CRISPR-Assisted Detection of RNA-Protein Interactions in Living Cells

Wenkai Yi[1,3,*], Jingyu Li[2,3,*], Xiaoxuan Zhu[2,3,*], Xi Wang[1,4,*], Ligang Fan[1,3], Wenju Sun[1], Linbu Liao[1], Jilin Zhang[5], Xiaoyu Li[1,3], Jing Ye[6], Fulin Chen[1], Jussi Taipale[5,7], Kui Ming Chan[2,3,#], Liang Zhang[2,3,#], Jian Yan[1,3,#]

1. School of Medicine, Northwest University, Xi'an, China
2. Key Laboratory of Biochip Technology, Biotech and Health Centre, Shenzhen Research Institute of City University of Hong Kong, Shenzhen, China
3. Department of Biomedical Sciences, City University of Hong Kong, Hong Kong SAR, China
4. Division of Theoretical Systems Biology, German Cancer Research Center, Heidelberg, Germany
5. Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden
6. Department of Pathology, The Fourth Military Medical University, Xi'an, China
7. Department of Biochemistry, University of Cambridge, Cambridge, UK

* These authors contributed equally to this work;

Correspondance should be addressed to: Jian Yan (jian.yan@cityu.edu.hk); Liang Zhang (liangzhang.28@cityu.edu.hk); Kui Ming Chan (ming.chan@cityu.edu.hk).

29

Abstract

We develop a CRISPR-Assisted RNA-Protein Interaction Detection method (CARPID), which leverages CRISPR/CasRx-based RNA targeting and proximity labeling to identify binding proteins of specific lncRNA in the native cellular context. Applied to the nuclear lncRNA XIST, CARPID captured a list of known interacting proteins and multiple previously uncharacterized binding proteins. We generalize CARPID to explore binders of lncRNA DANCR and MALAT1, revealing its wide applicability in identifying RNA binding proteins.

Main text

The long non-coding RNAs (lncRNA, defined as non-coding RNA of more than 200 nucleotides in length), have recently become widely recognized as important epigenetic regulating factors. The interplay with RNA binding proteins (RBPs) dictates the function and fate of RNA[1]. Despite their importance, significant technical limitations exist in elucidating lncRNA-protein interactions in living cells. Current methods mostly depend upon chemical- or UV-mediated crosslinking for efficient RNA-protein complex isolation[2,3], which may generate biases and mask physiological interactions[4]. Recently, a proximity-labeling-based methodology RaPID has been developed by integrating biotin ligase BASU with a navigation system of λN peptide recognizing RNA BoxB stem-loops[5]. Such system profoundly reduces non-specific background noise. However, the RNA of interest needs to be artificially engineered with a BoxB stem-loop and ectopically expressed, significantly curbing its broad application.

To circumvent the limitations and detect RBPs in living cells, we developed a method termed CRISPR-Assisted RNA-Protein Interaction Detection (CARPID) (**Fig. 1a**). Inspired by a strategy that utilized CRISPR/dCas9 to navigate biotin ligase to specific genomic loci[6], a nuclease-activity-free form of compact Type VI-D RNA CRISPR single-effector system dCasRx was employed for specific lncRNA targeting[7]. This dCasRx effector protein retained the capability of processing guide RNA (gRNA) arrays to two or more component gRNAs without cleaving targeted RNA transcripts. Taking this advantage, we designed a guide RNA array composed of two gRNA sequences spaced by a 30-nt direct repeat (DR) to target two adjacent loci on the same lncRNA transcript (**Extended Data Fig. 1a**; **Supplementary Table S1**), which in principle allows augmented targeting specificity with reduced background noise.

We fused the dCasRx with the engineered biotin ligase BASU followed by a self-cleaving T2A peptide and an eGFP to monitor their expression in living cells (**Extended Data Fig. 1b**). We tested various biotin-induction times to optimize the reaction duration. Either N- or C-terminal tagging of BASU to dCasRx did not lead to overt difference (**Extended Data Fig. 1c**). We chose BASU-dCasRx and treated the cells with 200 μM biotin for 15 min as the shortest but adequate reaction time in the subsequent analyses. We did not observe significantly altered gene expression in cells with over-expression of BASU-dCasRx and gRNAs, confirming that CARPID did not interfere the physiology of transfected cells (**Extended Data Fig. 1d**).

XIST is one of the well-studied mammalian lncRNA species and is expressed only from the inactive X chromosome (Xi) to regulate X chromosome inactivation (XCI) *in cis* in differentiated female cells[8,9]. Therefore, we focused on XIST to evaluate the performance of CARPID. We transfected HEK293T cells with BASU-dCasRx and three sets of distinct gRNAs that targeted different regions of XIST (**Extended Data Fig. 1a**). The specificity of these gRNA sets was confirmed by co-transfection with nuclease-active CasRx, which showed specific digestion of the targeted region without affecting the others (**Extended Data Fig. 2a**). Due to the highly ordered structure of XIST[10], we avoided targeting the predicted hairpin structures (**Extended Data Fig. 2b**; **Supplementary Data S1**). Three independent experiments were performed for each set of gRNAs to further "dilute" the non-specific noise from stochastic binding. To define the baseline of background biotinylation, we conducted control CARPID with an empty gRNA-expressing vector. Mass spectrometry (MS)-based protein identification showed that the majority of the detected proteins with at least two peptides (447 proteins) were shared among different gRNA sets and among the triplicates for

each set, demonstrating the robust reproducibility of CARPID (**Extended Data Fig. 3a**, **b**; **Supplementary Table S2**). For enrichment analysis, we applied label-free MS quantification and non-parametric rank product test with a cut-off of enrichment > 2-fold and adjusted p-value ≤ 0.05. As a result, 73 XIST-interacting proteins were significantly enriched by at least one set of gRNAs over controls (**Fig. 1b**; **Supplementary Table S3**). In addition, 23 proteins were discovered with at least two different sets of gRNAs, among which 13 proteins were shared by all three sets of gRNA pairs (**Extended Data Fig. 3c**, **d**; **Extended Data Fig. 4a**). More than a quarter of these strong XIST-interacting proteins (19/73) had been reported by previous studies[11] (**Supplementary Table S3**), including multiple functionally validated binders: Cohesin subunits (RAD21 and SMC1A), ATRX, and BRG1. We also noticed that some of the known XIST-interacting RBPs, e.g. SPEN and RBM15, did not pass our significance cut-off although their peptides were both detected. We reason that they may be hard to get enriched due to weak or dynamic bindings to XIST. The gene ontology (GO) analysis for the significantly enriched candidates showed that XIST-interacting proteins were largely involved in covalent chromatin modification and chromatin remodeling (**Extended Data Fig. 4b**).

In addition to known XIST interactors, CARPID also identified multiple novel factors, including a transcription initiation factor TFIID subunit TAF15 (**Fig. 1b**). TAF15 is a coactivator that recognizes TATA-box containing core promoters and facilitate transcription initiation. We first confirmed the association of TAF15 with XIST lncRNA using Western blotting (WB) and immunoFISH (**Fig. 1c**). TAF15 was reported as an RNA binding protein in mouse tissues[12]. Therefore, we re-explored the TAF15 CLIP-seq data[12] carried out in mouse brain and found that TAF15 indeed significantly bound to XIST lncRNA with over 9-fold enrichment of binding clusters than expected (**Extended Data Fig. 4c**,**d**). To validate the

binding in human cells, we also performed formaldehyde-assisted RIP-qPCR using antibody against TAF15. We detected over 4-fold enrichment of XIST transcript in TAF15 pulldown RNAs than IgG control, while no enrichment was shown for MALAT1, another abundant lncRNA in HEK293T cells (**Fig. 1d**). To investigate whether the binding of TAF15 depended on its biochemical binding affinity to the sequence feature of XIST, we carried out an HTR-SELEX experiment for TAF15 using a library containing 40-nt RNA fragments with randomized sequences (**Extended Data Fig. 5a**, **b**). The HTR-SELEX identified a significantly enriched RNA sequence motif of TAF15 which was similar to a previous report[12] (**Extended Data Fig. 5c**). Given that the abundant hairpin structure in XIST may cause dinucleotide interdependency, a position weight matrix model could not fully describe the impact of RNA sequence in TAF15 binding. Therefore, we employed a gapped k-mer based machine learning algorithm (see **Methods**) to model the RNA binding specificity of human TAF15 using HTR-SELEX data (**Extended Data Fig. 5d**; **Supplementary Data S2**). In concordance with both WB and MS results, the HTR-SELEX result further supported that TAF15 bound to XIST with higher affinity to locus 1/2 than locus 3 (**Fig. 1e**).

An ISWI (the imitation switch) chromatin remodeler[13], SNF2L was also identified in CARPID (**Fig. 1b**) and confirmed by WB and immunoFISH (**Extended Data Fig. 6a**). Consistently, the RIP-qPCR result showed that XIST was significantly associated with SNF2L (**Extended Data Fig. 6b**). IWSI has been known to associate with Cohesin complex in human cells[13]. In line with this, we also identified two Cohesin subunits SMC1A and RAD21 interacting with XIST **(Fig. 1b)**.

