US 20210340592A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2021/0340592 A1**

Zhang et al. (43) **Pub. Date: Nov. 4, 2021**

(54) **METHOD FOR DETERMINING LONG NON-CODING RIBONUCLEIC ACID INTERACTION PROTEINS**

(71) Applicant: **City University of Hong Kong**, Kowloon (HK)

(72) Inventors: **Liang Zhang**, New Territories (HK); **Jian Yan**, Kowloon (HK); **Jingyu Li**, Kowloon (HK); **Wenkai Yi**, New Territories (HK); **Ligang Fan**, Kowloon (HK)

(21) Appl. No.: **17/240,009**

(22) Filed: **Apr. 26, 2021**

(30) **Foreign Application Priority Data**

Apr. 30, 2020 (CN) 202010367970.0

Publication Classification

(51) **Int. Cl.**

C12Q 1/25	(2006.01)
C12N 9/00	(2006.01)
C12N 9/22	(2006.01)
C12N 15/11	(2006.01)
G16B 40/00	(2006.01)

(52) **U.S. Cl.**

CPC ***C12Q 1/25*** (2013.01); ***C12N 9/93*** (2013.01); ***C12N 9/22*** (2013.01); *G01N 2440/32* (2013.01); ***G16B 40/00*** (2019.02); *C12N 2310/20* (2017.05); ***C12N 15/11*** (2013.01)

(57) **ABSTRACT**

The present invention provides a novel method for determining a long-chain non-coding ribonucleic acid interaction protein. The present invention provides a fusion protein formed by BASU and dCasRx, a mammalian expression vector for expressing said fusion protein. The method for determining the lncRNA interaction protein according to the present invention comprises: co-transfecting a mammalian expression vector that expresses the fusion protein and a gRNA that specifically targets the target lncRNA into target cells, thereby BASU specifically biotin-labeling effector proteins nearby; isolating the biotinylated proteins by using a streptavidin affinity coupled magnetic bead and then eluting, and digesting by trypsin and quantitatively analyzing by a label-free mass spectrometry. The present invention can highly credibly determine the proteins that interact with lncRNA.

Specification includes a Sequence Listing.

