole cell lysates demonstrated a specific interaction with the CIZ1 protein (**Figure 4.9**). This was in contrast with results obtained from RIP, where despite an efficient pulldown of the CIZ1 protein, *XIST* RNA was not consistently found enriched in the elution (**Figure 3.17**). Given the same type of cells served as input for both experiments and the same antibody was used for CIZ1 detection, it's likely this difference reflects technical shortcomings of individual protocols with regards to lysate preparation or pulldown efficiency.

Since several protein partners were known to interact with mouse and human *XIST* repeat A, performing mass spectrometry would allow the detection of all these at the same time from the same sample. This would not have been possible with western blots since these proteins have similar sizes and antibodies suitable for their detection via western blotting were not available. Analysis of TMT-MS yielded a total of 40 proteins as specifically interacting with bovine *XIST* repeat A in bovine stromal cells (**Figure 4.15** and **Table 4.3**). The use of GO term analysis on this list of proteins revealed the majority of those had a function related to the cytosol and were either present in the cytosol or could translocate between cytosol and nucleus (**Figure 4.17**). This could be related to the preparation of bovine stromal lysates as whole cell lysates, reflected by subcellular markers (**Figure 4.6**)(Mili and Steitz, 2004). Among the identified proteins were hnRNPU, TOP1 and PRPF3. The first two have been found to interact with mouse *Xist* in the literature: hnRNPU in mESCs (Chu et al., 2015, McHugh et al., 2015) and Neuro2A (Hasegawa et al., 2010) and TOP1 in MEFs (Minajigi et al., 2015). PRPF3 was shown to interact with human

XIST in HEK293T cells (Graindorge et al., 2019). This protein is a homolog of yeast Prp3, which has been characterised to be a core component of the spliceosome and its depletion, lack of ubiquitination or SUMOylation leads to reduced splicing and assembly of precatalytic spliceosomes (Pozzi et al., 2017). In fact, hnRNPU has also been shown to interact with human *XIST* in a human erythroleukemic cell line (K562) (Hendrickson et al., 2016, Lu et al., 2020a, Yu et al., 2021) and primary human lung fibroblasts (Kolpa et al., 2016). The predominant localisation of hnRNPU is nuclear, albeit it can shuttle between the nucleus and the cytoplasm upon inflammatory stress (Zhao et al., 2012). Aside from its role in splicing, hnRNPU has been shown to aid in anchoring mouse *Xist* to the Xi, facilitating its gene silencing function in Neuro2A cells (Hasegawa et al., 2010, Yamada et al., 2015). Moreover, TOP1 has a strict nuclear localisation and functions as a topoisomerase, relieving torsional stress from DNA supercoiling (Minajigi et al., 2015), albeit no role has been described for it in XCI yet.

A protein that had not been previously described in the literature as an *XIST* partner includes TAF3, a component of the core promoter-recognition complex transcription factor II D (TFIID), which here associated with bovine *XIST* repeat A (**Table 4.3**). Human TFIID is composed of the TATA-binding protein and at least 13 associated factors that are involved in the formation of the transcription pre-initiation complex. The identification of this protein as a potential interactor of bovine *XIST* repeat A implicates repeat A in the regulation of transcription initiation of X-linked genes. Previously, TAF15, another component of the core promoter-recognition complex TFIID was shown to specifically and directly interact with human *XIST* with a preference for positions near the B repeat (position ~2 kbp) and D repeat (position ~6 kbp)(Yi et al., 2020). The TAF15-*XIST* interaction was shown to be involved in XCI owed to selective X-linked, but not autosomal, gene upregulation upon TAF15 depletion in female MEFs (Yi et al., 2020). Thus, a potential XCI-related role of such an interaction would serve to sequester transcription factors away from X-linked genes.

Another potential interactor of bovine *XIST* repeat A found here is Nudix Hydrolase 21 (NUDT21)(**Table 4.3**). NUDT21 is one subunit of the tetrameric cleavage factor Im (CFIm) complex which is involved in 3'-end cleavage and polyadenylation of RNA

and can lead to alternatively polyadenylated transcript isoforms with different 3'-ends (Zhu et al., 2018). Such a protein could process bovine *XIST* and generate several isoforms. Alternatively, competitive binding between NUDT21 and hnRNPK could ensure variable polyadenylated isoforms are not produced, instead forcing the generation of a specific isoform, as seen with the *NEAT1* lncRNA (Naganuma et al., 2012). Although hnRNPK was detected in the cow dataset, it was not found to have a log2FC>1 or pass the statistical significance cut-off.

Superkiller Viralicidic Activity 2-Like 2 (SKIV2L2) (also known as Mtr4 Exosome RNA Helicase, MTREX) has not been reported to associate with *XIST* in any species, but was bound to bovine *XIST* repeat A here (**Table 4.3**). This protein is a subunit of several complexes such as the nuclear exosome targeting (NEXT) which can associate with RNAs possessing a 5'-cap via interactions with ZC3H18 (Garland and Jensen, 2020). The NEXT complex has been demonstrated to be a degradation pathway for lncRNAs and other non-coding RNAs and that inhibiting or depleting this complex results in stabilisation of lncRNAs. The ZC3H18 protein was also found to bind bovine *XIST* repeat A with an average log_2FC of 1.43, albeit it was not statistically significant. SKIV2L2/MTREX is also a subunit of the polyA exosome targeting (PAXT) complex which can target polyadenylated transcripts for decay via associations with polyadenylate-binding nuclear protein 1 (PABPN1) (Garland and Jensen, 2020). PABPN1 was detected as a statistically significant interactor of bovine *XIST* repeat A (Student's t-test, p-value <0.05), albeit it did not pass the log_2FC>1 cut-off (log_2FC for PABPN1 was 0.77).

The human *XIST* repeat A was shown to interact with SPEN and WTAP (in Chapter 3) and an *XIST*-RBM15 interaction was shown here (**Figure 4.8**). No interaction of bovine *XIST* with SPEN, RBM15 or WTAP was evident from RIP in bovine stromal cells. In line with this, none of these three proteins were found in the high-confidence list of interactors (log_2FC >1 and p-value < 0.05 cut-offs) from the TMT-MS data of bovine *XIST* repeat A pulldowns in bovine stromal cells. A lack of a bovine *XIST* interaction with cow SPEN, RBM15 and WTAP proteins is in direct conflict with their binding profiles in mouse and human datasets (**Section 1.6.2**). Although signal for SPEN and WTAP proteins was not picked up at all with TMT-MS, RBM15 was detected in this dataset (with 11 unique peptides, albeit it was not statistically

significant and did not have a log_2FC>1). The lack of signal for SPEN and WTAP proteins could be an issue related to a low amount of input material used combined with potentially modest expression levels in the cow. However, the strength of TMT-MS as a quantitative MS technology is that it enables the detection of even low abundance proteins, via peptide tagging. An inability to detect a protein could arise from a small protein size, the peptides of which might not be sufficiently different from those of other proteins in order to differentiate them. A bias in peptide fragmentation could also be introduced by the use of trypsin as the protease due to almost half of the peptides generated being under 6 aa (Swaney et al., 2010). The use of additional, alternative proteases has been shown to enhance the amount of proteins identified by mass spectrometry by up to ~20% in yeast (Swaney et al., 2010).

A plausible explanation for these observations could be that SPEN, RBM15 and WTAP proteins exhibit cell-type specific binding to *XIST* in ~~other~~ cow tissues. Human *XIST*'s protein interactome has been recently shown to vary depending on the stage of cell differentiation and thus XCI. More specifically, comparing the protein partners of *XIST* across human B cells and ESCs, a previous study reported a 57.8% overlap in the proteins bound to *XIST* across these different XCI states (although this could also be cell-type specific)(Yu et al., 2021). SPEN and RBM15 were present in both datasets whereas WTAP was not present in the dataset from differentiated cells. The roles of WTAP and SPEN are important in the early stages of *XIST* post-transcriptional processing (WTAP) and onset of X-linked gene silencing (SPEN). One could speculate that if these proteins exhibited dynamic binding to *XIST*, their interaction with *XIST* after the establishment of XCI could be dispensable for maintaining a repressive chromatin environment on the Xi. This would be consistent with bovine stromal cells representing terminally differentiated cells with an established XCI landscape. Nonetheless, mutating the RNA-binding domains of SPEN in a differentiated B cell line resulted in X-linked gene de-repression from the Xi, further extending a role for SPEN in XCI maintenance in human B cells (Yu et al., 2021). It will be necessary to extend such observations in other human cell types to distinguish a cell-type specific role of SPEN in maintenance from a general somatic-cell-wide one.

An alternative explanation for a lack of *XIST* binding seen for cow SPEN, RBM15 and WTAP could be that these proteins have acquired different functions in the cow, no longer participating in XCI, therefore rendering an interaction with *XIST* weaker or redundant. One approach to predict whether a protein has maintained the same function across species is to examine signatures of selection acting on its sequence. Using comparative proteomic analyses powered by codon-based models of evolution and maximum likelihood, no signs of positive selection were detected on either WTAP (Wu et al., 2016) or SPEN proteins (Carter et al., 2020). This implies that both proteins have been under purifying selection to preserve their functions across species. More specifically, the RNA-binding domains of SPEN (RRM domain) have been shown not to have evolved much across human, mouse and cow (Carter et al., 2020). Future work could take such analysis a step further by probing the bovine *XIST* repeat A region for human SPEN RRM binding motifs available from CLIP data (Carter et al., 2020), to establish the degree of motif conservation as a proxy for likelihood of binding. Taken together, work in this chapter expanded the *XIST* protein interactome that is shared across placental mammals as well as introduced previously uncharacterised protein partners of bovine *XIST*.

4.4.3. Cross-species pulldown of bovine *XIST* in human cells reveals shared proteins partners with human and mouse but not bovine *XIST*

Differences in *XIST*-protein interactions across species may arise from differences in the lncRNA sequence, in the protein, or in both. To examine whether a biochemical lncRNA-protein interaction may be preserved across species due to a conserved lncRNA sequence, bovine *XIST* repeats E and A were used for pulldowns in human lysates. *XIST* repeat E displayed an affinity towards the CIZ1 protein in both human (**Figure 3.10C**) and cow (**Figure 4.9**). Nevertheless, when *in vitro* transcribed and biotinylated bovine *XIST* repeat E was mixed with human lysates, specific binding was not detected. It is possible an interaction may have not been detected due to an insufficient amount of protein available, consistent with a weak CIZ1 protein signal observed in elution samples (**Figure 3.17**). Yet, a more likely explanation could be the lack of *XIST* repeat E and CIZ1 conservation between human and cow. The similarity of *XIST* repeat E between human and cow was estimated to be ~54.6% (**Table 2.3**) whereas the CIZ1 protein was ~80% similar between the two species

(**Table 2.4**). Despite this high CIZ1 protein similarity however, there are 14 amino acid mismatches interspersed across the Matrin/Zinc-finger domain of CIZ1, responsible for RNA binding (**Figure 2.7**). It would be worth for future investigations to examine whether CIZ1-binding motifs overlap the repeat E region, whether these are present in both human and cow and compare their similarity. Overall, despite CIZ1 binds *XIST* across human, mouse and cow, the sequence of repeat E as well as of the CIZ1 protein in each species varies sufficiently to prevent cross-species redundancy.

Bovine *XIST* repeat A is ~85% similar to its human counterpart (**Table 2.3**) and some of the proteins that bind to human *XIST* repeat A are over 80% similar between human and cow (SPEN >80% and RBM15 >94%; **Table 2.4**). The similarity between human and cow SPEN and RBM15 proteins also spans their RNA binding domains. Identifying proteins that can bind the *XIST* RNA from two different species could offer insight into the *XIST* RNA and protein partner co-evolution. To test whether the sequence of bovine *XIST* repeat A was similar enough to pulldown known protein partners of human *XIST*, bovine *XIST* repeat A was *in vitro* transcribed and used in pulldowns with human lysates. A total of 33 significantly enriched proteins bound to bovine *XIST* repeat A transcript were observed in sense over the antisense transcript via MS ($log_2FC>1$) (**Figure 4.25**). Contrary to what was seen when bovine *XIST* repeat A was mixed with cow lysates, GO term analysis for biological process and cellular component revealed the majority of proteins identified here were involved in processes in the nucleus (**Figure 4.26A&B**). One explanation could be the difference in generation of lysates i.e. nuclear-enriched (human) versus whole cell (cow). The subcellular fractionation conditions used here were suitable for the preparation of a nuclear-enriched ISHIKAWA cell lysate (**Figure 4.7**), as seen by the cytoplasmic and nuclear markers. However, lysis conditions were not optimal for the separation of the cytoplasmic components from the nucleus, resulting in a whole cell lysate for bovine stromal cells (**Figure 4.6**). Therefore, the bovine *XIST* repeat A sequence could be sequestrated by abundant cytoplasmic proteins in a whole cell bovine stromal lysate, which could stoichiometrically inhibit the association of sequence- and structure-specific binding proteins from the nucleus. Consistently, a large proportion of the proteins pulled down when bovine *XIST* repeat A was mixed with whole cell bovine stromal lysates were proteins that have not been described to

occupy both compartments or shuttle between them i.e. 60S ribosomal proteins and mitochondrial components.

Of these 33 enriched proteins identified by pulldown of bovine *XIST* repeat A in human endometrial cells, five had been previously found to interact with both mouse (Chu et al., 2015b, Pintacuda et al., 2017a, Dossin et al., 2020) and human *XIST* (Graindorge et al., 2019, Yu et al., 2021): hnRNPC, hnRNPCL1, RALYL (a paralog of RALY), SRSF1 and MATR3. Interestingly, YY1 was also seen enriched following bovine *XIST* repeat A pulldown in bovine stromal cells (**Table 4.3**). An interaction of YY1 with *XIST* RNA has only been reported once in the literature where authors His-tagged YY1 proteins on magnetic beads and allowed them to interact with total RNA from female mouse embryonic fibroblasts, and *Xist* binding was detected by RT-qPCR (Jeon and Lee, 2011b). Binding was shown to occur at the C-repeat region of mouse *Xist*, located ~3 kbp downstream of the 5' end of the *Xist* RNA. However, more recent screens for mouse or human *XIST* protein partners have not reported YY1 as an interactor of the *XIST* RNA (Lu et al., 2017). Irrespective of binding the *XIST* RNA, a YY1 interaction with the *XIST* promoter was linked to *XIST* transcriptional activation in both female human lung fibroblasts (IMR-90 cell line)(Chapman et al., 2014, Makhlouf et al., 2014) and mESCs (Makhlouf et al., 2014). Depleting YY1 levels did not significantly affect *XIST* RNA expression in naïve or mouse activated female T cells, though (Wang et al., 2016). However, expression of 266 X-linked genes decreased following *ex vivo* YY1 deletion in female mouse activated splenic B cells (Syrett et al., 2017). In another study using YY1 ChIP-Seq in lymphoblastoid cells (GM12878 cell line), a bias towards YY1 docking on the Xi was found, specifically highlighting bindings sites on exon 1 of XIST, which was correlated with higher expression from FANTOM CAGE data (Chen et al., 2016a). Consistent with its ability to bind both DNA and RNA, depleting YY1 has been shown to induce loss of *XIST* localisation from the Xi in female human lung fibroblasts (IMR-90 cell line) (Chapman et al., 2014), mouse embryonic fibroblasts (Jeon and Lee, 2011b), human activated T cells and mouse activated B cells (Wang et al., 2016).

Furthermore, among the proteins that were found to be statistically significant and enriched were proteins not previously described as interactors of *XIST* such as ZCCHC10, WDR89, POLR1B, RBM5 and splicing factors (SNRPB2, SNRPE,

SNRPF, SNRPD1, SNRPD2, SNRPD3, SNRPN). Some splicing factors are known to bind within introns. However, exonic sequences of *XIST* were used here for pulldowns due to the fact that repeats A and E do not span exon-intron regions. Therefore, the splicing factors identified here probably bind exonic regions.

Pulldowns of bovine *XIST* repeat A were performed in both cow and human lysates with the aim to identify which proteins would have preserved their sequences sufficiently to enable cross-species interactions. No common proteins from the high-confidence lists (statistically significant, p-value <0.05 and $log_2FC>1$) were observed (compare **Table 4.3** and **Table 4.4**). This could be because of a lack of optimised input material added in each pulldown reaction and different binding efficiencies for these proteins across species. To decrease detection stringency on account of these reasons, the p-value cut-off was removed, but the $log_2FC>1$ cut-off was kept to ensure the detection of specific proteins. This enabled 41 proteins to be highlighted as common across the three replicates from both datasets, including hnRNPU, HNRNPA2B1, HNRNPA0, SRSF1, BRD2, BRD3, TAF3, PNN, RNPS1, ACIN1 and ZC3H18 (**Table 4.5**). More specifically, 24 of the 41 (~58.5%) common proteins had previously been identified by at least another study in human or mouse (indicated in **Table 4.5**). Pinin (PNN), Acinus (ACIN1) and RNA-binding protein with serine-rich domain 1 (RNPS1) are interactors of the exon-junction complex (EJC) and are involved in preventing exon skipping, ensuring faithful splicing (Boehm et al., 2018). Whilst bromodomain-containing proteins 2 and 3 (BRD2 and BRD3) have not been shown to bind human or mouse *Xist*, BRD4 was shown to bind mouse Xist DNA regions and regulate its transcription upon differentiation (Wu et al., 2015). BRD2 and BRD3 are known transcriptional regulators which bind to acetylated histones and have been identified to interact with mouse *Malat1* from pulldown experiments in a neuroblastoma cell line (NSC-34)(Scherer et al., 2020).

In trying to bridge the observation of a lack of significant overlap across the cow and human datasets, we can focus on hnRNPU. Although direct expression levels cannot be compared across human and cow, the pattern of hnRNPU expression levels both at the transcript and protein level were similar across human and cow (**Figure 2.10** and **2.11**). Therefore, a lack of detection based on a low abundance of hnRNPU can be eliminated as it's one of the most highly expressed proteins in this dataset.

Although, given its strict nuclear localisation, hnRNPU abundance in the bovine stromal lysate could have been decreased due to lysates being whole-cell preparations (**Figure 4.6**). Another possibility for a lack of an overlap in proteins that were statistically significant with a $log_2FC>1$ could be the use of different gene/protein annotation between human and cow. Despite the few proteins listed above having the same nomenclature, one protein was listed in both cow and human datasets albeit under a different name (SKIV2L2 in cow and MTREX in human)(**Table 4.5**), albeit this protein did not pass the statistical significance cut-off. Finally, issues with proteomic detection of (low abundance or other) proteins cannot be ruled out.

However, a lack of statistically significant shared interactors for bovine *XIST* repeat A, need not necessarily mean that an interaction is not conserved from one species to the other or that the *XIST* lncRNA and its partners are not co-evolving. It could also mean that whilst the sequence similarity of *XIST* and of the protein partners appears high enough to substitute for one another across different species, binding motifs could have diverged. Alternatively, there could be additional co-factors needed for efficient binding to occur in one species but not another. Although there is no evidence for such co-factors yet, such an example would be strong evidence for lncRNA-protein co-evolution whereby the protein in one species would have had to adapt to a rapidly evolving *XIST* lncRNA sequence by requiring synergistic binding with other proteins.

Previously cow SPEN, RBM15 and WTAP were not found among the proteins that were pulled down with bovine *XIST* repeat A in bovine stromal cells (**Table 4.3**). Whilst RBM15 was detected in the dataset with 11 unique peptides (albeit not passing the $log_2FC>1$ or p-value < 0.05 cut-offs), there was no signal at all for SPEN and WTAP proteins. In the human dataset here, RBM15 was detected with 10 unique peptides and an average log_2FC score of 0.99 across the three replicates. Additionally, RBM15B, a paralog of RBM15 previously shown to be required for efficient X-linked gene silencing (Patil et al., 2016), was found to be enriched at 1.12-fold on average (with 9 unique peptides), albeit it did not reach statistical significance (Student's t-test, $p<0.05$). Similarly, SPEN was detected with 5 unique peptides and

an average log_2FC score of 0.92 but it did not reach statistical significance (Student's t-test, $p<0.05$). In contrast, WTAP was not detected at all here either.

Collectively, work from Chapter 3 is complemented here, whereby RNA-centric pulldowns demonstrated an association between human *XIST* and the RBM15 protein (albeit concordant WTAP binding could not be robustly shown). The potential for bovine *XIST* to interact with similar protein partners as human *XIST* was validated here by *in vitro* transcription of bovine *XIST* repeat E coupled to pulldown establishing CIZ1 as a shared protein partner across human and cow. This was further consolidated by pulldowns of *in vitro* transcription of bovine *XIST* repeat A in bovine stromal lysates coupled to TMT-MS, highlighting hnRNPU and TOP1, as shared interactors. Among novel, cow-specific partners, TAF3 and PRPF3 were seen. This chapter also presents evidence of the *XIST* RNA showing the ability to pulldown proteins from different species, perhaps highlighting an example of lncRNA-protein co-evolution with potentially promising protein partners of *XIST* critical for XCI function. More specifically, TMT-MS following pulldowns of bovine *XIST* repeat A when mixed in human lysates highlighted human hnRNPC, hnRNPCL1, RALYL, SRSF1, MATR3, YY1 and RBM5 among others, as proteins that are poised for cross-species interactions across human and cow. However, when comparing proteins bound to bovine *XIST* repeat A from cow vs human lysates, there was no overlap between high-confidence lists (p-value <0.05 and log_2FC>1). When only proteins passing the log_2FC>1 cut-off were considered, there was an overlap of 40 proteins, among which were hnRNPU and SRSF1. The next chapter will examine whether differential binding of a small subset of *XIST* protein partners observed across human and cow could be explained by signatures of selective pressure acting on protein sequences. This will also allow for the study of a potential XIST-protein partner co-evolution across human, cow and perhaps other placental mammals.

5. Chapter 5: Examination of selective pressure variation acting on *XIST* protein partners across human, mouse, cow and pig

5.1. Introduction

Following on from the exploration of specific protein partners of mouse *Xist* in chapters 2 and 3, the focus of this chapter is to determine if the selective constraints on these *XIST* protein partners are such that they suggest conserved function in other placental mammals such as cow and pig. In chapters 3 and 4, it was demonstrated that whilst a high degree of amino acid conservation exists across each whole protein (including functional domains such as RNA-binding domains) in mouse, human, cow and pig, these proteins could not be robustly shown to bind cow *XIST*. This could be explained by technical limitations or cell-type specific RNA-protein interactions, or it could be biologically significant. This chapter will assess whether a small number of sites in the protein coding regions of these genes are under positive selective pressure and whether this may have altered the binding profiles/interacting partners of *XIST* in these different mammal species. This will facilitate the prediction of a conserved or divergent *XIST* protein partner interactome across placental mammals. Analysis of compensatory mutations resulting from divergent *XIST* sequence across placental mammals can provide information on co-evolution of *XIST* and its protein partners.

Examining selective pressure variation in homologous protein coding sequences between species allows us to identify regions under different selection regimes, i.e. neutral evolution, purifying or positive selection (described in Section 1.12.1). Identifying signatures of positive selection enables the understanding of which parts of a specific protein coding region of the genome have substitutions which increased the fitness of that individual in its environment. There are also several examples where biochemical studies of positively selected residues illustrate the impact of these residues on protein functional shift (**Section 1.13**. There are several approaches that have been developed for the assessment of selective pressure variation across sites and across lineages (**Section 1.12.1-1.12.3**). Here, codon-based models of evolution that assess selective pressure variation in a likelihood framework are applied. The models used are implemented in CodeML, part of the

PAML programme (Yang, 2007) and the packages employed to streamline the process are implemented in VESPA (Webb et al., 2017) and Vespasian (*Constantinides B, Webb AE, and O'Connell MJ.; manuscript in preparation)*. Here selective pressure variation acting on *XIST* partner protein coding sequences is probed for, which might contribute to differential *XIST* protein partner interactions observed across human and cow observed in Chapters 3 & 4.

Substitutions accumulating on a region of a sequence neighbouring positively or negatively selected sites, could alter the level of constraint enforced on those sites. For instance, mutating glycine at position 38 in rat RNase to aspartate enables it to interact with lysine at position 41, disturbing the active site [discussed in (Fitch and Markowitz, 1970)], whilst mutating serine at position 39 to arginine is non-detrimental (Fay et al., 2002). However, in bovine RNase, the amino acids at positions 38 and 39 are indeed aspartate and arginine, respectively, with no loss of catalytic activity. This can be explained due to the arginine having a positive charge, which offsets aspartate's negative charge, again preventing an interaction with lysine at position 41. To reconstruct the order in which these mutations must have arisen, it was then deduced that arginine at 39 (responsible for charge neutralisation) needed to be present in the genome first, followed by aspartate at 38 (which would have its charged nullified). Following fixation of both residues, the loss of arginine at 39 (responsible for charge neutralisation) to a different amino acid, would not be tolerated. These observations by Wyckoff and others (Fares and Travers, 2006) demonstrated that even neutral substitutions can exercise an effect on a protein coding sequence, i.e. sites that can tolerate substitutions may change as substitutions are fixed. Compensation by fixing substitutions in response to altered constraint is a mode of 'co-evolution' (Pazos and Valencia, 2008).

The vast majority of protein coding regions are evolving under constraint or 'purifying selection'. This involves the purging of substitutions that negatively influence organismal fitness, e.g. a crucial splice site, a structure that affects function or a binding site for an interacting partner (Ngandu et al., 2008). In case of an interacting partner, if the interaction plays a key role in cellular development or homeostasis, it is likely the amino acid sequence of the partner will also remain invariable. This indicates a selective pressure to maintain the interaction, meaning that they both

evolve under purifying selection. For example, a 9 bp region in the *rev response element* within the *env* gene (forming the viral envelope) of HIV-1 coding sequences is evolving under purifying selection (Ngandu et al., 2008). The rev response element is bound by the rev protein and involves the nuclear export of *HIV-1* mRNAs to the cytoplasm. The 9 bp region contains sites where the rev protein binds the *HIV-1* mRNA most strongly. Thus, co-evolution of interacting sequences maintains the function resulting from their interaction.

Another mode of co-evolution is ‘co-adaptation’ (Pazos and Valencia, 2008). In the context of two partners interacting, this would imply that changes in the sequence of one partner would be compensated by changes in the sequence of the other, in order to maintain the interaction. One such example of co-evolution can be seen in the case of human spliceosomal proteins U1A/U2B’’ and their target transcripts U1/U2 snRNAs, which together form small interacting nuclear ribonucleoproteins (snRNPs) (Strange et al., 2016). In Drosophila, there is a single SNF protein that associates with both U1 and U2 snRNAs whereas in human there are two paralogs (U1A/U2B’’), which exhibit >70% sequence conservation across the three proteins in the two species. The Drosophila SNF interacts strongly and specifically with stemloop II of the U1 snRNA (SLII) and with a lower affinity to stemloop IV of the U2 snRNA (SLIV) (Strange et al., 2016). Binding in both cases is mediated by the N-terminal RRM domain of the protein. This interaction is important for *cis*-splicing, although some organisms, including nematodes, can employ *trans*-splicing, whereby nascently generated RNAs are joined to form mRNA, without the need for the SNF-U1 interaction (Strange et al., 2016). Assessing for selective pressure variation across SNF proteins in nematodes and arthropods revealed that although both lineages were evolving under purifying selection, trans-splicing nematodes (ω= 0.04560) were less constrained compared to cis-splicing arthropods (ω= 0.02414) (Strange et al., 2016). An amino acid substitution was found in position 53 of the N-terminal RRM domain of SNF, which represents a phenylalanine in arthropods but a histidine in nematodes. Given the occurrence of the substitution in a domain responsible for RNA binding, it was predicted to interfere with the interaction of SNF with its respective snRNAs. Examination of the SLII part of U1 snRNA revealed a loop of 10 nucleotides in arthropods lacking a clear base preference at position 9. In contrast, an 11-nt loop exists in the nematode SLII region of U1 snRNA, with a

conserved U at position 9 (Strange et al., 2016). An increased U1 snRNA loop size has been predicted to decrease stability of the interaction with SNF which could result in lower affinity. Equally, the SLIV region of nematode U2 snRNA contains a 12 nt loop with no conservation of a specific base at position 2 whereas in arthropods, the loop is 11 nt long with a preference for an A, which improves binding affinity, according to *in vitro* binding assays. Altogether, this study predicted compensatory mutations in the SNF proteins and stem loops of U1/U2 snRNAs in nematodes in order to maintain an interaction (Strange et al., 2016). Considering these changes are likely to result in a reduced affinity of the SNF-U1 interaction, these results correlate well with an increased adoption of a trans-splicing mechanism in nematodes, which would rely less on a SNF-U1 interaction.

The molecular co-operation of RNA and protein is central to gene expression regulation from early development and throughout in life, upholding cellular homeostasis. An interdependence between these two molecules can be traced to the beginning of life, according to the Ribonucleopeptide World hypothesis (Di Giulio, 1997), whereby RNA and protein evolved together, in order to stabilise one another (Frenkel-Pinter et al., 2020). These interactions could result in expanded half-lives, increased likelihood of further synergy with other ligands, and/or catalysis (Frenkel-Pinter et al., 2020). Among previous studies focused on the co-evolution of RNA and interacting proteins are examples from the ribosome (Mallik et al., 2015) and the spliceosome (Strange et al., 2016). However, at the time of writing I am not aware of studies that have been carried out to assess the co-evolution between lncRNAs and their protein partners. Even though there is only a small percentage of lncRNAs that have been shown to be functional, the ones that are functional tend to regulate key processes together with their protein partners and they tend to be involved in major biochemical pathways in eukaryotic cells (Section 1.1). Typically, lncRNAs lack extended conservation across the full length of their sequence, instead displaying short, interspersed regions of high conservation (**Section 1.2**). The potential co-evolution of lncRNAs interactions with partners has consequences for the maintenance of these interactions and by extension the resulting function. For instance, if one protein partner's binding surface or a lncRNA's protein docking site diverge rapidly, in the absence of compensatory mutations, the specificity of the

interaction, and subsequently the function of the lncRNA-protein complex can be lost, as seen for RNA-protein interactions (Strange et al., 2016) and protein-protein interactions (Dey et al., 2012). In the case of a macromolecular assembly such as the ribosome, molecular co-evolution has also been observed across components of the assembly (i.e. proteins that make up the multi-subunit protein ribosomal complex) and interacting partners, whether that involves RNA-protein or protein-protein interactions (Mallik et al., 2015).

The *XIST* lncRNA, together with its protein partners, orchestrate XCI, which is a crucial function for developmental competency of female embryos, indispensable for life (**Section 1.4**). Studies of mouse *Xist* lncRNA demonstrate interactions with at least 81 proteins. If these interactions all occurred in parallel, this would constitute an RNA-protein assembly analogous to the ribosome, if not larger in size (Lu et al., 2020a). A subset of these proteins has been functionally characterised, and have been shown to play key roles in mediating X-linked gene silencing or aiding in *Xist* localisation to the Xi in the mouse, highlighting the biological relevance of their interaction with *Xist* (**Section 1.6.2**). Here, the focus of the investigation was placed on the aforementioned subset of functionally characterised mouse *Xist* partners, which display differential binding to human and cow *XIST* in pulldowns (Chapters 3 and 4). These were split into three categories:
a) proteins which showed binding to human but not cow *XIST*, i.e. SPEN, WTAP and hnRNPK,
b) proteins which bound both human and cow *XIST*, i.e. CIZ1 and hnRNPU,
and, c) proteins that were shown to bind human and mouse *Xist* in other studies but not here, i.e. RBM15, PTBP1, MATR3 and LBR.

Proteins that only bound cow but not human had not been identified by that point, as those experiments happened later (after COVID-19 restrictions were lifted and laboratory access was granted) – essentially mass-spec analysis and cow-specific protein identification happened whilst writing the thesis and 1.5-2 months before submission

Implantation strategies, early pregnancy events, and timing of XCI hallmarks vary across the placental mammals (**Section 1.7**). Thus, XIST-protein partner co-

evolution could have been influenced by unique features pertaining to reproductive morphologies across placental mammals. Reasoning that loss of RNA-protein interaction partners can arise from a functional shift directly linked to amino acid sites under positive selection, it was hypothesised that putative protein partners of *XIST* could have functionally diverged across human, mouse, cow or pig due to perturbations in protein domains involved in RNA-binding or other XCI-related functions. To this end a comparative genetic analysis was carried out across a subset of proteins that are known interacting partners of mouse and human *XIST*. It was assessed whether any of these proteins showed substitution patterns indicative of positive selection in human, mouse, cow and pig and indeed whether these signatures coincide with regions known to interact with *XIST*. With the exception of WTAP (Wu et al., 2016) and SPEN (Carter et al., 2020), such analyses had not been previously performed on these proteins.

5.2. Materials and Methods

Placental mammal genomes included in this work were chosen based on the quality of the genome available and their phylogenetic placement**.** In total 16 placental mammal species were used, one marsupial and one monotreme (**Table 5.1**).