To validate the biological role of these two novel RBPs in XCI, we employed a female mouse embryonic fibroblast (MEF) cell line bearing an Xi-linked GFP transgene[14]. The GFP

transcript (**Fig. 1f**) or fluorescent signal cannot be observed (**Extended Data Fig. 7a**) due to XCI. However, when suppressing the DNA methylation with 5'-aza-cytosine (5-aza), GFP mRNA and fluorescent signals were both dramatically increased. To elucidate their functional significance in XCI, we depleted TAF15 and SNF2L in the presence of 5-aza treatment (**Extended Data Fig. 7b**). Strikingly, the 5-aza boosted GFP level was significantly re-diminished upon TAF15 silencing, which could be partially rescued by an ectopically expressed TAF15 clone resistant to RNAi knockdown (**Extended Data Fig. 7c**, **d**). We then asked whether the downregulation of GFP was ascribed to potential global reduction of transcription induced by TAF15 depletion. We examined five randomly selected autosomal genes on different chromosomes and found that their expression levels were virtually unaffected by TAF15 knockdown (**Extended Data Fig. 7e**). This supports the notion that the suppression of Xi-linked GFP is resulted from a specific role of TAF15 in XCI. We also conducted RNA-seq following the depletion of TAF15 in a female MEF cell line with hybrid genetic background[15]. Allelic expression of genes can be determined by the availability of SNPs and indels (insertion or deletion) between the two different genetic backgrounds. As expected, genes on X chromosome displayed stronger allelic depletion than autosomal genes (**Extended Data Fig. 7f**), confirming a role of TAF15 in antagonizing XCI. By contrast, the knockdown of SNF2L resulted in enhanced de-repression of GFP (**Fig. 1f**, **Extended Data Fig. 7a**). XIST likely recruits SNF2L to X chromosome and promotes XCI since SNF2L is known to condense the chromatin high-order structure and downregulate transcription[13]. Our finding of both TAF15 and SNF2L as XIST binders supports a model that recruiting repressive factors (e.g. SNF2L) and evicting transcription activators (e.g. TAF15) could both confer XIST-mediated XCI, consistent with previous models[8,16] (**Extended Data Fig. 7g**).

To generalize the application of CARPID for non-nuclear lncRNA, we designed gRNA sets to target lncRNA DANCR (differentiation antagonizing nonprotein coding RNA)[17] and MALAT1 (metastasis associated lung adenocarcinoma transcript 1)[18] **(Fig. 2a**,**b)**. It is important to note that DANCR is of ~1000 nucleotides long, much shorter than XIST. This, together with its low abundance (**Extended Data Fig. 8a**), makes it technically challenging to investigate with methods such as ChIRP-MS that requires dozens of different RNA probes for sufficient capture. Applying CARPID with the two gRNA sets targeting DANCR in HEK293T cells (**Extended Data Fig. 8b**), we detected 640 associating proteins (≥2 peptides; **Supplementary Table S4**) with 35 and 26 proteins significantly enriched at locus 1 and 2, respectively (**Fig. 2a**, **Extended Data Fig. 8c**, **d**; **Supplementary Table S5a-c**). Notably, GO analysis revealed that the vast majority of DANCR-binding proteins were associated with extracellular exosomes, denoting the presence of DANCR in this specialized cellular compartment (**Supplementary Table S5d**). We then examined RNA content from both exosomes and whole cell lysates of HEK293T cells (**Extended Data Fig. 9a**, **b**). Indeed, RT-qPCR analysis revealed an approximately 5-fold enrichment of DANCR in exosomes compared with the cell lysates (**Fig. 2c**). By contrast, XIST was drastically depleted in exosomes (**Extended Data Fig. 9c**). We also noticed one interesting DANCR binding protein Ezrin, a membrane bound cytoskeleton linker protein associated with poor prognosis in a number of cancers[19], was enriched for both DANCR gRNA targeted loci. RIP-qPCR revealed an approximately 2-fold enrichment of DANCR lncRNA in Ezrin pulldown over IgG control, which was not observed for MALAT1 lncRNA (**Fig. 2d**).

MALAT1 is known to be present both in nuclei and cytosol[20]. Two different gRNA sets enabled the capture of 484 proteins (≥2 peptide; **Supplementary Table S6**), out of which 43 were significantly enriched (**Fig. 2b**, **Extended Data Fig. 10a,b**; **Supplementary Table S7**).

Comparison of the CARPID results among the three lncRNAs with partially shared subcellular distribution displayed virtually no overlap (**Extended Data Fig. 10c**), demonstrating the high specificity and applicability of the CARPID method for lncRNA of different lengths and expression levels in various subcellular localizations.

We realize that CARPID has room for improvement, e.g. biotin label-based pulldown remains unavoidable. Technologies, such as XRNAX[21], OOPS[22], and PTex[23], utilize the physiochemical property to allow isolation and avoid affinity capture. However, these methods do not allow recognition of individual RNA-protein pairs. Therefore, CARPID and these methods can be complementary in investigating RBP-lncRNA interactions.

Acknowledgements

We are grateful to Prof. B. Ren (UC San Diego) and Dr. D. Leung (HKUST) for insightful comments during manuscript preparation. This work was supported by the City University of Hong Kong (7200595, 7005314, 9667188 and 9610424 to J.Yan), the National Natural Science Foundation of China (81873642 to J.Yan and 31900443 to W.S.), Research Grants Council of Hong Kong (21100615, 11102118, 11101919, C7007-17GF to K.M.C., and 21101917, 11103318 to L.Z.), the Shenzhen Science and Technology Fund Program (JCYJ20170818104203065, JCYJ20180307124019360 to K.M.C., JCYJ20170413141047772, JCYJ20180507181659781 to L.Z.); Opening Foundation of Key Laboratory of Resource Biology and Biotechnology in Western China (Northwest University), the Chinese Ministry of Education (J. Ye); and the Hong Kong Epigenomics Project of the EpiHK consortium (Lo Ka Chung Charitable Foundation) (J.Yan and K.M.C.).

Author contributions

W.Y., J.L., X.Z., K.M.C., L.Z. and J.Yan conceived the project. W.Y., J.L., X.Z., L.F., X.L. and L.Z. carried out experiments. X.W., W.S., L.L., J.Z., J.T., F.C., K.M.C., L.Z. and J.Ye performed data analysis. J.Yan, K.M.C. and L.Z. wrote the manuscript.

Ethics Declaration

The authors declare no conflict of competing interest. Correspondence and materials requests should be addressed to J.Yan, L.Z. or K.M.C. No human or animal subject is included in this study.

Reference

1 Castello, A. *et al.* Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. *Cell* **149**, 1393-1406, doi:10.1016/j.cell.2012.04.031 (2012).

2 Licatalosi, D. D., Ye, X. & Jankowsky, E. Approaches for measuring the dynamics of RNA-protein interactions. *Wiley Interdiscip Rev RNA* **11**, e1565, doi:10.1002/wrna.1565 (2020).

3 Ramanathan, M., Porter, D. F. & Khavari, P. A. Methods to study RNA-protein interactions. *Nat Methods* **16**, 225-234, doi:10.1038/s41592-019-0330-1 (2019).

4 Baranello, L., Kouzine, F., Sanford, S. & Levens, D. ChIP bias as a function of cross-linking time. *Chromosome Res* **24**, 175-181, doi:10.1007/s10577-015-9509-1 (2016).

5 Ramanathan, M. *et al.* RNA-protein interaction detection in living cells. *Nat Methods* **15**, 207-212, doi:10.1038/nmeth.4601 (2018).

6 Myers, S. A. *et al.* Discovery of proteins associated with a predefined genomic locus via dCas9-APEX-mediated proximity labeling. *Nat Methods* **15**, 437-439, doi:10.1038/s41592-018-0007-1 (2018).

7 Konermann, S. *et al.* Transcriptome Engineering with RNA-Targeting Type VI-D CRISPR Effectors. *Cell* **173**, 665-676 e614, doi:10.1016/j.cell.2018.02.033 (2018).

8 Minajigi, A. *et al.* Chromosomes. A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation. *Science* **349**, doi:10.1126/science.aab2276 aab2276 (2015).

9 Lee, J. T. Epigenetic regulation by long noncoding RNAs. *Science* **338**, 1435-1439, doi:10.1126/science.1231776 (2012).

10 Lu, Z. *et al.* RNA Duplex Map in Living Cells Reveals Higher-Order Transcriptome Structure. *Cell* **165**, 1267-1279, doi:10.1016/j.cell.2016.04.028 (2016).

245 11 Mira-Bontenbal, H. & Gribnau, J. New Xist-Interacting Proteins in X-Chromosome
246 Inactivation. *Curr Biol* **26**, 1383, doi:10.1016/j.cub.2016.05.009 (2016).
247 12 Kapeli, K. *et al.* Distinct and shared functions of ALS-associated proteins TDP-43,
248 FUS and TAF15 revealed by multisystem analyses. *Nat Commun* **7**, 12143,
249 doi:10.1038/ncomms12143 (2016).
250 13 Hakimi, M. A. *et al.* A chromatin remodelling complex that loads cohesin onto human
251 chromosomes. *Nature* **418**, 994-998, doi:10.1038/nature01024 (2002).
252 14 Chan, K. M., Zhang, H., Malureanu, L., van Deursen, J. & Zhang, Z. Diverse factors
253 are involved in maintaining X chromosome inactivation. *Proc Natl Acad Sci U S A*
254 **108**, 16699-16704, doi:10.1073/pnas.1107616108 (2011).
255 15 Yildirim, E., Sadreyev, R. I., Pinter, S. F. & Lee, J. T. X-chromosome hyperactivation
256 in mammals via nonlinear relationships between chromatin states and transcription.
257 *Nat Struct Mol Biol* **19**, 56-61, doi:10.1038/nsmb.2195 (2011).
258 16 Jegu, T. *et al.* Xist RNA antagonizes the SWI/SNF chromatin remodeler BRG1 on the
259 inactive X chromosome. *Nat Struct Mol Biol* **26**, 96-109, doi:10.1038/s41594-018-
260 0176-8 (2019).
261 17 Kretz, M. *et al.* Suppression of progenitor differentiation requires the long noncoding
262 RNA ANCR. *Genes Dev* **26**, 338-343, doi:10.1101/gad.182121.111 (2012).
263 18 Hutchinson, J. N. *et al.* A screen for nuclear transcripts identifies two linked
264 noncoding RNAs associated with SC35 splicing domains. *BMC Genomics* **8**, 39,
265 doi:10.1186/1471-2164-8-39 (2007).
266 19 Li, J. *et al.* Prognostic Value of Ezrin in Various Cancers: A Systematic Review and
267 Updated Meta-analysis. *Sci Rep* **5**, 17903, doi:10.1038/srep17903 (2015).

20 Wilusz, J. E., Freier, S. M. & Spector, D. L. 3' end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. *Cell* **135**, 919-932, doi:10.1016/j.cell.2008.10.012 (2008).