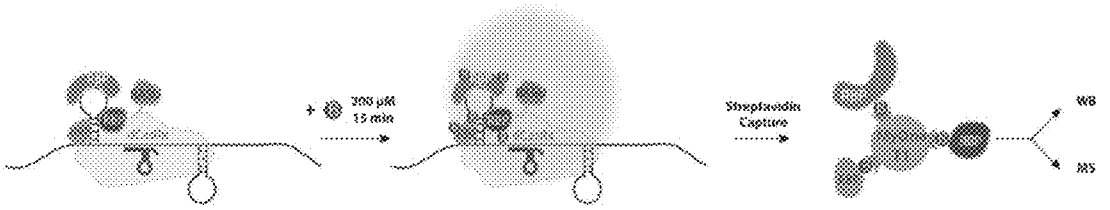

Figure 1a

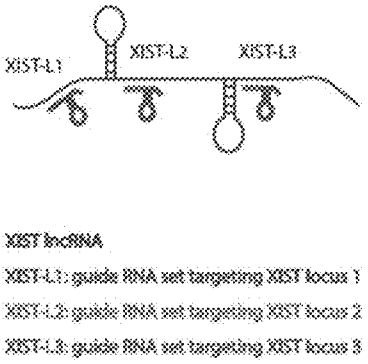

Figure 1b

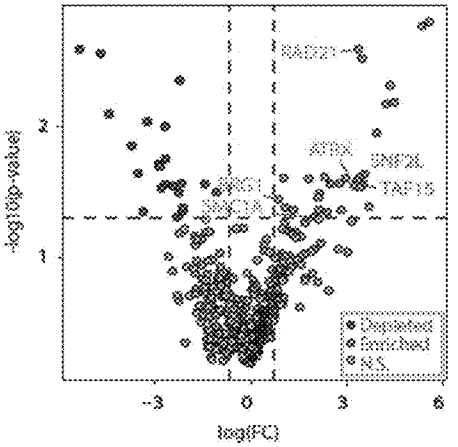

Figure 1c

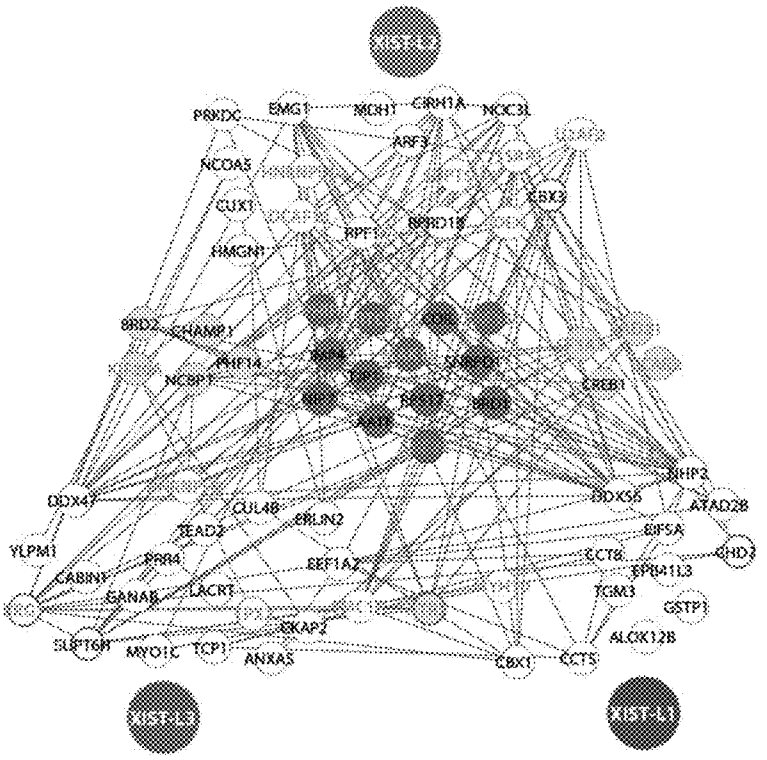

Figure 1d

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

Figure 1e

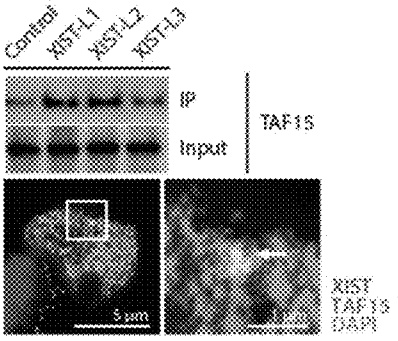

Figure 2a

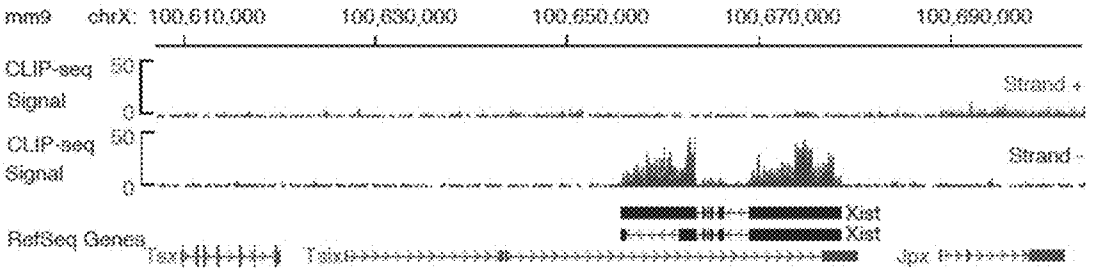