5.2.1. Data assembly

The protein-coding genes selected here were SPEN, WTAP, RBM15, CIZ1, hnRNPK, hnRNPU, PTBP1, MATR3 and LBR, as elaborated earlier (**Section 5.1.2**). Not all proteins described in the Introduction were examined here since the selective pressure variation analysis described here was performed prior to the RNA pulldowns described in the previous Chapter (Chapter 4). Coding DNA sequences (CDSs) for the 9 protein coding genes were retrieved from Ensembl genome browser release 102 (Yates et al., 2020). The genes of interest were listed in the "gene_list.txt" file. To establish there was no hidden paralogy, the Orthologous Matrix (OMA) database (Altenhoff et al., 2018) was queried for every gene to check for the presence of 1:1 orthology. Single gene orthologs were then downloaded from the Ensembl API using the command line shown below. The list of species formed the "species_list.txt" file. The command line shown below, kindly donated by David Orr, (and associated input files) successfully extracted all transcripts of the protein coding gene from Ensembl and placed them into a file called "seqs_out". The source code is available at https://github.com/david0rr/ensembl-pull-seqs

```
python ensemble_seq_run.py -genes gene_lists.txt -species
species_lists.txt -type symbol
```

Code box 1.

The file for each gene was a FASTA formatted nucleotide file, and output from this stage can be found in E-appendix 5.2.

5.2.2. Preparation of orthologous gene sets for selective pressure analyses

In an attempt to semi-automate selective pressure variation analysis whilst reducing the introduction of errors, the already existing VESPA pipeline was used (Webb et al., 2017). The end goal of using VESPA was to generate with files in the appropriate format required to carry out selective pressure variation analysis:

i) a multiple sequence alignment of the amino acid sequence for each gene,
ii) a gene tree describing the relationship between species selected based on the evolution of each assayed gene and
iii) a species tree describing the relationship between species selected based on the current phylogeny of mammals according to the literature

This work was undertaken on ARC3, part of the High Performance Computing (HPC) facilities at the University of Leeds, UK.

After having retrieved coding sequences from ensembl (www.ensembl.org), the first step involved retaining only the longest transcript for each orthologous gene to be used in downstream processing steps. To this end, the following VESPA command was called:

```
vespa.py clean -input=inputname.fa
```

Code box 2.

This generated a file name with the prefix “Cleaned_(input name)”. Output from this file processing step can be found in E-appendix 5.2.2. Then, header names of these FastA files were made compatible for downstream processing steps by trimming the Ensembl gene identifiers.

To convert Latin species names to common species names, the following custom script was run, which scans every line in all files with a “.fa” extension to identify the Latin name and replace it with the common species names.

```
for i in *.fa
do
sed -i 's/homo_sapiens/Human/g' *.fa
sed -i 's/ornithorhynchus_anatinus/Platypus/g' *.fa
sed -i 's/monodelphis_domestica/Opossum/g' *.fa
sed -i 's/dasypus_novemcinctus/Armadillo/g' *.fa
sed -i 's/loxodonta_africana/Elephant/g' *.fa
sed -i 's/ovis_aries/Sheep/g' *.fa
sed -i 's/canis_lupus_familiarisgreatdane/Dog/g' *.fa
sed -i 's/tursiops_truncatus/Dolphin/g' *.fa
sed -i 's/bos_taurus/Cow/g' *.fa
sed -i 's/sus_scrofa/Pig/g' *.fa
sed -i 's/myotis_lucifugus/Microbat/g' *.fa
sed -i 's/equus_caballus/Horse/g' *.fa
sed -i 's/felis_catus/Cat/g' *.fa
sed -i 's/mus_musculus/Mouse/g' *.fa
sed -i 's/oryctolagus_cuniculus/Rabbit/g' *.fa
sed -i 's/macaca_mulatta/Rhesus_macaque/g' *.fa
sed -i 's/pongo_abelii/Bornean_orangutan/g' *.fa
sed -i 's/gorilla_gorilla/Gorilla/g' *.fa
sed -i 's/pan_troglodytes/Chimpanzee/g' *.fa
done
```

Code box 3.

After this conversion, the header of each sequence in the files containing the coding sequences for each gene looked like this:

```
>Human|SPEN
ATGGTCCGGGAAACCAGGCATCTCTGGGTGGGCAACTTACCCGAGAACGTGCGGGAAGAG
```

File output 1. Result from Code box 2. Only the first two lines of the output are shown for human. This output contained several lines in the format shown, with a header listing the species and gene name followed in the next line by the nucleotide sequence of the same gene in different species.

Using the “translate” function in the VESPA package (Webb et al., 2017) all nucleic acid sequences were translated into their corresponding amino acid sequences as follows:

```
vespa.py translate -cleave_terminal=False
-input=Cleaned_inputname.fa
```

Code box 4.

This generated a file with a prefix “Translated_(input name)” which looked like this:

```
>Human|SPEN
MVRETRHLWVGNLPENVREEKIIEHFKRYGRVESVKILPKRGSEGGVAAFVDFVDIKSAQ
```

File output 2. Result from Code box 3. Only the first two lines of the output are shown for human. This output contained several lines in the format shown, with a header listing the species and protein name followed in the next line by the amino acid sequence of the same protein in different species.

Output from this file processing step can be found in E-Appendix 5.2.2. A two-step quality control step was undertaken to ensure the length of each sequence was a multiple of 3 and that no internal stop codons were present (https://web.expasy.org/translate/)(Duvaud et al., 2021).

5.2.3. Multiple sequence alignment of orthologous gene sets

To detect sequence homology across multiple species and create a robust phylogenetic tree to visualise evolutionary relationships between species, amino acid sequences were aligned using multiple sequence alignment tools. The multiple sequence alignment tool MAFFT (Katoh and Standley, 2013) was used to align orthologous protein sequences. MView (Brown et al., 1998) was used to visually inspect alignments. The output files would look like the following (also found in E-appendix 5.2.2.):

```
>Human|SPEN
------------------------------------------------------------
------------------------------------------------------------
-----------------------------------MVRETRHL-----------
---WVGNLPENVREEKIIEHFKRYGRVESVKILPKRGSEGGVAAFVDFVDIKSAQKAHNS
VNKMGDRDLRTDYNEPGTIPSAARGLDDTVSIASRSREVSGFRGGGGGPAYGPPPSLHAR
```

File output 5. Only the first six lines of the output are shown for human. This output contained several lines in the format shown, with a header listing the species and protein name followed in the next line by the multiple sequence alignment of the same protein across all the species used as described by **Table 5.1**.

Table 5.1. Species included in selective pressure variation analyses.

Common Name	Species name	Genome Coverage	Genome Version (Ensembl v102)
Gorilla	*Gorilla gorilla*	x80	gorGor4
Human	*Homo sapiens*	Deep	GRCh38.p13
Pig	*Sus scrofa*	x65	Sscrofa11.1
Rabbit	*Oryctolagus cuniculus*	x7.48	oryCun2
Orangutan	*Pongo pygmaeus*	x6	PPYG2
Mouse	*Mus musculus*	Deep	GRCm38.p6
Microbat	*Myotis lucifugus*	x7	myoLuc2
Macaque	*Macaca mulatta*	Deep	MMUL 10
Horse	*Equus caballus*	x88	Equ Cab 3
Elephant	*Loxodonta africana*	x7	Loxafr3.0
Dolphin	*Tursiops truncatus*	x2.59	turTru1
Cow	*Bos taurus*	x80	ARS-UCD1.2
Dog	*Canis familiaris*	50x	UMICH_Zoey_3.1 great dane
Chimpanzee	*Pan troglodytes*	55x	Pan_tro_3
Cat	*Felis catus*	x72	Felis_Catus_9
Armadillo	*Dasypus novemcinctus*	x6	dasNov3
Platypus	*Ornithorhynchus anatinus*	x58.8	mOrnAna1.p.v1
Opossum	*Monodelphis domestica*	N/A	ASM229v1

A permutation tail probability test (yaptp) from Clann software package (Creevey and McInerney, 2005) was performed for each alignment to assess whether the alignment contained signal better than random using the following command:

```
clann -ln -c genetree_inputname.txt

yaptp

quit
```

Code box 5.

5.2.4. Assessing phylogenetic signal contained in multiple sequence alignments via quartet puzzling

To determine whether the alignments contained phylogenetic signal, maximum likelihood mapping by quartet puzzling was performed (Strimmer and von Haeseler, 1996). Quartet puzzling breaks down each dataset into all possible sets of four sequences (or quartets), and estimates the maximum likelihood of each quartet belonging to one of three possible unrooted phylogenies (Strimmer and von Haeseler, 1996). The maximum likelihood values can be mapped onto an equilateral triangle as dots, where each corner of the triangle indicates a potential resolved phylogeny (Strimmer and von Haeseler, 1997). The position of each dot on the triangle can therefore inform us about support for distinct phylogenies, e.g. **Figure 5.1**. If maximum likelihood values cluster predominantly at corners of the triangle, then a specific phylogeny can be inferred to have more support over another (**Figure 5.1A**). However, it is not always the case that a single phylogeny is favoured. Short or very similar sequences could make it difficult to distinguish clear support for a definite phylogeny, leaving support values either in the edges or the middle of the triangle (**Figure 5.1B**). Since support values found at corners of the triangle represent well-defined phylogenies, if there was equal probability for each of the three outcomes for a quartet, such a probability would appear in the middle of the triangle (**Figure 5.1C**). This would be indicative of an unresolved phylogeny. A strict cut-off is set for these analyses, i.e. if the cumulative probability percentage of partly resolved and unresolved regions is higher than 10%, the data are inferred to have

insufficient phylogenetic signal. In these cases, the gene level alignment does not have adequate phylogenetic signal to reconstruct a robust phylogeny and a species level phylogeny (made from many different genes) may need to be considered. To proceed with maximum likelihood mapping, multiple sequence alignments were converted from fasta to phylip interleaved format via EMBOSS seqret (https://www.ebi.ac.uk/Tools/sfc/emboss_seqret/)(Madeira et al., 2019). Approximate maximum likelihood mapping was performed by using TREE-PUZZLE (http://www.tree-puzzle.de/; (Schmidt et al., 2002) using a neighbour-joining tree parameter estimation, JTT model of substitution and a uniform rate of heterogeneity.

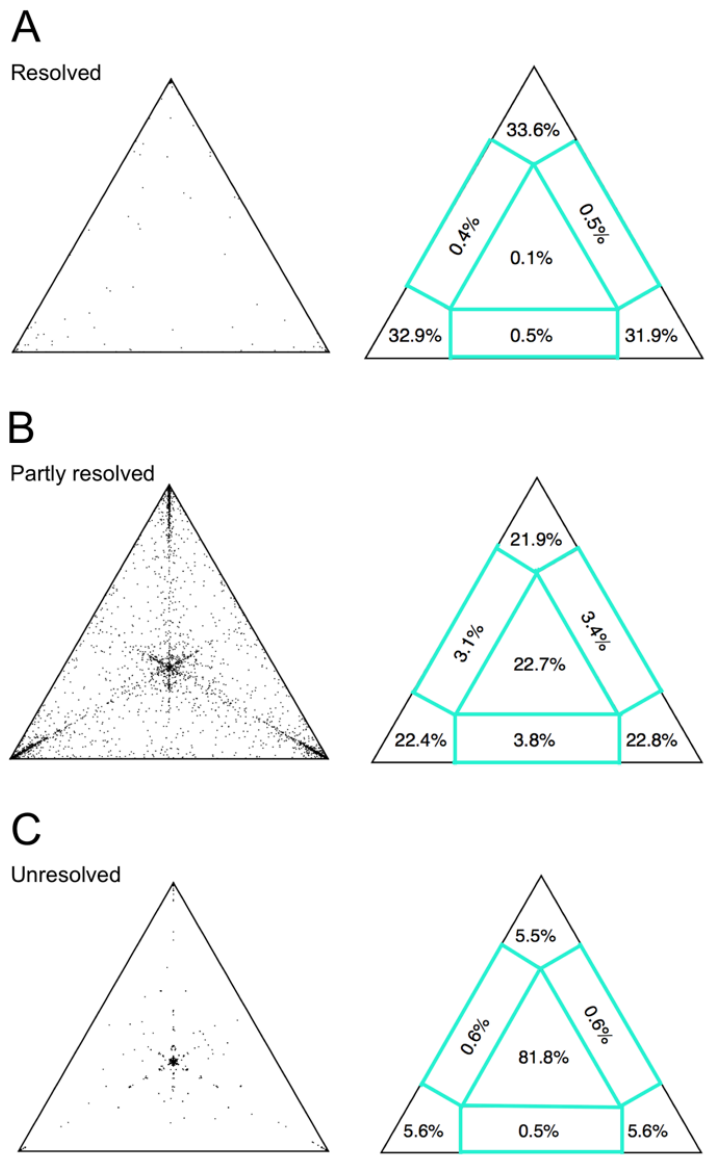


Figure 5.1. Representative images of Quartet Puzzling approach to identifying basins of attraction demonstrating the degree of phylogenetic signal contained in alignment files

Dots indicate the maximum likelihood for three possible trees estimated from four sequences in an equilateral triangle (*left*). Dots falling in defined areas of the triangle are counted and shown as a percentage of the total number of dots (*right*). **A**) In case most dots fall within corners, there is support for a specific phylogeny by the maximum likelihood values. Example given here is the SPEN gene. **B**) When, however, more than 10% of dots are found outside of corners (in the edges and centre of triangle marked in teal), there is a lack of clear distinction between phylogenies. **C**) If all three phylogeny combinations are equally likely, dots fall within the middle area of the triangle (also known as a “star” or completely unresolved phylogeny).

5.2.5. Reconstructing gene trees from multiple sequence alignments and assessment of the resulting trees

To examine whether the evolutionary relationships between species could be recapitulated based on amino acid sequence alignments of single gene orthologs from 18 species, IQ-TREE (Trifinopoulos et al., 2016) was used to generate gene trees for each amino acid alignment based on maximum likelihood. Default settings were used in each case which include the use of ModelFinder (Kalyaanamoorthy et al., 2017) to detect the substitution model of best-fit for each dataset. Ultrafast bootstrapping was performed with a maximum of 1,000 iterations (for gene tree outputs and associated bootstrap values see E-appendix 5.2.3). IQ-TREE automatically performed tree topology evaluation using the approximately unbiased test of phylogenetic tree selection (Shimodaira, 2002) with 1,000 replicates. This test assigns confidence values to trees tested which reflect the probability of each tree being the true tree. The approximately unbiased test is preferred over other tests because it controls for i) selection bias, i.e. overestimating the likelihood value when many comparisons are made, and ii) false positives, i.e. having a probability of erroneously rejecting the null hypothesis greater than 5% (Shimodaira, 2002). The test works by implementing several rounds of comparisons, with each round changing the length of amino acid sequences and checking for how many of those variations there is support for the hypothesis. Based on this count, confidence in each tree is provided by values between 0 and 1, whereby the closer a value is to 1, the more likely it represents the true tree (Shimodaira, 2002). Gene trees were visualised using the interactive tree of life website, iTOL (Letunic and Bork, 2019).

Additionally, a species tree in Newick format was constructed based on the current phylogeny of mammals according to the literature (Jebb et al., 2020) (**Figure 1.5**)(newick formatted tree in E-appendix 5.2.3):

To examine how similar the two trees (i.e. the gene tree and known species tree) were, Clann (Creevey and McInerney, 2005) was used to estimate the Robinson-Foulds distance score (Robinson and Foulds, 1981). The RF distance serves as a proxy for similarity where the smaller the Robinson-Foulds distance, the more similar two trees would be.

```
clann -ln -c tree_inputname.txt

rfdists output=vector

quit
```

Code box 6.

5.2.6.1. Selective pressure variation analysis in VESPA and Vespasian

To carry out a selective pressure analysis, three files are required: (1) the alignment in nucleotide (codon) format, (2) the corresponding tree that describes how the taxa in that gene family are related to one another (either a gene or species tree depending on your analysis), and (3) a control file that describes the precise settings that code for the appropriate model of evolution to be assessed.

5.2.6.2. Preparing the nucleotide alignments

To investigate selective pressure variation across protein-coding genes the standard assessment of dN/dS (or ω) was used. The amino acid sequence alignments were used to guide the alignment of the original nucleotide sequence files using the following commands in VESPA:

```
vespa.py create_database -input=Cleaned_inputname.fa

mkdir Map_Gaps_Translated_Cleaned_inputname

scp -r Translated_Cleaned_inputname.fa.mft
Map_Gaps_Translated_Cleaned_inputname/Translated_Cleaned_inputname
.fa.mft

vespa.py map_alignments -
input=Translated_Cleaned_inputname.fa.mft -database=database.fas
```

Code box 7.

5.2.6.3. Preparing the gene tree

To ‘prune’ branches of the species tree containing species absent from the gene tree, the vespasian software was used (https://github.com/bede/vespasian).

```
vespasian infer-gene-trees data
species_tree.txt
```

Code box 8.

where “data” is a directory that contains a nucleotide alignment with a **.fasta** suffix.

It is well known that for likelihood-based estimations of selective pressure variation more good quality data provides more robust results (Anisimova et al., 2002). Previous assessments of the sensitivity of these approaches indicated that more than six species are required to the power to detect sites under positive selection (Anisimova et al., 2002).

5.2.6.3. Setting up the files for the appropriate models of evolution

Two categories of selective pressure variation models were considered:
(i) Site-specific models enable the investigation of sites that have been under various selective pressures across all species regardless of lineage. And (ii) lineage-specific models assess selective pressure variation across sites and specific branches (i.e. lineages) in the phylogeny as compared to other branches. Here, besides testing for site-specific models of evolution, lineage-specific models of evolution were considered for human, mouse, cow and pig. To specify all the codon-based models of evolution to be tested, vespasian (https://github.com/bede/vespasian) (Webb et al., 2017) was used to run CodeML (part of the PAML package):

```
vespasian codeml-setup data gene-trees --branches
label_table.txt
```

Code box 9.

where “data” is a directory that contains a nucleotide gap alignment with a **.fasta** suffix and “label_table.txt” a .txt file that specifies which species to be considered as foreground for lineage-specific models of evolution. CodeML data structures generated here can be found in E-appendix 5.2.4.

CodeML models were tested by using the following command with vespasian:

```
cd codeml && snakemake --jobs 300 --cluster "qsub -cwd
-V -l h_rt=48:00:00" --max-status-checks-per-second
0.1
```

Code box 10.

To differentiate which models of evolution best fit the data, likelihood ratio tests (LRTs) have been shown to be better powered to detect sites under positive selection when seven species were used per tree (Anisimova and Yang, 2007). Hence, 18 species were used in this analysis. To determine the model of best fit to the data from all models tested, log-likelihood values (lnL) were generated for all models. By performing LRTs, statistical significance was estimated for appropriate comparisons of the site-specific and lineage-specific models (shown in **Table 5.2**).

Table 5.2. Likelihood Ratio Test (LRT) Calculations.

Models eligible for comparison are displayed with the associated degrees of freedom (df) and chi-squared thresholds required to reach significance. χ^2 values determine whether the alternative hypothesis is true based on random chance <5% of the time (unless otherwise stated). Adapted from (Morgan et al., 2012).

Comparison	df	Δl	Critical χ2 values
M0 v M3k2	2	X2	≥ 5.99
M3Dk2 v M3Dk3	-	X1	≥ 1.00
M1a v M2a	2	X2	≥ 5.99
M7 v M8	2	X2	≥ 5.99
M8 v M8a	1	X2	≥ 2.71 (@5%) ≥ 5.41 (@1%)
M1a v Model A	2	X2	≥ 5.99
Model A v Model A null	1	X2	≥ 3.84 (@5%)

Likelihood ratio tests (LRT) and summary tables of models tested and associated confidence values were generated with vespasian by using the following command:

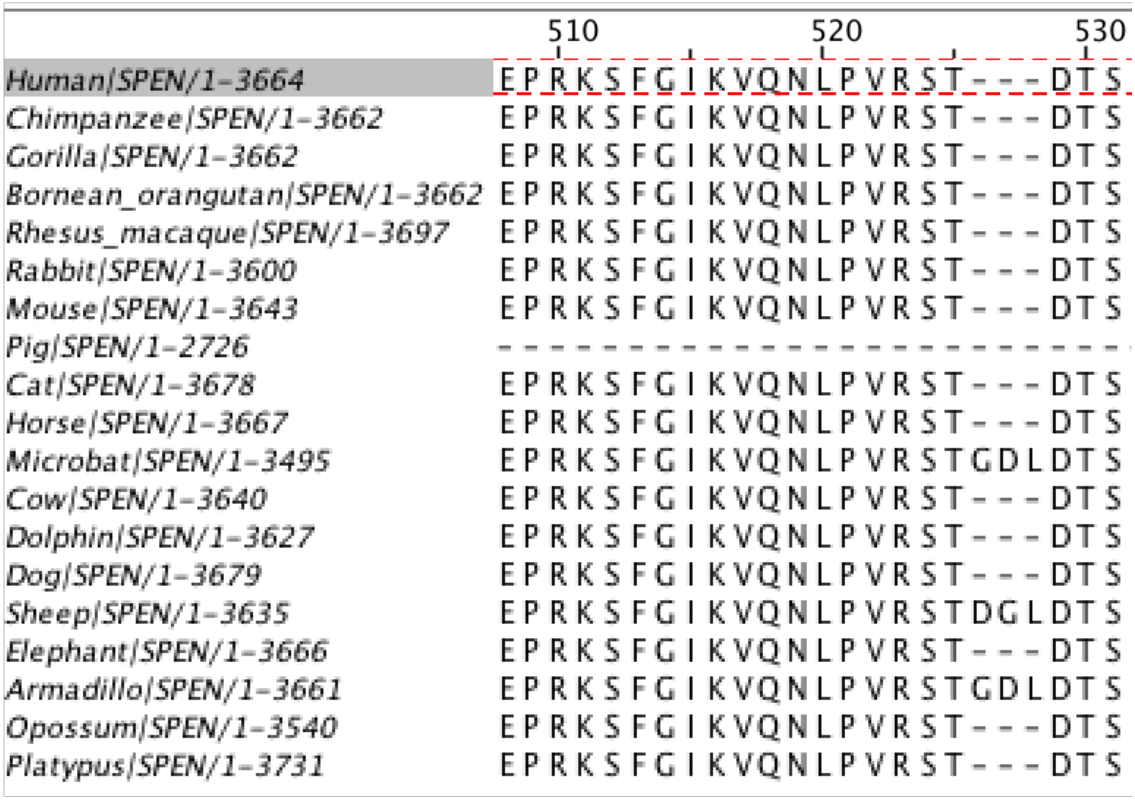
```
cd .. && vespasian report --progress codeml --hide -p
```

Code box 11.

In all cases where models allowed for the estimation of positively selected sites, the contributions of individual sites to that signal for positive selection was estimated using a Bayes Empirical Bayes (BEB) approach. The Naive Empirical Bayes (NEB) estimation of posterior probability does not account for sampling errors which can exponentially affect large data sets (Anisimova et al., 2002) and therefore BEB scores are preferred. BEB estimates for each site on each gene were reported if the posterior probability was greater than 0.5. To obtain a set of identified sites with higher confidence, two user-defined posterior probability (PP) cut-offs were set at PP=0.95 and PP=0.99.

To examine in more detail what potential impact the positively selected sites may have on the function/s of the protein we employed UniProt annotations. In cases where there was evidence of lineage-specific positive selection on a protein with poor UniProt annotation, the location of positively selected sites was mapped onto the respective amino acid sequence from humans in Uniprot (The UniProt, 2021). The human ortholog was chosen due to a superior annotation of protein functional domains in UniProt. It is worth noting that the positions of residues predicted by CodeML are based on the alignment file. Thus, to determine the position of those sites on the human ortholog, gaps inserted during the alignment process must be factored in. This was done by initially counting the number of gaps in the amino acid multiple sequence alignment for the human gene and subtracting that from the predicted position of the positively selected sites (for a graphical representation, see **Figure 5.3**). To check whether predicted sites overlapped any functional protein domains, UniProt was scanned for existing information on protein domains. If available, these were plotted on top of the location of positively selected site predictions. In case protein domain information was not available from UniProt, InterPro (Blum et al., 2021) was used to computationally predict functional domains on the human genes.

A



B

CodeML site	gaps_preceding	actual_position_human
530	-180	350

Figure 5.2. Mapping CodeML sites on the human protein ortholog.

A) Amino acid alignment of the SPEN protein, **B**) Position of amino acid that would be predicted to be under positive selection (left), number of gaps preceding the amino acid reported by CodeML (middle) and actual position of residue on the human SPEN amino acid sequence (right). Amino acid positions for sites predicted to be under positive selection by CodeML are reported based on their position in the alignment file. During the alignment process gaps are inserted (denoted by a dash shown in **A**). Accounting for the number of gaps inserted during the alignment allows for an accurate identification of positively selected sites on select species, e.g. CodeML site 530 refers to the amino acid at position 350 in the human SPEN protein sequence. Note this example is not based on actual data for SPEN and only serves illustrative purposes.

5.3 Results

To study selective pressure variation on putative protein partners of *XIST*, candidate genes were selected based on the propensity of mouse *Xist* to interact with mouse orthologs of these proteins and play roles in *Xist* processing, X chromosome localisation or X-linked gene silencing. Hence, SPEN, CIZ1, RBM15, WTAP, LBR, hnRNPK, hnRNPU, MATR3 and PTBP1 were chosen for this analysis. These candidates have been described in the main introduction (**Section 1.6.2**).

5.3.1. Confidence in multiple sequence alignments containing non-random signal

Random data produce a normal distribution with a skewness score of 0, therefore the skewness of the distribution of probability scores assigned to alignments can be informative. The results of the yaptp test performed on each alignment showed that most genes tested display a negative skewness ranging from -3.1 (for SPEN) to -0.05 (hnRNPK) (**Table 5.3**), arguing for non-random signal. Therefore, the alignments are non-random and contain structure that indicates they are most likely homologous sequences. Another way to assess signal is a qualitative assessment by visual inspection of multiple sequence alignment files, in all cases multiple regions of conservation were observed indicating non-random data (i.e. alignment of homologous regions was evident from visual inspection) (E-appendix 5.2.2).

Table 5.3. Summary table of permutation tail probability tests.

Gene trees created were imported to Clann (Creevey and McInerney, 2005) and the 'yaptp' command was used with default settings. These included equiprobably randomization of 100 repetitions where taxa from branches were unlabeled and shuffled based on sub tree pruning and re-grafting. Any deviation from the normal distribution (skewness of 0) indicates non-randomness of data. SD, standard deviation.

Gene	Mean	Variance	SD	**Skewness**
SPEN	4.4086	0.3041	0.55146	**-3.1545**
WTAP	4.3668	0.3262	0.57119	**-1.4945**
CIZ1	4.3701	0.3451	0.5875	**-1.0778**
LBR	4.0931	0.2872	0.5359	**-0.8398**
RBM15	4.371	0.23626	0.486	**-0.462**
PTBP1	4.0919	0.2349	0.4847	**-0.4336**
hnRNPU	4.3615	0.2238	0.473	**-0.1534**
hnRNPK	4.0838	0.1577	0.3971	**-0.0568**
MATR3	4.3564	0.2412	0.491155	**0.0261**

5.3.2. Insufficient phylogenetic signal in multiple sequence alignments to establish phylogeny

To quantify the amount of phylogenetic signal present in the multiple sequence alignments, we performed quartet puzzling. In this way it was determined whether sequences in alignments are too closely related or too divergent to reliably reconstruct a phylogeny. Most genes used here lack support for a resolved phylogeny (**Figure 5.4**). SPEN is the only gene with sufficient phylogenetic signal, given a cumulative percentage of unresolved quartets less than 10% (**Figure 5.4A**). CIZ1, WTAP and RBM15 were found just above the 10% cut-off threshold (**Figure 5.4B-D**). Consistently, PTBP1, hnRNPU, MATR3 were found to have between 19-30% of cumulatively unresolved quartets, arguing against an adequate quantity of phylogenetic signal (**Figure 5.4E-G**). LBR exhibited a cumulative probability of partly resolved and unresolved as high as 33% (**Figure 5.4H**). hnRNPK had most quartets featuring an equal likelihood for all phylogenies, characterized by 83.5% of quartets accumulating in the centre of the triangle (**Figure 5.4I**). This is indicative of insufficient phylogenetic signal in most genes to infer evolutionary relationships between species, except for the SPEN gene. Often genes with strong selective constraints and low mutational rates do not have good levels of phylogenetic signal (Blomberg and Garland Jr, 2002, Lanier et al., 2014).

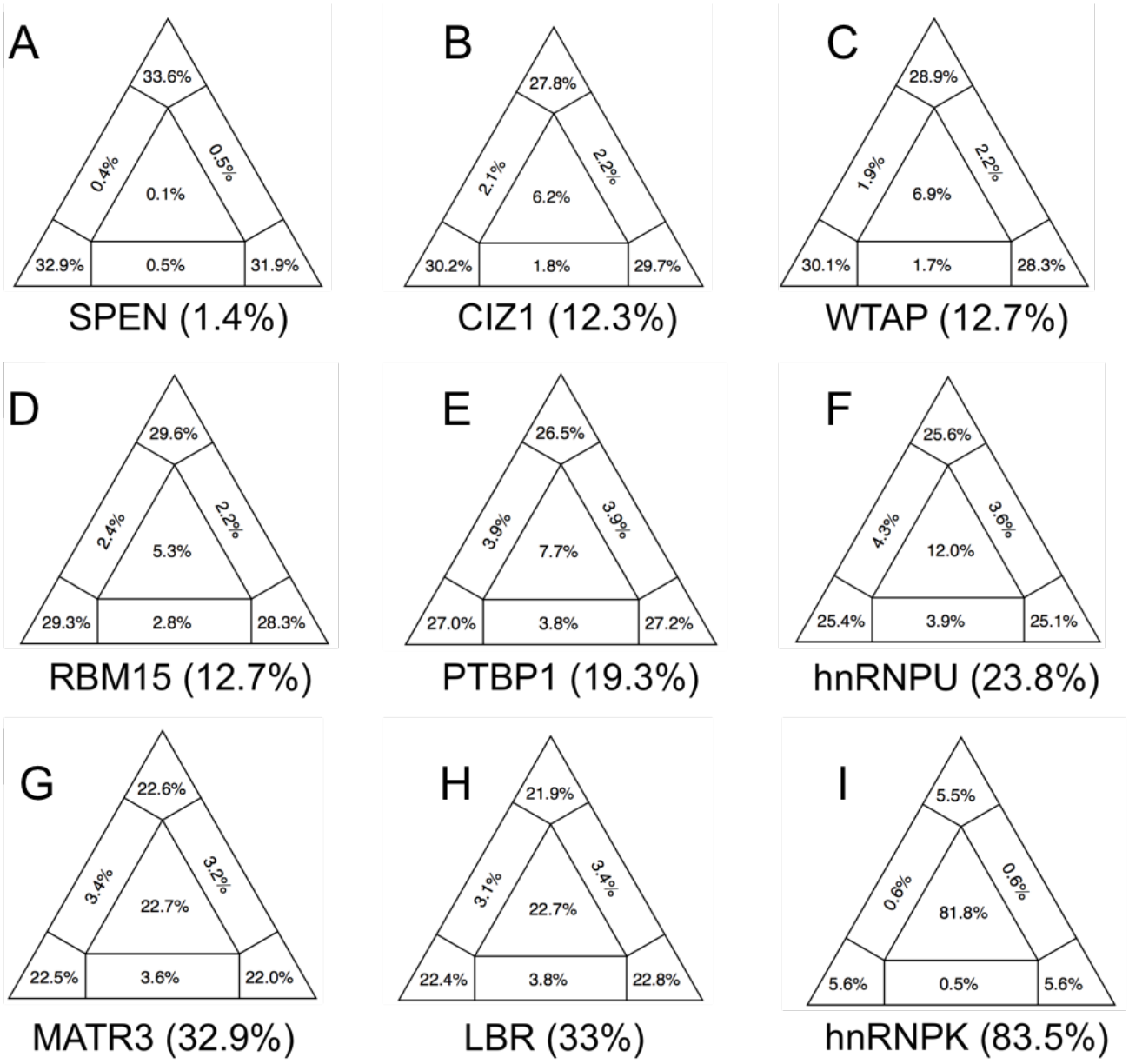


Figure 5.3. Estimation of phylogenetic signal contained in amino acid alignments.

Likelihood mapping and quartet puzzling for **A**) SPEN, **B**) CIZ1, **C**) WTAP, **D**) RBM15, **E**) PTBP1, **F**) hnRNPU, **G**) MATR3, **H**) LBR and **I**) hnRNPK. Numbers in brackets denote cumulative percentages of quartets from areas in the middle and edges of the triangles. Likelihood mapping and basins of attraction were created based on amino acid alignments using TREE-PUZZLE (Schmidt et al., 2002) with default settings. Namely, approximate maximum likelihood was used with all possible quartets (3,876), the neighbor-joining tree parameter, JTT substitution model and a uniform rate of heterogeneity. Amino acid frequencies were estimated from the data.

5.3.3. Incongruence between gene trees and placental mammal species tree

Gene trees were generated in IQ-TREE (5.2.3 and E-appendix 5.2.2). As described in the main introduction to this thesis, gene and species trees for the same set of taxa are distinct, have different applications and are not always in agreement (see Section 1.1.2). Confidence in trees generated, reflecting the probability of trees being true, was obtained using the AU two tree test. Confidence in each tree is provided by values between 0 and 1, whereby the closer a value is to 1, the more likely it represents the true tree (Shimodaira, 2002). For genes examined here, a gene tree close to the species tree could be obtained with confidence scores over 0.9 (**Table 5.4**). Confidence in the tree for WTAP was just below 0.9 (~0.88) whereas trees for SPEN and hnRNPK were not as reliable (0.62 and 0.37, respectively).