21 Trendel, J. *et al.* The Human RNA-Binding Proteome and Its Dynamics during Translational Arrest. *Cell* **176**, 391-403 e319, doi:10.1016/j.cell.2018.11.004 (2019).

22 Queiroz, R. M. L. *et al.* Comprehensive identification of RNA-protein interactions in any organism using orthogonal organic phase separation (OOPS). *Nat Biotechnol* **37**, 169-178, doi:10.1038/s41587-018-0001-2 (2019).

23 Urdaneta, E. C. *et al.* Purification of cross-linked RNA-protein complexes by phenol-toluol extraction. *Nat Commun* **10**, 990, doi:10.1038/s41467-019-08942-3 (2019).

Figure Legends

Fig. 1 |CARPID identifies lncRNA XIST-associated proteins in living cells.

(a) Scheme of the CARPID workflow.

(b) Volcano plot of XIST-associated proteins identified by CARPID. Significantly enriched proteins are labelled as orange dots. Proteins previously validated to interact with XIST are in orange font (n=3 and 9 independent experiments for control and XIST group, respectively).

(c) Top: WB detection of TAF15 in input and streptavidin IP samples of control (C) and three XIST gRNA sets (L1, L2 and L3). Bottom: immunoFISH images of XIST and TAF15 in HEK293T cells. The white boxed region on the left is magnified and shown on the right. Three independent experiments were carried out with similar results and a representative result is shown.

(d) Validation of XIST-TAF15 interaction using RIP assay (mean ± SEM, n=3 independent experiments, two-sided paired Student's t-test).

(e) RNA binding specificity of TAF15 using HTR-SELEX. The blue curve shows the predicted binding affinity of TAF15 along XIST RNA, compared with averaged value from 1,000 randomly selected genomic fragments (orange).

(f) X-linked GFP de-repression under depletion of TAF15 (shTaf15-07 and shTaf15-44) and SNF2L (shSnf2l-29 and shSnf2l-31) in iMEF, using SmcHD1 (shSmcHD1) as a positive control (mean ± SD, n=3 independent experiments, two-sided unpaired Student's t-test using NT+5-aza as controls).

Fig. 2 | Identification of lncRNA DANCR/MALAT1-associated proteins in living cells.

(a,**b)** The volcano plots show the enrichment of DANCR **(a)** and MALAT1 **(b)** associated proteins in HEK293T cells. Significantly enriched proteins are labelled as orange dots (n=3 and 6 independent experiments for control and lncRNA group, respectively).

(c) Comparison of DANCR abundance in cell lysates and exosomes. Two sets of qPCR primers were used (**Extended Data Fig. 8b**; mean ± SD, n=3 independent experiments with 3 technical replicates respectively, two-sided unpaired Student's t test).

(d) Validation of XIST-Ezrin interaction using formaldehyde-assisted RIP assay (mean ± SEM, n = 4 independent experiments for DANCR and n=3 independent experiments for MALAT1, two-sided paired Student's t test). Inset shows the abundance of Ezrin in Input sample and IP samples.

Online Methods

A step-by-step protocol is available as a Supplementary Protocol (**Supplementary Information**) and an open resource in Protocol Exchange[24].

Plasmid Cloning

BASU was sub-cloned from RaPID plasmid (Addgene #107250)[5] into the backbone of an EF1a-dCasRx-2A-EGFP plasmid (Addgene #109050)[7] to generate BASU-dCasRx or dCasRx-BASU constructs with a Gibson assembly kit (NEB).

For RNAi-based knockdown assay in iMEF cells, shRNA oligos were synthesized and cloned into pLKO.1-puro (Addgene #8453). Two different shRNAs were used for TAF15 (shTaf15-44 and shTaf15-07) and SNF2L (shSnf2l-29 and shSnf2l-31), respectively. One shRNA was used for SmcHD1 (shSmcHD1) as a positive control for X-linked GFP de-repression. For the rescue experiment of TAF15 knockdown assay, full-length TAF15 coding sequence (CDS) was cloned and site-directed mutagenesis was performed to generate shRNA resistant construct. For RNAi-based knockdown assay, siRNA oligos were purchased from Qiagen (SI03650325; SI04915806). For HTR-SELEX assay, the human TAF15 full-length ORF was sub-cloned into pETG20A-SBP for *E.coli* expression with a Thioredoxin+6×His-tag at N-terminus and an SBP-tag at C-terminus. All cloning primers sequences were included in **Supplementary Table S8**.

Design of CRISPR/CasRx gRNA sets

We generated gRNA sets (Addgene #109054) composed of two gRNAs spaced by 30-nt direct repeats (DRs) to target two adjacent loci on the same lncRNA. The gRNA sequences were chosen at GC%=40%~60% and filtered for off-target. The uniquely aligned sequences were further aligned to the transcriptome. The spacing between the pairs will be calculated. The output structure was predicted based on the result of RNAfold[25]. The higher value of

"DotPercent" (RNAfold) suggests a lower probability of forming hairpin structure. The sequence pairs with expected RNAfold structure ratio (column F and J in **Supplementary Table S1**) and spacing (column K: 16~20nt) were manually selected from these results. For XIST gRNA design, we extracted its secondary structural data from PARIS analysis[10] to further inspect the potential of forming stem loops at the gRNA targeting loci. None of gRNA pairs targeted any hairpin-rich regions. Paired guide RNA sets of lncRNA XIST, DANCR, and MALAT1 were also listed in **Supplementary Table S1**.

Cell Culture and Transfection

HEK293T cells and iMEF E2C4 cells were cultured in DMEM supplemented with 10% FBS and 1% penicillin-streptomycin (all from Life Technologies) at 37°C with 5% CO_2. The hybrid iMEF cell line was a gift from Dr. Jeannie T Lee[15] in which inactive X chromosome is of *Mus musculus* (129S1) origin and active X chromosome is of *Mus castaneus* (CAST/EiJ) origin.

HEK293T cells were transfected with 0.1% PEI (Polysciences). Note that 1.5 μg of gRNA expressing plasmid and 1.5 μg of BASU-dCasRx constructs were co-transfected with 9 μg of PEI. For qPCR assays, transfections were performed in 6-well plates and 0.5×10^6 cells were seeded in each well 24-h before transfection.

Lentiviral Packaging and Infection

The lentivirus packaging was performed following a previously established protocol[14]. Virus-containing cell culture medium was collected at 48-h post-transfection. The infected iMEF cells were cultured in selection medium (culture medium with 1.5 μg/ml of puromycin) and collected for the downstream analysis at 72-h post-infection. For the rescue experiment, overexpression-virus was mixed with shRNA-virus before infecting cells. For 5-aza treatment,

10 mM 5-aza stock was prepared in DMSO and freshly diluted with culture medium before use. A working concentration of 0.2 μM was used for 72-h before collection and the 5-aza containing medium was refreshed every 24-h. An equal amount of DMSO was added to the control cells.

CARPID-Western Blot

Cells were washed twice with cold PBS (ThermoFisher) and lysed with 1 ml lysis buffer (50 mM Tris-HCl (pH 7.4); 150 mM NaCl; 0.5% TritonX-100; 1 mM EDTA supplemented with fresh protease inhibitors (Roche)) at 4°C for 10 min with rotating. Then, lysate was spun down at 15,000 rpm for 10 min, at 4°C. The supernatant was quantified and normalized for protein concentration, which was sampled for Input. Biotinylated proteins were enriched with MyOne T1 streptavidin beads (ThermoFisher) after 2-h incubation at 4°C with rotation and three washes were performed with 1 ml ice cold lysis buffer. Proteins were eluted from the beads into elution buffer by incubation for 10 min at 95°C and labeled as IP. The signals were visualized using the Bio-Rad ChemiDoc Imaging System. ImageJ software (version 1.8.0_172) was used to quantify the WB signals. The antibodies used in this study are listed in **Supplementary Table S9**.

Mass Spectrometry Analysis

For on-beads digestion, enriched streptavidin beads were washed with 50 mM ammonium bicarbonate (pH 8.0) (Sigma) at 4°C for three times. Beads were re-suspended in 50 μl Elution buffer I (50 mM Tris-HCl pH 8.0; 2 M urea; 10 μg/ml Sequencing Grade Trypsin (ThermoFisher); 1 mM DTT) and mixed at 400 rpm, 30°C for 60 min. The supernatant was collected into a fresh vial. The beads were additional eluted twice with 25 μl Elution buffer II (50 mM Tris-HCl pH 8.0; 2 M urea; 5 mM iodoacetamide) and all three elutes were combined. An additional 0.25 μg trypsin was added to the combined elutes, followed by

incubation at 37°C overnight. The digestion was quenched by adding 10% formic acid solution (FA) at a ratio of 1:25 (v/v). The digested samples were then desalted using C18 tips (ThermoFisher) following manufacturer's instruction and reconstituted in 20 µl 0.1% FA. The LC-MS/MS analysis was performed using an Easy-nLC 1200 system coupled to a Q Exactive HF mass spectrometry (ThermoFisher).

LC-MS/MS Data Analysis

Raw files created by XCalibur 4.0.27 (ThermoFisher) software were analyzed using the Proteome Discoverer 2.2 software (ThermoFisher), against the UniProt human protein database in Sequest HT node. The precursor and fragment mass tolerances were set to 10 ppm and 0.02 Da, respectively. The maximum of two missed cleavage sites of trypsin was allowed. Carbamidomethylation (C) was set as static modification, and oxidation (M) and acetyl (protein N-terminal) were set as variable modifications. False discovery rate (FDR) of peptide spectrum matches (PSMs) and peptide identification were determined using the Percolator algorithm at 1% based on q-value. For label-free quantification (LFQ), the Minora Feature Detector node was used in the processing workflow, and the Precursor Ions Quantifier node and the Feature Mapper node in the consensus workflow.