Figure 2b

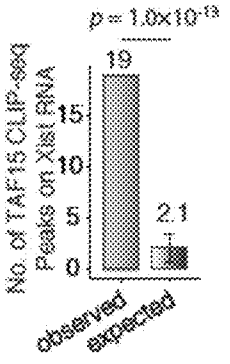

Figure 2c

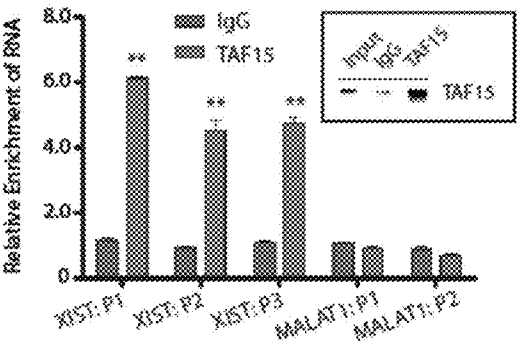

Figure 2d

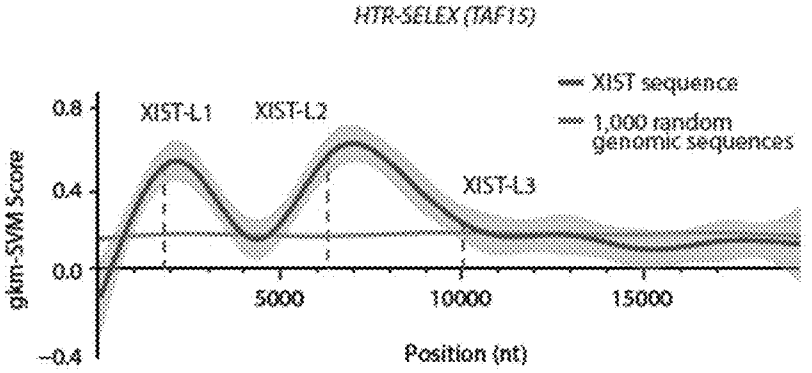

Figure 2e

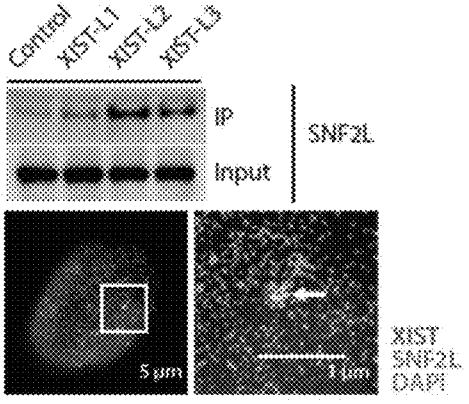

Figure 2f

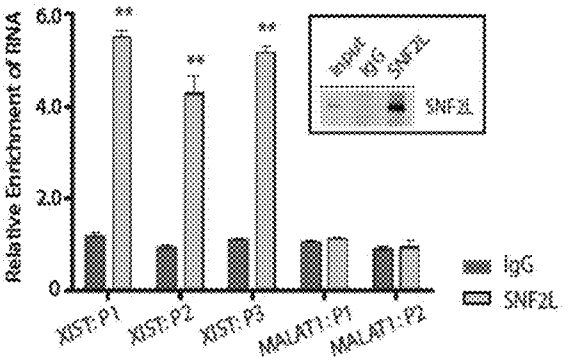

Figure 2g

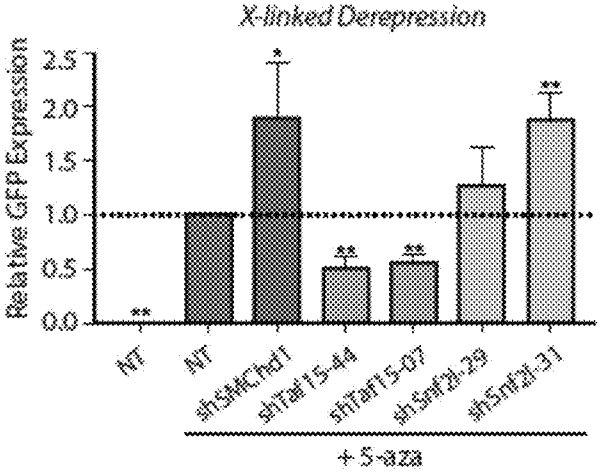

Figure 2h

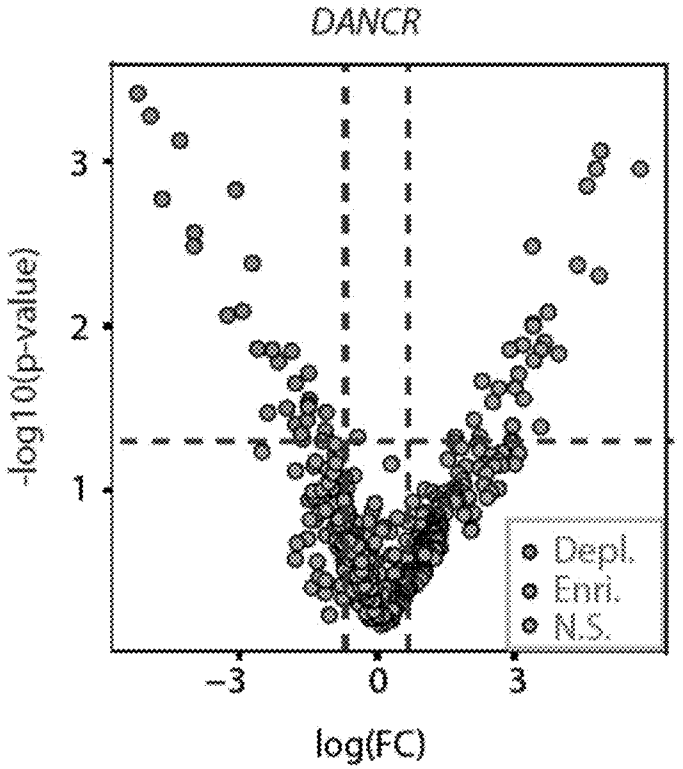