Table 5.4. Summary table of AU test performed on gene versus species trees.

Gene trees were obtained from IQ-TREE and confidence values from the AU test (Trifinopoulos et al., 2016).

AU test results

Gene	Gene tree confidence values (larger values represent higher confidence)
hnRNPU	1
PTBP1	0.996
RBM15	0.0987
CIZ1	0.968
LBR	0.935
MATR3	0.931
WTAP	0.885
SPEN	0.623
hnRNPK	0.374

As previously described (Section 1.8), when gene trees differ from species tree, it can lead to systematic biases in phylogeny reconstruction and generate incorrect relationships between species. To examine the congruence between gene tree and species tree for each gene, the RF distance was calculated. The gene trees for the SPEN and CIZ1 genes appeared to be similar to the species tree with RF values of 0.125 and 0.25, respectively (**Table 5.5**). A medium degree of incongruence between the gene and species trees was observed for LBR, WTAP and RBM15, with RF scores between 0.4 and 0.5. In contrast, gene trees for PTBP1, hnRNPU, MATR3 and hnRNPK were highly incongruent with the species tree, resulting in RF scores greater than 0.7. Therefore, with the exception of SPEN and CIZ1 genes, gene trees for remaining genes did not match the species tree. This observation suggests that the specific genes assayed here (with the exception of SPEN and CIZ1 genes) have evolved differently to the way the species have evolved. Because of that, the species phylogeny was imposed before proceeding with selective pressure variation. Any species not present in the amino acid alignments for each gene were removed from the species tree (tree ‘pruning’). Essentially, this makes both gene tree and species tree have the same set of species and the same topology (‘forcing’ a species phylogeny).

Table 5.5. Summary table of phylogenetic distance between gene trees and species tree.

Each gene phylogenetic tree was compared to a species tree containing the same taxa. The Robinson-Foulds distance was inferred from those comparisons using Clann (Creevey and McInerney, 2005) with default settings. Note that the Robinson-Foulds distance serves as a proxy for the similarity of the gene tree to the species tree, hence, the smaller the distance, the more similar the trees.

Gene	Robinson-Foulds distance (smaller is better)
SPEN	0.125
CIZ1	0.25
LBR	0.4
WTAP	0.4375
RBM15	0.5
PTBP1	0.733
hnRNPU	0.75
MATR3	0.8125
hnRNPK	0.933

5.3.3. Assessing evolutionary forces underlying selected genes

Results presented so far argue that multiple alignments for genes selected show signal that is better than random. The phylogenetic signal contained in these alignments however is not high enough to allow for confident inferences of the relationships of placental mammals based on these genes. This was reflected by high RF scores, indicating incongruence between gene and species trees. Because the current placental mammal phylogeny is known, the species phylogeny was imposed before proceeding with selective pressure variation.

To examine the nature of selective pressure acting on a set of putative *XIST* protein partners, both site- and lineage-specific models of evolution were tested per gene using the VESPA wrapper for the CodeML program (PAML package)(Yang, 2007). In order to probe for lineage-specific positive selection of each gene, analyses focused on the human, mouse, cow and pig lineages (used as foreground) for each of the 9 interacting partners of *XIST*.

As an example, site-specific models of selection tested for the CIZ1 gene have been summarised in **Table 5.6** (models for all genes can be found in E-appendix 5.2.5). According to model M0, the average ratio of dN/dS across all amino acid sites in all species denotes that CIZ1 is under purifying selection with an ω value of ~0.28. From M1 it could be inferred that ~64.5% of all amino acid sites evolve under very strong purifying selection with ω=0 (fixed), whereas ~35.5% of all sites evolve neutrally. The M2 model posits that 64% of all sites evolve under strong purifying selection of ω=0.12, ~35% of all sites evolve neutrally and 0.0067% of all sites evolve under very strong positive selection with ω>>1 (ω_2=5.4). Namely, there were nine sites predicted to be evolving under positive selection (PP>0.5), 2 with PP>0.95 and 0 sites with PP>0.99. Model M3K2 denotes 58% of all sites evolve under strong purifying selection of ω=0.095, whereas ~42% of all sites evolve under purifying selection with ω=0.73. Results from M3K3 indicate ~21% of all sites evolve under strong purifying selection of ω =0.014, 49% of sites evolve under purifying selection of ω=0.19 and ~30% of sites are evolving neutrally with ω=0.92. According to M7, all sites fall within the distribution dictated by p and q with ω estimates between 0 and 1 (purifying selection and near neutral evolution, p=0.52812 q=0.98596). p and q

symbolise the beta distribution shape parameters, essentially allowing the shape of the distribution to vary depending on values given, which make for a flexible and effective model. In M8, ~90% of sites were estimated to fall within the distribution dictated by p and q (p=0.52812 q=0.98596) and ~10% of the sites are found under positive selection with ω=1.4. From M8a, it would be inferred that ~79% of sites fall within the distribution dictated by p and q (p=0.83985, q=3.35200) and ~21% of the sites are evolving neutrally.

Table 5.6. Summary of site-specific selective pressure models tested for CIZ1.

The number of sites predicted to be under positive selection are shown in the last column, separated by posterior probability cutoffs. Negative/purifying selection (ω=0-0.9), neutral evolution (ω=0.9-1.1) and positive selection (ω>1.1) (Zhang et al., 2005). P = number of free parameters estimated in the model; lnL = natural log likelihood; p_n = proportion of sites with a particular ω value; ω or Dn/Ds= the ratio of non-synonymous substitutions per nonsynonymous sites (Dn) compared to the number of synonymous substitutions per synonymous site (Ds). BEB = Bayes Empirical Bayes estimations.

Site specific analysis

Model	P	lnL	Estimates of parameters	Positively selected sites (BEB)
M0: One Ratio	1	-72412	ω = 0.27745	Not allowed
M1: Neutral	1	-71210	p0= 0.81, p1=0.1897, ω0=0.08479, ω1=1	Not allowed
M2: Selection	3	-71208	p0= 0.81, p1=0.1895, p2=0.00036, ω0=0.08486, ω1=1, ω2=5.089	9>0.5,2>0.95,0>0.99
M3: Discrete (K=2)	3	-71048	p0=0.71, p1=0.2897, ω0=0.05096, ω1=0.5682	None
M3: Discrete (K=3)	5	-70997	p0=0.572, p1=0.317, p2=0.11, ω0=0.0284, ω1=0.288, ω2=0.878	None
M7: Beta	2	-71010	p=0.335, q=1.303	Not allowed
M8: Beta&Omega>1	4	-70995	p0=0.942, p=0.422, q=2.244, p1=0.0577, ω=1.027	18>0.5,2>0.95,0>0.99
M8a: Beta&Omega=1	4	-70995	p0=0.937, p=0.428, q=2.327, p1=0.062, ω=1	Not allowed

To determine which of the site-specific models of selection tested above best fit the data, eligible models were compared using likelihood ratio tests (LRTs). LRT results are summarised in **Table 5.7** (E-appendix 5.2.5 for LRTs on all genes). Based on the lnL values seen for site-specific models of CIZ1 in **Table 5.6** and LRT results for these models in **Table 5.7**, model M8 was the model of best fit for the CIZ1 gene. This would mean that ~90% of sites were estimated to fall within the distribution dictated by p and q (p=0.52812 q=0.98596), with the majority of sites being highly conserved. The remaining ~10% of amino acid sites were found to be evolving with an ω ratio of 1.027, indicating positive selection. More specifically, this corresponded to a total of 18 sites with a PP greater than 0.5 (also indicated as 18>0.5 in **Table 5.6**). Of those 18 sites, two exhibited a PP greater than 0.95, but none were above 0.99

Table 5.7. Summary of LRT tests for the determination of site-specific model of best fit for CIZ1.

Tree	Null_model	Alt_model	Null_lnl	Alt_lnl	LRT	p-value	Critical_value	Null_rejected
CIZ1_nuc_gap	m0	m3Discrtk2	-21696.715	-21267.181	859.06733	2.86E-187	5.991	TRUE
CIZ1_nuc_gap	m1Neutral	m2Selection	-21281.599	-21273.965	15.269372	0.00048339	5.991	TRUE
CIZ1_nuc_gap	m3Discrtk2	m3Discrtk3	-21267.181	-21230.911	36.270632	-	1	TRUE
CIZ1_nuc_gap	m7	m8	-21245.019	-21228.128	33.78203	4.62E-08	5.991	TRUE
CIZ1_nuc_gap	m8a	m8	-21230.642	-21228.128	5.028764	0.0809129	2.706	TRUE

As an example, lineage-specific models of selection tested for the CIZ1 gene have been summarised in **Table 5.8** (models for all genes can be found in E-appendix 5.2.5). Selective pressure variation on the CIZ1 gene was assessed using two lineage-specific models for each of the four lineages tested (human, mouse, cow and pig). When assessing selective pressure variation on the CIZ1 gene using lineage model A with the mouse lineage as foreground, it was calculated that ~60.5% of all sites evolve under strong purifying selection of ω=0.11, whereas ~32% of all sites evolve neutrally. Furthermore, 2.5% of all sites in the mouse lineage were estimated to evolve under positive selection (ω_2=4.7). Using lineage model A null, it was calculated that ~55% of all sites evolve under strong purifying selection of ω=0.11, whereas ~30% of all sites evolve neutrally. Moreover, ~10.5% of all sites in the mouse lineage were found to evolve under neutral selection. Besides the mouse lineage, the human, cow and pig lineages were also investigated with the same method (**Table 5.8**). To determine which of the lineage-specific models of selection tested above best fit the data, the two models were compared using LRTs. LRT results are summarised in **Table 5.9** (E-appendix 5.2.5). Based on the lnL values seen for lineage-specific models of CIZ1 in **Table 5.8** and LRT results for these models in **Table 5.9**, lineage model A was the model of best fit for the CIZ1 gene in the mouse lineage. This would mean that ~2.5% (p_2) of all the sites in the mouse protein (foreground) were predicted to be under positive selection with an ω ratio of >>1 (ω_2=4.7). Namely, there were 25 amino acid sites with a PP greater than 0.5, two with a PP greater than 0.95, but none above 0.99 (**Table 5.8**). Moreover, there was no evidence of positive selection in the human, cow or pig lineages tested since lineage model A null was the model of best fit for CIZ1 in these lineages. Overall, there was only evidence for positive selection acting on CIZ1 in the mouse lineage across the four lineages tested. To assess if positively selected sites on the CIZ1 protein alignment overlap any known (or predicted) functional domains, CodeML-predicted amino acid sites were mapped on the amino acid sequence of mouse Ciz1. Four out of the 25 amino acids predicted to be under positive selection (702N, 703P, 704S and 713R) spanned the zinc-finger domain of mouse Ciz1 (702-733 aa in mouse Ciz1; **Figure 5.4)**.Zinc-finger domains have been shown to have RNA binding capacity (Klug, 1999, Brown, 2005, Hall, 2005).

Table 5.8. Summary of lineage-specific selective pressure models tested for CIZ1.

The number of sites predicted to be under positive selection are shown in the last column, separated by posterior probability cutoffs. Negative/purifying selection (ω=0-0.9), neutral evolution (ω=0.9-1.1) and positive selection (ω>1.1) (Zhang et al., 2005). P, number of free parameters estimated in the model; lnL, log-likelihood values; p_n proportion of sites under a particular ω value; ω, the ratio of non-synonymous to synonymous nucleotide substitution (dN/dS). BEB, Bayes Empirical Bayes estimations.

Lineage-specific analyses	P	lnL	Estimates of parameters	Positively selected sites (BEB)
Human				
modelA	7	-21282	p0=0.64457 p1=0.35543 p2=0.0 p3=0.0 ω0=0.12072 ω1=1.0 ω2=1.0	None
modelAnull	7	-21282	p0=0.64457 p1=0.35543 p2=0.0 p3=0.0 ω0=0.12072 ω1=1.0 ω2=1.0	None
Mouse				
modelA	7	-21262	p0=0.60542 p1=0.3226 p2=0.04696 p3=0.02502 ω0=0.11444 ω1=1.0 ω2=4.72081	25>0.5, 2>0.95, 0>0.99
modelAnull	7	-21266	p0=0.54575 p1=0.29239 p2=0.1054 p3=0.05646 ω0=0.11103 ω1=1.0 ω2=1.0	None
Cow				
modelA	7	-21280	p0=0.64022 p1=0.35284 p2=0.00447 p3=0.00247 ω0=0.11993 ω1=1.0 ω2=24.3891	1>0.5, 0 >0.95, 0>0.99
modelAnull	7	-21281	p0=0.62061 p1=0.34203 p2=0.02409 p3=0.01328 ω0=0.12029 ω1=1.0 ω2=1.0	None
Pig				
modelA	7	-21282	p0=0.64457 p1=0.35543 p2=0.0 p3=0.0 ω0=0.12072 ω1=1.0 ω2=1.0	None
modelAnull	7	-21282	p0=0.64457 p1=0.35543 p2=0.0 p3=0.0 ω0=0.12072 ω1=1.0 ω2=1.0	None

Table 5.9. Summary of LRT tests for the determination of lineage-specific model of best fit for CIZ1.

Tree	Null_model	Alt_model	Null_lnl	Alt_lnl	LRT	p-value	Critical_value	Null_rejected
CIZ1_nuc_gap_Cow	m1Neutral	modelA	-21281.599	-21279.994	3.210782	0.20081102	5.991	FALSE
CIZ1_nuc_gap_Cow	modelAnull	modelA	-21281.451	-21279.994	2.9146	0.23286416	3.841	FALSE
CIZ1_nuc_gap_Human	m1Neutral	modelA	-21281.599	-21281.599	-2.00E-06	1	5.991	FALSE
CIZ1_nuc_gap_Human	modelAnull	modelA	-21281.599	-21281.599	2.00E-06	0.999999	3.841	FALSE
CIZ1_nuc_gap_Mouse	m1Neutral	modelA	-21281.599	-21261.82	39.558794	2.57E-09	5.991	TRUE
CIZ1_nuc_gap_Mouse	modelAnull	modelA	-21266.454	-21261.82	9.267534	0.00971808	3.841	TRUE
CIZ1_nuc_gap_Pig	m1Neutral	modelA	-21281.599	-21281.599	0	1	5.991	FALSE
CIZ1_nuc_gap_Pig	modelAnull	modelA	-21281.599	-21281.599	2.00E-06	0.999999	3.841	FALSE

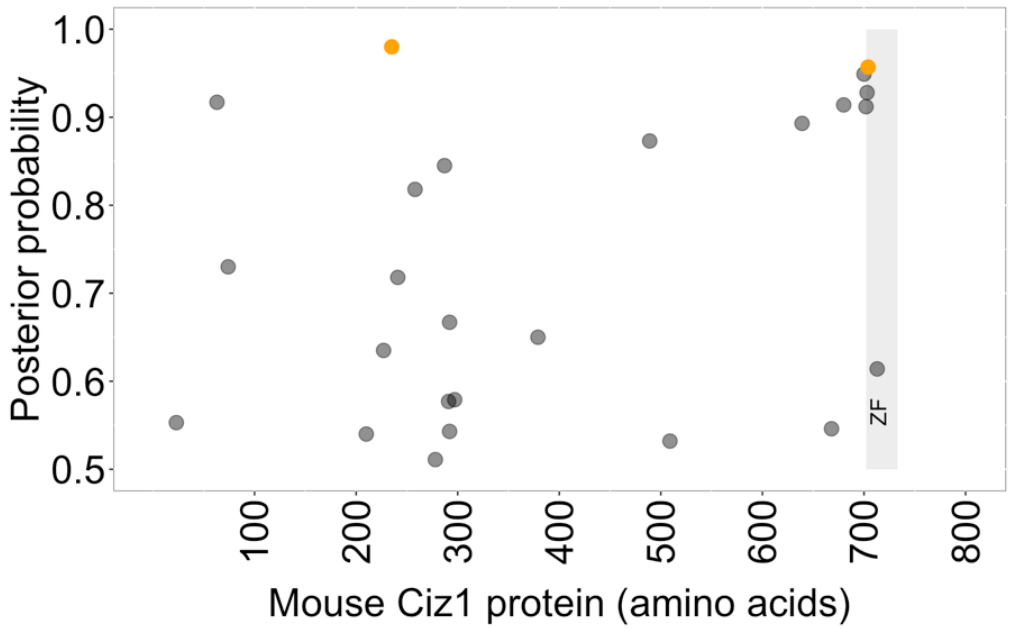


Figure 5.4. Ciz1 shows evidence of positive selection in the mouse lineage.

The x-axis represents the amino acid sequence of the Ciz1 gene from the first to the last amino acid residue. The y-axis denotes posterior probability, a confidence value for the prediction that a site is under positive selection, ranging from 0.5 to 1 (50-100%). Vertical bars symbolise the protein's functional domains according to their location. Dots in black indicate positively selected sites with a posterior probability score of 0.5 to <0.95 whereas dots in orange have a posterior probability score of 0.95 to 1. The zinc-finger domain (ZF) is highlighted in grey panel and corresponds to positions 702 to 733 of the amino acid sequence.

Both site and lineage-site models were applied because the site models are the null for the lineage-site models and are therefore necessary to justify the application of the lineage parameters. However, the results most relevant to the question of interest in this chapter are based on the lineage-site results as these are the models that allow us to determine if there are selective pressure variations detectable between species. In summary, lineage-specific models allow us to assess whether positive selection could (perhaps in part at least) explain *XIST* protein partner differences observed across placental mammals. Results from the LRTs and the comparisons between models were shown, but in the outcomes of lineage-specific models for the remaining genes examined are more relevant for the question of interest here. A summary table of lineage-specific selective pressure models where positive selection was found in selected lineages is shown in **Table 5.10**. To understand the impact of these sites under positive selective pressure in the context of the association of these proteins with *XIST*, the location of positively selected sites was reviewed with regards to each protein's functional domains. The aim of this was to identify sites whose positive selection could potentially have an impact on protein function through for e.g. modification of RNA binding.

When assessing for selective pressure variation on the SPEN gene, there was evidence of positive selection exclusively in the pig lineage (**Table 5.10**). In the pig protein, ~0.2% of all amino acid sites were predicted to be under positive selection with an ω ratio of >>1 (ω_2=459). Namely, there were four amino acid sites with a PP greater than 0.5. Of those four sites, three exhibited a PP greater than 0.95, and two were above 0.99. Mapping CodeML-predicted positively selected sites from the SPEN amino acid alignment on pig SPEN, none of the four sites (1173M, 2711L, 2712Q, 2714Q, 2715Q) were found to overlap functional domains as described by UniProt (RRM1 35-110 aa and RRM2 114-186 aa; **Figure 5.5A**). Given a superior annotation of the human amino acid sequence in UniProt, CodeML-predicted positively selected sites in the pig lineage from the SPEN amino acid alignment were next mapped on the human SPEN amino acid sequence. Once again, none (1590M, 3128L, 3129Q, 3131Q, 3132Q) were found to overlap functional domains as described by UniProt (RRM1 6-81, RRM2 335-415, RRM3 438-513, RRM4 517-589, RID 2201-2707 and SPOC 3498-3664 aa; **Figure 5.5B**). This would imply that the

RNA- binding domains of this protein have been evolving slowly, remaining unchanged between the species examined here.

Table 5.10. Summary of lineage-specific selective pressure models of best fit tested for all genes.

The number of sites predicted to be under positive selection are shown in the last column, separated by posterior probability cutoffs. Negative/purifying selection (ω=0-0.9), neutral evolution (ω=0.9-1.1) and positive selection (ω>1.1) (Zhang et al., 2005). P, number of free parameters estimated in the model; lnL, log-likelihood values; p_n proportion of sites under a particular ω value; ω, the ratio of non-synonymous to synonymous nucleotide substitution (dN/dS). BEB, Bayes Empirical Bayes estimations.

Gene: Foreground Lineage	P	Estimates of parameters	Positively selected sites (BEB)
CIZ 1: Mouse			
modelA	7	p0=0.60542 p1=0.3226 p2=0.04696 p3=0.02502 ω0=0.11444 ω1=1.0 ω2=4.72081	25>0.5, 2>0.95, 0>0.99
SPEN: Pig			
modelA	7	p0=0.80872, p1=0.18868, p2=0.00211, p3=0.00049, ω0=0.08452, ω1=1.0, ω2=459.0714	4 >0.5, 3>0.95, 2>0.99
RBM15: Cow			
modelA	7	p0=0.90387 p1=0.09105 p2=0.00462 p3=0.00047 ω0=0.03298 ω1=1.0 ω2=935.93937	3>0.5, 0>0.95, 0>0.99
MATR3: Cow			
modelA	7	p0=0.94955 p1=0.04663 p2=0.00365 p3=0.00018 ω0=0.04935 ω1=1.0 ω2=999.0	3>0.5, 3>0.95, 2>0.99
PTBP1: Human			
modelA	7	p0=0.85415 p1=0.13817 p2=0.00662 p3=0.00107 ω0=0.03019 ω1=1.0 ω2=999.0	4>0.5, 1>0.95, 1>0.99
PTBP1: Pig			
modelA	7	p0=0.8475 p1=0.12035 p2=0.02815 p3=0.004 ω0=0.03025 ω1=1.0 ω2=31.99978	6>0.5, 3>0.95, 2>0.99

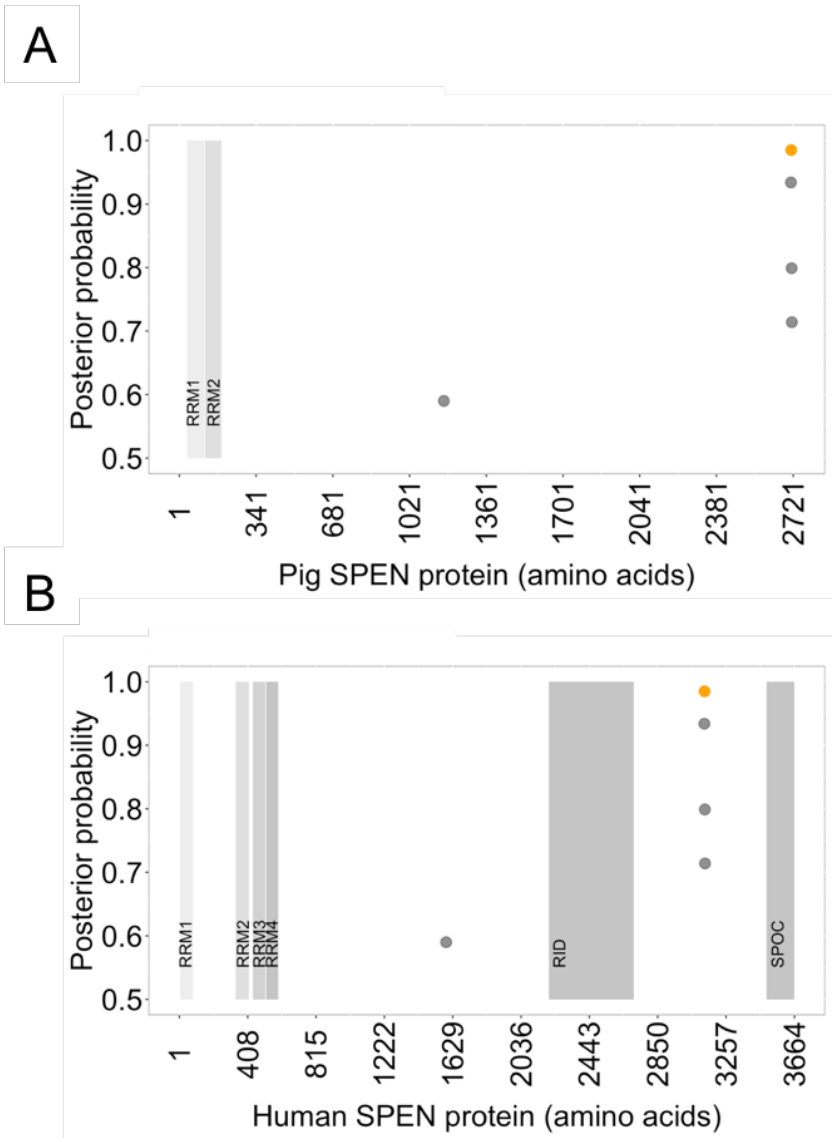


Figure 5.5. SPEN shows evidence of positive selection in the pig lineage.

CodeML predicted sites identified in the pig lineage are here mapped on the **A**) pig and **B**) human sequence. The x-axis represents the amino acid sequence of the SPEN gene from the first to the last amino acid residue. The y-axis denotes posterior probability, a confidence value for the prediction that a site is under positive selection, ranging from 0.5 to 1 (50-100%). Vertical bars symbolise the protein's functional domains according to their location, reading from left to right these are RRM1 (aa positions 6 to 81), RRM2 (aa positions 335 to 415), RRM3 (aa positions 438 to 513), RRM4 (aa positions 517 to 589), RID (aa positions 2201 to 2707) and SPOC (aa positions 3498 to 3664). Dots in black indicate positively selected sites with a posterior probability score of 0.5 to <0.95 whereas dots in orange have a posterior probability score of 0.95 to 1.

Similarly, there was only evidence for positive selection of RBM15 in the cow lineage (**Table 5.10**). When focusing on the cow protein, ~0.46% of all amino acid sites were predicted to be under positive selection with an ω ratio of >>1 (ω_2=935.6). More specifically, there were three amino acid sites with a PP greater than 0.5, none of which exceeded the more stringent PP cut-off of 0.95. Mapping CodeML-predicted positively selected sites identified in the cow lineage from the RBM15 amino acid alignment on the cow RBM15 amino acid sequence, one out of the three sites (956F) was found to overlap the SPOC domain as described by UniProt (SPOC 778-957 aa; **Figure 5.6**). The SPOC domain has been shown to mediate protein-protein interactions (**Section 1.6.2**).

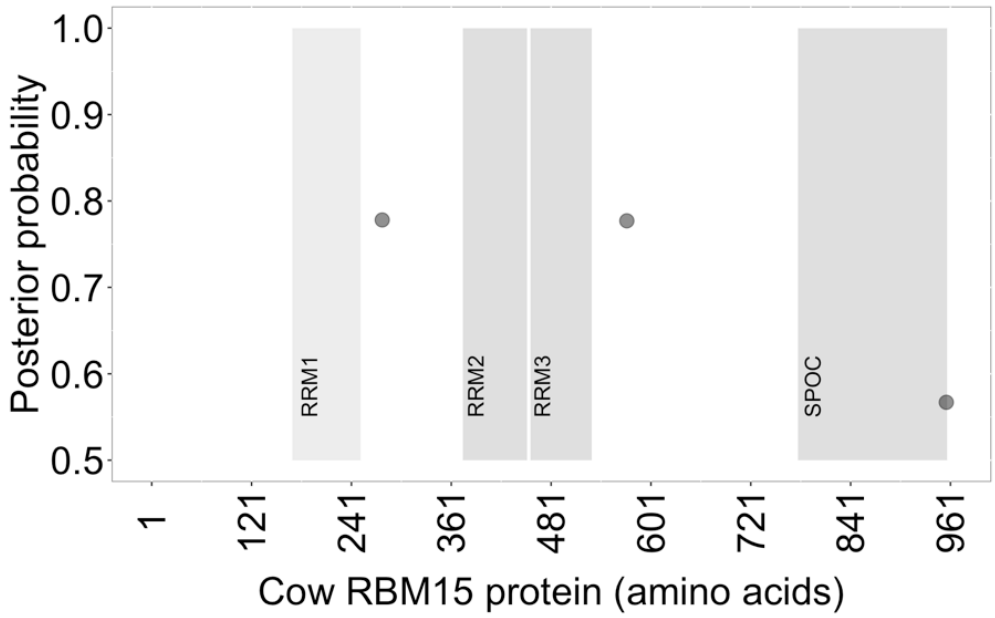


Figure 5.6. RBM15 shows evidence of positive selection in the cow lineage.

The x-axis represents the amino acid sequence of RBM15 from start to end. The y-axis denotes posterior probability, a confidence value for the prediction that a site is under positive selection, ranging from 0.5 to 1 (50-100%). Vertical bars symbolise the protein's functional domains according to their location, reading from left to right these are RRM1 (aa positions 170 to 252), RRM2 (aa positions 374 to 451), RRM3 (aa positions 455 to 529), and SPOC (aa positions 778 to 957). Dots in black indicate positively selected sites with a posterior probability score of 0.5 to <0.95 whereas dots in orange have a posterior probability score of 0.95 to 1.

In the MATR3 gene, there was evidence of positive selection in the cow lineage (**Table 5.10**). Within the cow MATR3 protein, ~0.37% (p_2) of all amino acid sites were predicted to be under positive selection with an ω ratio of >>1 (ω_2=999). In particular, three amino acid sites had PP>0.95, two of which had PP> 0.99. Mapping CodeML-predicted positively selected sites from the MATR3 amino acid alignment on the cow MATR3 amino acid sequence, no sites (842G, 844D, 845Y) were found to overlap functional domains as described by UniProt (RRM1 398-473, RRM2 496-571 aa; **Figure 5.7A**). The same was true when positively selected sites identified for MATR3 in the cow lineage were mapped onto the human MATR3 amino acid sequence (**Figure 5.7B**).

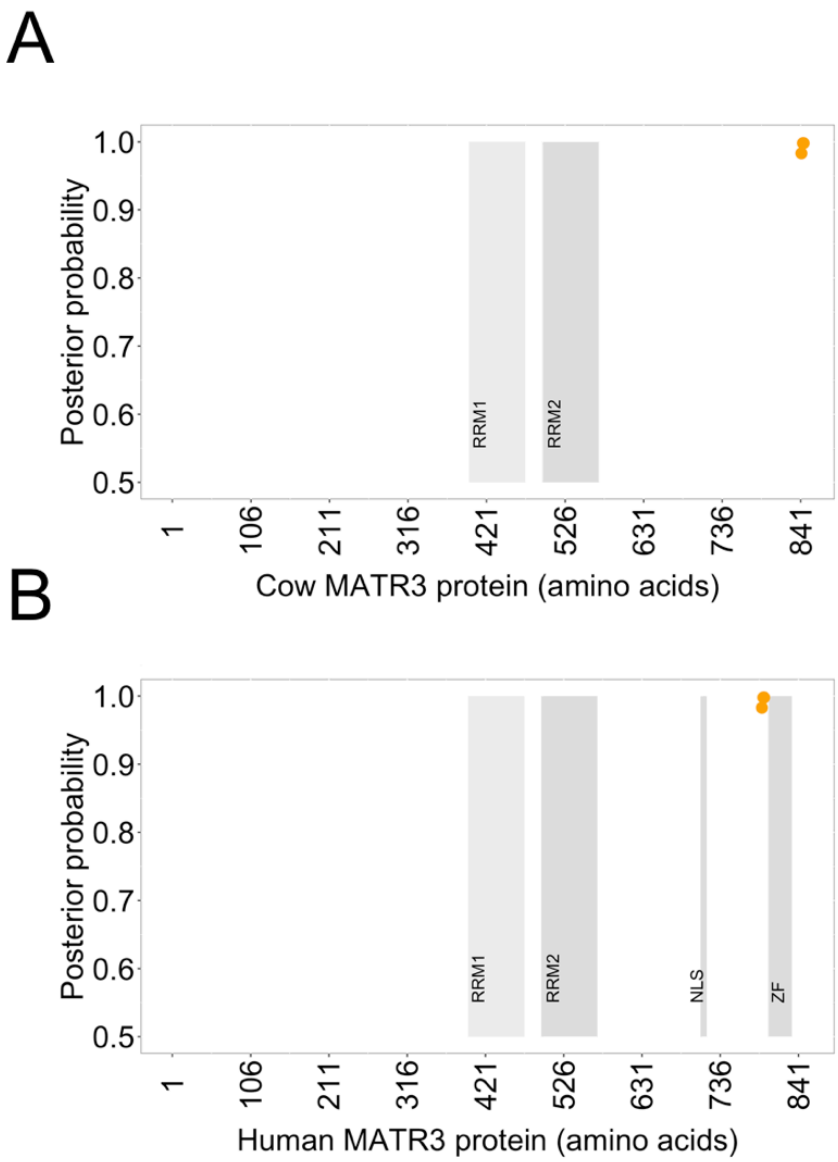


Figure 5.7. MATR3 shows evidence of positive selection in the cow lineage.

CodeML predicted sites identified in MATR3 from the cow lineage were mapped on the **A**) cow or **B**) human sequence. The x-axis represents the amino acid sequence of RBM15 from start to end. The y-axis denotes posterior probability, a confidence value for the prediction that a site is under positive selection, ranging from 0.5 to 1 (50-100%). Vertical bars symbolise the protein's functional domains according to their location, reading from left to right these are RRM1 (aa positions 398 to 473), RRM2 (aa positions 496 to 571), NLS (aa positions 710 to 718) and ZF (aa positions 801 to 832). The UniProt entry for bovine MATR3 did not describe coordinates for a ZF domain or an NLS. Dots in black indicate positively selected sites with a posterior probability score of 0.5 to <0.95 whereas dots in orange have a posterior probability score of 0.95 to 1.