Enrichment analyses were applied for proteins identified with ≥2 peptides. Human keratins were included in all analyses but were not included in the figures. The LFQ abundances were normalized across the pulldowns and logarithmized. Missing values were imputed with values representing the detection limit of the mass spectrometer. The rank products test[26] was employed to determine proteins statistically enriched in the gRNA expressing samples against empty vector controls. Proteins with an adjusted p-value of ≤ 0.05 and over 2-fold change of abundance were considered statistically enriched. In all volcano plots, the x-axis indicates the log2 fold change of protein levels in CARPID combining all gRNAs relative to

control. The y-axis shows the negative logarithm transformed BH-adjusted p-values (two-sided non-parametric test conducted by rank product method).

Fluorescence Microscopy

The GFP detection was conducted following a previously established protocol[14]. Briefly, the infected iMEF cells were seeded on chamber slides. Cells were fixed with 3% paraformaldehyde (PFA) at room temperature (RT) for 12 minutes and then counter-stained with DAPI. Slides were examined under Eclipse Ni-E upright fluorescence microscope (Nikon).

ImmunoFISH was performed following a previously published protocol[27]. Briefly, HEK293T cells were grown on 2 cm × 2 cm glass coverslips. Cells were washed twice with pre-warmed PBS before being fixed in 4% PFA for 10 min at RT and permeabilized with cold PBS containing 0.5% Triton X-100 (Sigma). Cells were then blocked with 5% BSA for 60 min at RT. After three washes with PBST, cells were further incubated with primary antibodies at 4°C overnight and washed three times with PBST. The following procedures were conducted in darkness at all times. Cells were incubated with secondary antibodies conjugated with CF 488A (2 mg/ml, Biotium) for 1 hour at RT. After removing the solution, cells were incubated with 10% formamide in 2×SSC (Life Technologies) for 5 min at RT. Cells were then incubated with 100 µl lncRNA XIST oligonucleotide probes (100 nM) with 5'-amino Cy3 modification (Sangon, Shanghai) for 1 hour at RT before washing with 10% formamide in 2×SSC for three times (5 min per wash). Cells were counter-stained with DAPI and mounted with antifade (ThermoFisher). The slides were examined with a confocal microscope (Zeiss LSM 880). The FISH probe sequences are in **Supplementary Table S8**.

HTR-SELEX experimental procedure

Selection reactions were performed as following: ~200 ng of TAF15 was mixed with ~1 µg RNA selection ligands in 20 µl Promega buffer at 4°C (50 mM NaCl, 1 mM $MgCl_2$, 0.5 mM Na_2EDTA and 4% glycerol in 50 mM Tris-HCl, pH 7.5). The reaction was conducted with incubation for 15 min at 37°C followed by additional 15 min at RT. Then, 50 µl of 1:50 diluted paramagnetic nickel beads (GE-Healthcare) blocked with 0.1% Tween 20 and 0.1 µg/µl BSA (NEB) were added in. TAF15-RNA complex was incubated with beads for 2-h, and then the unbound ligands were removed. After washes, the beads were suspended in elution buffer (0.5 µM RT-primer, 1 mM EDTA and 0.1% Tween20 in 10 mM Tris-HCl buffer, pH 7.0) and heated for 5 minutes at 70°C to denature the proteins. The retro-transcription primers were annealed, followed by retro-transcription and PCR amplification for library construction and sequencing. The efficiency of the selection process was evaluated by running a qPCR reaction in parallel with the standard PCR reaction.

Using HTR-SELEX Data to Predict Binding Potential of TAF15 on XIST RNA

To obtain the Position Weight Matrix model, Autoseed with default parameters was applied to the TAF15 HTR-SELEX data to identify the seed sequence and multinomial algorithm was employed to generate the motif[28]. We also adopted a supervised machine learning approach, gkm-SVM[29], to model the binding preference of TAF15 on RNA sequences based on HTR-SELEX data, and then applied the trained model to predict the binding potential of TAF15 on XIST. For model training, we took HTR-SELEX reads in cycles 2-4 as the positive sequences and other sequences in cycle 0 as negative sequences. Both positive and negative sequences were randomly down-sampled to 100,000 sequences due to computational capacity. To search for the best model, we considered three key parameters of gkm-SVM: *l*, the whole word length including gaps; *k*, the number of informative (i.e. non-gapped) positions in each word; and *d*, the maximum number of mismatches allowed. We used a 5-fold cross-validation for parameter combinatorial search, and when l=7, k=3, d=4 the highest cross-validation

accuracy of 87.3% was achieved. Finally, we used the best model to score all 7-mers occurring in the XIST RNA and plotted the smoothed gkm-SVM prediction scores along the XIST transcript.

Formaldehyde assisted RNA Immunoprecipitation Assay (RIP)

RIP was conducted according to a previous published protocol with small modifications[16]. Briefly, ~10 million HEK293T cells were grown on 10-cm plates and cross-linked with 1% formaldehyde for 10 min at RT. Crosslinking was then quenched by 0.125M glycine at RT for 5 min. After three washes with 10 ml ice-cold PBS, cells were scraped and spun down at 4°C. Cell pellets were resuspended in RIPA lysis buffer supplemented with fresh 0.5 mM DTT, protease inhibitors (Roche) and SUPERase-In RNase Inhibitor (ThermoFisher) and incubated with rotation for 15 min at 4°C. Cell lysates were sonicated with Covaris sonicator (M220), followed by spin-down at 16,000×g at 4°C for 10 min. Each sample was diluted to 1 ml with RIP binding buffer (25 mM Tris pH 7.5, 150 mM KCl, 0.5% NP40, 5 mM EDTA, 0.5 mM DTT supplemented with protease inhibitors and SUPERase-In RNase Inhibitor). The cell lysates were pre-cleared with 50 µl of Protein G Sepharose beads (ThermoFisher) at 4°C for 3-h. After taking 50 µl sample as an INPUT, the remaining lysate was subjected to immunoprecipitation process with 5 µg IgG control or specific antibodies (**Supplementary Table S9**) and incubated overnight at 4°C. Then, 50 µl of Protein G Sepharose beads blocked with RIPA+0.2% BSA was added to each sample and further incubated for 2-h. Beads were washed with 1 ml RIP binding buffer for three times followed by elution with 150 µl extraction buffer (1% SDS in TE buffer) for 15 min at 37°C. Protein digestion and decrosslinking were conducted by adding 100 µg of Proteinase K and incubating at 55°C for 1-h and 65°C for 1-h. RNA was isolated by TRIzol (ThermoFisher), followed by DNase I digestion. RNA was extracted again and reverse-transcribed with MiniBEST Universal RNA Extraction Kit (TAKARA) and PrimeScript™ RT Master Mix (Clontech), respectively.

Quantitative Real Time PCR was performed using TB Green™ Premix Ex Taq™ II (TAKARA). The primers used are designated in **Supplementary Table S8**.

lncRNA associated Protein-Protein Interaction Network Analysis

Protein-protein interaction information is extracted from STRING database[30]. LncRNA associated proteins were subjected to STRING and interaction scores ("Combined Score" from STRING) were retrieved and used for network establishment. The network is built with Cytoscape[31] and edges between proteins were drawn when interaction scores were over 0.40. The width of the edge is in proportion to the value of interaction score.

Gene Ontology Analysis

Gene Ontology (GO) analysis on biological process was performed with DAVID database (v6.8)[32,33].

RNA-seq, LM-seq and RT-qPCR Analyses

Total RNA was isolated using MiniBEST Universal RNA Extraction Kit (TAKARA) following manufacturer's instruction. RNA-seq library was prepared using LM-seq library prep protocol[34]. First, mRNA was isolated from total RNA with Next Poly A+ Isolation Kit (NEB) and followed by retro-transcription (Clontech SmartScribe kit). cDNA was purified with AMPure XP beads (Beckman) for the following adaptor ligation and PCR amplification. PCR product was then purified and prepared for further sequencing (BGI, Shenzhen). For allele-specific RNA-seq, total RNA was first processed to remove rRNA by NEBNext rRNA depletion kit (NEB) and then proceed to prepare RNA-seq library using NEBNext Ultra™ II directional RNA library prep kit for Illumina (NEB). Paired-end reads were mapped to the human reference genome (hg19) with HISAT2 (v2.1.0)[35,36] with default parameters. The gene expression values were calculated by gfold (v1.1.4)[37], gene expression results were

incorporated to a matrix and filtered using an in-house program (genes that had RPKM ≥ 0.1 in at least 60% of the samples were retained). Pearson correlation analysis between samples was performed using R-package (v3.5.1). For RT-qPCR, cDNAs were generated from total RNA with PrimeScript™ RT Master Mix (Clontech), and RT-qPCR was performed using TB Green™ Premix Ex Taq™ II (TAKARA). All RT-qPCR primer sequences were included in **Supplementary Table S8**. All RNA-seq sequencing reads are available in GEO with an accession number GSE137556.

Allelic expression analysis of RNA-seq data

We performed RNA-seq on a hybrid mouse embryonic fibroblast cell line with *mus* and *cast* hybrid genetic background to further confirm the role of TAF15 on XCI. The hybrid mouse was a crossing of a *Mus musculus* (129S1/SvImJ) female mouse and a *Mus castaneus* (CAST/EiJ) male mouse. To generate the reference genomes of 129S1/SvImJ and CAST/EiJ, we downloaded from the Mouse Genomes Project (v6) VCF files containing SNPs and indels of 129S1/SvImJ and CAST/EiJ. Thereafter, the vcf2diploid tool (version 0.2.6) in the AlleleSeq pipeline was used to reconstruct the genomes by incorporating the SNPs and indels into the C57BL/6J genome[38]. The transcriptome reference sequences were obtained using the liftOver tool based on the Ensembl gene annotation. RNA-seq reads were mapped to both 129S1/SvImJ and CAST/EiJ transcriptome reference sequences using Bowtie2 (version 2.1.0). Concordantly mapped read pairs were then assigned to the 129S1/SvImJ or CAST/EiJ allele by comparing the mapping errors.

To identify genes exhibiting significant allelic expression changes upon 5-aza treatment, a bootstrapping strategy[39] was adapted to achieve a bootstrap distribution of the allelic expression change in each gene, which was then summarized as a mean and a standard deviation. A Z-score was computed to represent the folds of standard deviation (SD) the

bootstrapping mean deviated from zero. The Z-scores were then converted to p-values, and adjusted for FDR using BH correction procedure. At FDR = 0.05, we identified 3 genes on the X chromosome and 37 autosomal genes with significant allelic changes after 5-aza treatment.