Figure 3a

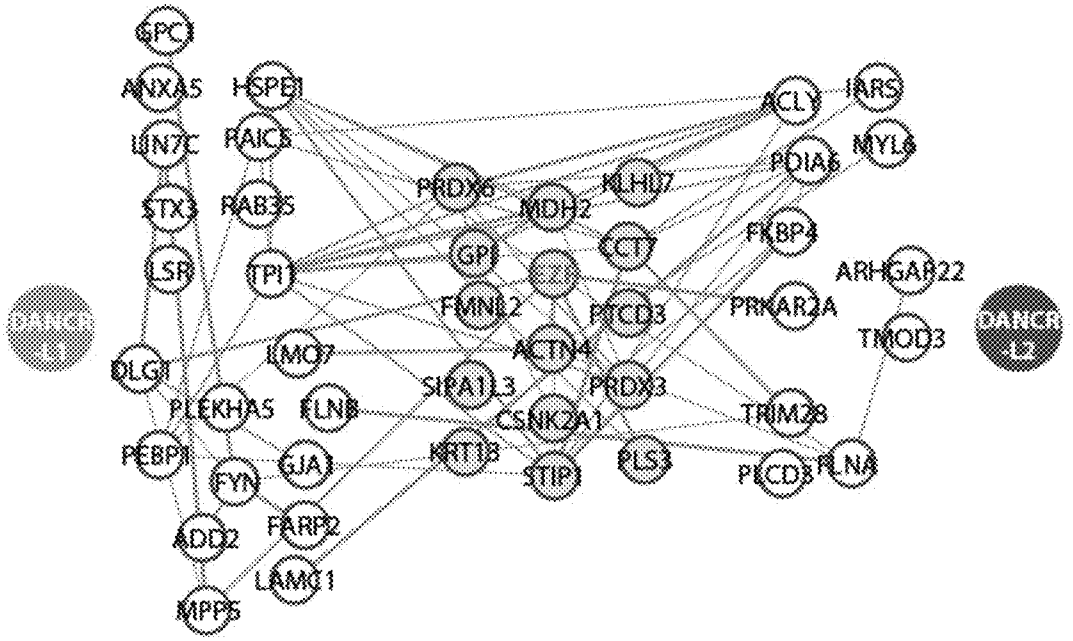

Figure 3b

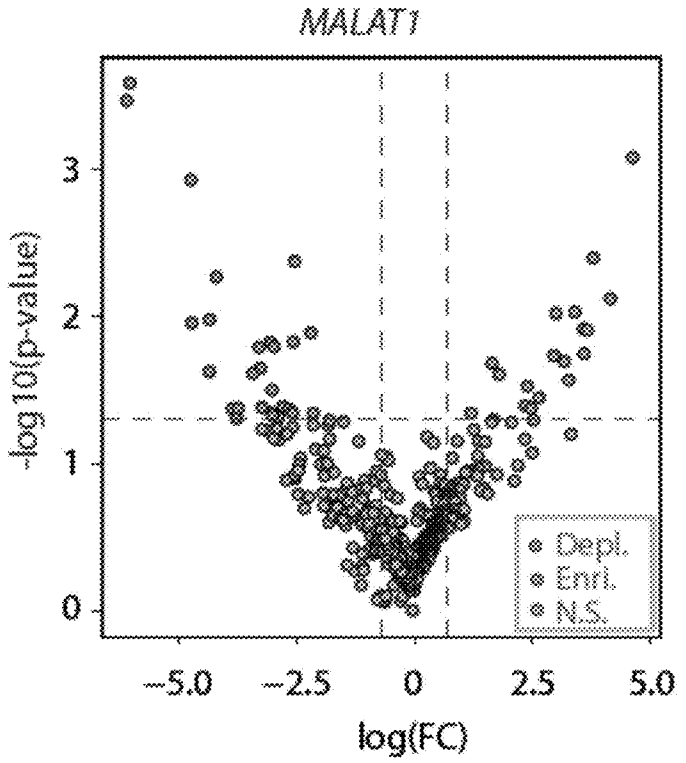

Figure 3c

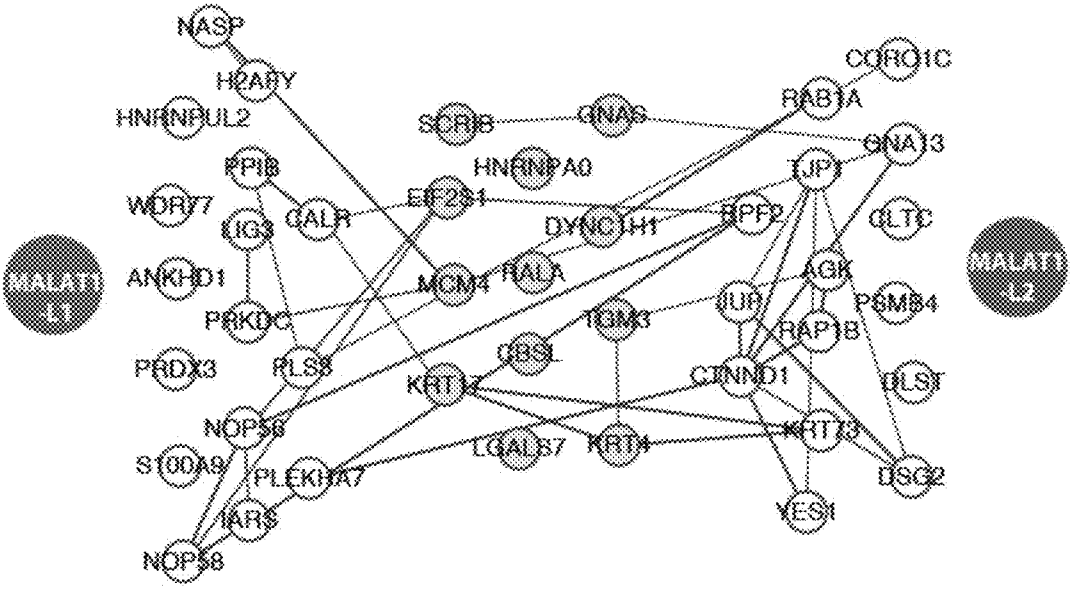

Figure 3d

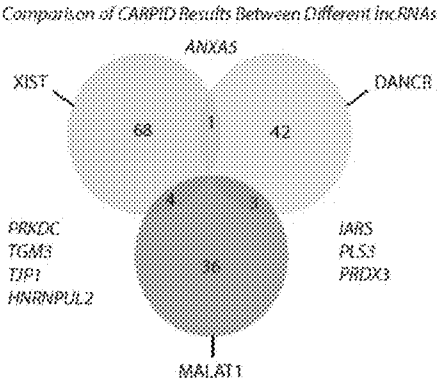

Figure 3e

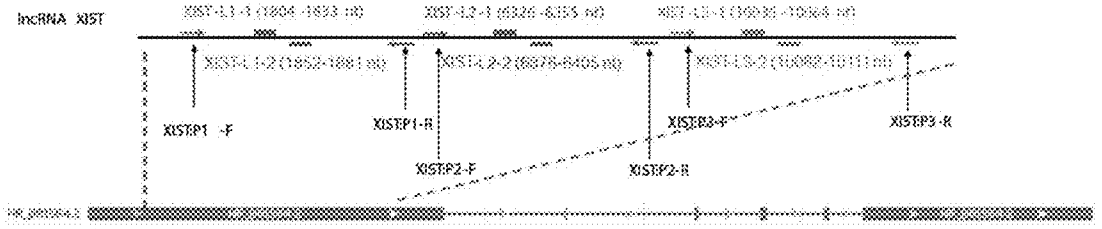

Figure 4a

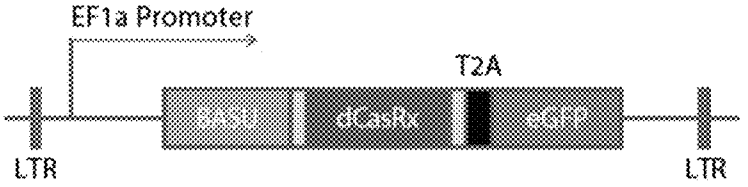

Figure 4b

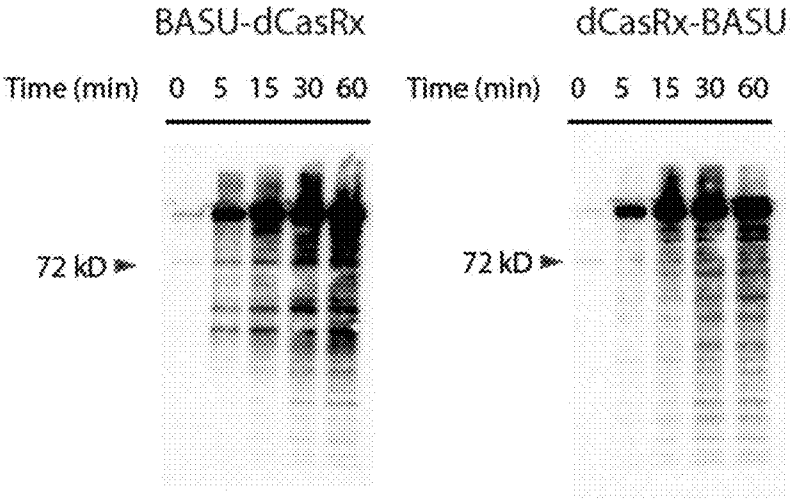