Examining selective pressure variation on the PTBP1 gene, there was evidence of positive selection in both the human and pig lineages (**Table 5.10**). In human, ~0.675 of all amino acid sites were predicted to be under positive selection with an ω ratio of >>1 (ω_2=999), which translated into four sites with a PP greater than 0.5, one of which exceeded the most stringent PP cut-off of 0.99. CodeML-predicted positively selected sites from the PTBP1 amino acid alignment identified from the human lineage were mapped on the human PTBP1 amino acid sequence. Four out of four sites (351P, 378Q, 379S, 385G) were found to overlap the RRM3 domain as described by UniProt (RRM1 59-143, RRM2 184-260, RRM3 337-411 and RRM4 454-529 aa; **Figure 5.8A**). Since these sites overlap an RNA-binding domain, this could have implications for the RNA binding capacity of this protein. In the pig lineage, ~2.8% of all sites were predicted to be under positive selection with an ω ratio of >>1 (ω_2=~32), resulting in six amino acids sites with PP> 0.5, three of which had PP > 0.95, and two with PP> 0.99. The amino acid sequence for pig PTBP1 generated from the analysis here was longer than the one listed by UniProt. Due to this disparity, positively sites were mapped on the human PTBP1 sequence directly. When CodeML-predicted positively selected sites found in the pig lineage were mapped on the human PTBP1 amino acid sequence, no sites (4I, 6P, 8I, 9A, 11G, 12T) were found to overlap any functional domains as described by UniProt (RRM1 59-143, RRM2 184-260, RRM3 337-411 and RRM4 454-529 aa; **Figure 5.8B**).

Complete protein structures for these proteins could not be found in The Protein Data Bank (PDB, rcsb.org) (Berman et al., 2000). Partial protein 3D structures based on homology could be traced from SWISS-MODEL (Waterhouse et al., 2018) for human SPEN, CIZ1, PTBP1, MATR3, RBM15 proteins. However, experimental evidence for structure was only available for single domains of human SPEN (SPOC domain) and PTBP1 (RRM2 domain).

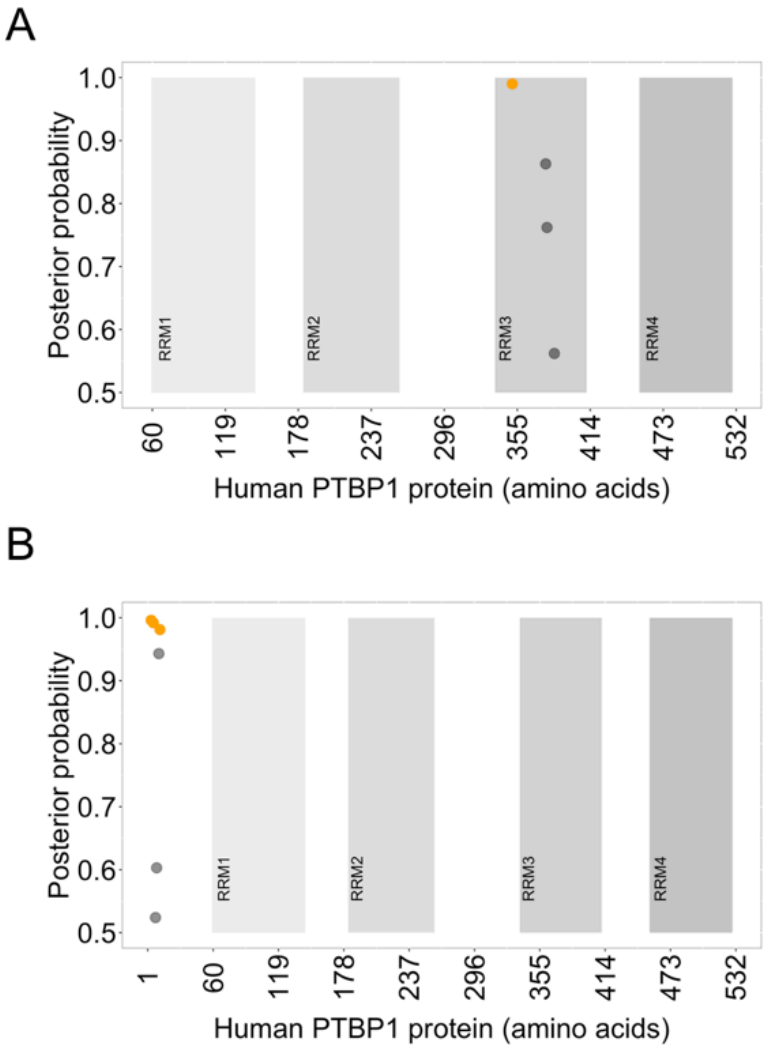


Figure 5.8. PTBP1 shows evidence of positive selection in the human and pig lineage.

CodeML predicted sites identified in the **A**) human or **B**) pig lineage are shown. In B) sites predicted in the pig lineage were plotted on the human sequence due to the disparity between the PTBP1 sequence generated here compared to the one listed in UniProt. The x-axis represents the amino acid sequence of the PTBP1 gene from the first to the last amino acid residue. The y-axis denotes posterior probability, a confidence value for the prediction that a site is under positive selection, ranging from 0.5 to 1 (50-100%). Vertical bars symbolise the protein's functional domains according to their location, reading from left to right these are RRM1 (aa positions 59 to 143), RRM2 (aa positions 182 to 260), RRM3 (aa positions 337 to 411) and RRM4 (aa positions 454 to 529). Dots in black indicate positively selected sites with a posterior probability score of 0.5 to <0.95 whereas dots in orange have a posterior probability score of 0.95 to 1.

5.4 Discussion

Experimental evidence from the previous chapters (Chapters 3 and 4) pointed to different protein partners of XIST across human and cow, which represent species of placental mammals with different implantation strategies. The hypothesis explored here was that positive selection has tailored the amino acid sequence and thus structure-function relationship of *XIST* protein partners in each placental mammal species, providing grounds for a functional shift, which might account for differential binding observed. This would imply that select proteins might have acquired mutations in functional domains mediating binding to *XIST*, resulting in interactions with altered affinity, or even loss of the *XIST* interaction across placental mammals.

5.4.1. Insufficient phylogenetic signal measured for most genes tested could be related to a high sequence conservation

Results presented in this chapter argue that multiple alignments for genes selected contain homologs and show signal that is better than random. The phylogenetic signal contained in these alignments however is not high enough to allow for confident inferences of the relationships of placental mammals based on these genes. Having protein-coding sequences for an alignment be too closely related or too divergent across species, can make it difficult to quantify the phylogenetic signal contained in alignments as there will either not be a sufficient number of changes present or too many changes could cloud any underlying relationship(s) (Jeffares et al., 2015). Genes that are highly similar may imply constraint to remain unchanged. The fact that the gene tree does not match the species trees can result in distortions of true phylogenetic relationships. Given phylogenetic relationships for species selected are known, based on the current placental mammal phylogeny, this enables the use of this information for our selective pressure variation analyses (forcing a species phylogeny). A possible reason for discordance could be hidden paralogy. However, each gene was checked against the OMA database, ensuring single gene ortholog genes were included in the analysis.

5.4.2. Positive selection detected in RNA-binding domains of mouse Ciz1 and human PTBP1

Overall, examining lineage-specific models of evolution that fit the data, evidence of positive selection could be substantiated for PTBP1 (human), Ciz1 (mouse), RBM15 and MATR3 (both in cow) and SPEN, PTBP1 (both in pig). Furthermore, no evidence of positive selection was detected for WTAP, LBR, hnRNPK or hnRNPU in any of the four lineages examined, implying amino acid residues are evolving neutrally or are under purifying selection in these species.

More specifically, there was support for positive selection acting on the Ciz1 gene solely in the mouse lineage (**Table 5.10**). Four out of 25 amino acid residues predicted to be under positive selection are within the zinc-finger domain of mouse Ciz1 (**Figure 5.5).** Zinc-finger domains have been shown to have RNA-binding potential (Klug, 1999, Brown, 2005, Hall, 2005), implicating this Ciz1 domain in *XIST* recognition and binding. The lack of experimental evidence for a mouse Ciz1 structure makes it difficult to infer whether amino acid residues identified here would be linked to a functional shift. Based on work presented here (**Figures 3.11, 4.7, & 4.9**) and that of others (Ridings-Figueroa et al., 2017, Sunwoo et al., 2017, Yue et al., 2017, Pandya-Jones et al., 2020, Sofi et al., 2020), CIZ1 has been demonstrated to be able to bind *XIST* in human, cow and mouse. Additionally, the role of this protein in XCI has been shown to be shared between human and mouse (**Section 1.6.2**). Therefore, future work could dissect whether the effect of positive selection observed points to an interaction with increased or decreased affinity in mouse. Despite the lack of available literature on the capacity of pig CIZ1 protein to interact with *XIST* to date, it would be interesting to speculate such an interaction could take place. This is based on the observed 80% sequence identity between pig and human CIZ1 (**Table 2.4**) and the expression of both pig CIZ1 protein and pig *XIST* at the same time in pig endometrium (**Figures 2.9 & 2.11**).

For the SPEN gene, there was evidence of positive selection solely in the pig lineage (**Table 5.10**). However, positively selected amino acid sites identified do not fall within described functional domains on the pig SPEN protein (or the human ortholog) (**Figure 5.6**). However, from work presented here (**Figures 2.9 & 2.11**) and

proteomics and transcriptomics data available in the TISSUES database (Palasca et al., 2018), pig SPEN mRNA and protein, respectively, are expressed in the pig (*Sus scrofa*). Whether the pig SPEN protein is capable of interacting with pig XIST has not yet been elucidated. In line with these findings, a previous study which performed selective pressure variation analyses across SPEN's functional domains using CodeML on 10 species also reported a lack of positive selection, with no evidence of eutherian lineage-specific evolution (Carter et al., 2020).

With regards to the RBM15 gene, positive selection was detected solely in the cow lineage (**Table 5.10**). One of three amino acid residues found to be under positive selection was part of the SPOC domain in cow RBM15 (**Figure 5.7**). Proteins with SPOC domains typically act as transcriptional regulators of gene expression during development (Kuang et al., 2000). The SPOC domain is thought to mediate protein-protein interactions (Section 1.6.2). The lack of experimental evidence for cow or human RBM15 structure makes it difficult to assess whether this particular residue might be linked to a functional shift. RBM15 has been demonstrated to interact with human (Patil et al., 2016) and mouse *Xist* (Moindrot et al., 2015) but there are no studies on cow or pig RBM15. Despite establishing that cow RBM15 is coordinately expressed with cow *XIST* in the endometrium (**Figures 2.9 & 2.11**), pulldown experiments were unable to confirm this interaction in bovine stromal cells (**Figure 3.18 & Table 4.4**).

Likewise, evidence of positive selection for the MATR3 gene was only predicted in the cow lineage (**Table 5.10**). Nevertheless, none of the amino acids identified as positively selected were located within a functional domain region when mapped on the cow or human MATR3 protein sequence (**Figure 5.8**). MATR3 was recently shown to interact with human (Graindorge et al., 2019) and mouse *Xist* repeat A (Chu et al., 2015b, Pandya-Jones et al., 2020), and play a role in ensuring proper mouse *Xist* localisation to the Xi (Pandya-Jones et al., 2020). MATR3 was also one of the proteins that co-immunoprecipitated with SPEN following RIP in ISHIKAWA cells (**Figure 3.16**). According to proteomics and transcriptomics data available in the TISSUES database (Palasca et al., 2018), there is evidence for the presence of MATR3 in cow and pig tissues. Nonetheless, there are no reports so far showing an association of *XIST* with MATR3 outside of human and mouse. Pulldown of cow

XIST repeat A failed to enrich for MATR3 to an average log2FC>1 and at a statistically significant level (MATR3 average log2FC = 0.71).

When assessing for selective pressure variation on the PTBP1 gene, positive selection was detected in the human and pig lineages (**Table 5.9**). All four amino acid sites found to contribute to positive selection overlapped the RRM3 domain in human PTBP1 (**Figure 5.9A**). Conversely, none of the sites predicted in the pig lineage were located on functional domains when these were mapped on the pig or human PTBP1 protein (**Figure 5.9B**). Due to the absence of a fully resolved human PTBP1 protein structure, an assessment of whether the predicted set of residues could affect the RNA binding activity of human PTBP1 was not undertaken. Whilst PTBP1 was not experimentally tested for its ability to interact with human or pig *XIST* as part of this work, others have shown that PTBP1 binds *XIST* in both human (Lu et al., 2020a, Pandya-Jones et al., 2020) and mouse (Smola et al., 2016, Vuong et al., 2016). The biological relevance of PTBP1 binding to human *XIST* has not been elucidated yet, but Ptbp1 has been demonstrated to regulate splicing of mouse *Xist* during development (Stork et al., 2019) and also contribute to *Xist* localisation to the Xi (Pandya-Jones et al., 2020). There is no evidence as of yet for PTBP1 binding to cow or pig *XIST*.

Lastly, the lack of positive selection seen here for the LBR, WTAP, hnRNPK and hnRNPU genes across human, mouse, cow and pig lineages does not definitively address the potential of these proteins to associate with XIST. A lack of positive selection seen for WTAP was consistent with a previous study where authors used the same codon-based models of evolution as here (CodeML), across primate, rodent and teleost lineages and found no sites in WTAP to be under positive selection (Wu et al., 2016). Taken together, data presented in this chapter argue against positive selection as a reason accounting for the differential binding of *XIST* observed for these proteins in human and cow (Chapters 3 & 4).

In Chapter 2, multiple sequence alignments of *XIST* protein partners across human, mouse, cow and pig were performed, which showed a high sequence similarity. The results from this chapter confirm and extend this high sequence similarity to other eutherian mammals, evident from manually inspecting the alignments or the quartet

puzzling data (**Fig. 5.1**). Additional evidence that *XIST* protein partners have evolved slowly across eutherian mammals can also be observed from a general lack of positive selection seen, with purifying selection along the majority of the proteins' functional domains. This could imply that the proteins are mostly under purifying selection (with the reported cases of positive selection as the exceptions) or that there is co-evolution of both proteins and XIST's sequence. In the latter case, rapid divergence of the *XIST* sequence would not necessarily result in a loss of an interaction with its partners as it could be compensated by substitutions in the protein partner sequence. This seems as a more attractive model whereby *XIST* repeat regions can evolve, expanding in size allowing for the recruitment of additional proteins, improving the overall X-linked gene silencing efficiency compared to the proto-*XIST* transcript (Brockdorff, 2018).

5.4.3. Assumptions of selective pressure variation analyses by CodeML

It is worth noting at this point that there are a number of assumptions that dN/dS estimations make. The accuracy of selective pressure variation prediction relies on these assumptions, which if violated, could influence the outcome of these results and hence our hypothesis. Pitfalls in CodeML analyses which can lead to false positives or false negatives can be broken down into two categories: i) assumptions on quality of input and ii) interpretation caveats. For instance, the software selected for multiple sequence alignment could be a determining factor of the number of positive selection false-positive occurrences (Blackburne and Whelan, 2012, Jordan and Goldman, 2012). Multiple sequence alignments and subsequent phylogenetic trees constructed could be of poor quality. Indeed, a previous study compared the performance of multiple sequence alignment tools in detecting positive selection, showing that some tools are better than others (Fletcher and Yang, 2010), highlighting CodeML analyses are susceptible to errors if alignments are poor. Multiple sequence alignments can further be affected by a lack of positional homology (hidden paralogy), which would compound selective pressure variation analyses. A phylogenetic tree could also be misleading when synonymous substitutions within a codon occur very frequently such as a when sequences diverge (leading to synonymous substitution saturation), which could bias dN/dS calculations making the dS difficult to measure with any degree of certainty (Seo and

Kishino, 2008, Wachter and Hill, 2016). Accurately estimating dS values can be confounded by differences in underlying mutational processes and the frequency with which each mutation occurs (Van den Eynden and Larsson, 2017). It has been demonstrated that synonymous mutations (or silent substitutions) are not neutral and can be selected for (Hurst and Pál, 2001, Plotkin and Kudla, 2011). For instance, certain genes that are highly expressed (Jeffares et al., 2015) could display selection on silent sites at splice sites (Hurst and Pál, 2001, Plotkin and Kudla, 2011).

There are also pitfalls when it comes to distinguishing between modes of evolution acting on genes. For genes where purifying selection was not detected ($0.9<\omega<1.1$), this does not necessarily imply the gene is neutrally evolving as this is unlikely, although can rarely happen. It could simply mean that there are changes in amino acid sites constantly being trialled for fitness, where some are retained, others removed and others are not 'visible' to selection if they are neither beneficial or deleterious. Essentially, a dN/dS score of 0.9-1.1 does not allow for rejection of the hypothesis for neutrality. The upper cut-off was set to 1.1 instead of 1 as a more stringent filter, to ensure sites evolving under neutrality are avoided. Another possible scenario that could be occurring but would be missed by the approaches used here would be if positive selection was taking place in a few sites across several genes (polygenic adaptation) which could be governing binding specificity displayed by proteins assayed here (Pritchard et al., 2010). For instance, polygenic adaptation could explain a shift in protein binding affinity or structure to facilitate several protein partners interacting with XIST at the same time. There could also be several sites that might be important for binding of both XIST and other RNA interactors of the proteins assayed here (XIST is not the only target RNA of the proteins examined and these proteins have roles outside of XCI).

Ultimately, obtaining an ω ratio higher than 1.1 may still not be definite proof in favour of a meaningful change in a protein's function, regardless of posterior probability confidence values obtained for any site. The ω ratio could be skewed if there was a population increase (usually seen in species with high population sizes) (Fay et al., 2002, Fay, 2011, Booker et al., 2017)(**also Section 1.11**). Additionally, relaxation of purifying selection, which could translate into a local build-up of non-synonymous mutations being tolerated, can be mistaken for positive selection

(Björnerfeldt et al., 2006, Shen et al., 2009, Calderoni et al., 2016). This could derive from the absence of a force imposing constraint or increased genetic drift mentioned earlier from a small N_e (Wertheim et al., 2015). Several amino acid replacements may seem as nearly neutral, but could additively have a small effect on protein function (Daub et al., 2013). An ω ratio higher than 1.1 could be due to alternative codon usage at a particular site, and previous studies have shown that codon usage could be modified to influence expression levels, splicing or mRNA stability (Chamary et al., 2006). All in all, it is important to remember that these are computational predictions. In the absence of experimental verification their biological relevance is not proven (Hughes, 2007).

6. Main discussion

Ensuring balanced gene expression from different sets of sex chromosomes between sexes, is vital for the progression of normal embryogenesis. Gene dosage compensation occurs via XCI in placental mammals. The *XIST* lncRNA orchestrates XCI by recruiting a multitude of proteins and *XIST*-RNP complexes to achieve chromosomal-wide gene silencing and 3-D compaction of one of the two X chromosomes in females. Perturbations of the XIST gene or its transcriptional regulators has been shown to result in aberrant developmental growth and embryonic lethality as early as day 10 post-coitum in mice (Takagi and Abe, 1990, Marahrens et al., 1997). Equally, depleting or knocking out key XIST partners can cause aberrant X-linked gene reactivation (**Section 1.7.4**).

Most lncRNAs display rapid evolution across species, and therefore *XIST* is an atypical lncRNA as it displays modest conservation across species (**Table 2.3**). The conservation of the *XIST* sequence may reflect functional constraints to retain protein binding sites and interactors, enabling *XIST* to faithfully elicit XCI across species. However, placental mammals display different early pregnancy events such as timing of embryonic genome activation, *XIST* expression, XCI timing, degree of implantation, and placental morphologies. Such differences could also have provided opportunities and/or pressures for *XIST* to evolve different interacting protein-partners across placental mammals, in a model where the sequence of *XIST* would presumably vary between species.

Our knowledge about the identity of *XIST*'s protein partners as well as their precise binding locations on the *XIST* sequence derives from studies of the mouse and human. Given the lack in our knowledge about which proteins interact with *XIST* outside of human and mouse, this thesis aimed to (i) investigate lncRNA-protein partner co-evolution using *XIST* as a model, and (ii) to address whether protein partners of *XIST* in endometrial cells are shared across two placental mammal species with different timing of EGA, XIST expression, XCI timing, implantation and placentation in human and cow. The rationale for focusing on the endometrium was that XCI is a female-specific process and differences in protein partners found across placental mammals could offer insight into gestation evolution.

6.1. Strong conservation and co-ordinate presence of *XIST* protein partners across human, mouse, cow and pig in uterine tissue/cells

The first aim of this thesis was to determine the conservation of *XIST* and its putative protein partners, and assess whether they are present in reproductive tissues from placental mammals with different implantation strategies.

Using Clustal-ω alignments, full-length *XIST* was shown to be >61% similar across human, mouse, cow and pig, and included short regions of higher conservation (**Table 2.4**). These short regions corresponded to previously documented repeats A, B and D on the *XIST* sequence. Repeats in the 5' end of XIST are known to be conserved (Brockdorff et al., 1992, Brown et al., 1992). These repeats showed higher levels of sequence conservation than full-length *XIST* in human, mouse cow and pig (78-90%, 62-91%, 54-86% similar, respectively), whilst repeat E was the least conserved across these species (~50% similar). This is consistent with a trend of lncRNA conservation in short stretches rather than across their whole sequence (**Section 1.3**), which could represent conserved binding sites for maintenance of critical RNA-protein interactions. Repeats A and E were more similar across human, cow and pig compared to mouse, which could suggest that human *XIST* protein interactors are more likely to be shared across cow and pig, than interactors of other regions (e.g. repeats B/C).

Estimating sequence conservation of protein partner orthologs of mouse *Xist* (i.e. Spen, Ciz1, Hnrnpk, Rbm15, Wtap, Lbr and Hnrnpu) using amino acid sequences, demonstrated a high percentage of amino acid conservation (>70%) across human, mouse, cow and pig. All of the above proteins except Spen (where no suitable antibody was found) could be detected in uterine tissue/cells by western blotting, making this the first report with uterine-specific expression data for putative *XIST* protein partners across human, mouse, cow and pig. This dataset could constitute a starting point for researchers interested in studying these proteins (and the processes they are involved in) outside of established model systems. The same tissues/cells used for RT-qPCR of the *XIST* RNA and mRNA of its protein partners were also used for western blotting, confirming a co-ordinate presence of *XIST* and of its putative protein partners in the same tissues/cells at the same time. This

observation coupled with a high sequence conservation spanning functional domains, suggests these proteins may be capable of biochemically interacting with *XIST* from other species besides the mouse. More specifically, *XIST* repeat A was ~85% similar between human and cow and the amino acid similarity of proteins previously shown to bind mouse *Xist* repeat A, Spen and Rbm15, was 88% and 99%, respectively (**Table 2.5**). According to a high-throughput RBP screen, if two proteins share more than 70% similarity in their RNA-binding domain, they will likely also share similar binding sites on an RNA (Ray et al., 2013). Therefore, this would predict SPEN and RBM15 as likely candidate interactors of bovine *XIST*. Conversely, a 54% similarity of *XIST* repeat E between human and cow despite an ~80% similarity in the CIZ1 amino acid sequence might intuitively not predict a likely interaction of CIZ1 with *XIST* in cow.

6.2. CIZ1 and hnRNPU associate with *XIST* in human and cow

The protein interactome of mouse *Xist* was originally determined in pluripotent embryonic cells where mouse *Xist* expression was induced, controlling the onset of XCI (**Section 1.6.2**). Thus, these experiments were designed to identify protein partners of *Xist* in the early stages of XCI. More recent studies have examined the protein partners of human *XIST* in differentiated cells, corresponding to the later stages of XCI. Protein partners of human *XIST* could vary up to ~40% across cell types but also across stages of XCI (Yu et al., 2021), according to a recent study employing pulldowns of human *XIST* in differentiated cell lines and assessing for overlap of protein interactors. Additionally, only one other recent study has elucidated the interactome of different human *XIST* repeats A, C and F but not E, in HEK293T cells (Graindorge et al., 2019), indicating a complete picture of *XIST* interactors in differentiated cells is lacking.

Chapters 3 and 4 addressed the protein interactome of human and bovine *XIST* in uterine-derived cells. Furthermore, Chapters 3 and 4 examined the roles of bovine *XIST* repeat A and E in mediating protein binding, expanding the collection of placental mammal *XIST* interactors and including cell types that are relevant to placental mammal reproduction. Additionally, the potential of bovine *XIST* to interact with human proteins was assessed in a cross-species pulldown experiment in

Chapter 4. Whilst the RIP approach taken in Chapter 3 has the potential to confirm specific protein partners of *XIST*, it is not as sensitive as the RNA pulldown approach taken in Chapter 4 (*in vitro* transcription pulldown). The latter pulldown approach will miss protein partners that are related to RNA modifications or structural motifs present when the whole sequence is assembled *in vivo* due to processing. Another limitation of both approaches is the lack of UV crosslinking prior to cell lysis, increasing the chances of indirect interactions being detected. The RIP approach allows the characterisation of RNA-protein interactions with relatively low sample input, which here would allow for the use of primary bovine stromal cells enabling the characterisation of bovine *XIST* protein partners outside of an immortalised cell line. Moreover, *in vitro* transcription pulldowns enable the assessment of protein binding to specific parts of an RNA (instead of the full-length transcript), to provide specific information on where the protein binds a useful feature for bridging predictions about local region conservation and RNA-protein co-evolution. Hence, employing both methods serves to orthogonally validate one another and allows filtering of a consensus list of interactors.

To experimentally test whether interactions occur between *XIST* and its putative protein partners (previously shown to be co-ordinately expressed in the same tissue **Sections 2.2** and **2.3**), RIP was performed in a human endometrial cell line (ISHIKAWA) and primary bovine stromal cells of the endometrium. Interactions of CIZ1, WTAP, hnRNPK and SPEN were demonstrated with human *XIST* (**Figures 3.10-3.15**), albeit none of these proteins could be confirmed to bind bovine *XIST* in bovine stromal cells (**Figures 3.17-3.20**). It is important to bear in mind that success in RIP is particularly dependent on good quality antibodies that can recover their target protein with a high affinity in pulldowns, which is different to an antibody being able to detect a protein in western blot. Antibodies used for pulldown of bovine SPEN, RBM15 and WTAP did not display high levels of protein enrichment in elution samples, and therefore, an interaction between these proteins and bovine *XIST* could not be formally addressed (**Figures 3.17-3.19**). Conversely, the antibody for hnRNPK was suitable for pulldowns in bovine cells but hnRNPK did not demonstrate an interaction with bovine *XIST* via RIP (**Figure 3.20**).

To circumvent difficulty in protein enrichment for putative *XIST* partners in cow due to unsuitable antibodies, the repeat regions of the *XIST* RNA, where protein binding was predicted to occur based on studies in mice, were *in vitro* transcribed and used in pulldowns. *In vitro* transcribed *XIST* repeats were mixed with uterine-derived lysates from human or cow and interactions were assessed via pulldowns. Using *in vitro* transcribed *XIST* repeat E, an interaction between CIZ1 and *XIST* repeat E was demonstrated in three independent biological replicates via western blotting in both human and cow (**Figures 4.8 and 4.9**). This is consistent with the high similarity of the amino acid sequence of the CIZ1 protein across human and cow (80%; **Table 2.4**), despite a 5% (14/285 aa) amino acid difference in the zinc finger domain of CIZ1 (**Figure 2.7**), which likely has RNA-binding capacity as shown by previous studies of this domain in other proteins (Klug, 1999, Brown, 2005, Hall, 2005).

Using *in vitro* transcription pulldowns coupled to TMT-MS, protein partners of bovine *XIST* repeat A were identified. Among the high-confidence candidate interactors (statistically significant, p-value <0.05 and $\log_2$FC>1) were hnRNPU, hnRNPA2B1, hnRNPA0 and TOP1 proteins, which were previously shown to interact with *XIST* repeat A from pulldowns in mouse and human (**Section 4.3.7** and **Figure 6**). hnRNPU displayed ~98% similarity in its amino acid sequence across human, mouse, cow and pig (**Table 2.4**). Putative protein partner hnRNPU was previously shown to exhibit broad binding across the whole length of mouse *Xist*, with multiple eCLIP peaks spanning exon 1 (repeats B, C, D and F) and exon 6 (outside of repeat E) (Lu et al., 2020a). Repeats B, C and F display >62% similarity across species (~54% for repeat D), therefore it is reasonable to expect hnRNPU-*XIST* binding in other placental mammals besides mouse.

SPEN, RBM15 and WTAP proteins could not be robustly assessed for an interaction with bovine *XIST* using RIP due to a poor antibody performance in pulldowns. In addition, these proteins were not found to be among statistically significant candidates with a $\log_2$FC>1 following pulldown of *XIST* repeat A in bovine stromal cells. This was surprising given a high (>88%) amino acid sequence conservation of these proteins between human and cow (**Table 2.4**) which also overlaps their functional domains (**Figures 2.1-2.3**) and the high level of *XIST* repeat A conservation (**Table 2.3**). On the other hand, these differences could account for

differential binding observed. Notably, even though RBM15 was not included in the high-confidence list of interactors, it was present in the dataset with an average log_2FC score of 0.8 (not statistically significant). In summary, SPEN, RPM15, and WTAP proteins may be specific interactors of human *XIST* in the system used but were missed due to technical limitations of the assays used (**Figure 6**).

The credibility of the high-confidence list of proteins identified from TMT-MS as interacting with *XIST* repeat A from cow was reinforced by an overlap with a few candidates also found by other studies of human and mouse *Xist* (such as hnRNPU, hnRNPA2B1, hnRNPA0 and TOP1). The consensus bovine *XIST* repeat A interactome from three replicates (~40 proteins) is probably an underestimate given what has been observed from studies of human and mouse. Possible reasons contributing to replicate variability and thus a lack of more proteins passing the p-value and log_2FC cut-offs, could be related to lysate preparation, amount of input material and pulldown efficiencies. More specifically, lysis conditions in the human cell line (ISHIKAWA) permitted the generation of a nuclear-enriched lysate whereas in bovine cells, only whole cell lysates could be generated (**Figure 4.6**). A higher amount of whole cell lysate could thus be required to achieve a sufficiently high concentration of nuclear proteins. Binding affinities with which protein interact with XIST could vary in the different species, hence, incubation buffer stringency could influence which partners are pulled down.

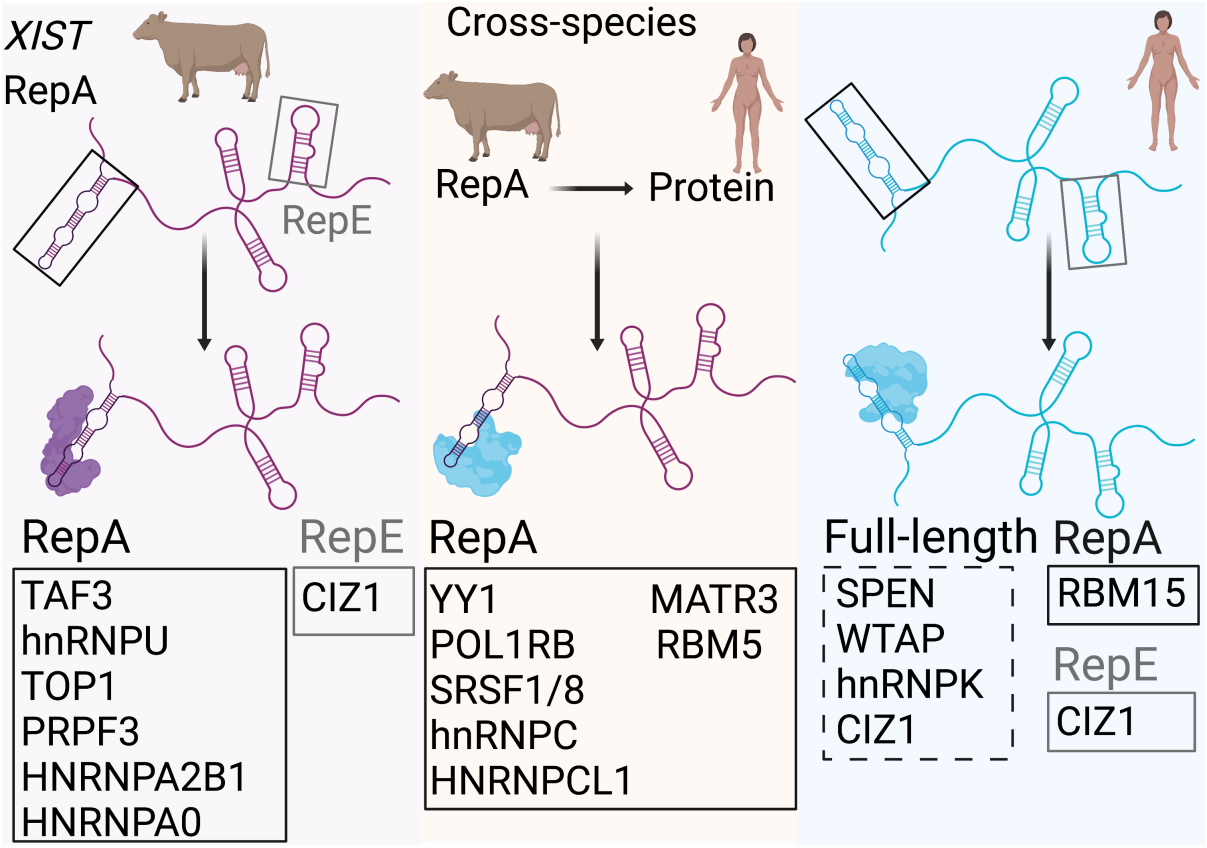


Figure 6.1. Summary schematic from pulldowns of human and bovine XIST.