Conditional Media and Exosomes Isolation

Before making the conditional media, the FBS was ultra-centrifuged at 120,000×g for 18-h at 4°C using a SW32Ti rotor (Beckman) to deplete existing extracellular vesicles (EVs). The supernatant was filtered through a 0.22 μm membrane. Five million HEK293T cells were seeded onto 10 cm plates in 10 ml EV-free medium. Conditional media were collected after 48-h culture. Isolation of exosomes was conducted based on a previously published protocol[40]. Briefly, the procedures involve sequential ultracentrifugation at 300×g for 10 min, 2,000×g for 15 min, 10,000×g for 70 min, and 120,000×g for 2 hr. The pellets were washed once with PBS and precipitated at 120,000×g for 70 min. Purified exosomes were further processed for downstream protein or RNA analyses.

Statistical Analysis

Statistical analyses were performed using GraphPad Prism and R.

Data availability

Sequencing data have been deposited in Gene Expression Omnibus database under the accession number GSE137556. The secondary structure prediction data of XIST lncRNA from PARIS analysis used in **Extended Data Fig. 2** were from a previous study (GEO accession: GSE74353)[10] and also deposited to **Supplementary Data S1** after additional processing. The raw sequence reads for TAF15 HTR-SELEX are deposited to **Supplementary Data S2**. Mass Spectrometry raw data have been deposited to iProX database under the project ID number IPX0001797000. The full scan of WB and corresponding quantification are included in **Source Data**. The TAF15 CLIP-seq data used in **Extended Data Fig. 4c**,**d** were from a previous study (GEO accession: GSE77700)[12].

Methods-only References

24 Yi, W. *et al.* CRISPR-Assisted Detection of RNA-Protein Interactions in Living Cells. *Protocol Exchange*, doi:10.21203/rs.3.pex-904/v1 (2020).

25 Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**, 26, doi:10.1186/1748-7188-6-26 (2011).

26 Breitling, R., Armengaud, P., Amtmann, A. & Herzyk, P. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett* **573**, 83-92, doi:10.1016/j.febslet.2004.07.055 (2004).

27 Yue, M., Charles Richard, J. L., Yamada, N., Ogawa, A. & Ogawa, Y. Quick fluorescent in situ hybridization protocol for Xist RNA combined with immunofluorescence of histone modification in X-chromosome inactivation. *J Vis Exp*, e52053, doi:10.3791/52053 (2014).

28 Arttu Jolma, J. Z., Estefania Mondragón, Teemu Kivioja, Yimeng Yin, Fangjie Zhu, Quaid Morris, Timothy R. Hughes, L. James Maher III, Jussi Taipale. Binding specificities of human RNA binding proteins towards structured and linear RNA sequences. *bioRxiv*, doi:10.1101/317909 (2019).

29 Ghandi, M., Lee, D., Mohammad-Noori, M. & Beer, M. A. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput Biol* **10**, e1003711, doi:10.1371/journal.pcbi.1003711 (2014).

30 Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* **47**, D607-D613, doi:10.1093/nar/gky1131 (2019).

31 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**, 2498-2504, doi:10.1101/gr.1239303 (2003).

32 Huang da, W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57, doi:10.1038/nprot.2008.211 (2009).

33 Huang da, W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**, 1-13, doi:10.1093/nar/gkn923 (2009).

34 Hou, Z. *et al.* A cost-effective RNA sequencing protocol for large-scale gene expression studies. *Sci Rep* **5**, 9570, doi:10.1038/srep09570 (2015).

35 Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* **12**, 357-360, doi:10.1038/nmeth.3317 (2015).

36 Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37**, 907-915, doi:10.1038/s41587-019-0201-4 (2019).

37 Feng, J. *et al.* GFOLD: a generalized fold change for ranking differentially expressed genes from RNA-seq data. *Bioinformatics* **28**, 2782-2788, doi:10.1093/bioinformatics/bts515 (2012).

38 Rozowsky, J. *et al.* AlleleSeq: analysis of allele-specific expression and binding in a network framework. *Mol Syst Biol* **7**, 522, doi:10.1038/msb.2011.54 (2011).

39 Hou, J. *et al.* Extensive allele-specific translational regulation in hybrid mice. *Mol Syst Biol* **11**, 825, doi:10.15252/msb.156240 (2015).

40 Luga, V. *et al.* Exosomes mediate stromal mobilization of autocrine Wnt-PCP signaling in breast cancer cell migration. *Cell* **151**, 1542-1556, doi:10.1016/j.cell.2012.11.024 (2012).

fig. 1

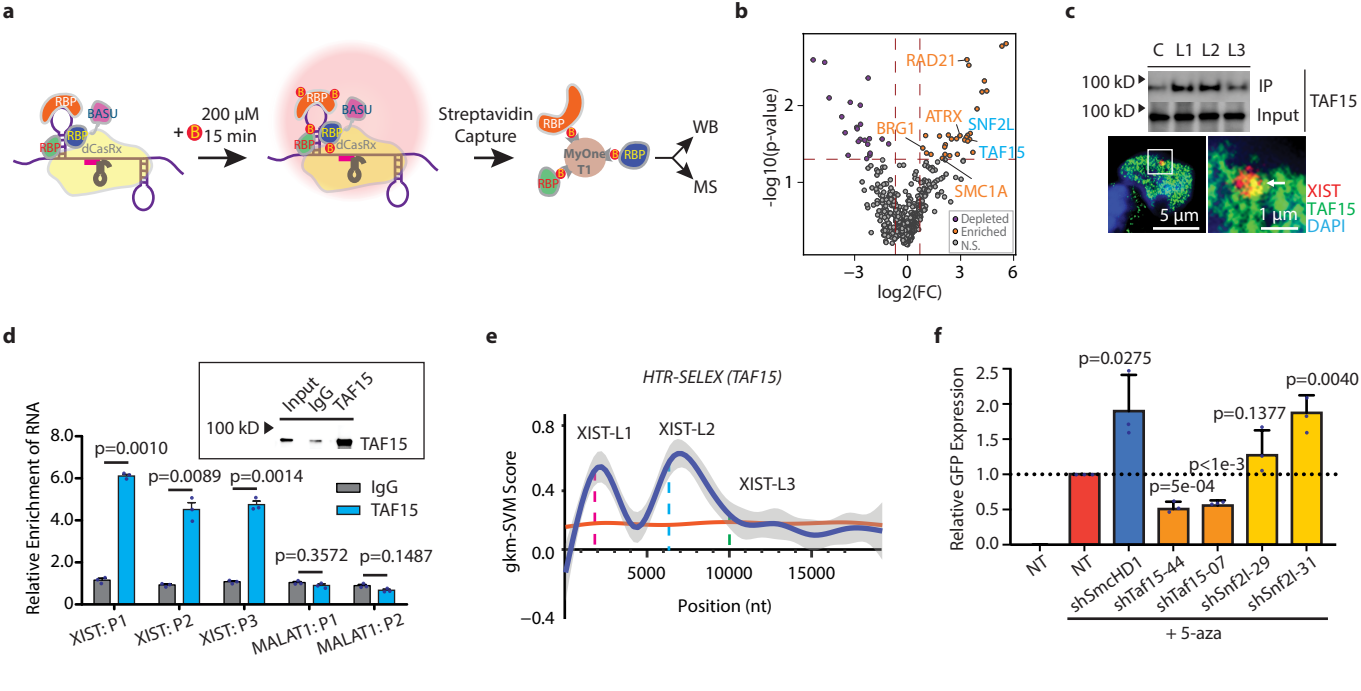
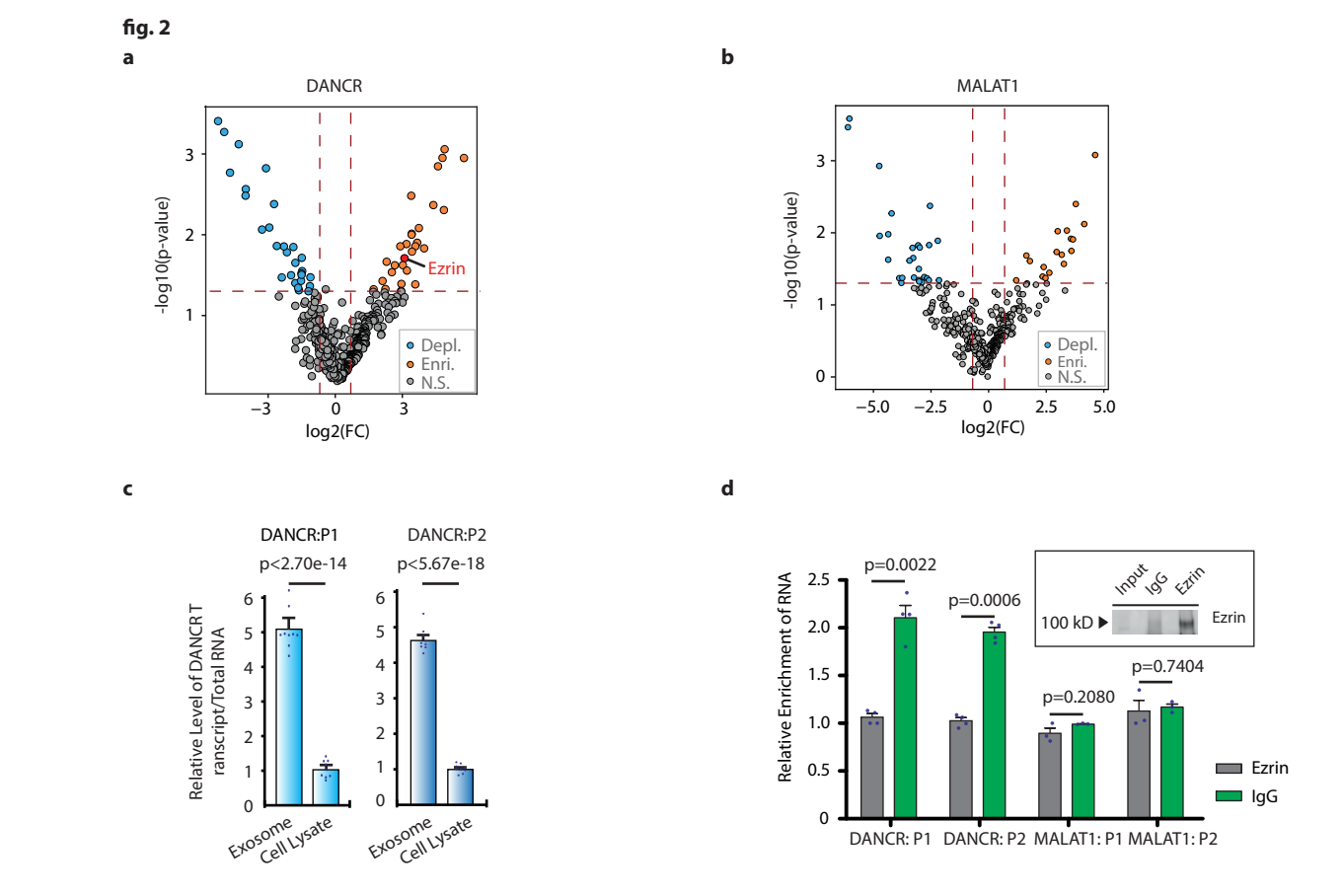
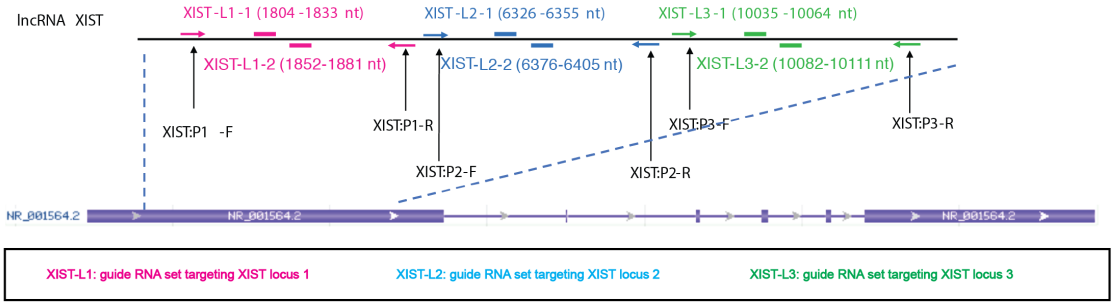