Figure 4c

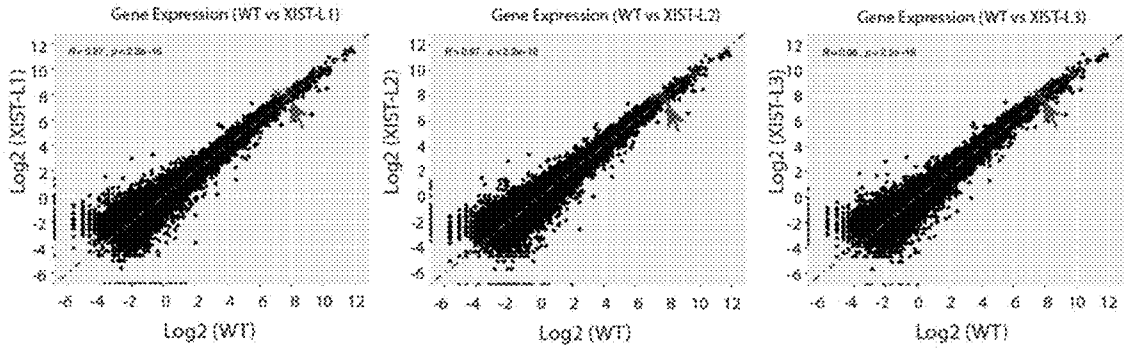

Figure 4d

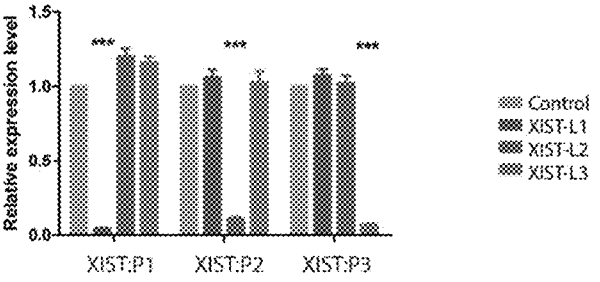

Figure 4e

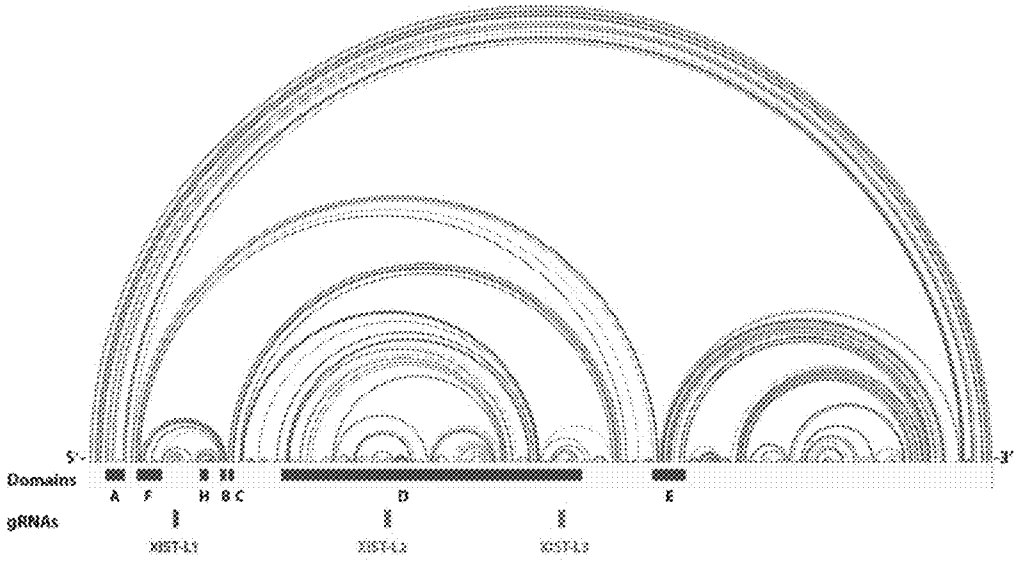

Figure 5

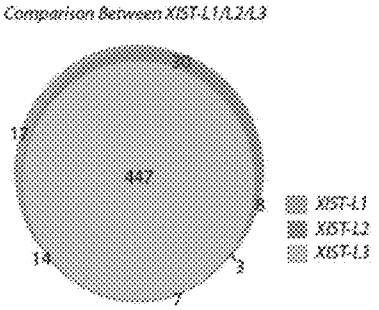

Figure 6a

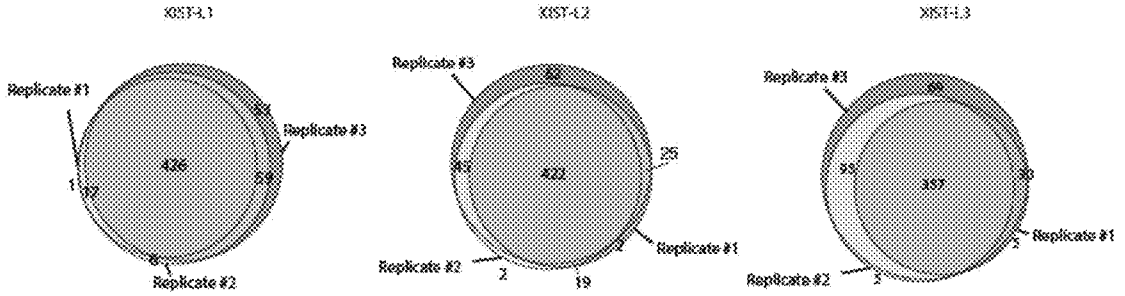

Figure 6b

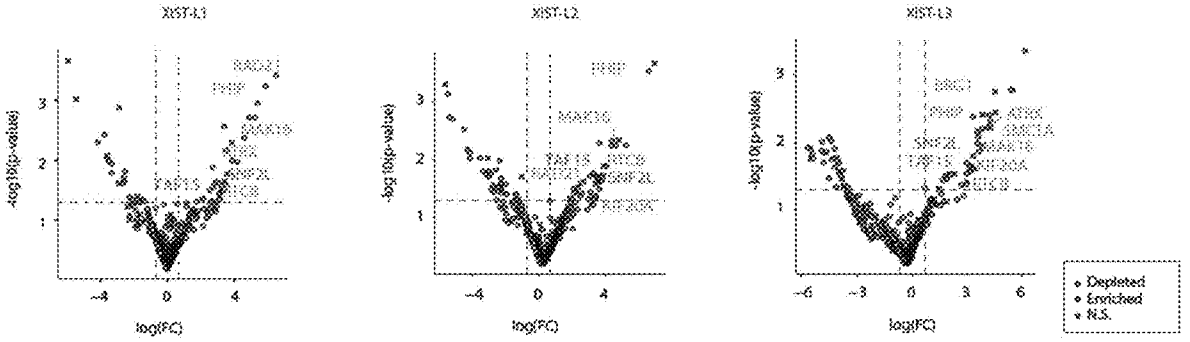

Figure 6c