The left column depicts representative bovine *XIST* partners out of 40 proteins identified from TMT-MS following RNA pulldowns with *in vitro* transcribed bovine *XIST* repeat A in primary stromal cells from bovine endometrium. RNA pulldowns with bovine *XIST* repeat E also highlighted CIZ1 as an interactor. The middle column highlights a few of the 33 potentially shared protein partners between human and cow, identified from TMT-MS following RNA pulldowns with *in vitro* transcribed bovine *XIST* repeat A in a human endometrial cell line (ISHIKAWA). The column on the right shows protein partners found to interact with human *XIST* from RIP and RNA pulldowns with *in vitro* transcribed human *XIST* repeat A and E in ISHIKAWA cells. Created with BioRender.com.

6.3. Bovine *XIST* pulldowns in human lysates highlight novel and conserved *XIST* protein partners

The evolution of the *XIST* RNA across placental mammals was coincident with the gain, expansion and exaptation as well as the loss of repetitive elements of various origin (i.e. retroviral or TE elements) across its sequence (Yen et al., 2007, Elisaphenko et al., 2008, Carlevaro-Fita et al., 2019). Some of these repetitive elements harbour binding sites for *XIST* protein interactors with critical roles for the onset and maintenance of XCI, as seen from studies of mouse and human (**Section 1.7.4**). Additionally, repetitive regions in *XIST* differ across placental mammals both in length of a repeat monomer and in the number of times a monomer repeats in tandem (**Section 1.5**). To establish whether *XIST* sequences from one species can substitute for *XIST* in a different species (as a proxy for divergence) and thus interact with the same set of proteins, pulldowns were performed with bovine *XIST* repeats A and E in a human endometrial cell line (ISHIKAWA). Using *XIST* from one species to pulldown interaction partners in a different species could provide several hints on how XIST binding partners have evolved differently, according to the context of each species. A cross-species XIST protein interaction could demonstrate that in both species the interaction depends on features that have been preserved on both the XIST sequence and the protein sequence or that simply the region harbouring the interaction interface is evolving slowly. Proteins that interact with XIST in individual species, but not in a cross-species fashion, could indicate that XIST has evolved to require a cofactor in some species but not in others, given the high protein aa conservation observed spanned functional domains. Both of these scenarios would reveal placental-mammal conserved XIST binding partners. Conversely, cross-species pulldowns where XIST from one species does not interact with any of the same proteins as XIST from another species, would indicate species-specific XIST binding partners. Hence, this approach enables in vitro testing of how different XIST interactors have evolved across species and can be customised to specific repeat regions, elucidating species-specific binding preferences of XIST's protein partners and/or species-specific XIST RNA sequence divergence.

An interaction between bovine *XIST* repeat E and human CIZ1 was not seen (**Figure 4.18**). This was despite an overall 80% similarity of the CIZ1 amino acid sequence (**Table 2.4**), with ~94% similarity (14/248 aa difference) in the zinc finger domain of

CIZ1 (responsible for RNA-binding) between human and cow (**Table 2.7**). A lack of an interaction could be observed due to low abundance of either the *XIST* fragment or CIZ1, the interaction being transient or the interaction being weak. XIST repeat E was introduced at an amount of 10 µg, which would be considered an overexpression, compared to its native abundance in the ISHIKAWA cell line. The expression of CIZ1 was not altered and therefore was considered within physiological levels, given nuclear-enriched lysates were used and the fact that CIZ1 predominantly localises to the nucleus. It was previously demonstrated that mouse Ciz1 tethers *Xist* to the Xi and that multiple Ciz1 proteins are recruited to enhance this tethering, via the formation of a mesh-like structure (Sofi et al., 2020). This is consistent with a concentrated presence of Ciz1 on *Xist* (compared to other regions in the nucleus) as determined by super-resolution microscopy. Given Xist remains attached to the Xi from the moment it is expressed, this argues against a transient interaction (Clemson et al., 1996). The strength of an interaction (avidity) could be a result of the binding affinity, valency of the proteins (how many binding sites are available on the RNA) and potential conformational changes in the tertiary structure of either protein that could help or limit interactions between the two proteins. A relatively high affinity might not be necessary for CIZ1 binding *in vivo*, given at high *in vitro* concentrations, mouse CIZ1 could bind repeat A, antisense repeat E and *Gapdh*, in addition to sense repeat E, as shown by electrophoretic mobility shift assays (Sofi et al., 2020). *XIST* Repeat E in cow is smaller in size compared to human (Yen et al., 2007)(**Figure 1.1**) and the sequence similarity between human and cow was estimated to be ~54.5% (**Table 2.3**), which could suggest a loss of binding sites in cow or a gain in humans. An alternative explanation could be that *XIST* repeat E in cow contains a denser cluster of CIZ1 binding sites than its human equivalent. Such a scenario cannot be ruled out in the absence of data with well-defined CIZ1 binding motifs. This hypothesis could also probably be tested by bioinformatically investigating repetitive regions spanning bovine *XIST* repeat E and comparing them with those from mouse and human *XIST* repeat E. One explanation is that a lack of cross-species interaction could imply that human CIZ1 requires co-operative binding with other, unknown factors, that are not necessary or present in cow.

TMT-MS following pulldowns of bovine *XIST* repeat A in human lysates identified 33 high-confidence protein partners (statistically significant, p-value <0.05 and log_2FC>1). Among them were proteins previously described in other proteomics-coupled Xist pulldown approaches in both mouse and human, such as YY1, hnRNPC, hnRNPCL1, MATR3 and various splicing factors (e.g. SRSF and SNRP proteins)(**Section 4.3.8**). In addition, proteins that had not been seen interacting with *XIST* before were found, such as RBM5, RALYL, POLR1B, ZNF43, WDR89, and ZCCHC10. None of the 33 proteins in the high-confidence set were common with the proteins found when bovine *XIST* repeat A was incubated in cow lysate. Relaxing the threshold stringency to include proteins with no statistical significance (but with a log_2FC>1), highlighted 41 proteins found across all three replicates from both cow and human datasets (**Table 4.5**). 24 of the 41 (58.5%) overlapping proteins had previously been identified by at least another study in human or mouse (indicated in **Table 4.5**), including hnRNPU, HNRNPA2B1, HNRNPA0, SRSF1, PNN, RNPS1, ACIN1 and ZC3H18 (**Table 4.5**). In agreement with pulldowns of bovine *XIST* repeat A in cow lysates, SPEN, RBM15 or WTAP were not among the list of 33 high-confidence interactors of the cross-species pulldowns with *XIST* repeat A from cow in human lysates. Surprisingly, RBM15 and SPEN were detected in the dataset (not statistically significant or log_2FC>1) with at least 2 unique peptides whereas WTAP was not detected at all in the human lysate dataset.

Proteins that were found to bind bovine *XIST* repeat A in both cow and human lysates are likely to display aa sequence conservation across placental mammals, specifically in their RNA-binding domains, as seen for hnRNPU (**Figure 2.5**). Proteins observed bound to bovine *XIST* repeat A in human lysates but not to bovine stromal lysates could be explained by technical limitations. This observation could be the result of lysate starting material, i.e. the assumption being a higher amount of protein would be required to capture the interaction. Lysate purity was different across the two, with human lysates representing a more nuclear-enriched fraction compared to whole-cell bovine lysates (**Figure 4.6**). Additionally, the stringency of the incubation buffer could influence pulldown efficiency, resulting in only observing high affinity interactions. Finally, differences between protein partners of bovine *XIST* repeat A found in cow but not in human lysates could be related to the divergence of cow and human, whereby novel proteins could have arisen in the cow to take up

XCI-related roles or that the same protein has acquired a new function in one species compared to the other (functional shift). As discussed earlier (**Section 1.7**) these differences could pertain to i) different implantation timing, ii) embryonic genome activation, iii) timing of *Xist* expression and onset of XCI or iv) placental morphology. A summary of some of the protein partners identified from the different pulldown approaches employed across chapters is depicted in **Figure 6**.

6.3. Positive selection does not account for variation in *XIST* protein partners across placental mammals

We wished to test whether specific variation in protein interactors identified for human and mouse *XIST* have evolved due to positive selective pressure in e.g. cow or pig. The rationale was that if signatures of positive selection (indicative of protein functional shift) were found overlapping *XIST*-binding regions in a protein from cow, this could explain why such an interaction would not be seen in human compared to cow (and *vice versa*). To this end we carried out an analysis of selective pressure variation on a subset of differentially enriched proteins across human and cow pulldowns. These analyses enable the prediction of amino acid sites which could be evolving under positive selection across lineages. Each predicted site comes with a posterior probability (PP) which can be used as a confidence filter for deciding which sites to experimentally validate at a later point. For this purpose, the PP cut-offs used here were 0.95 and 0.99.

Signatures of positive selection in RNA-binding domains (regions with a strong potential for a functional shift) were identified in mouse CIZ1 and human PTBP1. A single residue was found to be under positive selection in mouse CIZ1 (704S) with PP=0.957 (**Figure 5.5**) but not in other lineages. Four sites were found to be evolving under positive selection in human PTBP1 (351P, 378Q, 379S, 385G), out of which 351P had a posterior probability of 0.99 (the rest were below 0.9) (**Figure 5.9**). Nevertheless, the potential of these residues to alter the structure sufficiently to cause a shift in RNA-binding could not be assessed due to the lack of resolved protein structure models. A functional shift could denote a change in the affinity of the interaction of a putative protein partner with XIST (tighter or looser interaction) or

could result in the loss of the interaction. One line of evidence against a loss of an XIST-CIZ1 interaction across mouse, human or cow is the demonstrable interaction of CIZ1 with *XIST* in mouse (**Section 1.6.2**), human (**Figure 3.10**) and cow (**Figure 4.9**). PTBP1 has previously been shown to interact with *XIST* repeat E in mouse and human by other studies (**Section 1.6.2**). Although PTBP1 was not identified as a high-confidence interactor of bovine *XIST* repeat A, it did display a log_2FC average of 0.84 and 1.5 in the cow and human datasets, respectively (not statistically significant in either dataset). This may suggest PTBP1 can differentiate between sense and antisense bovine *XIST* repeat A transcripts *in vitro*, emphasizing its retained *XIST* RNA-binding capacity. Besides a role in splicing of mouse *Xist* (Stork et al., 2019), PTBP1 has also been shown to be involved in human *XIST* localisation to the Xi (Pandya-Jones et al., 2020), supporting a biological relevance for this interaction.

Additional sites predicted to be evolving under positive selection were further identified for the RBM15 gene in cow, where one (956F) out of the three sites (956F, 572L, 278G) was found to overlap the SPOC domain (the other two were not spanning any functional domains). The SPOC domain in RBM15 is a homolog of the SPOC domain in the SPEN protein, conserved from the spen gene in *Drosophila melanogaster* (Ma et al., 2007). SPEN's SPOC domain has been shown to facilitate protein-protein interactions (Ariyoshi and Schwabe, 2003). RBM15 was demonstrated to recruit the m6A methylation machinery, METTL3, to *XIST* via interactions with the WTAP protein (Patil et al., 2016). Although not formally addressed, it is likely the SPOC domain of RBM15 is responsible for mediating these protein-protein interactions, as suggested by the function of this domain in the SPEN protein. Lastly, sites predicted to be evolving under positive selection were predicted for the SPEN gene in pig (1173M, 2711L, 2712Q, 2714Q, 2715Q), none of which were found to overlap any known functional domains (**Figure 5.6**). Equally, no sites predicted to be under positive selection in MATR3 in pig (842G, 844D, 845Y) were found to overlap functional domains (**Figure 5.8**).

No sites were predicted to be under positive selection for LBR, WTAP, hnRNPK and hnRNPU genes across human, mouse, cow and pig lineages, suggesting these genes are under purifying selection. The lack of predicted sites under positive

selection observed for WTAP is consistent with previous findings of analyses using codon-based models of evolution across primate, rodent and teleost lineages (Wu et al., 2016). Based on selective pressure variation analyses it is unlikely a functional shift brought about by positive selection can account for differences in *XIST* protein partners observed from pulldowns in human and cow (Chapters 3 & 4).

Notably, with increasing technological advancements, genomic data for a larger number of species will become available. Incorporating a more diverse and larger number of species in selective pressure variation analyses can increase the accuracy and power of such studies as simulations have shown in the past (Anisimova et al., 2002). Expanding the collection of species may also allow for different selection signatures to be identified for different taxa, which may translate to different lncRNA-protein evolutionary relationships, providing novel hypotheses to be tested via pulldown assays.

Furthermore, performing phylogenetic analyses on XIST across placental mammals has the potential to contribute to the identification of lncRNA-protein relationships. Performing selective pressure variation analyses on XIST has not been explored much given the lack of species for which the XIST sequences is known. The inclusion of more mammalian species in such phylogenetic analyses would permit more precise mapping of XIST repetitive regions, i.e. tracking which repeats have been conserved, expanded, shrunk or selected against in different placental mammals. Given repetitive regions are known to harbour secondary structure or binding sites, this, in turn could reveal potential binding partner conservation, gain or loss for XIST partners. One hypothesis that has been put forward is that expanding the number of repeat-A monomers on XIST could contribute to more SPEN proteins binding (Brockdorff, 2018). In theory, this could translate to more efficient XCI gene silencing, given SPEN recruits protein complexes to halt gene transcription on X-linked genes. Another example of how a repeat expansion could lead to more effective XCI is that of the E-repeat (Brockdorff, 2018). The E-repeat harbours binding sites for proteins that associate with the nuclear matrix (such as CIZ1) and this association has been shown to be important for proper XIST localisation. Therefore, an increasing number of CIZ1 binding sites on XIST could potentially contribute to more stable binding of XIST by CIZ1, enabling it to spend more time

anchored to the nuclear lamina, instead of diffusing in the nucleus. In turn this would ensure proper XIST localisation to the Xi as well as physical separation from the active X chromosome.

Altogether, this work expands on the knowledge around XIST and its protein partner interactions in eutherian mammals as well as their co-evolution since the inception of the project. More specifically, this project is the first report that describes the expression of a collection of *XIST*'s protein partners in the endometrium of four placental mammals, including cow and pig. Therefore, data presented here can serve as the foundation for future researchers to investigate the roles of these proteins in XCI in a different context. Through robust biochemical pulldown approaches and mass spectrometry, this thesis reports over 70 protein partners of *XIST* across repeats A and E between human and cow uterine-derived cells. Besides confirming interactions with previously characterised protein partners of XIST, novel interactors are also described. This expands the repertoire of XIST's partners which can potentially reveal different biological processes that XIST is involved in post-differentiation. To my knowledge, this is also the first report where protein partners of bovine *XIST* are characterised and perhaps more importantly, this is the first instance where cross-species pulldowns are performed in cow using human XIST RNA. This approach allowed the characterisation of lncRNA-protein interactions outside of their native environment, potentially informing us which proteins may require cofactors for binding. Results here hint at lncRNA-protein partner co-evolution and biochemical assays highlight CIZ1 and hnRNPU as XIST interactors which have been retained in the eutherian lineage (human, mouse and cow in this report). Previous studies have examined whether there is any evidence of selection in two of XIST's protein partners (i.e. SPEN and WTAP). In this project, selective pressure variation analyses included SPEN, LBR, WTAP, RBM15, CIZ1, PTPBP1, hnRNPK, hnRNPU, expanding the number of partners for which we have data available for their evolutionary trajectory.

6.4. Conclusions and future perspectives

Data presented in this thesis supports the hypothesis that XIST is a lncRNA with conserved repetitive regions, and exhibits a dynamic protein interactome with a multitude of proteins binding across different parts of the *XIST* RNA. *XIST* RNA-protein partners were found to vary across species. Data obtained from RIP and *in vitro* transcription pulldown experiments in human and cow endometrial-derived cells, highlighted the protein interactome of *XIST* at the late stages of XCI. Some of the proteins identified in both cow and human datasets have well-characterised roles in XCI from mouse and human studies. Whilst a great degree of overlap was not seen in protein partners of *XIST* between human and cow, this could be a direct result of inconsistencies observed with replicates, lysate preparation and stringency of washing conditions. Future experiments should aim to generate more replicates where a higher amount of nuclear-enriched starting lysate is utilised in pulldown assays to ensure more proteins are detected as well as generate a consensus across replicates (which will improve statistical power).

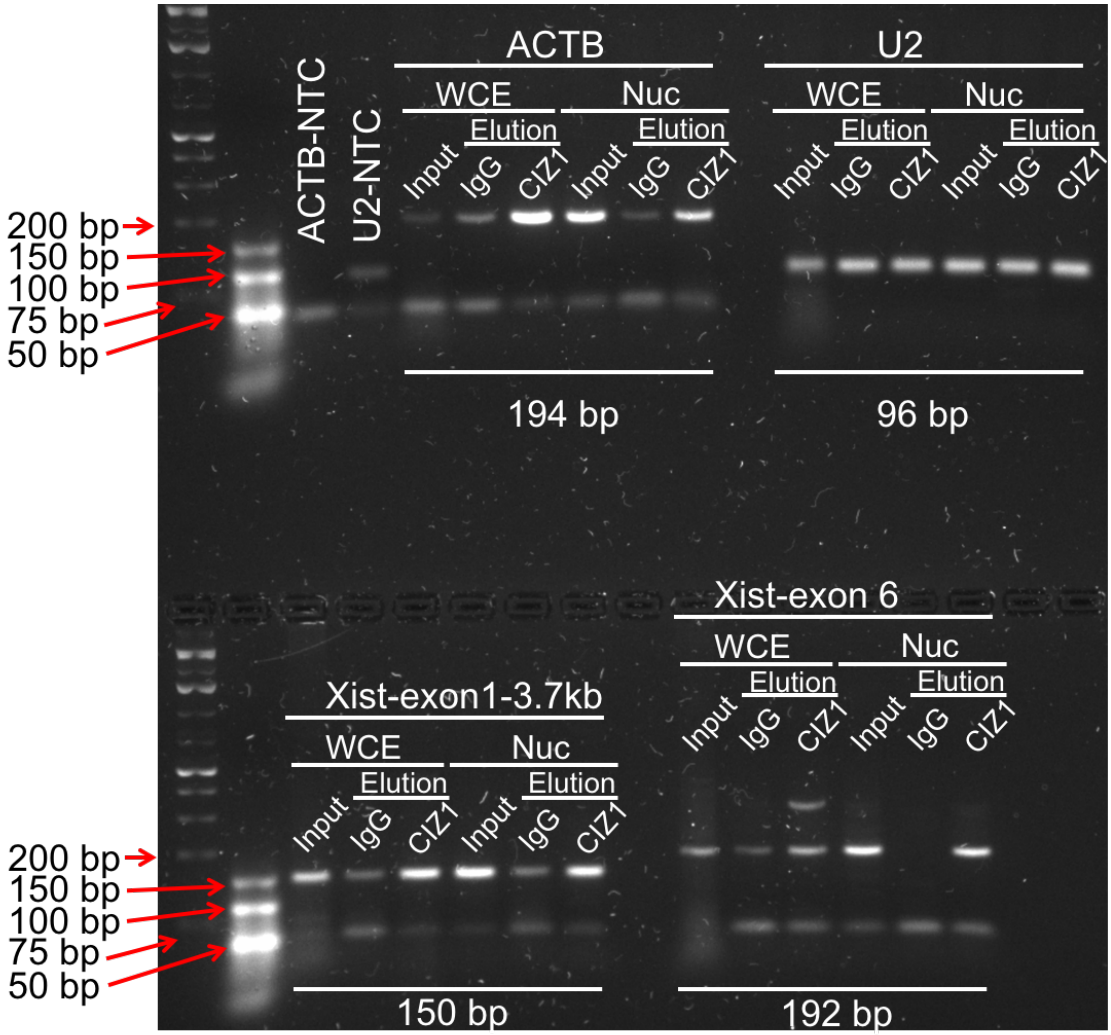
An important question this thesis aimed to address was how a lncRNA rapidly evolving across species can maintain protein partners with a crucial biological function, paramount for embryonic development. *XIST* was used as a model to investigate the co-evolution of functional lncRNA-protein interactions, given it exists across many species and performs the same function across these species. The combination of pulldown approaches and selective pressure variation analyses across the mouse, human, cow and pig lineages revealed a few sites that could be under positive selection in mouse CIZ1, human PTBP1, cow RBM15 and pig SPEN and MATR3. hnRNPK, hnRNPU, and LBR proteins were found to be under purifying selection. Future work to model these changes in a 3-D protein structure, could link between structure to function and address whether these residues are responsible for a functional shift. Future work could also focus on examining whether there are differences in structures that the *XIST* RNA adopts across species and whether structures are also maintained, as these could also have implications in lncRNA-protein partner co-evolution. This could be carried out either computationally via predicting structures with a software such as CROSS (Delli Ponti et al., 2017) or experimentally as has been done for mouse *Xist* via SHAPE (Smola et al., 2016), Psoralen Analysis of RNA Interactions and Structures (PARIS)(Lu et al., 2016) or DMS-seq (Fang et al., 2015).

Exploring whether the role of XCI could have played a role in the divergent evolution of pregnancy is a novel concept, with a clear gap in the literature. As a first step to address this question one would need to first establish whether an XCI-related influence on pregnancy would be causative or correlative and whether XCI could influence pregnancy or vice versa. For instance, following experiments to verify the role of bovine *XIST* protein partners in XCI, the necessity of those proteins for proper embryo development (developmental competency) could be shown via protein loss-of-function and implantation assays or abnormal blastocyst morphological phenotypes.

Data generated in Chapter 4 could be expanded in two ways, one of which would be to expand the investigation to document protein partners of other regions of *XIST* such as repeats B, C and D. No studies so far across placental mammals have examined the role of *XIST* repeat D in XCI. Moreover, some of the proteins found in the cow and human datasets have not been investigated for whether they have both a necessary and sufficient role in an aspect of XCI such as X-linked silencing or *XIST* localisation, despite their consistent characterisation as XIST-interacting. Prior to that, validation of those candidates as true interactors of *XIST* could be carried out via UV-RIP in bovine stromal cells using tagged versions of the proteins or their RNA-binding domains. Therefore, data presented here can serve as the foundation for future researchers to investigate the roles of these proteins firstly in XCI, and secondly in gestation evolution across placental mammals.

XIST is a lncRNA that is indispensable for proper embryogenesis across placental mammals as it orchestrates XCI. This thesis reports over 70 protein partners of XIST across repeats A and E between human and cow uterine-derived cells. To my knowledge, this is the first report where protein partners of bovine *XIST* are characterised and the first instance where XIST protein partners are presented in a uterine-derived cells and tissues. A combination of RNA and protein pulldown approaches employed found CIZ1 and hnRNPU to be interacting with XIST in human and cow pointing to potentially conserved protein partners of *XIST* across placental mammals. Selective pressure variation analyses undertaken for nine putative protein partners of *XIST* determined that differential binding observed across human and

bovine *XIST* is not attributed to a functional loss of *XIST* RNA-binding in the proteins as a result of positive selection. More work can be focused on exploring the relationship between positively selected site predictions and their effect on protein structure models, perhaps examining binding kinetics. Finally, cross-species pulldowns of bovine *XIST* in a human endometrial cell line identified several proteins which could be co-evolving to maintain an interaction between human and bovine *XIST*.



<u>Appendix I.</u> RT-PCR of CIZ1 RIP from whole cell and nuclear extract of ISHIKAWA cells. IgG is a non-specific control in RIP experiments. *ACTB* serves as a specific interacting transcript positive control whereas *U2* served as a non-specific transcript negative control. N=1 biological replicate. Numbers given indicate expected PCR product size (bp). RIP, RNA immunoprecipitation; RT-qPCR, reverse transcription quantitative PCR; NTC, no template (water) control.

Supplementary Information

s

References

Almeida, M., Pintacuda, G., Masui, O., Koseki, Y., Gdula, M., Cerase, A., Brown, D., Mould, A., Innocent, C., Nakayama, M., Schermelleh, L., Nesterova, T. B., Koseki, H. & Brockdorff, N. (2017). PCGF3/5-PRC1 initiates Polycomb recruitment in X chromosome inactivation. *Science,* 356**,** 1081-1084.

Altenhoff, A. M., Glover, N. M., Train, C.-M., Kaleb, K., Warwick vesztrocy, A., Dylus, D., De farias, T. M., Zile, K., Stevenson, C., Long, J., Redestig, H., Gonnet, G. H. & Dessimoz, C. (2018). The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Research,* 46**,** D477-D485.

Amodio, N., Raimondi, L., Juli, G., Stamato, M. A., Caracciolo, D., Tagliaferri, P. & Tassone, P. (2018). MALAT1: a druggable long non-coding RNA for targeted anti-cancer approaches. *Journal of Hematology & Oncology,* 11**,** 63.

Anisimova, M., Bielawski, J. P. & Yang, Z. (2001). Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol Biol Evol,* 18**,** 1585-92.

Anisimova, M., Bielawski, J. P. & Yang, Z. (2002). Accuracy and Power of Bayes Prediction of Amino Acid Sites Under Positive Selection. *Molecular Biology and Evolution,* 19**,** 950-958.

Anisimova, M. & Yang, Z. (2007). Multiple Hypothesis Testing to Detect Lineages under Positive Selection that Affects Only a Few Sites. *Molecular Biology and Evolution,* 24**,** 1219-1228.

Ard, R., Allshire, R. C. & Marquardt, S. (2017). Emerging Properties and Functional Consequences of Noncoding Transcription. *Genetics,* 207**,** 357-367.

Ariyoshi, M. & Schwabe, J. W. R. (2003). A conserved structural motif reveals the essential transcriptional repression function of Spen proteins and their role in developmental signaling. *Genes & development,* 17**,** 1909-1920.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* 25**,** 25-9.

Azam, S., Hou, S., Zhu, B., Wang, W., Hao, T., Bu, X., Khan, M. & Lei, H. (2019). Nuclear retention element recruits U1 snRNP components to restrain spliced lncRNAs in the nucleus. *RNA Biol,* 16**,** 1001-1009.

Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. (2007). Quantitative mass spectrometry in proteomics: a critical review. *Analytical and Bioanalytical Chemistry,* 389**,** 1017-1031.

Barra, J., Gaidosh, G. S., Blumenthal, E., Beckedorff, F., Tayari, M. M., Kirstein, N., Karakach, T. K., Jensen, T. H., Impens, F., Gevaert, K., Leucci, E., Shiekhattar, R. & Marine, J.-C. (2020). Integrator restrains paraspeckles assembly by promoting isoform switching of the lncRNA <em>NEAT1</em&gt. *Science Advances,* 6**,** eaaz9072.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A. A., Kim, S., Wilson, C. J., Lehár, J., Kryukov, G. V., Sonkin, D., Reddy, A., Liu, M., Murray, L., Berger, M. F., Monahan, J. E., Morais, P., Meltzer, J., Korejwa, A., Jané-Valbuena, J., Mapa, F. A., Thibault, J., Bric-Furlong, E., Raman, P., Shipway, A., Engels, I. H., Cheng, J., Yu, G. K.,

Yu, J., Aspesi, P., De Silva, M., Jagtap, K., Jones, M. D., Wang, L., Hatton, C., Palescandolo, E., Gupta, S., Mahan, S., Sougnez, C., Onofrio, R. C., Liefeld, T., Macconaill, L., Winckler, W., Reich, M., Li, N., Mesirov, J. P., Gabriel, S. B., Getz, G., Ardlie, K., Chan, V., Myer, V. E., Weber, B. L., Porter, J., Warmuth, M., Finan, P., Harris, J. L., Meyerson, M., Golub, T. R., Morrissey, M. P., Sellers, W. R., Schlegel, R. & Garraway, L. A. 2012. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* [Online], 483. Available: https://doi.org/10.1038/nature11003 [Accessed 2012/03//].

Berg, D. K., Van Leeuwen, J., Beaumont, S., Berg, M. & Pfeffer, P. L. (2010). Embryo loss in cattle between Days 7 and 16 of pregnancy. *Theriogenology,* 73**,** 250-260.

Berg, D. K., Smith, C. S., Pearton, D. J., Wells, D. N., Broadhurst, R., Donnison, M. & Pfeffer, P. L. (2011). Trophectoderm Lineage Determination in Cattle. *Developmental Cell,* 20**,** 244-255.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research,* 28**,** 235-242.

Björnerfeldt, S., Webster, M. T. & Vilà, C. (2006). Relaxation of selective constraint on dog mitochondrial DNA following domestication. *Genome research,* 16**,** 990-994.

Blackburne, B. P. & Whelan, S. (2012). Measuring the distance between multiple sequence alignments. *Bioinformatics,* 28**,** 495-502.

Blomberg, S. P. & Garland Jr, T. (2002). Tempo and mode in evolution: phylogenetic inertia, adaptation and comparative methods. *Journal of Evolutionary Biology,* 15**,** 899-910.

Blum, M., Chang, H. Y., Chuguransky, S., Grego, T., Kandasaamy, S., Mitchell, A., Nuka, G., Paysan-Lafosse, T., Qureshi, M., Raj, S., Richardson, L., Salazar, G. A., Williams, L., Bork, P., Bridge, A., Gough, J., Haft, D. H., Letunic, I., Marchler-Bauer, A., Mi, H., Natale, D. A., Necci, M., Orengo, C. A., Pandurangan, A. P., Rivoire, C., Sigrist, C. J. A., Sillitoe, I., Thanki, N., Thomas, P. D., Tosatto, S. C. E., Wu, C. H., Bateman, A. & Finn, R. D. (2021). The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res,* 49**,** D344-d354.

Boehm, V., Britto-Borges, T., Steckelberg, A.-L., Singh, K. K., Gerbracht, J. V., Gueney, E., Blazquez, L., Altmüller, J., Dieterich, C. & Gehring, N. H. (2018). Exon Junction Complexes Suppress Spurious Splice Sites to Safeguard Transcriptome Integrity. *Molecular Cell,* 72**,** 482-495.e7.

Bofkin, L. & Goldman, N. (2007). Variation in Evolutionary Processes at Different Codon Positions. *Molecular Biology and Evolution,* 24**,** 513-521.

Booker, T. R., Jackson, B. C. & Keightley, P. D. (2017). Detecting positive selection in the genome. *BMC Biology,* 15**,** 98.

Bose, D. A., Donahue, G., Reinberg, D., Shiekhattar, R., Bonasio, R. & Berger, S. L. (2017). RNA Binding to CBP Stimulates Histone Acetylation and Transcription. *Cell,* 168**,** 135-149.e22.

Bou, G., Liu, S., Sun, M., Zhu, J., Xue, B., Guo, J., Zhao, Y., Qu, B., Weng, X., Wei, Y., Lei, L. & Liu, Z. (2017). CDX2 is essential for cell proliferation and polarity in porcine blastocysts. *Development,* 144**,** 1296-1306.

Bousard, A., Raposo, A. C., Żylicz, J. J., Picard, C., Pires, V. B., Qi, Y., Gil, C., Syx, L., Chang, H. Y., Heard, E. & Da Rocha, S. T. (2019). The role of Xist-mediated Polycomb recruitment in the initiation of X-chromosome inactivation. *EMBO reports,* 20**,** e48019.

Brimacombe, R., Stiege, W., Kyriatsoulis, A. & Maly, P. (1988). Intra-RNA and RNA-protein cross-linking techniques in Escherichia coli ribosomes. *Methods in enzymology,* 164**,** 287-309.

Brockdorff, N., Ashworth, A., Kay, G. F., Mccabe, V. M., Norris, D. P., Cooper, P. J., Swift, S. & Rastan, S. (1992). The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. *Cell,* 71**,** 515-26.

Brockdorff, N. (2018). Local Tandem Repeat Expansion in Xist RNA as a Model for the Functionalisation of ncRNA. *Noncoding RNA,* 4.

Bromham, L. & Penny, D. (2003). The modern molecular clock. *Nat Rev Genet,* 4**,** 216-24.

Brown, C. J., Hendrich, B. D., Rupert, J. L., Lafreniere, R. G., Xing, Y., Lawrence, J. & Willard, H. F. (1992). The human XIST gene: analysis of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell,* 71**,** 527-42.

Brown, C. J. & Baldry, S. E. (1996). Evidence that heteronuclear proteins interact with XIST RNA in vitro. *Somat Cell Mol Genet,* 22**,** 403-17.

Brown, J. A., Valenstein, M. L., Yario, T. A., Tycowski, K. T. & Steitz, J. A. (2012). Formation of triple-helical structures by the 3'-end sequences of MALAT1 and MENβ noncoding RNAs. *Proceedings of the National Academy of Sciences of the United States of America,* 109**,** 19202-19207.

Brown, N. P., Leroy, C. & Sander, C. (1998). MView: a web-compatible database search or multiple alignment viewer. *Bioinformatics,* 14**,** 380-381.

Brown, R. S. (2005). Zinc finger proteins: getting a grip on RNA. *Current Opinion in Structural Biology,* 15**,** 94-98.

Brown, W. M., Prager, E. M., Wang, A. & Wilson, A. C. (1982). Mitochondrial DNA sequences of primates: tempo and mode of evolution. *J Mol Evol,* 18**,** 225-39.