a
RBP
BASU
RBP
dCasRx
200 μM
+ B 15 min
Streptavidin
Capture
MyOne T1
WB
MS
b
-log10(p-value)
log2(FC)
RAD21
ATRX
BRG1
SNF2L
TAF15
SMC1A
Depleted
Enriched
N.S.
c
C L1 L2 L3
100 kD
IP
Input
TAF15
5 μm
1 μm
XIST
TAF15
DAPI
d
Relative Enrichment of RNA
Input
IgG
TAF15
100 kD
TAF15
p=0.0010
p=0.0089
p=0.0014
p=0.3572
p=0.1487
IgG
TAF15
XIST: P1
XIST: P2
XIST: P3
MALAT1: P1
MALAT1: P2
e
HTR-SELEX (TAF15)
gkm-SVM Score
XIST-L1
XIST-L2
XIST-L3
Position (nt)
f
Relative GFP Expression
p=0.0275
p=0.0040
p=0.1377
p<1e-3
p=5e-04
NT
NT
shSmcHD1
shTaf15-44
shTaf15-07
shSnf2l-29
shSnf2l-31
+ 5-aza

fig. 2

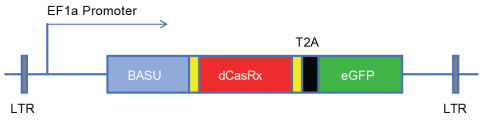

a
DANCR
-log10(p-value)
log2(FC)
Ezrin
Depl.
Enri.
N.S.
b
MALAT1
-log10(p-value)
log2(FC)
Depl.
Enri.
N.S.
c
DANCR:P1
p<2.70e-14
DANCR:P2
p<5.67e-18
Relative Level of DANCR Transcript/Total RNA
Exosome
Cell Lysate
d
Relative Enrichment of RNA
p=0.0022
p=0.0006
p=0.2080
p=0.7404
Input
IgG
Ezrin
100 kD
Ezrin
Ezrin
IgG
DANCR: P1
DANCR: P2
MALAT1: P1
MALAT1: P2