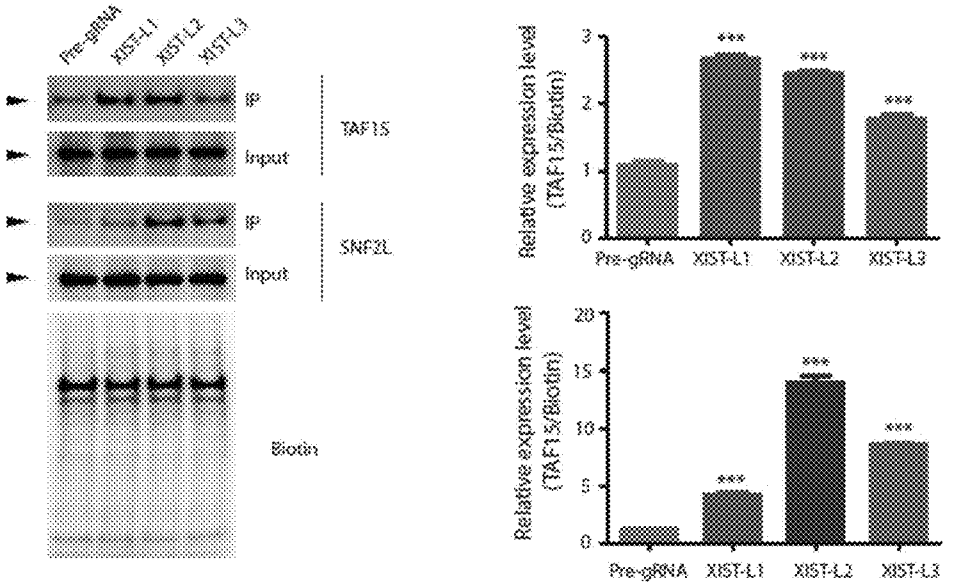

Figure 7

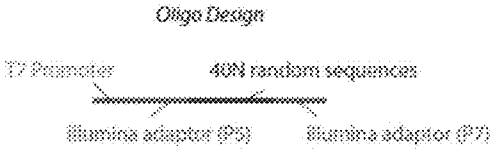

Figure 8a

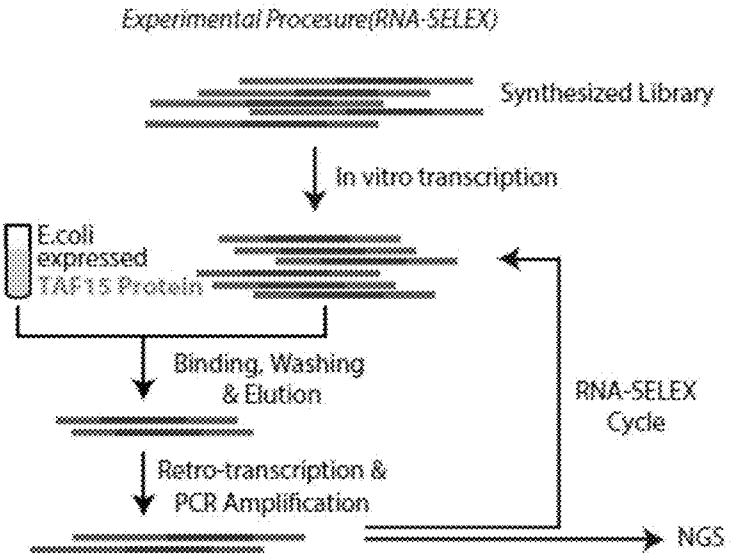

Figure 8b

Figure 8c

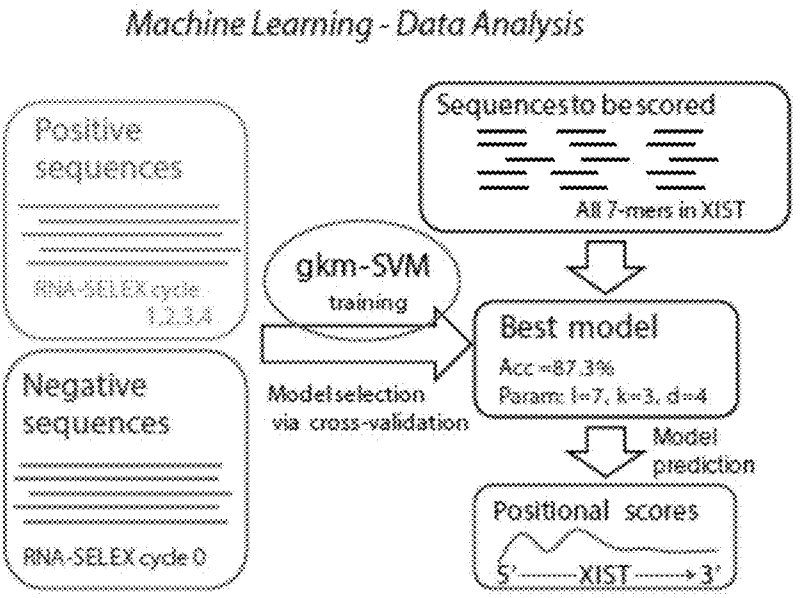

Figure 8d

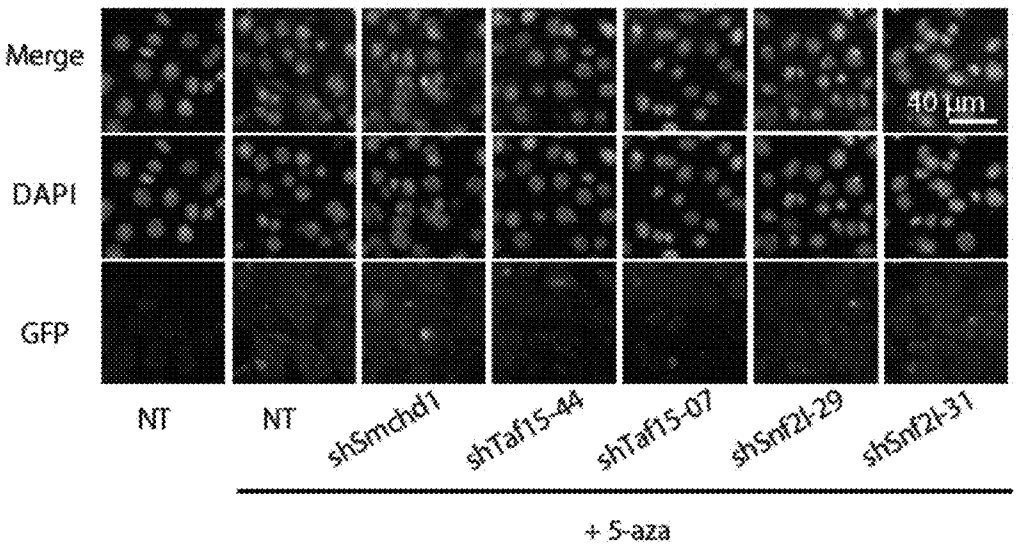

Figure 9a

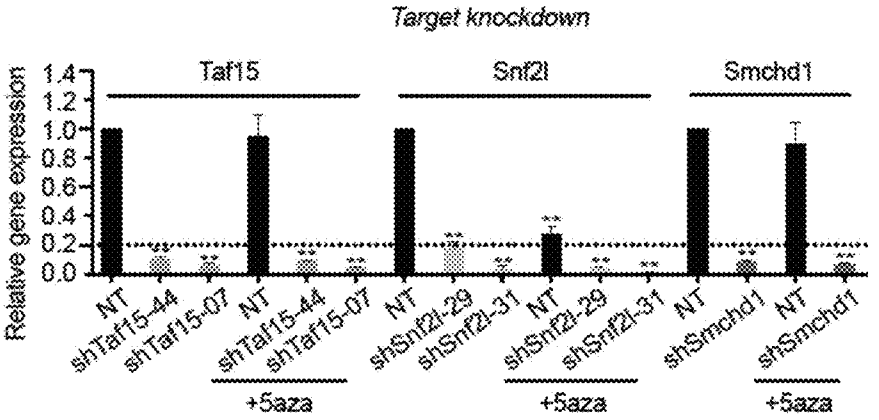

Figure 9b

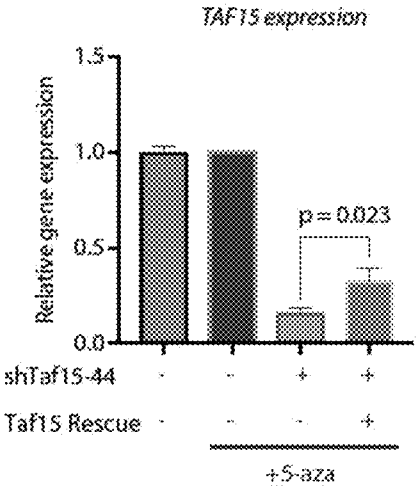

Figure 9c