Bustamante, C. D., Fledel-Alon, A., Williamson, S., Nielsen, R., Todd Hubisz, M., Glanowski, S., Tanenbaum, D. M., White, T. J., Sninsky, J. J., Hernandez, R. D., Civello, D., Adams, M. D., Cargill, M. & Clark, A. G. (2005). Natural selection on protein-coding genes in the human genome. *Nature,* 437**,** 1153-1157.

Cabili, M. N., Trapnell, C., Goff, L., Koziol, M., Tazon-Vega, B., Regev, A. & Rinn, J. L. (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes & development,* 25**,** 1915-1927.

Calderoni, L., Rota-Stabelli, O., Frigato, E., Panziera, A., Kirchner, S., Foulkes, N. S., Kruckenhauser, L., Bertolucci, C. & Fuselli, S. (2016). Relaxed selective constraints drove functional modifications in peripheral photoreception of the cavefish P. andruzzii and provide insight into the time of cave colonization. *Heredity,* 117**,** 383-392.

Cao, H., Wahlestedt, C. & Kapranov, P. (2018). Strategies to Annotate and Characterize Long Noncoding RNAs: Advantages and Pitfalls. *Trends Genet,* 34**,** 704-721.

Caparros, M. L., Alexiou, M., Webster, Z. & Brockdorff, N. (2002). Functional analysis of the highly conserved exon IV of Xist RNA. *Cytogenetic and Genome Research,* 99**,** 99-105.

Carlevaro-Fita, J. & Johnson, R. (2019). Global Positioning System: Understanding Long Noncoding RNAs through Subcellular Localization. *Mol Cell,* 73**,** 869-883.

Carlevaro-Fita, J., Polidori, T., Das, M., Navarro, C., Zoller, T. I. & Johnson, R. (2019). Ancient exapted transposable elements promote nuclear enrichment of human long noncoding RNAs. *Genome research,* 29**,** 208-222.

Carlevaro-Fita, J., Lanzós, A., Feuerbach, L., Hong, C., Mas-Ponte, D., Pedersen, J. S., Drivers, P., Functional Interpretation, G., Johnson, R. & Consortium, P. (2020). Cancer LncRNA Census reveals evidence for deep functional conservation of long noncoding RNAs in tumorigenesis. *Communications biology,* 3**,** 56-56.

Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C., Kodzius, R., Shimokawa, K., Bajic, V. B., Brenner, S. E., Batalov, S., Forrest, A. R., Zavolan, M., Davis, M. J., Wilming, L. G., Aidinis, V., Allen, J. E., Ambesi-Impiombato, A., Apweiler, R., Aturaliya, R. N., Bailey, T. L., Bansal, M., Baxter, L., Beisel, K. W., Bersano, T., Bono, H., Chalk, A. M., Chiu, K. P., Choudhary, V., Christoffels, A., Clutterbuck, D. R., Crowe, M. L., Dalla, E., Dalrymple, B. P., De Bono, B., Della Gatta, G., Di Bernardo, D., Down, T., Engstrom, P., Fagiolini, M., Faulkner, G., Fletcher, C. F., Fukushima, T., Furuno, M., Futaki, S., Gariboldi, M., Georgii-Hemming, P., Gingeras, T. R., Gojobori, T., Green, R. E., Gustincich, S., Harbers, M., Hayashi, Y., Hensch, T. K., Hirokawa, N., Hill, D., Huminiecki, L., Iacono, M., Ikeo, K., Iwama, A., Ishikawa, T., Jakt, M., Kanapin, A., Katoh, M., Kawasawa, Y., Kelso, J., Kitamura, H., Kitano, H., Kollias, G., Krishnan, S. P., Kruger, A., Kummerfeld, S. K., Kurochkin, I. V., Lareau, L. F., Lazarevic, D., Lipovich, L., Liu, J., Liuni, S., Mcwilliam, S., Madan Babu, M., Madera, M., Marchionni, L., Matsuda, H., Matsuzawa, S., Miki, H., Mignone, F., Miyake, S., Morris, K., Mottagui-Tabar, S., Mulder, N., Nakano, N., Nakauchi, H., Ng, P., Nilsson, R., Nishiguchi, S., Nishikawa, S., et al. (2005). The transcriptional landscape of the mammalian genome. *Science,* 309**,** 1559-63.

Carrieri, C., Cimatti, L., Biagioli, M., Beugnet, A., Zucchelli, S., Fedele, S., Pesce, E., Ferrer, I., Collavin, L., Santoro, C., Forrest, A. R., Carninci, P., Biffo, S., Stupka, E. & Gustincich, S. (2012). Long non-coding antisense RNA controls Uchl1 translation through an embedded SINEB2 repeat. *Nature,* 491**,** 454-7.

Carter, A. C., Xu, J., Nakamoto, M. Y., Wei, Y., Zarnegar, B. J., Shi, Q., Broughton, J. P., Ransom, R. C., Salhotra, A., Nagaraja, S. D., Li, R., Dou, D. R., Yost, K. E., Cho, S.-W., Mistry, A., Longaker, M. T., Khavari, P. A., Batey, R. T., Wuttke, D. S. & Chang, H. Y. (2020). Spen links RNA-mediated endogenous retrovirus silencing and X chromosome inactivation. *eLife,* 9**,** e54508.

Chamary, J. V., Parmley, J. L. & Hurst, L. D. (2006). Hearing silence: non-neutral evolution at synonymous sites in mammals. *Nature Reviews Genetics,* 7**,** 98-108.

Chao, J. A., Patskovsky, Y., Patel, V., Levy, M., Almo, S. C. & Singer, R. H. (2010). ZBP1 recognition of beta-actin zipcode induces RNA looping. *Genes & development,* 24**,** 148-158.

Chapman, A. G., Cotton, A. M., Kelsey, A. D. & Brown, C. J. (2014). Differentially methylated CpG island within human XIST mediates alternative P2 transcription and YY1 binding. *BMC Genetics,* 15**,** 89.

Chen, C.-Y., Shi, W., Balaton, B. P., Matthews, A. M., Li, Y., Arenillas, D. J., Mathelier, A., Itoh, M., Kawaji, H., Lassmann, T., Hayashizaki, Y., Carninci, P., Forrest, A. R. R., Brown, C. J. & Wasserman, W. W. (2016a). YY1 binding association with sex-biased transcription revealed through X-linked transcript levels and allelic binding analyses. *Scientific Reports,* 6**,** 37324.

Chen, C. K., Blanco, M., Jackson, C., Aznauryan, E., Ollikainen, N., Surka, C., Chow, A., Cerase, A., Mcdonel, P. & Guttman, M. (2016b). Xist recruits the X chromosome to the nuclear lamina to enable chromosome-wide silencing. *Science,* 354**,** 468-472.

Chen, G. I. & Gingras, A.-C. (2007). Affinity-purification mass spectrometry (AP-MS) of serine/threonine phosphatases. *Methods,* 42**,** 298-305.

Chen, M.-Y., Liang, D. & Zhang, P. (2017). Phylogenomic Resolution of the Phylogeny of Laurasiatherian Mammals: Exploring Phylogenetic Signals within Coding and Noncoding Sequences. *Genome Biology and Evolution,* 9**,** 1998-2012.

Chen, Z., Hagen, D. E., Wang, J., Elsik, C. G., Ji, T., Siqueira, L. G., Hansen, P. J. & Rivera, R. M. (2016c). Global assessment of imprinted gene expression in the bovine conceptus by next generation sequencing. *Epigenetics,* 11**,** 501-16.

Chodroff, R. A., Goodstadt, L., Sirey, T. M., Oliver, P. L., Davies, K. E., Green, E. D., Molnár, Z. & Ponting, C. P. (2010). Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. *Genome biology,* 11**,** R72-R72.

Chu, C., Zhang, Qiangfeng c., Da rocha, Simão t., Flynn, Ryan a., Bharadwaj, M., Calabrese, J. M., Magnuson, T., Heard, E. & Chang, Howard y. (2015a). Systematic Discovery of Xist RNA Binding Proteins. *Cell,* 161**,** 404-416.

Chu, C., Zhang, Q. C., Da Rocha, S. T., Flynn, R. A., Bharadwaj, M., Calabrese, J. M., Magnuson, T., Heard, E. & Chang, H. Y. (2015b). Systematic discovery of Xist RNA binding proteins. *Cell,* 161**,** 404-16.

Cirillo, D., Blanco, M., Armaos, A., Buness, A., Avner, P., Guttman, M., Cerase, A. & Tartaglia, G. G. (2016). Quantitative predictions of protein interactions with long noncoding RNAs. *Nat Methods,* 14**,** 5-6.

Clark, M. B., Johnston, R. L., Inostroza-Ponta, M., Fox, A. H., Fortini, E., Moscato, P., Dinger, M. E. & Mattick, J. S. (2012). Genome-wide analysis of long noncoding RNA stability. *Genome research,* 22**,** 885-898.

Clemson, C. M., Mcneil, J. A., Willard, H. F. & Lawrence, J. B. (1996). XIST RNA paints the inactive X chromosome at interphase: evidence for a novel RNA involved in nuclear/chromosome structure. *The Journal of cell biology,* 132**,** 259-275.

Coelho, M. B., Attig, J., Bellora, N., König, J., Hallegger, M., Kayikci, M., Eyras, E., Ule, J. & Smith, C. W. (2015). Nuclear matrix protein Matrin3 regulates alternative splicing and forms overlapping regulatory networks with PTB. *Embo j,* 34**,** 653-68.

Cohen, H. R. & Panning, B. (2007). XIST RNA exhibits nuclear retention and exhibits reduced association with the export factor TAP/NXF1. *Chromosoma,* 116**,** 373-383.

Coker, H., Wei, G., Moindrot, B., Mohammed, S., Nesterova, T. & Brockdorff, N. (2020). The role of the Xist 5? m6A region and RBM15 in X chromosome inactivation [version 1; peer review: 1 approved, 2 approved with reservations]. *Wellcome Open Research,* 5.

Comings, D. E. (1968). The rationale for an ordered arrangement of chromatin in the interphase nucleus. *American journal of human genetics,* 20**,** 440-460.

Couldrey, C., Johnson, T., Lopdell, T., Zhang, I. L., Littlejohn, M. D., Keehan, M., Sherlock, R. G., Tiplady, K., Scott, A., Davis, S. R. & Spelman, R. J. (2017). Bovine mammary gland X chromosome inactivation. *Journal of Dairy Science,* 100**,** 5491-5500.

Cranston, K. A., Hurwitz, B., Ware, D., Stein, L. & Wing, R. A. (2009). Species Trees from Highly Incongruent Gene Trees in Rice. *Systematic Biology,* 58**,** 489-500.

Creamer, K. M. & Lawrence, J. B. (2017). XIST RNA: a window into the broader role of RNA in nuclear chromosome architecture. *Philos Trans R Soc Lond B Biol Sci,* 372.

Creevey, C. J. & Mcinerney, J. O. (2005). Clann: investigating phylogenetic information through supertree analyses. *Bioinformatics,* 21**,** 390-2.

Csankovszki, G., Panning, B., Bates, B., Pehrson, J. R. & Jaenisch, R. (1999). Conditional deletion of Xist disrupts histone macroH2A localization but not maintenance of X inactivation. *Nat Genet,* 22**,** 323-4.

Daigneault, B. W., Rajput, S., Smith, G. W. & Ross, P. J. (2018). Embryonic POU5F1 is Required for Expanded Bovine Blastocyst Formation. *Scientific Reports,* 8**,** 7753.

Darnell, R. B. (2010). HITS-CLIP: panoramic views of protein-RNA regulation in living cells. *Wiley Interdiscip Rev RNA,* 1**,** 266-86.

Daub, J. T., Hofer, T., Cutivet, E., Dupanloup, I., Quintana-Murci, L., Robinson-Rechavi, M. & Excoffier, L. (2013). Evidence for Polygenic Adaptation to Pathogens in the Human Genome. *Molecular Biology and Evolution,* 30**,** 1544-1558.

Dayon, L., Hainard, A., Licker, V., Turck, N., Kuhn, K., Hochstrasser, D. F., Burkhard, P. R. & Sanchez, J.-C. (2008). Relative Quantification of Proteins in Human Cerebrospinal Fluids by MS/MS Using 6-Plex Isobaric Tags. *Analytical Chemistry,* 80**,** 2921-2931.

De La Fuente, R., Hahnel, A., Basrur, P. K. & King, W. A. (1999). X inactive-specific transcript (Xist) expression and X chromosome inactivation in the preattachment bovine embryo. *Biol Reprod,* 60**,** 769-75.

Delli Ponti, R., Marti, S., Armaos, A. & Tartaglia, G. G. (2017). A high-throughput approach to profile RNA structure. *Nucleic Acids Res,* 45**,** e35.

Deng, X., Berletch, J. B., Nguyen, D. K. & Disteche, C. M. (2014). X chromosome regulation: diverse patterns in development, tissues and disease. *Nature reviews. Genetics,* 15**,** 367-378.

Derrien, T., Johnson, R., Bussotti, G., Tanzer, A., Djebali, S., Tilgner, H., Guernec, G., Martin, D., Merkel, A., Knowles, D. G., Lagarde, J., Veeravalli, L., Ruan, X., Ruan, Y., Lassmann, T., Carninci, P., Brown, J. B., Lipovich, L., Gonzalez, J. M., Thomas, M., Davis, C. A., Shiekhattar, R., Gingeras, T. R., Hubbard, T. J., Notredame, C., Harrow, J. & Guigo, R. (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res,* 22**,** 1775-89.

Deuve, J. L., Bonnet-Garnier, A., Beaujean, N., Avner, P. & Morey, C. (2015). Antagonist Xist and Tsix co-transcription during mouse oogenesis and maternal Xist expression during pre-implantation development calls into question the nature of the maternal imprint on the X chromosome. *Epigenetics,* 10**,** 931-942.

Dey, A., Adithi, V. R. & Chatterji, D. (2012). Co-evolution of RNA polymerase with RbpA in the phylum Actinobacteria. *Applied & translational genomics,* 1**,** 9-20.

Di Giulio, M. (1997). On the RNA World: Evidence in Favor of an Early Ribonucleopeptide World. *Journal of Molecular Evolution,* 45**,** 571-578.

Dindot, S. V., Kent, K. C., Evers, B., Loskutoff, N., Womack, J. & Piedrahita, J. A. (2004). Conservation of genomic imprinting at the XIST, IGF2, and GTL2 loci in the bovine. *Mammalian Genome,* 15**,** 966-974.

Dixon-Mcdougall, T. & Brown, C. J. (2021). Independent domains for recruitment of PRC1 and PRC2 by human XIST. *PLOS Genetics,* 17**,** e1009123.

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Bar, N. S., Batut, P., Bell, K., Bell, I., Chakrabortty, S., Chen, X., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Falconnet, E., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Gunawardena, H., Howald, C., Jha,

S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Luo, O. J., Park, E., Persaud, K., Preall, J. B., Ribeca, P., Risk, B., Robyr, D., Sammeth, M., Schaffer, L., See, L.-H., Shahab, A., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, H., Wrobel, J., Yu, Y., Ruan, X., Hayashizaki, Y., Harrow, J., Gerstein, M., Hubbard, T., Reymond, A., Antonarakis, S. E., Hannon, G., Giddings, M. C., Ruan, Y., Wold, B., Carninci, P., Guigó, R. & Gingeras, T. R. (2012). Landscape of transcription in human cells. *Nature,* 489**,** 101-108.

Dominguez, D., Freese, P., Alexis, M. S., Su, A., Hochman, M., Palden, T., Bazile, C., Lambert, N. J., Van Nostrand, E. L., Pratt, G. A., Yeo, G. W., Graveley, B. R. & Burge, C. B. (2018). Sequence, Structure, and Context Preferences of Human RNA Binding Proteins. *Molecular cell,* 70**,** 854-867.e9.

Dossin, F., Pinheiro, I., Żylicz, J. J., Roensch, J., Collombet, S., Le Saux, A., Chelmicki, T., Attia, M., Kapoor, V., Zhan, Y., Dingli, F., Loew, D., Mercher, T., Dekker, J. & Heard, E. (2020). SPEN integrates transcriptional and epigenetic control of X-inactivation. *Nature,* 578**,** 455-460.

Dunham, I., Kundaje, A., Aldred, S. F., Collins, P. J., Davis, C. A., Doyle, F., Epstein, C. B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B. R., Landt, S. G., Lee, B.-K., Pauli, F., Rosenbloom, K. R., Sabo, P., Safi, A., Sanyal, A., Shoresh, N., Simon, J. M., Song, L., Trinklein, N. D., Altshuler, R. C., Birney, E., Brown, J. B., Cheng, C., Djebali, S., Dong, X., Dunham, I., Ernst, J., Furey, T. S., Gerstein, M., Giardine, B., Greven, M., Hardison, R. C., Harris, R. S., Herrero, J., Hoffman, M. M., Iyer, S., Kellis, M., Khatun, J., Kheradpour, P., Kundaje, A., Lassmann, T., Li, Q., Lin, X., Marinov, G. K., Merkel, A., Mortazavi, A., Parker, S. C. J., Reddy, T. E., Rozowsky, J., Schlesinger, F., Thurman, R. E., Wang, J., Ward, L. D., Whitfield, T. W., Wilder, S. P., Wu, W., Xi, H. S., Yip, K. Y., Zhuang, J., Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., Snyder, M., Pazin, M. J., Lowdon, R. F., Dillon, L. a. L., Adams, L. B., Kelly, C. J., Zhang, J., Wexler, J. R., Green, E. D., Good, P. J., Feingold, E. A., Bernstein, B. E., Birney, E., Crawford, G. E., Dekker, J., Elnitski, L., Farnham, P. J., Gerstein, M., Giddings, M. C., Gingeras, T. R., Green, E. D., Guigó, R., Hardison, R. C., Hubbard, T. J., Kellis, M., Kent, W. J., Lieb, J. D., Margulies, E. H., Myers, R. M., Snyder, M., Stamatoyannopoulos, J. A., Tenenbaum, S. A., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature,* 489**,** 57-74.

Duret, L., Chureau, C., Samain, S., Weissenbach, J. & Avner, P. (2006). The Xist RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science,* 312**,** 1653-5.

Duszczyk, M. M., Wutz, A., Rybin, V. & Sattler, M. (2011). The Xist RNA A-repeat comprises a novel AUCG tetraloop fold and a platform for multimerization. *Rna,* 17**,** 1973-82.

Duvaud, S., Gabella, C., Lisacek, F., Stockinger, H., Ioannidis, V. & Durinx, C. (2021). Expasy, the Swiss Bioinformatics Resource Portal, as designed by its users. *Nucleic Acids Research,* 49**,** W216-W227.

Dyer, K. A., Canfield, T. K. & Gartler, S. M. (1989). Molecular cytological differentiation of active from inactive X domains in interphase: implications for X chromosome inactivation. *Cytogenet Cell Genet,* 50**,** 116-20.

Elisaphenko, E. A., Kolesnikov, N. N., Shevchenko, A. I., Rogozin, I. B., Nesterova, T. B., Brockdorff, N. & Zakian, S. M. (2008). A dual origin of the Xist gene from a protein-coding gene and a set of transposable elements. *PLoS One,* 3**,** e2521.

Elsik, C. G., Tellam, R. L., Worley, K. C., Gibbs, R. A., Muzny, D. M., Weinstock, G. M., Adelson, D. L., Eichler, E. E., Elnitski, L., Guigó, R., Hamernik, D. L., Kappes, S. M., Lewin, H. A., Lynn, D. J., Nicholas, F. W., Reymond, A., Rijnkels, M., Skow, L. C., Zdobnov, E. M., Schook, L., Womack, J., Alioto, T., Antonarakis, S. E., Astashyn, A., Chapple, C. E., Chen, H. C., Chrast, J., Câmara, F., Ermolaeva, O., Henrichsen, C. N., Hlavina, W., Kapustin, Y., Kiryutin, B., Kitts, P., Kokocinski, F., Landrum, M., Maglott, D., Pruitt, K., Sapojnikov, V., Searle, S. M., Solovyev, V., Souvorov, A., Ucla, C., Wyss, C., Anzola, J. M., Gerlach, D., Elhaik, E., Graur, D., Reese, J. T., Edgar, R. C., Mcewan, J. C., Payne, G. M., Raison, J. M., Junier, T., Kriventseva, E. V., Eyras, E., Plass, M., Donthu, R., Larkin, D. M., Reecy, J., Yang, M. Q., Chen, L., Cheng, Z., Chitko-Mckown, C. G., Liu, G. E., Matukumalli, L. K., Song, J., Zhu, B., Bradley, D. G., Brinkman, F. S., Lau, L. P., Whiteside, M. D., Walker, A., Wheeler, T. T., Casey, T., German, J. B., Lemay, D. G., Maqbool, N. J., Molenaar, A. J., Seo, S., Stothard, P., Baldwin, C. L., Baxter, R., Brinkmeyer-Langford, C. L., Brown, W. C., Childers, C. P., Connelley, T., Ellis, S. A., Fritz, K., Glass, E. J., Herzig, C. T., Iivanainen, A., Lahmers, K. K., Bennett, A. K., Dickens, C. M., Gilbert, J. G., Hagen, D. E., Salih, H., Aerts, J., Caetano, A. R., et al. (2009). The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science,* 324**,** 522-8.

Endo, T., Ikeo, K. & Gojobori, T. (1996). Large-scale search for genes on which positive selection may operate. *Mol Biol Evol,* 13**,** 685-90.

Engreitz, J. M., Haines, J. E., Perez, E. M., Munson, G., Chen, J., Kane, M., Mcdonel, P. E., Guttman, M. & Lander, E. S. (2016). Local regulation of gene expression by lncRNA promoters, transcription and splicing. *Nature,* 539**,** 452-455.

Faghihi, M. A., Modarresi, F., Khalil, A. M., Wood, D. E., Sahagan, B. G., Morgan, T. E., Finch, C. E., St Laurent, G., 3rd, Kenny, P. J. & Wahlestedt, C. (2008). Expression of a noncoding RNA is elevated in Alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nature medicine,* 14**,** 723-730.

Faghihi, M. A., Zhang, M., Huang, J., Modarresi, F., Van Der Brug, M. P., Nalls, M. A., Cookson, M. R., St-Laurent, G., 3rd & Wahlestedt, C. (2010). Evidence for natural antisense transcript-mediated inhibition of microRNA function. *Genome biology,* 11**,** R56-R56.

Fang, R., Moss, W. N., Rutenberg-Schoenberg, M. & Simon, M. D. (2015). Probing Xist RNA Structure in Cells Using Targeted Structure-Seq. *PLOS Genetics,* 11**,** e1005668.

Fares, M. A., Elena, S. F., Ortiz, J., Moya, A. & Barrio, E. (2002). A Sliding Window-Based Method to Detect Selective Constraints in Protein-Coding Genes and Its Application to RNA Viruses. *Journal of Molecular Evolution,* 55**,** 509-521.

Fares, M. A. & Travers, S. a. A. (2006). A novel method for detecting intramolecular coevolution: adding a further dimension to selective constraints analyses. *Genetics,* 173**,** 9-23.

Fay, J. C., Wyckoff, G. J. & Wu, C.-I. (2002). Testing the neutral theory of molecular evolution with genomic data from Drosophila. *Nature,* 415**,** 1024-1026.

Fay, J. C. (2011). Weighing the evidence for adaptation at the molecular level. *Trends in genetics : TIG,* 27**,** 343-349.

Fecko, C. J., Munson, K. M., Saunders, A., Sun, G., Begley, T. P., Lis, J. T. & Webb, W. W. (2007). Comparison of femtosecond laser and continuous wave UV sources for protein-nucleic acid crosslinking. *Photochem Photobiol,* 83**,** 1394-404.

Fitch, W. M. & Markowitz, E. (1970). An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochemical Genetics,* 4**,** 579-593.

Fletcher, W. & Yang, Z. (2010). The Effect of Insertions, Deletions, and Alignment Errors on the Branch-Site Test of Positive Selection. *Molecular Biology and Evolution,* 27**,** 2257-2267.

Fogarty, N. M. E., Mccarthy, A., Snijders, K. E., Powell, B. E., Kubikova, N., Blakeley, P., Lea, R., Elder, K., Wamaitha, S. E., Kim, D., Maciulyte, V., Kleinjung, J., Kim, J.-S., Wells, D., Vallier, L., Bertero, A., Turner, J. M. A. & Niakan, K. K. (2017). Genome editing reveals a role for OCT4 in human embryogenesis. *Nature,* 550**,** 67.

Foote, A. D., Liu, Y., Thomas, G. W. C., Vinař, T., Alföldi, J., Deng, J., Dugan, S., Van Elk, C. E., Hunter, M. E., Joshi, V., Khan, Z., Kovar, C., Lee, S. L., Lindblad-Toh, K., Mancia, A., Nielsen, R., Qin, X., Qu, J., Raney, B. J., Vijay, N., Wolf, J. B. W., Hahn, M. W., Muzny, D. M., Worley, K. C., Gilbert, M. T. P. & Gibbs, R. A. (2015). Convergent evolution of the genomes of marine mammals. *Nature Genetics,* 47**,** 272-275.

Forde, N., Maillo, V., O'gaora, P., Simintiras, C. A., Sturmey, R. G., Ealy, A. D., Spencer, T. E., Gutierrez-Adan, A., Rizos, D. & Lonergan, P. (2016). Sexually Dimorphic Gene Expression in Bovine Conceptuses at the Initiation of Implantation. *Biology of reproduction,* 95**,** 92-92.

Frank, S., Ahuja, G., Bartsch, D., Russ, N., Yao, W., Kuo, J. C., Derks, J. P., Akhade, V. S., Kargapolova, Y., Georgomanolis, T., Messling, J. E., Gramm, M., Brant, L., Rehimi, R., Vargas, N. E., Kuroczik, A., Yang, T. P., Sahito, R. G. A., Franzen, J., Hescheler, J., Sachinidis, A., Peifer, M., Rada-Iglesias, A., Kanduri, M., Costa, I. G., Kanduri, C., Papantonis, A. & Kurian, L. (2019). yylncT Defines a Class of Divergently Transcribed lncRNAs and Safeguards the T-mediated Mesodermal Commitment of Human PSCs. *Cell Stem Cell,* 24**,** 318-327.e8.

Frenkel-Pinter, M., Haynes, J. W., Mohyeldin, A. M., C, M., Sargon, A. B., Petrov, A. S., Krishnamurthy, R., Hud, N. V., Williams, L. D. & Leman, L. J. (2020). Mutually stabilizing interactions between proto-peptides and RNA. *Nature communications,* 11**,** 3137-3137.

Garland, W. & Jensen, T. H. (2020). Nuclear sorting of RNA. *WIREs RNA,* 11**,** e1572.

Gerstberger, S., Hafner, M., Ascano, M. & Tuschl, T. (2014a). Evolutionary conservation and expression of human RNA-binding proteins and their role in human genetic disease. *Advances in experimental medicine and biology,* 825**,** 1-55.

Gerstberger, S., Hafner, M. & Tuschl, T. (2014b). A census of human RNA-binding proteins. *Nature Reviews Genetics,* 15**,** 829-845.

Gharib, W. H. & Robinson-Rechavi, M. (2013). The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in GC. *Molecular biology and evolution,* 30**,** 1675-1686.

Gil, N. & Ulitsky, I. (2020). Regulation of gene expression by cis-acting long non-coding RNAs. *Nature Reviews Genetics,* 21**,** 102-117.

Graindorge, A., Pinheiro, I., Nawrocka, A., Mallory, A. C., Tsvetkov, P., Gil, N., Carolis, C., Buchholz, F., Ulitsky, I., Heard, E., Taipale, M. & Shkumatava, A. (2019). In-cell identification and measurement of RNA-protein interactions. *Nat Commun,* 10**,** 5317.

Gundling, W. E. & Wildman, D. E. (2015). A review of inter- and intraspecific variation in the eutherian placenta. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 370.

Hall, T. M. T. (2005). Multiple modes of RNA recognition by zinc finger proteins. *Current Opinion in Structural Biology,* 15**,** 367-373.

Hallström, B. M., Kullberg, M., Nilsson, M. A. & Janke, A. (2007). Phylogenomic Data Analyses Provide Evidence that Xenarthra and Afrotheria Are Sister Groups. *Molecular Biology and Evolution,* 24**,** 2059-2068.

Harrison, C. A., Turner, D. H. & Hinkle, D. C. (1982). Laser crosslinking of E. coli RNA polymerase and T7 DNA. *Nucleic acids research,* 10**,** 2399-2414.

Hasegawa, Y., Brockdorff, N., Kawano, S., Tsutui, K., Tsutui, K. & Nakagawa, S. (2010). The matrix protein hnRNP U is required for chromosomal localization of Xist RNA. *Dev Cell,* 19**,** 469-76.

He, S., Gu, W., Li, Y. & Zhu, H. (2013). ANRIL/CDKN2B-AS shows two-stage clade-specific evolution and becomes conserved after transposon insertions in simians. *BMC evolutionary biology,* 13**,** 247-247.

Hendrickson, G. D., Kelley, D. R., Tenen, D., Bernstein, B. & Rinn, J. L. (2016). Widespread RNA binding by chromatin-associated proteins. *Genome Biology,* 17**,** 28.

Hezroni, H., Koppstein, D., Schwartz, M. G., Avrutin, A., Bartel, D. P. & Ulitsky, I. (2015). Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell Rep,* 11**,** 1110-22.

Hockensmith, J. W., Kubasek, W. L., Vorachek, W. R. & Von Hippel, P. H. (1986). Laser cross-linking of nucleic acids to proteins. Methodology and first applications to the phage T4 DNA replication system. *The Journal of biological chemistry,* 261**,** 3512-3518.

Holdt, L. M., Hoffmann, S., Sass, K., Langenberger, D., Scholz, M., Krohn, K., Finstermeier, K., Stahringer, A., Wilfert, W., Beutner, F., Gielen, S., Schuler, G., Gäbel, G., Bergert, H., Bechmann, I., Stadler, P. F., Thiery, J. & Teupser, D. (2013). Alu elements in ANRIL non-coding RNA at chromosome 9p21 modulate atherogenic cell functions through trans-regulation of gene networks. *PLoS genetics,* 9**,** e1003588-e1003588.

Horiuchi, K., Kawamura, T., Iwanari, H., Ohashi, R., Naito, M., Kodama, T. & Hamakubo, T. (2013). Identification of Wilms' tumor 1-associating protein complex and its role in alternative splicing and the cell cycle. *J Biol Chem,* 288**,** 33292-302.

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., Da rin fioretto, L., Davidson, C., Dodiya, K., El houdaigui, B., Fatima, R., Gall, A., Garcia giron, C., Grego, T., Guijarro-Clarke, C., Haggerty, L., Hemrom, A., Hourlier, T., Izuogu, O. G., Juettemann, T., Kaikala, V., Kay, M., Lavidas, I., Le, T., Lemos, D., Gonzalez martinez, J., Marugán, J. C., Maurel, T., Mcmahon, A. C., Mohanan, S., Moore, B., Muffato, M., Oheh, D. N., Paraschas, D., Parker, A., Parton, A., Prosovetskaia, I., Sakthivel, M. P., Salam, Ahamed i a., Schmitt, B. M., Schuilenburg, H., Sheppard, D., Steed, E., Szpak, M., Szuba, M., Taylor, K., Thormann, A., Threadgold, G., Walts, B., Winterbottom, A., Chakiachvili, M., Chaubal, A., De silva, N., Flint, B., Frankish, A., Hunt, S. E., Iisley, G. R., Langridge, N., Loveland, J. E., Martin, F. J., Mudge, J. M., Morales, J., Perry, E., Ruffier, M., Tate, J., Thybert, D., Trevanion, S. J., Cunningham, F., Yates, A. D., Zerbino, D. R. & Flicek, P. (2021). Ensembl 2021. *Nucleic Acids Research,* 49**,** D884-D891.

Hughes, A. L. (2007). Looking for Darwin in all the wrong places: the misguided quest for positive selection at the nucleotide sequence level. *Heredity,* 99**,** 364-373.

Hung, T., Wang, Y., Lin, M. F., Koegel, A. K., Kotake, Y., Grant, G. D., Horlings, H. M., Shah, N., Umbricht, C., Wang, P., Wang, Y., Kong, B., Langerød, A., Børresen-Dale, A.-L., Kim, S. K., Van De Vijver, M., Sukumar, S., Whitfield, M. L., Kellis, M., Xiong, Y., Wong, D. J. & Chang, H. Y. (2011). Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. *Nature genetics,* 43**,** 621-629.

Hurst, L. D. & Pál, C. (2001). Evidence for purifying selection acting on silent sites in BRCA1. *Trends Genet,* 17**,** 62-5.

Hwang, D. G. & Green, P. (2004). Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proceedings of the National Academy of Sciences of the United States of America,* 101**,** 13994.

Hwang, J. Y., Kim, E. B., Ka, H. & Lee, C.-K. (2013). Identification of the Porcine XIST Gene and Its Differential CpG Methylation Status in Male and Female Pig Cells. *PLOS ONE,* 8**,** e73677.

Hwang, J. Y., Oh, J.-N., Park, C.-H., Lee, D.-K. & Lee, C.-K. (2015). Dosage compensation of X-chromosome inactivation center-linked genes in porcine preimplantation embryos: Non-chromosome-wide initiation of X-chromosome inactivation in blastocysts. *Mechanisms of Development,* 138**,** 246-255.