Extended Data Fig. 1

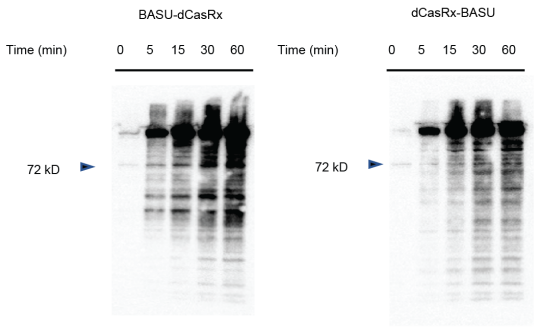
a



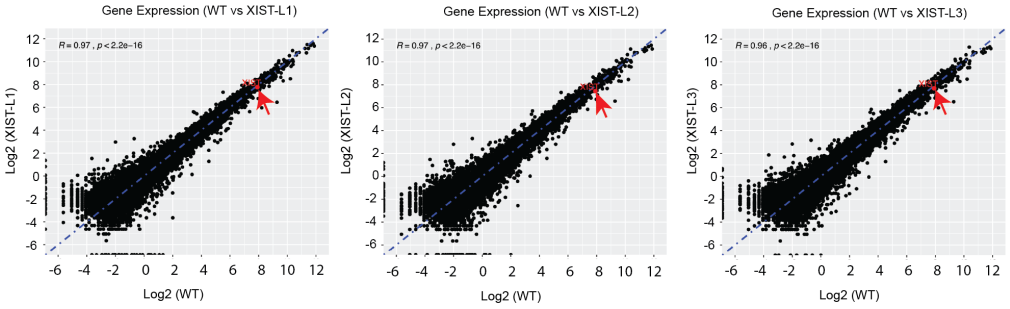
b



c

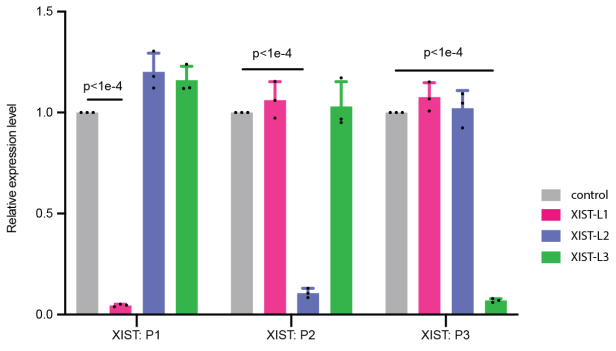


d

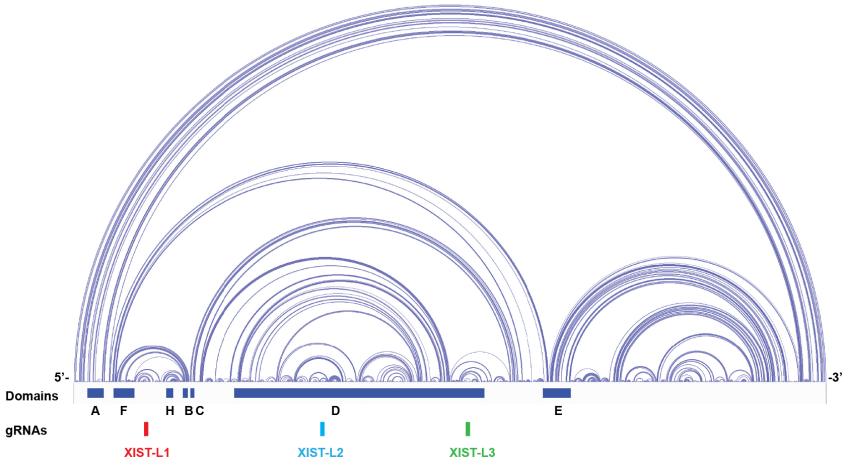


Extended Data Fig. 2

a

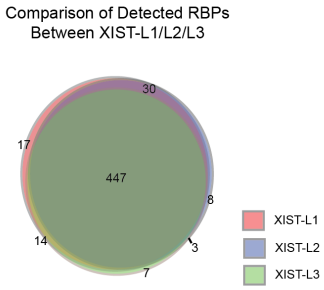


b

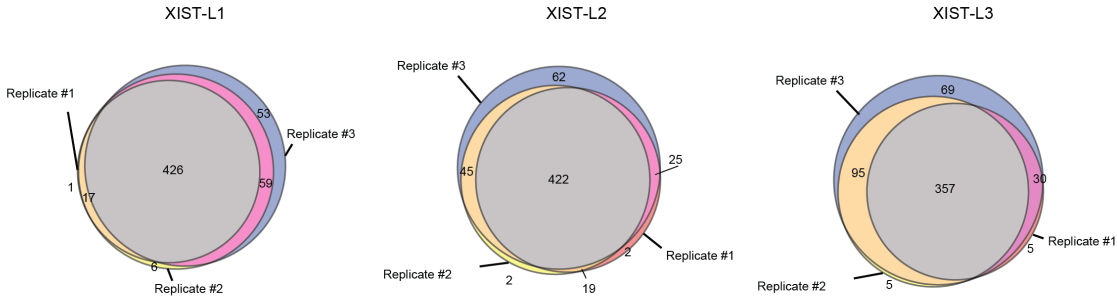


Extended Data Fig. 3

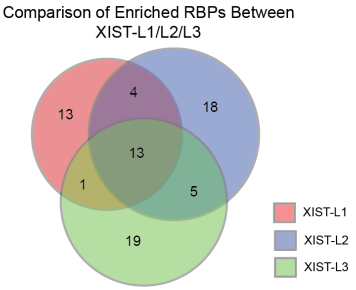
a



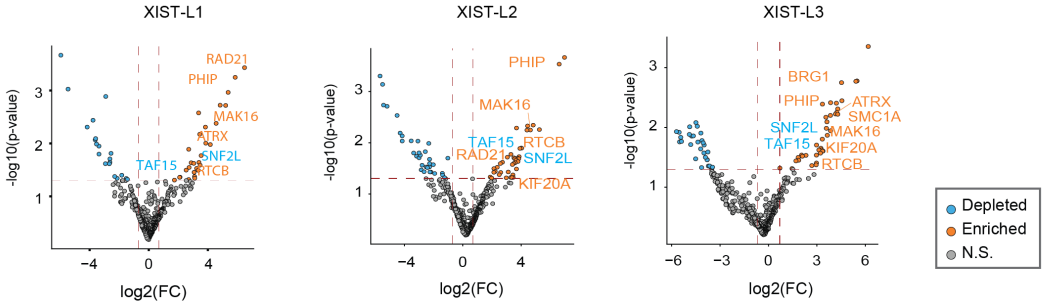
b



c

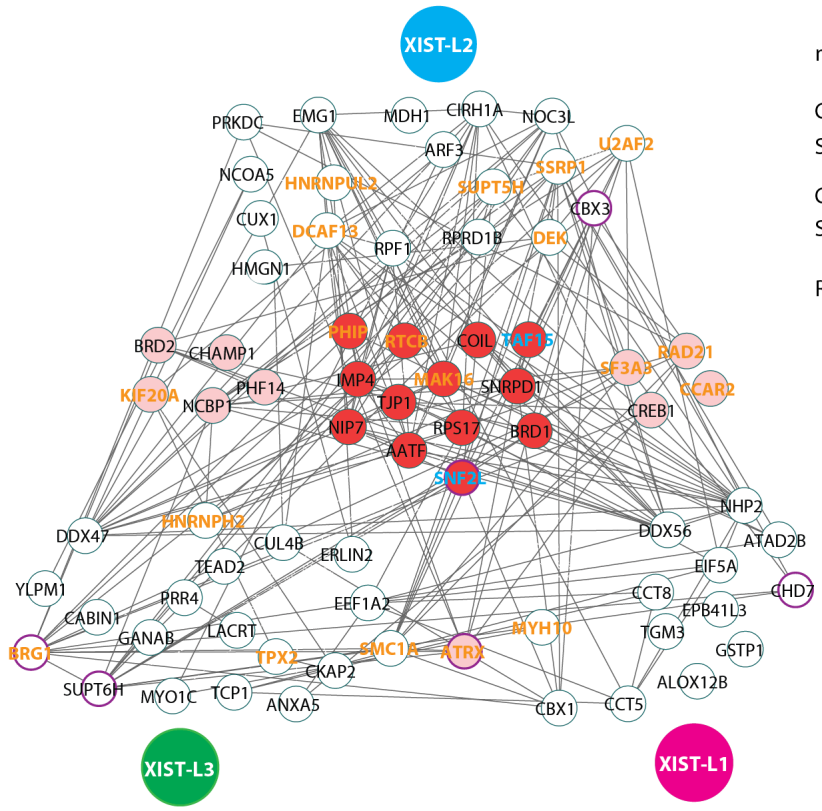


d

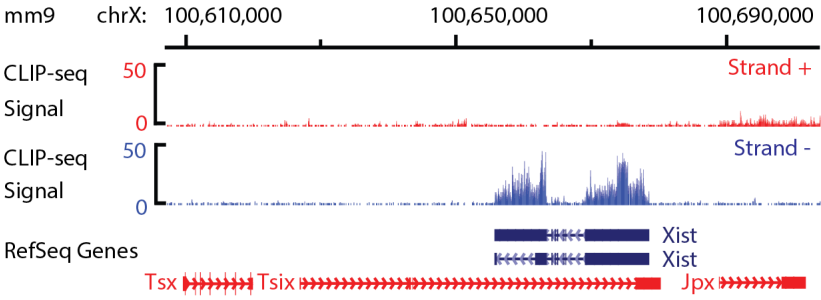


Extended Data Fig. 4

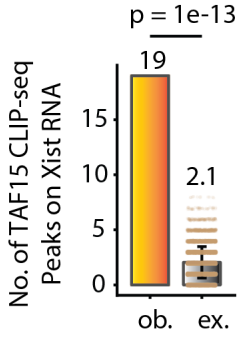
a



c



d

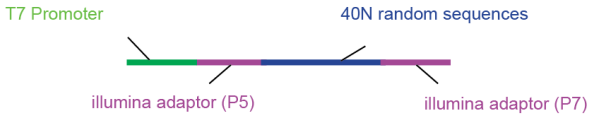


b

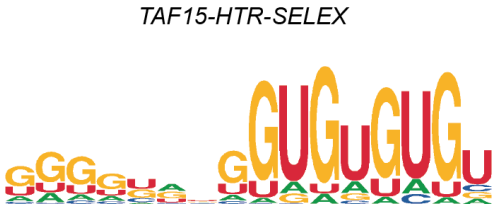
GO term	p Value	Genes Involved
Covalent Chromatin Modification	6.0e-5	ATRX, SNF2L, BRG1, BRD2, CHD7
Chromatin Remodeling	5.8e-4	ATRX, SNF2L, BRG1, CHD7
Positive Regulation of Transcription	2.1e-3	ATRX, RAD21, BRG1, AATF, CREB1, CHD7, CKAP2, PHIP
RNA splicing	3.8e-3	CCAR2, NCBP1, SNRPD1, SF3A3
RNA Maturation	2.8e-2	MAK16, NHP2