Ilik, I. A., Quinn, J. J., Georgiev, P., Tavares-Cadete, F., Maticzka, D., Toscano, S., Wan, Y., Spitale, R. C., Luscombe, N., Backofen, R., Chang, H. Y. & Akhtar, A. (2013). Tandem stem-loops in roX RNAs act together to mediate X chromosome dosage compensation in Drosophila. *Molecular cell,* 51**,** 156-173.

International Human Genome Sequencing, C. (2004). Finishing the euchromatic sequence of the human genome. *Nature,* 431**,** 931-945.

Ireland, J. J., Murphee, R. L. & Coulson, P. B. (1980). Accuracy of predicting stages of bovine estrous cycle by gross appearance of the corpus luteum. *J Dairy Sci,* 63**,** 155-60.

Jebb, D., Huang, Z., Pippel, M., Hughes, G. M., Lavrichenko, K., Devanna, P., Winkler, S., Jermiin, L. S., Skirmuntt, E. C., Katzourakis, A., Burkitt-Gray, L., Ray, D. A., Sullivan, K. a. M., Roscito, J. G., Kirilenko, B. M., Dávalos, L. M., Corthals, A. P., Power, M. L., Jones, G., Ransome, R. D., Dechmann, D. K. N., Locatelli, A. G., Puechmaille, S. J., Fedrigo, O., Jarvis, E. D., Hiller, M., Vernes, S. C., Myers, E. W. & Teeling, E. C. (2020). Six reference-quality genomes reveal evolution of bat adaptations. *Nature,* 583**,** 578-584.

Jeffares, D. C., Tomiczek, B., Sojo, V. & Dos Reis, M. 2015. A Beginners Guide to Estimating the Non-synonymous to Synonymous Rate Ratio of all Protein-Coding Genes in a Genome. *In:* PEACOCK, C. (ed.) *Parasite Genomics Protocols.* New York, NY: Springer New York.

Jegu, T., Aeby, E. & Lee, J. T. (2017). The X chromosome in space. *Nat Rev Genet,* 18**,** 377-389.

Jeon, Y. & Lee, J. T. (2011a). YY1 tethers Xist RNA to the inactive X nucleation center. *Cell,* 146**,** 119-33.

Jeon, Y. & Lee, J. T. (2011b). YY1 tethers Xist RNA to the inactive X nucleation center. *Cell,* 146**,** 119-133.

Ji, P., Diederichs, S., Wang, W., Böing, S., Metzger, R., Schneider, P. M., Tidow, N., Brandt, B., Buerger, H., Bulk, E., Thomas, M., Berdel, W. E., Serve, H. & Müller-Tidow, C. (2003).

MALAT-1, a novel noncoding RNA, and thymosin β4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene,* 22**,** 8031-8041.

Johnson, R. & Guigo, R. (2014). The RIDL hypothesis: transposable elements as functional domains of long noncoding RNAs. *Rna,* 20**,** 959-76.

Jonkers, I., Monkhorst, K., Rentmeester, E., Grootegoed, J. A., Grosveld, F. & Gribnau, J. (2008). Xist RNA is confined to the nuclear territory of the silenced X chromosome throughout the cell cycle. *Molecular and cellular biology,* 28**,** 5583-5594.

Jordan, G. & Goldman, N. (2012). The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol Biol Evol,* 29**,** 1125-39.

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermiin, L. S. (2017). ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods,* 14**,** 587-589.

Karpievitch, Y. V., Dabney, A. R. & Smith, R. D. (2012). Normalization and missing value imputation for label-free LC-MS analysis. *BMC bioinformatics,* 13 Suppl 16**,** S5-S5.

Katoh, K. & Standley, D. M. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution,* 30**,** 772-780.

Kelley, D. & Rinn, J. (2012). Transposable elements reveal a stem cell-specific class of long noncoding RNAs. *Genome Biol,* 13**,** R107.

Khan, M., Hou, S., Azam, S. & Lei, H. (2021). Sequence-dependent recruitment of SRSF1 and SRSF7 to intronless lncRNA NKILA promotes nuclear export via the TREX/TAP pathway. *Nucleic Acids Research,* 49**,** 6420-6436.

Kimura, M. (1968). Evolutionary Rate at the Molecular Level. *Nature,* 217**,** 624-626.

Kimura, M. & Ohta, T. (1974). On some principles governing molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America,* 71**,** 2848-2852.

Kino, T., Hurt, D. E., Ichijo, T., Nader, N. & Chrousos, G. P. (2010). Noncoding RNA Gas5 Is a Growth Arrest– and Starvation-Associated Repressor of the Glucocorticoid Receptor. *Science Signaling,* 3**,** ra8.

Kirk, J. M., Kim, S. O., Inoue, K., Smola, M. J., Lee, D. M., Schertzer, M. D., Wooten, J. S., Baker, A. R., Sprague, D., Collins, D. W., Horning, C. R., Wang, S., Chen, Q., Weeks, K. M., Mucha, P. J. & Calabrese, J. M. (2018). Functional classification of long non-coding RNAs by k-mer content. *Nat Genet,* 50**,** 1474-1482.

Kliman, R., Sheehy, B. & Schultz, J. (2008). Genetic Drift and Effective Population Size. *Nature Education,* 1**,** 3.

Klug, A. (1999). Zinc finger peptides for the regulation of gene expression. *Journal of Molecular Biology,* 293**,** 215-218.

Kolpa, H. J., Fackelmayer, F. O. & Lawrence, J. B. (2016). SAF-A Requirement in Anchoring XIST RNA to Chromatin Varies in Transformed and Primary Cells. *Dev Cell,* 39**,** 9-10.

Kosiol, C., Holmes, I. & Goldman, N. (2007). An Empirical Codon Model for Protein Sequence Evolution. *Molecular Biology and Evolution,* 24**,** 1464-1479.

Krchňáková, Z., Thakur, P. K., Krausová, M., Bieberstein, N., Haberman, N., Müller-Mcnicoll, M. & Staněk, D. (2019). Splicing of long non-coding RNAs primarily depends on polypyrimidine tract and 5′ splice-site sequences due to weak interactions with SR proteins. *Nucleic Acids Research,* 47**,** 911-928.

Kretz, M., Webster, D. E., Flockhart, R. J., Lee, C. S., Zehnder, A., Lopez-Pajares, V., Qu, K., Zheng, G. X. Y., Chow, J., Kim, G. E., Rinn, J. L., Chang, H. Y., Siprashvili, Z. & Khavari,

P. A. (2012). Suppression of progenitor differentiation requires the long noncoding RNA ANCR. *Genes & development,* 26**,** 338-343.

Kuang, B., Wu, S. C., Shin, Y., Luo, L. & Kolodziej, P. (2000). split ends encodes large nuclear proteins that regulate neuronal cell fate and axon extension in the Drosophila embryo. *Development,* 127**,** 1517-1529.

Lanier, H. C., Huang, H. & Knowles, L. L. (2014). How low can you go? The effects of mutation rate on the accuracy of species-tree estimation. *Molecular Phylogenetics and Evolution,* 70**,** 112-119.

Lanzós, A., Carlevaro-Fita, J., Mularoni, L., Reverter, F., Palumbo, E., Guigó, R. & Johnson, R. (2017). Discovery of Cancer Driver Long Noncoding RNAs across 1112 Tumour Genomes: New Candidates and Distinguishing Features. *Scientific Reports,* 7**,** 41544.

Lartillot, N. (2013). Phylogenetic Patterns of GC-Biased Gene Conversion in Placental Mammals and the Evolutionary Dynamics of Recombination Landscapes. *Molecular Biology and Evolution,* 30**,** 489-502.

Lasman, L., Krupalnik, V., Viukov, S., Mor, N., Aguilera-Castrejon, A., Schneir, D., Bayerl, J., Mizrahi, O., Peles, S., Tawil, S., Sathe, S., Nachshon, A., Shani, T., Zerbib, M., Kilimnik, I., Aigner, S., Shankar, A., Mueller, J. R., Schwartz, S., Stern-Ginossar, N., Yeo, G. W., Geula, S., Novershtern, N. & Hanna, J. H. (2020). Context-dependent functional compensation between Ythdf m(6)A reader proteins. *Genes Dev,* 34**,** 1373-1391.

Lauberth, S. M., Nakayama, T., Wu, X., Ferris, A. L., Tang, Z., Hughes, S. H. & Roeder, R. G. (2013). H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation. *Cell,* 152**,** 1021-1036.

Lee, E. S., Wolf, E. J., Smith, H. W., Emili, A. & Palazzo, A. F. (2019). TPR is required for the nuclear export of mRNAs and lncRNAs from intronless and intron-poor genes. *bioRxiv***,** 740498.

Lee, S., Kopp, F., Chang, T.-C., Sataluri, A., Chen, B., Sivakumar, S., Yu, H., Xie, Y. & Mendell, J. T. (2016). Noncoding RNA NORAD Regulates Genomic Stability by Sequestering PUMILIO Proteins. *Cell,* 164**,** 69-80.

Letunic, I. & Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res,* 44**,** W242-5.

Letunic, I. & Bork, P. (2019). Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research,* 47**,** W256-W259.

Leucci, E., Vendramin, R., Spinazzi, M., Laurette, P., Fiers, M., Wouters, J., Radaelli, E., Eyckerman, S., Leonelli, C., Vanderheyden, K., Rogiers, A., Hermans, E., Baatsen, P., Aerts, S., Amant, F., Van Aelst, S., Van Den Oord, J., De Strooper, B., Davidson, I., Lafontaine, D. L., Gevaert, K., Vandesompele, J., Mestdagh, P. & Marine, J. C. (2016). Melanoma addiction to the long non-coding RNA SAMMSON. *Nature,* 531**,** 518-22.

Levasseur, A., Gouret, P., Lesage-Meessen, L., Asther, M., Asther, M., Record, E. & Pontarotti, P. (2006). Tracking the connection between evolutionary and functional shifts using the fungal lipase/feruloyl esterase A family. *BMC Evolutionary Biology,* 6**,** 92.

Li, B., Chen, P., Qu, J., Shi, L., Zhuang, W., Fu, J., Li, J., Zhang, X., Sun, Y. & Zhuang, W. (2014). Activation of LTBP3 gene by a long noncoding RNA (lncRNA) MALAT1 transcript in mesenchymal stem cells from multiple myeloma. *J Biol Chem,* 289**,** 29365-75.

Li, J.-H., Liu, S., Zheng, L.-L., Wu, J., Sun, W.-J., Wang, Z.-L., Zhou, H., Qu, L.-H. & Yang, J.-H. (2015). Discovery of Protein–lncRNA Interactions by Integrating Large-Scale CLIP-Seq and RNA-Seq Datasets. *Frontiers in Bioengineering and Biotechnology,* 2.

Li, W. H., Wu, C. I. & Luo, C. C. (1985). A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Molecular Biology and Evolution,* 2**,** 150-174.

Liu, X., Li, D., Zhang, W., Guo, M. & Zhan, Q. (2012). Long non-coding RNA gadd7 interacts with TDP-43 and regulates Cdk6 mRNA decay. *The EMBO journal,* 31**,** 4415-4427.

Lockhart, P. J., Steel, M. A., Hendy, M. D. & Penny, D. (1994). Recovering evolutionary trees under a more realistic model of sequence evolution. *Mol Biol Evol,* 11**,** 605-12.

Long, Y., Wang, X., Youmans, D. T. & Cech, T. R. (2017). How do lncRNAs regulate transcription? *Science Advances,* 3**,** eaao2110.

Loughran, N. B., O'connor, B., O'fágáin, C. & O'connell, M. J. (2008). The phylogeny of the mammalian heme peroxidases and the evolution of their diverse functions. *BMC evolutionary biology,* 8**,** 101-101.

Loughran, N. B., Hinde, S., Mccormick-Hill, S., Leidal, K. G., Bloomberg, S., Loughran, S. T., O'connor, B., O'fágáin, C., Nauseef, W. M. & O'connell, M. J. (2012). Functional consequence of positive selection revealed through rational mutagenesis of human myeloperoxidase. *Molecular biology and evolution,* 29**,** 2039-2046.

Lu, Z., Zhang, Q. C., Lee, B., Flynn, R. A., Smith, M. A., Robinson, J. T., Davidovich, C., Gooding, A. R., Goodrich, K. J., Mattick, J. S., Mesirov, J. P., Cech, T. R. & Chang, H. Y. (2016). RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. *Cell,* 165**,** 1267-1279.

Lu, Z., Carter, A. C. & Chang, H. Y. (2017). Mechanistic insights in X-chromosome inactivation. *Philos Trans R Soc Lond B Biol Sci,* 372.

Lu, Z., Guo, J. K., Wei, Y., Dou, D. R., Zarnegar, B., Ma, Q., Li, R., Zhao, Y., Liu, F., Choudhry, H., Khavari, P. A. & Chang, H. Y. (2020a). Structural modularity of the XIST ribonucleoprotein complex. *Nature Communications,* 11**,** 6163.

Lu, Z., Guo, J. K., Wei, Y., Dou, D. R., Zarnegar, B., Ma, Q., Li, R., Zhao, Y., Liu, F., Choudhry, H., Khavari, P. A. & Chang, H. Y. (2020b). Structural modularity of the XIST ribonucleoprotein complex. *Nat Commun,* 11**,** 6163.

Lubelsky, Y. & Ulitsky, I. (2018). Sequences enriched in Alu repeats drive nuclear localization of long RNAs in human cells. *Nature,* 555**,** 107-111.

Lynch, M. (2007). The frailty of adaptive hypotheses for the origins of organismal complexity. *Proc Natl Acad Sci U S A,* 104 Suppl 1**,** 8597-604.

Lyon, M. F. (1962). Sex Chromatin and Gene Action in the Mammalian X-Chromosome. *Am J Hum Genet,* 14**,** 135-48.

Ma, X., Renda, M. J., Wang, L., Cheng, E. C., Niu, C., Morris, S. W., Chi, A. S. & Krause, D. S. (2007). Rbm15 modulates Notch-induced transcriptional activation and affects myeloid differentiation. *Mol Cell Biol,* 27**,** 3056-64.

Macdonald, W. A. & Mann, M. R. W. (2020). Long noncoding RNA functionality in imprinted domain regulation. *PLOS Genetics,* 16**,** e1008930.

Madeira, F., Park, Y. M., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A. R. N., Potter, S. C., Finn, R. D. & Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. *Nucleic acids research,* 47**,** W636-W641.

Makhlouf, M., Ouimette, J.-F., Oldfield, A., Navarro, P., Neuillet, D. & Rougeulle, C. (2014). A prominent and conserved role for YY1 in Xist transcriptional activation. *Nature communications,* 5**,** 4878-4878.

Mallam, A. L., Sae-Lee, W., Schaub, J. M., Tu, F., Battenhouse, A., Jang, Y. J., Kim, J., Wallingford, J. B., Finkelstein, I. J., Marcotte, E. M. & Drew, K. (2019). Systematic

Discovery of Endogenous Human Ribonucleoprotein Complexes. *Cell Reports,* 29**,** 1351-1368.e5.

Mallik, S., Akashi, H. & Kundu, S. (2015). Assembly constraints drive co-evolution among ribosomal constituents. *Nucleic acids research,* 43**,** 5352-5363.

Malovannaya, A., Lanz, R. B., Jung, S. Y., Bulynko, Y., Le, N. T., Chan, D. W., Ding, C., Shi, Y., Yucer, N., Krenciute, G., Kim, B. J., Li, C., Chen, R., Li, W., Wang, Y., O'malley, B. W. & Qin, J. (2011). Analysis of the human endogenous coregulator complexome. *Cell,* 145**,** 787-99.

Marahrens, Y., Panning, B., Dausman, J., Strauss, W. & Jaenisch, R. (1997). Xist-deficient mice are defective in dosage compensation but not spermatogenesis. *Genes Dev,* 11**,** 156-66.

Marais, G. (2003). Biased gene conversion: implications for genome and sex evolution. *Trends in Genetics,* 19**,** 330-338.

Mcgaughran, A. & Holland, B. R. (2010). Testing the Effect of Metabolic Rate on DNA Variability at the Intra-Specific Level. *PLOS ONE,* 5**,** e9686.

Mchugh, C. A., Russell, P. & Guttman, M. (2014). Methods for comprehensive experimental identification of RNA-protein interactions. *Genome Biology,* 15**,** 203.

Mchugh, C. A., Chen, C. K., Chow, A., Surka, C. F., Tran, C., Mcdonel, P., Pandya-Jones, A., Blanco, M., Burghard, C., Moradian, A., Sweredoski, M. J., Shishkin, A. A., Su, J., Lander, E. S., Hess, S., Plath, K. & Guttman, M. (2015). The Xist lncRNA interacts directly with SHARP to silence transcription through HDAC3. *Nature,* 521**,** 232-6.

Mchugh, C. A. & Guttman, M. (2018). RAP-MS: A Method to Identify Proteins that Interact Directly with a Specific RNA Molecule in Cells. *Methods Mol Biol,* 1649**,** 473-488.

Melé, M., Mattioli, K., Mallard, W., Shechner, D. M., Gerhardinger, C. & Rinn, J. L. (2017). Chromatin environment, transcriptional regulation, and splicing distinguish lincRNAs and mRNAs. *Genome Res,* 27**,** 27-37.

Mellacheruvu, D., Wright, Z., Couzens, A. L., Lambert, J.-P., St-Denis, N. A., Li, T., Miteva, Y. V., Hauri, S., Sardiu, M. E., Low, T. Y., Halim, V. A., Bagshaw, R. D., Hubner, N. C., Al-Hakim, A., Bouchard, A., Faubert, D., Fermin, D., Dunham, W. H., Goudreault, M., Lin, Z.-Y., Badillo, B. G., Pawson, T., Durocher, D., Coulombe, B., Aebersold, R., Superti-Furga, G., Colinge, J., Heck, A. J. R., Choi, H., Gstaiger, M., Mohammed, S., Cristea, I. M., Bennett, K. L., Washburn, M. P., Raught, B., Ewing, R. M., Gingras, A.-C. & Nesvizhskii, A. I. (2013). The CRAPome: a contaminant repository for affinity purification-mass spectrometry data. *Nature methods,* 10**,** 730-736.

Memili, E., Hong, Y.-K., Kim, D.-H., Ontiveros, S. D. & Strauss, W. M. (2001). Murine Xist RNA isoforms are different at their 3' ends: a role for differential polyadenylation. *Gene,* 266**,** 131-137.

Mendonca, A. D. S., Silveira, M. M., Rios, A. F. L., Mangiavacchi, P. M., Caetano, A. R., Dode, M. a. N. & Franco, M. M. (2019). DNA methylation and functional characterization of the XIST gene during in vitro early embryo development in cattle. *Epigenetics*.

Mercer, T. R., Dinger, M. E., Sunkin, S. M., Mehler, M. F. & Mattick, J. S. (2008). Specific expression of long noncoding RNAs in the mouse brain. *Proceedings of the National Academy of Sciences of the United States of America,* 105**,** 716-721.

Merck Millipore. 2009. *Magna RIP™ RNA-Binding Protein Immunoprecipitation Kit* [Online]. Available: https://www.merckmillipore.com/GB/en/product/Magna-RIP-RNA-Binding-Protein-Immunoprecipitation-Kit,MM_NF-17-700#anchor_UG [Accessed 03 August 2021].

Mess, A. (2014). Placental Evolution within the Supraordinal Clades of Eutheria with the Perspective of Alternative Animal Models for Human Placentation. *Advances in Biology,* 2014**,** 21.

Mi, H., Ebert, D., Muruganujan, A., Mills, C., Albou, L.-P., Mushayamaha, T. & Thomas, P. D. (2021). PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Research,* 49**,** D394-D403.

Mili, S. & Steitz, J. A. (2004). Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses. *RNA (New York, N.Y.),* 10**,** 1692-1694.

Minajigi, A., Froberg, J., Wei, C., Sunwoo, H., Kesner, B., Colognori, D., Lessing, D., Payer, B., Boukhali, M., Haas, W. & Lee, J. T. (2015). Chromosomes. A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation. *Science,* 349.

Minks, J. 2012. *Role of XIST RNA and its interacting protein partners in gene silencing.* Text.

Minks, J., Baldry, S. E. L., Yang, C., Cotton, A. M. & Brown, C. J. (2013). XIST-induced silencing of flanking genes is achieved by additive action of repeat a monomers in human somatic cells. *Epigenetics & Chromatin,* 6**,** 23.

Mitchell, S. F. & Parker, R. (2014). Principles and properties of eukaryotic mRNPs. *Mol Cell,* 54**,** 547-58.

Mitchell-Olds, T., Willis, J. H. & Goldstein, D. B. (2007). Which evolutionary processes influence natural genetic variation for phenotypic traits? *Nature Reviews Genetics,* 8**,** 845-856.

Moindrot, B., Cerase, A., Coker, H., Masui, O., Grijzenhout, A., Pintacuda, G., Schermelleh, L., Nesterova, T. B. & Brockdorff, N. (2015). A Pooled shRNA Screen Identifies Rbm15, Spen, and Wtap as Factors Required for Xist RNA-Mediated Silencing. *Cell Rep,* 12**,** 562-72.

Moller, K., Zwieb, C. & Brimacombe, R. (1978). Identification of the oligonucleotide and oligopeptide involved in an RNA--protein crosslink induced by ultraviolet irradiation of Escherichia coli 30 S ribosomal subunits. *J Mol Biol,* 126**,** 489-506.

Monfort, A., Di minin, G., Postlmayr, A., Freimann, R., Arieti, F., Thore, S. & Wutz, A. (2015). Identification of Spen as a Crucial Factor for Xist Function through Forward Genetic Screening in Haploid Embryonic Stem Cells. *Cell Reports,* 12**,** 554-561.

Morgan, C. C., Shakya, K., Webb, A., Walsh, T. A., Lynch, M., Loscher, C. E., Ruskin, H. J. & O'connell, M. J. (2012). Colon cancer associated genes exhibit signatures of positive selection at functionally significant positions. *BMC Evolutionary Biology,* 12**,** 114.

Morgan, C. C., Foster, P. G., Webb, A. E., Pisani, D., Mcinerney, J. O. & O'connell, M. J. (2013). Heterogeneous models place the root of the placental mammal phylogeny. *Molecular biology and evolution,* 30**,** 2145-2156.

Munschauer, M., Nguyen, C. T., Sirokman, K., Hartigan, C. R., Hogstrom, L., Engreitz, J. M., Ulirsch, J. C., Fulco, C. P., Subramanian, V., Chen, J., Schenone, M., Guttman, M., Carr, S. A. & Lander, E. S. (2018). The NORAD lncRNA assembles a topoisomerase complex critical for genome stability. *Nature,* 561**,** 132-136.

Murphy, W. J., Eizirik, E., Johnson, W. E., Zhang, Y. P., Ryder, O. A. & O'brien, S. J. (2001). Molecular phylogenetics and the origins of placental mammals. *Nature,* 409**,** 614-618.

Naganuma, T., Nakagawa, S., Tanigawa, A., Sasaki, Y. F., Goshima, N. & Hirose, T. (2012). Alternative 3'-end processing of long noncoding RNA initiates construction of nuclear paraspeckles. *The EMBO Journal,* 31**,** 4020-4034.

Nakamoto, M. Y., Lammer, N. C., Batey, R. T. & Wuttke, D. S. (2020). hnRNPK recognition of the B motif of Xist and other biological RNAs. *Nucleic Acids Research,* 48**,** 9320-9335.

Navarro, P., Pichard, S., Ciaudo, C., Avner, P. & Rougeulle, C. (2005). Tsix transcription across the Xist gene alters chromatin conformation without affecting Xist transcription: implications for X-chromosome inactivation. *Genes Dev,* 19**,** 1474-84.

Nesterova, T. B., Slobodyanyuk, S. Y., Elisaphenko, E. A., Shevchenko, A. I., Johnston, C., Pavlova, M. E., Rogozin, I. B., Kolesnikov, N. N., Brockdorff, N. & Zakian, S. M. (2001). Characterization of the genomic Xist locus in rodents reveals conservation of overall gene structure and tandem repeats but rapid evolution of unique sequence. *Genome Res,* 11**,** 833-49.

Nesterova, T. B., Wei, G., Coker, H., Pintacuda, G., Bowness, J. S., Zhang, T., Almeida, M., Bloechl, B., Moindrot, B., Carter, E. J., Alvarez Rodrigo, I., Pan, Q., Bi, Y., Song, C.-X. & Brockdorff, N. (2019). Systematic allelic analysis defines the interplay of key pathways in X chromosome inactivation. *Nature communications,* 10**,** 3129-3129.

New England Biolabs. 2021. *RNA Synthesis with Modified Nucleotides (E2040)* [Online]. Available: https://international.neb.com/protocols/0001/01/01/rna-synthesis-with-modified-nucleotides-e2040 [Accessed 12 August 2021].

Ngandu, N. K., Scheffler, K., Moore, P., Woodman, Z., Martin, D. & Seoighe, C. (2008). Extensive purifying selection acting on synonymous sites in HIV-1 Group M sequences. *Virology journal,* 5**,** 160-160.

Nguyen, E. D., Balas, M. M., Griffin, A. M., Roberts, J. T. & Johnson, A. M. (2018). Global profiling of hnRNP A2/B1-RNA binding on chromatin highlights LncRNA interactions. *RNA biology,* 15**,** 901-913.

Nishida, M. (2002). The Ishikawa cells from birth to the present. *Hum Cell,* 15**,** 104-17.

Nishihara, H., Okada, N. & Hasegawa, M. (2007). Rooting the eutherian tree: the power and pitfalls of phylogenomics. *Genome Biol,* 8**,** R199.

Nishihara, H., Maruyama, S. & Okada, N. (2009). Retroposon analysis and recent geological data suggest near-simultaneous divergence of the three superorders of mammals. *Proceedings of the National Academy of Sciences,* 106**,** 5235.

Ntini, E., Louloupi, A., Liz, J., Muino, J. M., Marsico, A. & Orom, U. a. V. (2018). Long ncRNA A-ROD activates its target gene DKK1 at its release from chromatin. *Nat Commun,* 9**,** 1636.

Okamoto, I., Patrat, C., Thepot, D., Peynot, N., Fauque, P., Daniel, N., Diabangouaya, P., Wolf, J. P., Renard, J. P., Duranthon, V. & Heard, E. (2011). Eutherian mammals use diverse strategies to initiate X-chromosome inactivation during development. *Nature,* 472**,** 370-4.

Ong, C.-T. & Corces, V. G. (2014). CTCF: an architectural protein bridging genome topology and function. *Nature Reviews Genetics,* 15**,** 234-246.

Ong, S. E. & Mann, M. (2006). A practical recipe for stable isotope labeling by amino acids in cell culture (SILAC). *Nat Protoc,* 1**,** 2650-60.

Palasca, O., Santos, A., Stolte, C., Gorodkin, J. & Jensen, L. J. (2018). TISSUES 2.0: an integrative web resource on mammalian tissue expression. *Database,* 2018.

Pamilo, P. & Nei, M. (1988). Relationships between gene trees and species trees. *Molecular Biology and Evolution,* 5**,** 568-583.

Pandya-Jones, A., Markaki, Y., Serizay, J., Chitiashvili, T., Mancia Leon, W. R., Damianov, A., Chronis, C., Papp, B., Chen, C.-K., Mckee, R., Wang, X.-J., Chau, A., Sabri, S., Leonhardt, H., Zheng, S., Guttman, M., Black, D. L. & Plath, K. (2020). A protein assembly mediates Xist localization and gene silencing. *Nature,* 587**,** 145-151.

Pang, K. C., Frith, M. C. & Mattick, J. S. (2006). Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet,* 22**,** 1-5.

Pappireddi, N., Martin, L. & Wühr, M. (2019). A Review on Quantitative Multiplexed Proteomics. *Chembiochem : a European journal of chemical biology,* 20**,** 1210-1224.

Pardo, M. & Choudhary, J. S. (2012). Assignment of Protein Interactions from Affinity Purification/Mass Spectrometry Data. *Journal of Proteome Research,* 11**,** 1462-1474.

Park, C.-H., Uh, K.-J., Mulligan, B. P., Jeung, E.-B., Hyun, S.-H., Shin, T., Ka, H. & Lee, C.-K. (2011). Analysis of Imprinted Gene Expression in Normal Fertilized and Uniparental Preimplantation Porcine Embryos. *PLOS ONE,* 6**,** e22216.

Park, S.-W., Kang, Y. I., Sypula, J. G., Choi, J., Oh, H. & Park, Y. (2007). An evolutionarily conserved domain of roX2 RNA is sufficient for induction of H4-Lys16 acetylation on the Drosophila X chromosome. *Genetics,* 177**,** 1429-1437.

Pashev, I. G., Dimitrov, S. I. & Angelov, D. (1991). Crosslinking proteins to nucleic acids by ultraviolet laser irradiation. *Trends Biochem Sci,* 16**,** 323-6.

Patil, D. P., Chen, C. K., Pickering, B. F., Chow, A., Jackson, C., Guttman, M. & Jaffrey, S. R. (2016). m(6)A RNA methylation promotes XIST-mediated transcriptional repression. *Nature,* 537**,** 369-373.

Pauli, A., Valen, E., Lin, M. F., Garber, M., Vastenhouw, N. L., Levin, J. Z., Fan, L., Sandelin, A., Rinn, J. L., Regev, A. & Schier, A. F. (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res,* 22**,** 577-91.

Pazos, F. & Valencia, A. (2008). Protein co-evolution, co-adaptation and interactions. *The EMBO Journal,* 27**,** 2648-2655.

Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S. & Brockdorff, N. (1996). Requirement for Xist in X chromosome inactivation. *Nature,* 379**,** 131-7.

Petropoulos, S., Edsgard, D., Reinius, B., Deng, Q., Panula, S. P., Codeluppi, S., Plaza Reyes, A., Linnarsson, S., Sandberg, R. & Lanner, F. (2016). Single-Cell RNA-Seq Reveals Lineage and X Chromosome Dynamics in Human Preimplantation Embryos. *Cell,* 165**,** 1012-26.

Petrov, D. A. & Hartl, D. L. (1999). Patterns of nucleotide substitution in Drosophila and mammalian genomes. *Proc Natl Acad Sci U S A,* 96**,** 1475-9.

Phung, T. N., Olney, K. C., Silasi, M., Perley, L., O'bryan, J., Kliman, H. J. & Wilson, M. A. (2021). X chromosome inactivation in the human placenta is patchy and distinct from adult tissues. *bioRxiv***,** 785105.

Ping, X.-L., Sun, B.-F., Wang, L., Xiao, W., Yang, X., Wang, W.-J., Adhikari, S., Shi, Y., Lv, Y., Chen, Y.-S., Zhao, X., Li, A., Yang, Y., Dahal, U., Lou, X.-M., Liu, X., Huang, J., Yuan, W.-P., Zhu, X.-F., Cheng, T., Zhao, Y.-L., Wang, X., Danielsen, J. M. R., Liu, F. & Yang, Y.-G. (2014). Mammalian WTAP is a regulatory subunit of the RNA N6-methyladenosine methyltransferase. *Cell Research,* 24**,** 177-189.

Pintacuda, G., Wei, G., Roustan, C., Kirmizitas, B. A., Solcan, N., Cerase, A., Castello, A., Mohammed, S., Moindrot, B., Nesterova, T. B. & Brockdorff, N. (2017a). hnRNPK Recruits PCGF3/5-PRC1 to the Xist RNA B-Repeat to Establish Polycomb-Mediated Chromosomal Silencing. *Mol Cell,* 68**,** 955-969.e10.

Pintacuda, G., Young, A. N. & Cerase, A. (2017b). Function by Structure: Spotlights on Xist Long Non-coding RNA. *Frontiers in Molecular Biosciences,* 4.

Pisani, G. & Baron, B. (2019). Nuclear paraspeckles function in mediating gene regulatory and apoptotic pathways. *Non-coding RNA Research,* 4**,** 128-134.

Plotkin, J. B. & Kudla, G. (2011). Synonymous but not the same: the causes and consequences of codon bias. *Nature reviews. Genetics,* 12**,** 32-42.

Ponjavic, J., Oliver, P. L., Lunter, G. & Ponting, C. P. (2009). Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. *PLoS Genet,* 5**,** e1000617.

Postepska-Igielska, A., Giwojna, A., Gasri-Plotnitsky, L., Schmitt, N., Dold, A., Ginsberg, D. & Grummt, I. (2015). LncRNA Khps1 Regulates Expression of the Proto-oncogene SPHK1 via Triplex-Mediated Changes in Chromatin Structure. *Mol Cell,* 60**,** 626-36.

Pozzi, B., Bragado, L., Will, C. L., Mammi, P., Risso, G., Urlaub, H., Lührmann, R. & Srebrow, A. (2017). SUMO conjugation to spliceosomal proteins is required for efficient pre-mRNA splicing. *Nucleic acids research,* 45**,** 6729-6745.

Pritchard, J. K., Pickrell, J. K. & Coop, G. (2010). The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current biology : CB,* 20**,** R208-R215.

Quinn, J. J., Zhang, Q. C., Georgiev, P., Ilik, I. A., Akhtar, A. & Chang, H. Y. (2016). Rapid evolutionary turnover underlies conserved lncRNA-genome interactions. *Genes & development,* 30**,** 191-207.

Ramanathan, M., Porter, D. F. & Khavari, P. A. (2019). Methods to study RNA-protein interactions. *Nature methods,* 16**,** 225-234.

Ray, D., Kazan, H., Cook, K. B., Weirauch, M. T., Najafabadi, H. S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., Na, H., Irimia, M., Matzat, L. H., Dale, R. K., Smith, S. A., Yarosh, C. A., Kelly, S. M., Nabet, B., Mecenas, D., Li, W., Laishram, R. S., Qiao, M., Lipshitz, H. D., Piano, F., Corbett, A. H., Carstens, R. P., Frey, B. J., Anderson, R. A., Lynch, K. W., Penalva, L. O. F., Lei, E. P., Fraser, A. G., Blencowe, B. J., Morris, Q. D. & Hughes, T. R. (2013). A compendium of RNA-binding motifs for decoding gene regulation. *Nature,* 499**,** 172-177.

Rees, J. S. & Lilley, K. S. (2011). Method for suppressing non-specific protein interactions observed with affinity resins. *Methods,* 54**,** 407-12.

Ridings-Figueroa, R., Stewart, E. R., Nesterova, T. B., Coker, H., Pintacuda, G., Godwin, J., Wilson, R., Haslam, A., Lilley, F., Ruigrok, R., Bageghni, S. A., Albadrani, G., Mansfield, W., Roulson, J. A., Brockdorff, N., Ainscough, J. F. X. & Coverley, D. (2017). The nuclear matrix protein CIZ1 facilitates localization of Xist RNA to the inactive X-chromosome territory. *Genes Dev,* 31**,** 876-888.

Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., Goodnough, L. H., Helms, J. A., Farnham, P. J., Segal, E. & Chang, H. Y. (2007a). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell,* 129**,** 1311-1323.

Rinn, J. L., Kertesz, M., Wang, J. K., Squazzo, S. L., Xu, X., Brugmann, S. A., Goodnough, L. H., Helms, J. A., Farnham, P. J., Segal, E. & Chang, H. Y. (2007b). Functional Demarcation of Active and Silent Chromatin Domains in Human HOX Loci by Noncoding RNAs. *Cell,* 129**,** 1311-1323.

Rivas, E., Clements, J. & Eddy, S. R. (2017). A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. *Nature methods,* 14**,** 45-48.

Rivas, E., Clements, J. & Eddy, S. R. (2020). Estimating the power of sequence covariation for detecting conserved RNA structure. *Bioinformatics,* 36**,** 3072-3076.

Rivas, E. & Eddy, S. R. (2020). Response to Tavares et al., "Covariation analysis with improved parameters reveals conservation in lncRNA structures". *bioRxiv***,** 2020.02.18.955047.

Rivas, E. (2021). Evolutionary conservation of RNA sequence and structure. *WIREs RNA,* 12**,** e1649.

Roberts, R. M., Green, J. A. & Schulz, L. C. (2016). The Evolution of the Placenta. *Reproduction (Cambridge, England),* 152**,** R179-R189.

Robinson, D. F. & Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical Biosciences,* 53**,** 131-147.

Ross, C. J., Rom, A., Spinrad, A., Gelbard-Solodkin, D., Degani, N. & Ulitsky, I. (2021). Uncovering deeply conserved motif combinations in rapidly evolving noncoding sequences. *Genome Biology,* 22**,** 29.

Roux, B. T., Heward, J. A., Donnelly, L. E., Jones, S. W. & Lindsay, M. A. (2017). Catalog of Differentially Expressed Long Non-Coding RNA following Activation of Human and Mouse Innate Immune Response. *Frontiers in immunology,* 8**,** 1038-1038.

Sacco, M. A., Koropacka, K., Grenier, E., Jaubert, M. J., Blanchard, A., Goverse, A., Smant, G. & Moffett, P. (2009). The Cyst Nematode SPRYSEC Protein RBP-1 Elicits Gpa2- and RanGAP2-Dependent Plant Cell Death. *PLOS Pathogens,* 5**,** e1000564.

Sahakyan, A., Kim, R., Chronis, C., Sabri, S., Bonora, G., Theunissen, T. W., Kuoy, E., Langerman, J., Clark, A. T., Jaenisch, R. & Plath, K. (2017). Human Naive Pluripotent Stem Cells Model X Chromosome Dampening and X Inactivation. *Cell stem cell,* 20**,** 87-101.

Sakaguchi, T., Hasegawa, Y., Brockdorff, N., Tsutsui, K., Tsutsui, K. M., Sado, T. & Nakagawa, S. (2016). Control of Chromosomal Localization of Xist by hnRNP U Family Molecules. *Dev Cell,* 39**,** 11-12.

Sarma, K., Levasseur, P., Aristarkhov, A. & Lee, J. T. (2010). Locked nucleic acids (LNAs) reveal sequence requirements and kinetics of Xist RNA localization to the X chromosome. *Proceedings of the National Academy of Sciences of the United States of America,* 107**,** 22196-22201.

Sasaki, Y. T. F., Ideue, T., Sano, M., Mituyama, T. & Hirose, T. (2009). MENepsilon/beta noncoding RNAs are essential for structural integrity of nuclear paraspeckles. *Proceedings of the National Academy of Sciences of the United States of America,* 106**,** 2525-2530.

Sayres, M. a. W., Venditti, C., Pagel, M. & Makova, K. D. (2011). DO VARIATIONS IN SUBSTITUTION RATES AND MALE MUTATION BIAS CORRELATE WITH LIFE-HISTORY TRAITS? A STUDY OF 32 MAMMALIAN GENOMES. *Evolution,* 65**,** 2800-2815.

Scherer, M., Levin, M., Butter, F. & Scheibe, M. (2020). Quantitative Proteomics to Identify Nuclear RNA-Binding Proteins of Malat1. *International journal of molecular sciences,* 21**,** 1166.

Schmid, K. & Yang, Z. (2008). The Trouble with Sliding Windows and the Selective Pressure in BRCA1. *PLOS ONE,* 3**,** e3746.

Schmidt, H. A., Strimmer, K., Vingron, M. & Von Haeseler, A. (2002). TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics,* 18**,** 502-504.

Seal, R. L., Chen, L.-L., Griffiths-Jones, S., Lowe, T. M., Mathews, M. B., O'reilly, D., Pierce, A. J., Stadler, P. F., Ulitsky, I., Wolin, S. L. & Bruford, E. A. (2020). A guide to naming human non-coding RNA genes. *The EMBO Journal,* 39**,** e103777.

Sela, N., Mersch, B., Hotz-Wagenblatt, A. & Ast, G. (2010). Characteristics of transposable element exonization within human and mouse. *PloS one,* 5**,** e10907-e10907.

Sella, G., Petrov, D. A., Przeworski, M. & Andolfatto, P. (2009). Pervasive Natural Selection in the Drosophila Genome? *PLOS Genetics,* 5**,** e1000495.

Seo, T.-K. & Kishino, H. (2008). Synonymous Substitutions Substantially Improve Evolutionary Inference from Highly Diverged Proteins. *Systematic Biology,* 57**,** 367-377.

Shen, Y.-Y., Shi, P., Sun, Y.-B. & Zhang, Y.-P. (2009). Relaxation of selective constraints on avian mitochondrial DNA following the degeneration of flight ability. *Genome research,* 19**,** 1760-1765.

Shevelyov, Y. Y., Lavrov, S. A., Mikhaylova, L. M., Nurminsky, I. D., Kulathinal, R. J., Egorova, K. S., Rozovsky, Y. M. & Nurminsky, D. I. (2009). The B-type lamin is required for somatic repression of testis-specific gene clusters. *Proceedings of the National Academy of Sciences,* 106**,** 3282.

Shi, Y., Downes, M., Xie, W., Kao, H.-Y., Ordentlich, P., Tsai, C.-C., Hon, M. & Evans, R. M. (2001). Sharp, an inducible cofactor that integrates nuclear receptor repression and activation. *Genes & Development,* 15**,** 1140-1151.

Shimodaira, H. (2002). An Approximately Unbiased Test of Phylogenetic Tree Selection. *Systematic Biology,* 51**,** 492-508.

Shukla, C. J., Mccorkindale, A. L., Gerhardinger, C., Korthauer, K. D., Cabili, M. N., Shechner, D. M., Irizarry, R. A., Maass, P. G. & Rinn, J. L. (2018). High-throughput identification of RNA nuclear enrichment sequences. *The EMBO journal,* 37**,** e98452.

Singh, G., Pratt, G., Yeo, G. W. & Moore, M. J. (2015). The Clothes Make the mRNA: Past and Present Trends in mRNP Fashion. *Annual review of biochemistry,* 84**,** 325-354.

Smith, J. M. & Smith, N. H. (1996). Synonymous nucleotide divergence: what is "saturation"? *Genetics,* 142**,** 1033-1036.

Smith, N. G. C. & Eyre-Walker, A. (2002). Adaptive protein evolution in Drosophila. *Nature,* 415**,** 1022-1024.

Smola, M. J., Christy, T. W., Inoue, K., Nicholson, C. O., Friedersdorf, M., Keene, J. D., Lee, D. M., Calabrese, J. M. & Weeks, K. M. (2016). SHAPE reveals transcript-wide interactions, complex structural domains, and protein interactions across the Xist lncRNA in living cells. *Proc Natl Acad Sci U S A,* 113**,** 10322-7.

Sofi, S., Williamson, L., Turvey, G. L., Scoynes, C., Hirst, C., Godwin, J., Brockdorff, N., Ainscough, J. & Coverley, D. (2020). A polyglutamine domain is required for de novo CIZ1 assembly formation at the inactive X chromosome. *bioRxiv***,** 2020.11.10.376558.

Som, A. (2015). Causes, consequences and solutions of phylogenetic incongruence. *Briefings in Bioinformatics,* 16**,** 536-548.

Sonnett, M., Yeung, E. & Wühr, M. (2018). Accurate, Sensitive, and Precise Multiplexed Proteomics Using the Complement Reporter Ion Cluster. *Analytical chemistry,* 90**,** 5032-5039.

Stewart, E. R., Turner, R. M. L., Newling, K., Ridings-Figueroa, R., Scott, V., Ashton, P. D., Ainscough, J. F. X. & Coverley, D. (2019). Maintenance of epigenetic landscape

requires CIZ1 and is corrupted in differentiated fibroblasts in long-term culture. *Nature Communications,* 10**,** 460.

Stork, C., Li, Z., Lin, L. & Zheng, S. (2019). Developmental Xist induction is mediated by enhanced splicing. *Nucleic Acids Research,* 47**,** 1532-1543.

Strange, R. M., Russelburg, L. P. & Delaney, K. J. (2016). Co-evolution of SNF spliceosomal proteins with their RNA targets in trans-splicing nematodes. *Genetica,* 144**,** 487-96.

Strein, C., Alleaume, A. M., Rothbauer, U., Hentze, M. W. & Castello, A. (2014). A versatile assay for RNA-binding proteins in living cells. *Rna,* 20**,** 721-31.

Strimmer, K. & Von Haeseler, A. (1996). Quartet Puzzling: A Quartet Maximum-Likelihood Method for Reconstructing Tree Topologies. *Molecular Biology and Evolution,* 13**,** 964-964.

Strimmer, K. & Von Haeseler, A. (1997). Likelihood-mapping: a simple method to visualize phylogenetic content of a sequence alignment. *Proc Natl Acad Sci U S A,* 94**,** 6815-9.

Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat Struct Mol Biol,* 14**,** 103-5.

Sunwoo, H., Colognori, D., Froberg, J. E., Jeon, Y. & Lee, J. T. (2017). Repeat E anchors Xist RNA to the inactive X chromosomal compartment through CDKN1A-interacting protein (CIZ1). *Proc Natl Acad Sci U S A,* 114**,** 10654-10659.

Swaney, D. L., Wenger, C. D. & Coon, J. J. (2010). Value of Using Multiple Proteases for Large-Scale Mass Spectrometry-Based Proteomics. *Journal of Proteome Research,* 9**,** 1323-1329.

Syrett, C. M., Sindhava, V., Hodawadekar, S., Myles, A., Liang, G., Zhang, Y., Nandi, S., Cancro, M., Atchison, M. & Anguera, M. C. (2017). Loss of Xist RNA from the inactive X during B cell development is restored in a dynamic YY1-dependent two-step process in activated B cells. *PLOS Genetics,* 13**,** e1007050.

Szöllősi, G. J., Tannier, E., Daubin, V. & Boussau, B. (2015). The Inference of Gene Trees with Species Trees. *Systematic Biology,* 64**,** e42-e62.

Takagi, N. & Abe, K. (1990). Detrimental effects of two active X chromosomes on early mouse development. *Development,* 109**,** 189-201.

Tamm-Rosenstein, K., Simm, J., Suhorutshenko, M., Salumets, A. & Metsis, M. (2013). Changes in the Transcriptome of the Human Endometrial Ishikawa Cancer Cell Line Induced by Estrogen, Progesterone, Tamoxifen, and Mifepristone (RU486) as Detected by RNA-Sequencing. *PLOS ONE,* 8**,** e68907.

Tarver, J. E., Dos Reis, M., Mirarab, S., Moran, R. J., Parker, S., O'reilly, J. E., King, B. L., O'connell, M. J., Asher, R. J., Warnow, T., Peterson, K. J., Donoghue, P. C. J. & Pisani, D. (2016). The Interrelationships of Placental Mammals and the Limits of Phylogenetic Inference. *Genome Biology and Evolution,* 8**,** 330-344.

Tavares, R. C. A., Pyle, A. M. & Somarowthu, S. (2019). Phylogenetic Analysis with Improved Parameters Reveals Conservation in lncRNA Structures. *Journal of Molecular Biology,* 431**,** 1592-1603.

Telford, N. A., Watson, A. J. & Schultz, G. A. (1990). Transition from maternal to embryonic control in early mammalian development: a comparison of several species. *Mol Reprod Dev,* 26**,** 90-100.

The Gene Ontology Consortium (2021). The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research,* 49**,** D325-D334.

The Uniprot, C. (2021). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research,* 49**,** D480-D489.

Thermofisher Scientific. 2016. *Dynabeads™ MyOne™ Streptavidin C1 manual* [Online]. Available: https://www.thermofisher.com/document-connect/document-connect.html?url=https%3A%2F%2Fassets.thermofisher.com%2FTFS-Assets%2FLSG%2Fmanuals%2Fdynabeads_myone_savC1_man.pdf&title=RHluYWJlYWRzIE15T25lIFN0cmVwdGF2aWRpbiBDMQ== [Accessed 03 August 2021].

Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T. & Hamon, C. (2003). Tandem Mass Tags: A Novel Quantification Strategy for Comparative Analysis of Complex Protein Mixtures by MS/MS. *Analytical Chemistry,* 75**,** 1895-1904.

Tichon, A., Gil, N., Lubelsky, Y., Havkin Solomon, T., Lemze, D., Itzkovitz, S., Stern-Ginossar, N. & Ulitsky, I. (2016). A conserved abundant cytoplasmic long noncoding RNA modulates repression by Pumilio proteins in human cells. *Nat Commun,* 7**,** 12209.

Tichon, A., Perry, R. B.-T., Stojic, L. & Ulitsky, I. (2018). SAM68 is required for regulation of Pumilio by the NORAD long noncoding RNA. *Genes & Development,* 32**,** 70-78.

Tilgner, H., Knowles, D. G., Johnson, R., Davis, C. A., Chakrabortty, S., Djebali, S., Curado, J., Snyder, M., Gingeras, T. R. & Guigó, R. (2012). Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome research,* 22**,** 1616-1625.

Tinning, H., Taylor, A., Wang, D., Constantinides, B., Sutton, R., Oikonomou, G., Velazquez, M. A., Thompson, P., Treumann, A., O'connell, M. J. & Forde, N. (2020). The role of CAPG in molecular communication between the embryo and the uterine endometrium: Is its function conserved in species with different implantation strategies? *Faseb j,* 34**,** 11015-11029.

Toki, N., Takahashi, H., Sharma, H., Valentine, M. N. Z., Rahman, F.-U. M., Zucchelli, S., Gustincich, S. & Carninci, P. (2020). SINEUP long non-coding RNA acts via PTBP1 and HNRNPK to promote translational initiation assemblies. *Nucleic acids research,* 48**,** 11626-11644.

Trifinopoulos, J., Nguyen, L.-T., Von haeseler, A. & Minh, B. Q. (2016). W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. *Nucleic Acids Research,* 44**,** W232-W235.

Tuck, A. C., Natarajan, K. N., Rice, G. M., Borawski, J., Mohn, F., Rankova, A., Flemr, M., Wenger, A., Nutiu, R., Teichmann, S. & Bühler, M. (2018). Distinctive features of lincRNA gene expression suggest widespread RNA-independent functions. *Life science alliance,* 1**,** e201800124-e201800124.

Tukiainen, T., Villani, A.-C., Yen, A., Rivas, M. A., Marshall, J. L., Satija, R., Aguirre, M., Gauthier, L., Fleharty, M., Kirby, A., Cummings, B. B., Castel, S. E., Karczewski, K. J., Aguet, F., Byrnes, A., Consortium, G. T., Lappalainen, T., Regev, A., Ardlie, K. G., Hacohen, N. & Macarthur, D. G. (2017). Landscape of X chromosome inactivation across human tissues. *Nature,* 550**,** 244.

Uhlén, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, Å., Kampf, C., Sjöstedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P. H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., Von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., Von Heijne, G., Nielsen, J. & Pontén, F. (2015). Proteomics. Tissue-based map of the human proteome. *Science,* 347**,** 1260419.

Ulitsky, I., Shkumatava, A., Jan, C. H., Sive, H. & Bartel, D. P. (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell,* 147**,** 1537-1550.

Ulitsky, I. & Bartel, D. P. (2013). lincRNAs: genomics, evolution, and mechanisms. *Cell,* 154**,** 26-46.

Ulitsky, I. (2016). Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. *Nat Rev Genet,* 17**,** 601-14.

Urdaneta, E. C., Vieira-Vieira, C. H., Hick, T., Wessels, H.-H., Figini, D., Moschall, R., Medenbach, J., Ohler, U., Granneman, S., Selbach, M. & Beckmann, B. M. (2019). Purification of cross-linked RNA-protein complexes by phenol-toluol extraction. *Nature Communications,* 10**,** 990.

Vallot, C., Ouimette, J. F. & Rougeulle, C. (2016). Establishment of X chromosome inactivation and epigenomic features of the inactive X depend on cellular contexts. *Bioessays,* 38**,** 869-80.

Van Den Berg, I. M., Laven, J. S., Stevens, M., Jonkers, I., Galjaard, R. J., Gribnau, J. & Van Doorninck, J. H. (2009). X chromosome inactivation is initiated in human preimplantation embryos. *Am J Hum Genet,* 84**,** 771-9.

Van Den Bussche, R. A., Baker, R. J., Huelsenbeck, J. P. & Hillis, D. M. (1998). Base Compositional Bias and Phylogenetic Analyses: A Test of the "Flying DNA" Hypothesis. *Molecular Phylogenetics and Evolution,* 10**,** 408-416.

Van Den Eynden, J. & Larsson, E. (2017). Mutational Signatures Are Critical for Proper Estimation of Purifying Selection Pressures in Cancer Somatic Mutation Data When Using the dN/dS Metric. *Frontiers in Genetics,* 8.

Van Der Lee, R., Wiel, L., Van Dam, T. J. P. & Huynen, M. A. (2017). Genome-scale detection of positive selection in nine primates predicts human-virus evolutionary conflicts. *Nucleic acids research,* 45**,** 10634-10648.

Van Nostrand, E. L., Pratt, G. A., Shishkin, A. A., Gelboin-Burkhart, C., Fang, M. Y., Sundararaman, B., Blue, S. M., Nguyen, T. B., Surka, C., Elkins, K., Stanton, R., Rigo, F., Guttman, M. & Yeo, G. W. (2016). Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nature methods,* 13**,** 508-514.

Viphakone, N., Sudbery, I., Griffith, L., Heath, C. G., Sims, D. & Wilson, S. A. (2019). Co-transcriptional Loading of RNA Export Factors Shapes the Human Transcriptome. *Molecular Cell,* 75**,** 310-323.e8.

Volders, P. J., Anckaert, J., Verheggen, K., Nuytens, J., Martens, L., Mestdagh, P. & Vandesompele, J. (2019). LNCipedia 5: towards a reference set of human long non-coding RNAs. *Nucleic Acids Res,* 47**,** D135-d139.

Vuong, J. K., Lin, C. H., Zhang, M., Chen, L., Black, D. L. & Zheng, S. (2016). PTBP1 and PTBP2 Serve Both Specific and Redundant Functions in Neuronal Pre-mRNA Splicing. *Cell Rep,* 17**,** 2766-2775.

Wachter, J. & Hill, S. (2016). Positive Selection Pressure Drives Variation on the Surface-Exposed Variable Proteins of the Pathogenic Neisseria. *PLOS ONE,* 11**,** e0161348.

Wang, J., Syrett, C. M., Kramer, M. C., Basu, A., Atchison, M. L. & Anguera, M. C. (2016). Unusual maintenance of X chromosome inactivation predisposes female lymphocytes for increased expression from the inactive X. *Proceedings of the National Academy of Sciences,* 113**,** E2029.

Wang, P., Xue, Y., Han, Y., Lin, L., Wu, C., Xu, S., Jiang, Z., Xu, J., Liu, Q. & Cao, X. (2014). The STAT3-binding long noncoding RNA lnc-DC controls human dendritic cell differentiation. *Science,* 344**,** 310-3.

Wang, X., Miller, D. C., Clark, A. G. & Antczak, D. F. (2012). Random X inactivation in the mule and horse placenta. *Genome research,* 22**,** 1855-1863.

Warder, D. E. & Keherly, M. J. (2003). Ciz1, Cip1 interacting zinc finger protein 1 binds the consensus DNA sequence ARYSR(0-2)YYAC. *J Biomed Sci,* 10**,** 406-17.

Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F. T., De beer, T. A p., Rempfer, C., Bordoli, L., Lepore, R. & Schwede, T. (2018). SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research,* 46**,** W296-W303.

Webb, A. E., Walsh, T. A. & O'connell, M. J. (2017). VESPA: Very large-scale Evolutionary and Selective Pressure Analyses. *PeerJ Computer Science,* 3**,** e118.

Wei, Q., Zhong, L., Zhang, S., Mu, H., Xiang, J., Yue, L., Dai, Y. & Han, J. (2017). Bovine lineage specification revealed by single-cell gene expression analysis from zygote to blastocyst†. *Biology of Reproduction,* 97**,** 5-17.

Wei, Q., Li, R., Zhong, L., Mu, H., Zhang, S., Yue, L., Xiang, J., Li, E., Zhi, M., Cao, S. & Han, J. (2018). Lineage specification revealed by single-cell gene expression analysis in porcine preimplantation embryos†. *Biology of Reproduction***,** ioy062-ioy062.

Weigand, H. & Leese, F. (2018). Detecting signatures of positive selection in non-model species using genomic data. *Zoological Journal of the Linnean Society,* 184**,** 528-583.

Welch, J. J. (2006). Estimating the genomewide rate of adaptive protein evolution in Drosophila. *Genetics,* 173**,** 821-837.

Werner, M. S. & Ruthenburg, A. J. (2015). Nuclear Fractionation Reveals Thousands of Chromatin-Tethered Noncoding RNAs Adjacent to Active Genes. *Cell Rep,* 12**,** 1089-98.

Wertheim, J. O., Murrell, B., Smith, M. D., Kosakovsky Pond, S. L. & Scheffler, K. (2015). RELAX: detecting relaxed selection in a phylogenetic framework. *Molecular biology and evolution,* 32**,** 820-832.

Wilusz, J. E. (2016). Long noncoding RNAs: Re-writing dogmas of RNA processing and stability. *Biochimica et biophysica acta,* 1859**,** 128-138.

Woo, J., Robertson David, L. & Lovell Simon, C. (2010). Constraints on HIV-1 Diversity from Protein Structure. *Journal of Virology,* 84**,** 12995-13003.

Woodhams, M. D., Lockhart, P. J. & Holland, B. R. (2016). Simulating and Summarizing Sources of Gene Tree Incongruence. *Genome Biology and Evolution,* 8**,** 1299-1315.

Wu, L. S., Qian, J. Y., Wang, M. & Yang, H. (2016). Identifying the role of Wilms tumor 1 associated protein in cancer prediction using integrative genomic analyses. *Mol Med Rep,* 14**,** 2823-2831.

Wu, T., Pinto, H. B., Kamikawa, Y. F. & Donohoe, M. E. (2015). The BET family member BRD4 interacts with OCT4 and regulates pluripotency gene expression. *Stem cell reports,* 4**,** 390-403.

Wurtmann, E. J. & Wolin, S. L. (2009). RNA under attack: cellular handling of RNA damage. *Crit Rev Biochem Mol Biol,* 44**,** 34-49.

Wutz, A., Rasmussen, T. P. & Jaenisch, R. (2002). Chromosomal silencing and localization are mediated by different domains of Xist RNA. *Nature Genetics,* 30**,** 167-174.

Xue, F., Tian, X. C., Du, F., Kubota, C., Taneja, M., Dinnyes, A., Dai, Y., Levine, H., Pereira, L. V. & Yang, X. (2002). Aberrant patterns of X chromosome inactivation in bovine clones. *Nature Genetics,* 31**,** 216.

Yamada, N., Hasegawa, Y., Yue, M., Hamada, T., Nakagawa, S. & Ogawa, Y. (2015). Xist Exon 7 Contributes to the Stable Localization of Xist RNA on the Inactive X-Chromosome. *PLOS Genetics,* 11**,** e1005430.

Yang, L., Kirby, J. E., Sunwoo, H. & Lee, J. T. (2016). Female mice lacking Xist RNA show partial dosage compensation and survive to term. *Genes Dev,* 30**,** 1747-60.

Yang, Z. & Nielsen, R. (2002). Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages. *Molecular Biology and Evolution,* 19**,** 908-917.

Yang, Z., Wong, W. S. W. & Nielsen, R. (2005). Bayes Empirical Bayes Inference of Amino Acid Sites Under Positive Selection. *Molecular Biology and Evolution,* 22**,** 1107-1118.

Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol,* 24**,** 1586-91.

Yang, Z. & Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nat Rev Genet,* 13**,** 303-14.

Yap, K. L., Li, S., Muñoz-Cabello, A. M., Raguz, S., Zeng, L., Mujtaba, S., Gil, J., Walsh, M. J. & Zhou, M.-M. (2010). Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. *Molecular cell,* 38**,** 662-674.

Yates, A. D., Achuthan, P., Akanni, W., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Marugán, J. C., Cummins, C., Davidson, C., Dodiya, K., Fatima, R., Gall, A., Giron, C. G., Gil, L., Grego, T., Haggerty, L., Haskell, E., Hourlier, T., Izuogu, O. G., Janacek, S. H., Juettemann, T., Kay, M., Lavidas, I., Le, T., Lemos, D., Martinez, J. G., Maurel, T., Mcdowall, M., Mcmahon, A., Mohanan, S., Moore, B., Nuhn, M., Oheh, D. N., Parker, A., Parton, A., Patricio, M., Sakthivel, M. P., Abdul salam, A. I., Schmitt, B. M., Schuilenburg, H., Sheppard, D., Sycheva, M., Szuba, M., Taylor, K., Thormann, A., Threadgold, G., Vullo, A., Walts, B., Winterbottom, A., Zadissa, A., Chakiachvili, M., Flint, B., Frankish, A., Hunt, S. E., Iisley, G., Kostadima, M., Langridge, N., Loveland, J. E., Martin, F. J., Morales, J., Mudge, J. M., Muffato, M., Perry, E., Ruffier, M., Trevanion, S. J., Cunningham, F., Howe, K. L., Zerbino, D. R. & Flicek, P. (2020). Ensembl 2020. *Nucleic Acids Research,* 48**,** D682-D688.

Ye, B., Liao, D., Zhang, X., Zhang, P., Dong, H. & Huganir, R. L. (2000). GRASP-1: a neuronal RasGEF associated with the AMPA receptor/GRIP complex. *Neuron,* 26**,** 603-17.

Yen, Z. C., Meyer, I. M., Karalic, S. & Brown, C. J. (2007). A cross-species comparison of X-chromosome inactivation in Eutheria. *Genomics,* 90**,** 453-63.

Yi, W., Li, J., Zhu, X., Wang, X., Fan, L., Sun, W., Liao, L., Zhang, J., Li, X., Ye, J., Chen, F., Taipale, J., Chan, K. M., Zhang, L. & Yan, J. (2020). CRISPR-assisted detection of RNA–protein interactions in living cells. *Nature Methods,* 17**,** 685-688.

Yin, Y., Lu, J. Y., Zhang, X., Shao, W., Xu, Y., Li, P., Hong, Y., Cui, L., Shan, G., Tian, B., Zhang, Q. C. & Shen, X. (2020). U1 snRNP regulates chromatin retention of noncoding RNAs. *Nature,* 580**,** 147-150.

Yu, B., Van Tol, H. T. A., Stout, T. a. E. & Roelen, B. a. J. (2020). Initiation of X Chromosome Inactivation during Bovine Embryo Development. *Cells,* 9**,** 1016.

Yu, B., Qi, Y., Li, R., Shi, Q., Satpathy, A. T. & Chang, H. Y. (2021). B cell-specific XIST complex enforces X-inactivation and restrains atypical B cells. *Cell,* 184**,** 1790-1803.e17.

Yue, M., Ogawa, A., Yamada, N., Charles Richard, J. L., Barski, A. & Ogawa, Y. (2017). Xist RNA repeat E is essential for ASH2L recruitment to the inactive X and regulates histone modifications and escape gene expression. *PLoS Genet,* 13**,** e1006890.

Yue, M. & Ogawa, Y. (2017). CRISPR/Cas9-mediated modulation of splicing efficiency reveals short splicing isoform of Xist RNA is sufficient to induce X-chromosome inactivation. *Nucleic Acids Res*.

Zhai, W., Nielsen, R., Goldman, N. & Yang, Z. (2012). Looking for Darwin in Genomic Sequences—Validity and Success of Statistical Methods. *Molecular Biology and Evolution,* 29**,** 2889-2893.

Zhang, B., Xing, X., Li, J., Lowdon, R. F., Zhou, Y., Lin, N., Zhang, B., Sundaram, V., Chiappinelli, K. B., Hagemann, I. S., Mutch, D. G., Goodfellow, P. J. & Wang, T. (2014). Comparative DNA methylome analysis of endometrial carcinoma reveals complex and distinct deregulation of cancer promoters and enhancers. *BMC Genomics,* 15**,** 868.

Zhang, J., Nielsen, R. & Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol Biol Evol,* 22**,** 2472-9.

Zhang, L., Tran, N.-T., Su, H., Wang, R., Lu, Y., Tang, H., Aoyagi, S., Guo, A., Khodadadi-Jamayran, A., Zhou, D., Qian, K., Hricik, T., Côté, J., Han, X., Zhou, W., Laha, S., Abdel-Wahab, O., Levine, R. L., Raffel, G., Liu, Y., Chen, D., Li, H., Townes, T., Wang, H., Deng, H., Zheng, Y. G., Leslie, C., Luo, M. & Zhao, X. (2015). Cross-talk between PRMT1-mediated methylation and ubiquitylation on RBM15 controls RNA splicing. *eLife,* 4**,** e07938.

Zhang, L.-F., Huynh, K. D. & Lee, J. T. (2007). Perinucleolar Targeting of the Inactive X during S Phase: Evidence for a Role in the Maintenance of Silencing. *Cell,* 129**,** 693-706.

Zhao, W., Wang, L., Zhang, M., Wang, P., Qi, J., Zhang, L. & Gao, C. (2012). Nuclear to cytoplasmic translocation of heterogeneous nuclear ribonucleoprotein U enhances TLR-induced proinflammatory cytokine production by stabilizing mRNAs in macrophages. *J Immunol,* 188**,** 3179-87.

Zhu, X., Wu, Y.-B., Zhou, J. & Kang, D.-M. (2016). Upregulation of lncRNA MEG3 promotes hepatic insulin resistance via increasing FoxO1 expression. *Biochemical and Biophysical Research Communications,* 469**,** 319-325.

Zhu, Y., Wang, X., Forouzmand, E., Jeong, J., Qiao, F., Sowd, G. A., Engelman, A. N., Xie, X., Hertel, K. J. & Shi, Y. (2018). Molecular Mechanisms for CFIm-Mediated Regulation of mRNA Alternative Polyadenylation. *Mol Cell,* 69**,** 62-74.e4.

Zou, H., Yu, D., Du, X., Wang, J., Chen, L., Wang, Y., Xu, H., Zhao, Y., Zhao, S., Pang, Y., Liu, Y., Hao, H., Zhao, X., Du, W., Dai, Y., Li, N., Wu, S. & Zhu, H. (2019). No imprinted XIST expression in pigs: biallelic XIST expression in early embryos and random X inactivation in placentas. *Cell Mol Life Sci,* 76**,** 4525-4538.

Zuckerman, B., Ron, M., Mikl, M., Segal, E. & Ulitsky, I. (2020). Gene Architecture and Sequence Composition Underpin Selective Dependency of Nuclear Export of Long RNAs on NXF1 and the TREX Complex. *Molecular Cell,* 79**,** 251-267.e6.