Extended Data Fig. 5

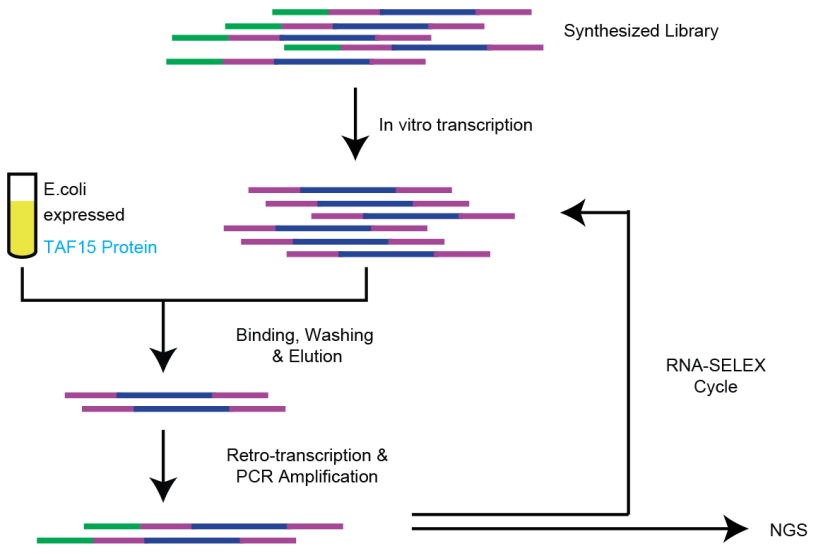
a



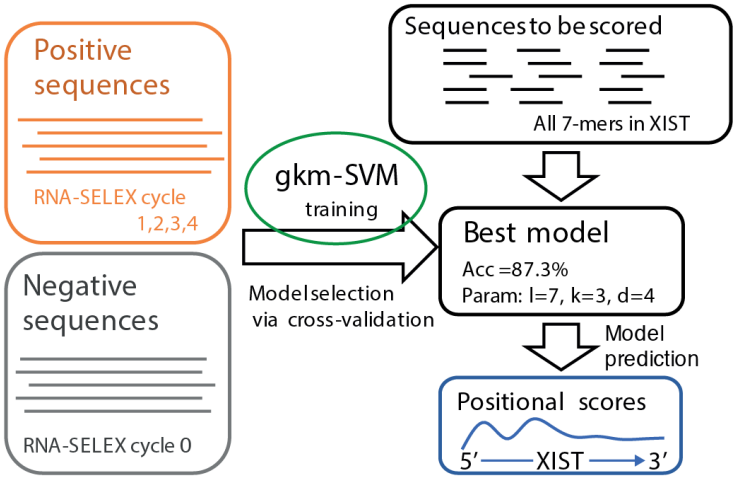
c



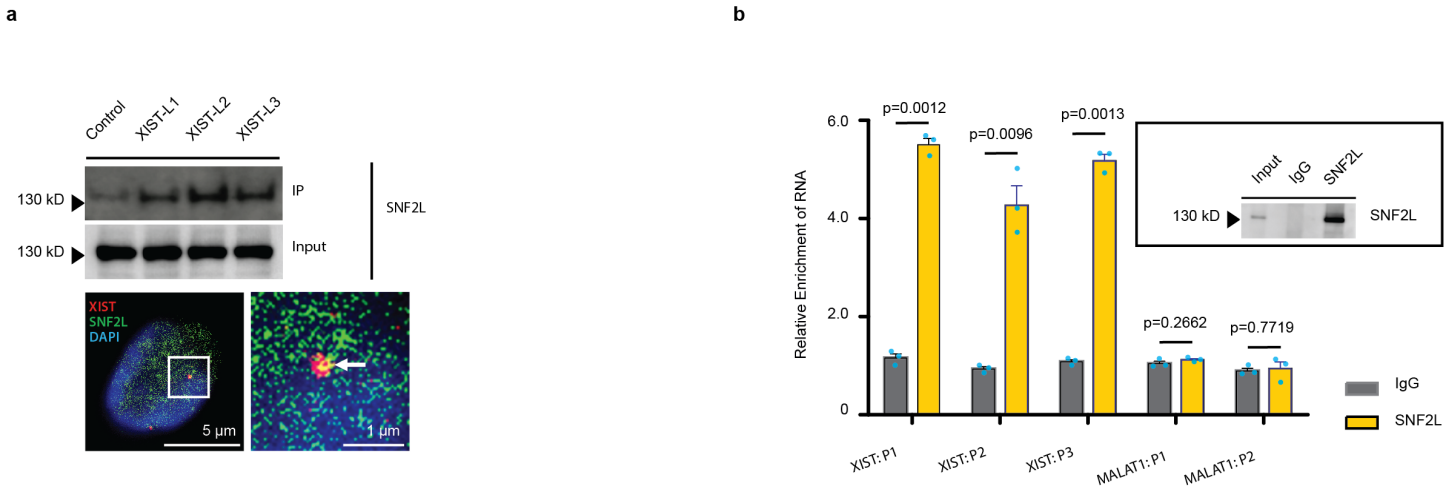
b



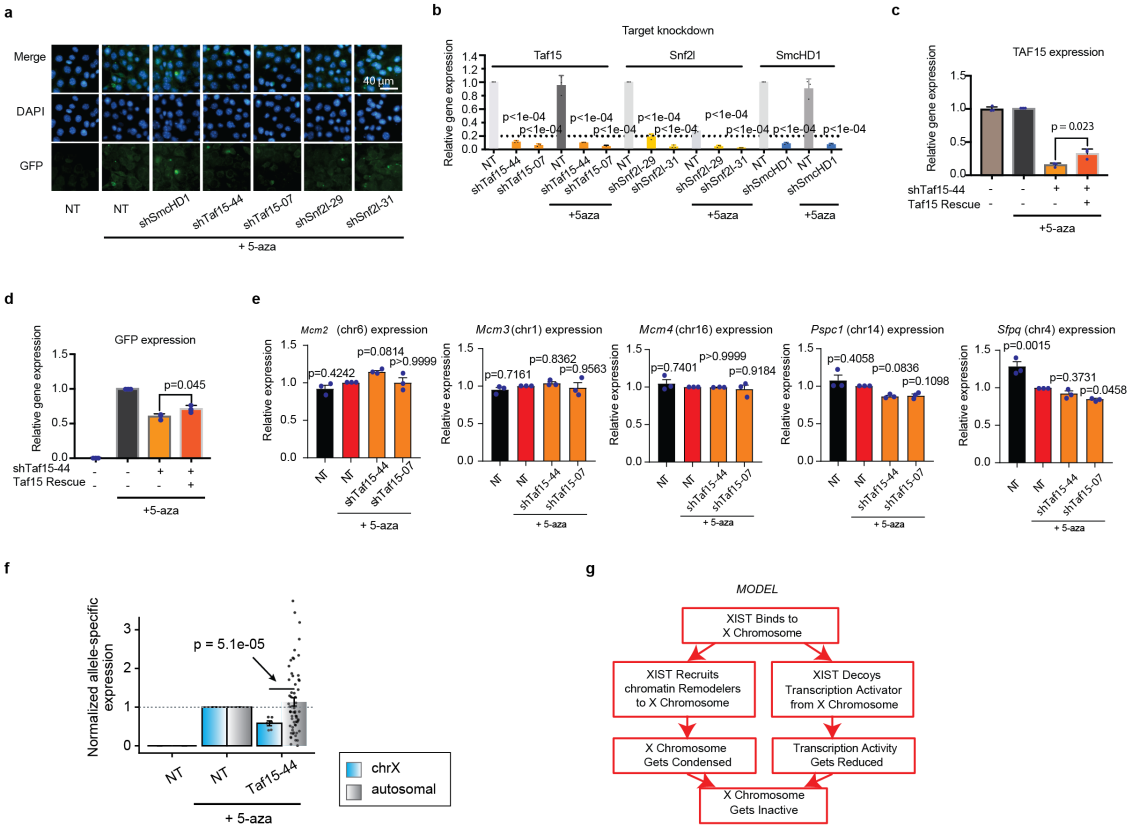
d



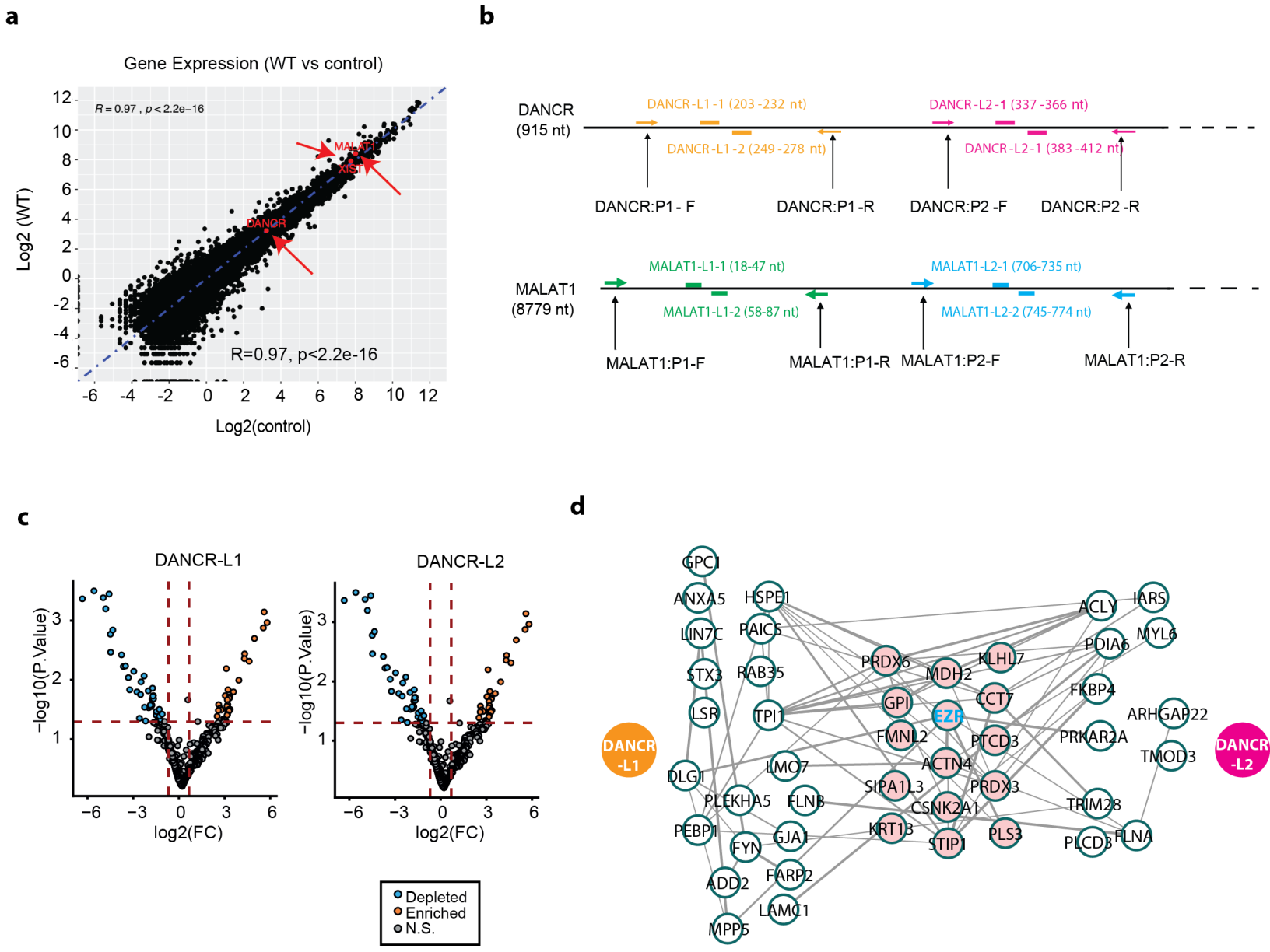
Extended Data Fig. 6



Extended Data Fig. 7

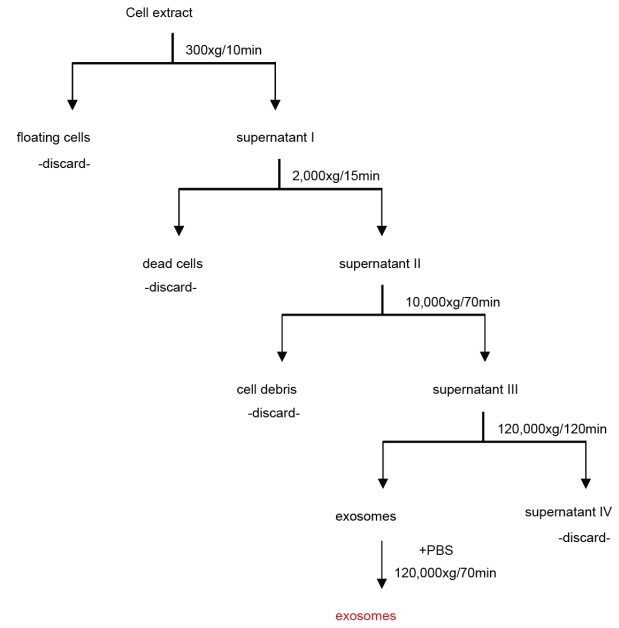


Extended Data Fig. 8

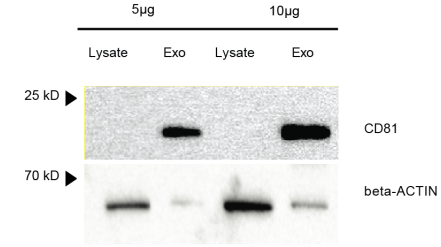


Extended Data Fig. 9

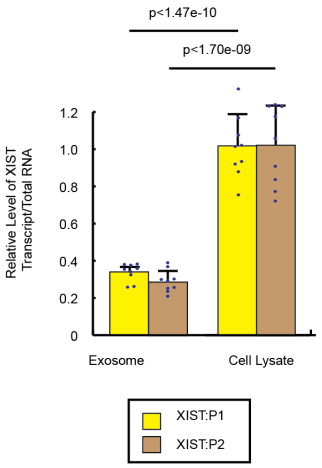
a



b



c



Extended Data Fig. 10