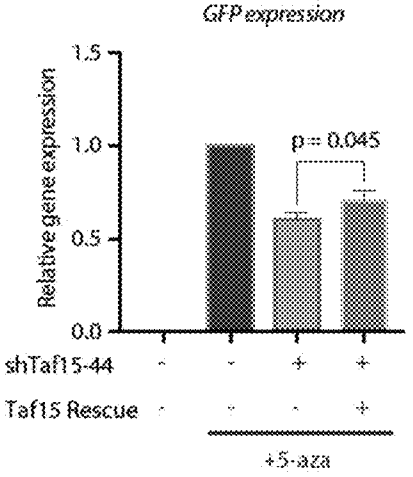

Figure 9d

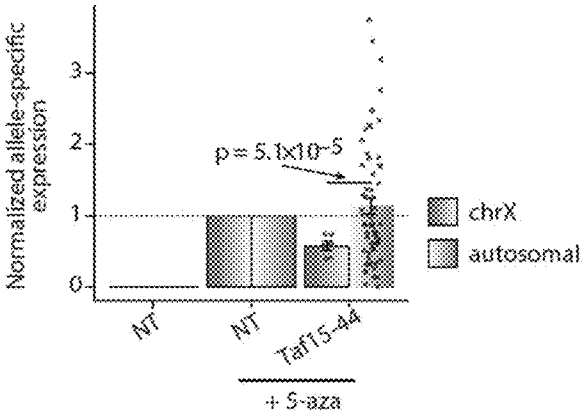

Figure 9e

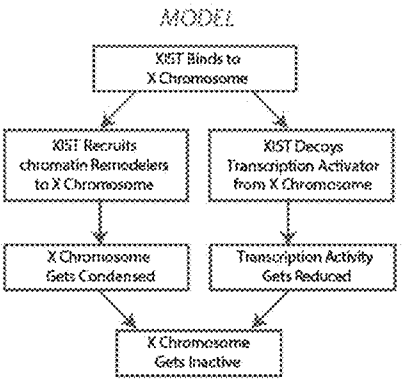

Figure 9f

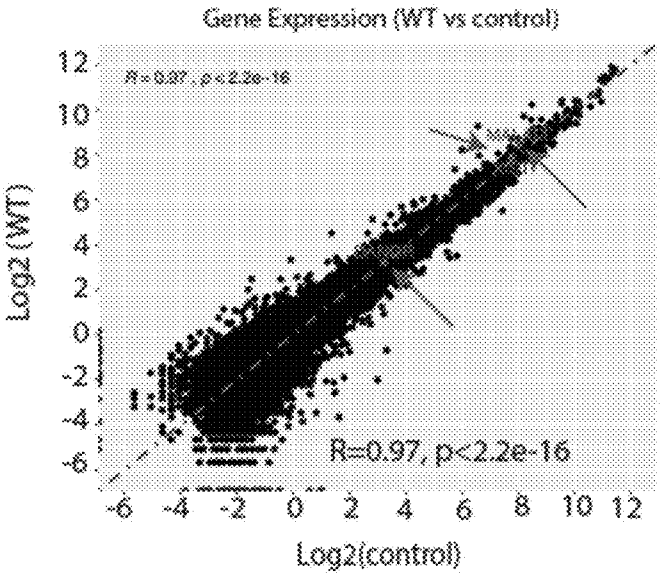

Figure 10a

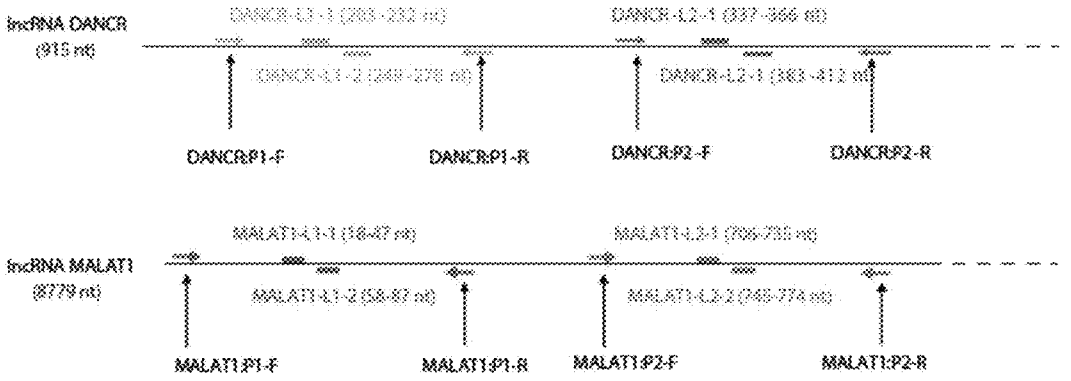

Figure 10b

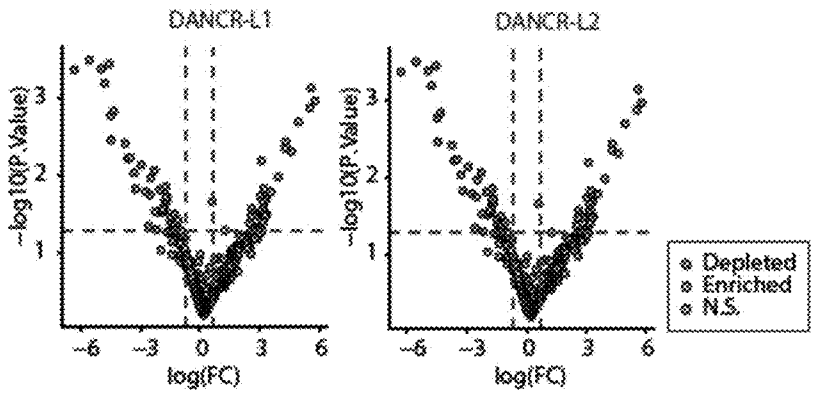

Figure 10c

GO term	p Value	Genes Involved
Extracellular Exosome	4.9e-5	FKBP4, ACTN4, ANXA5, CCT7, DLG1, EZR, FLNB, GJA1, GPI KRT13, MDH2, PEBP1, PAICS, PDIA6, STX3, TCP1, TPI1
Stress Fibre	3.0e-3	LIMA1, ACTN4, FLNB, SIPA1L3
Brush Border	2.3e-3	LIMA1, ACTN4, EZR, FLNB
Myelin Sheath	3.1e-3	EZR, GPI, MDH2, STIP1, TCP1
Focal Adhesion	4.8e-3	LIMA1, ACTN4, ANXA5, EZR, FLNB, GJA1
Actin Filament	3.3e-2	FYN, EZR, PLS3

Figure 10d

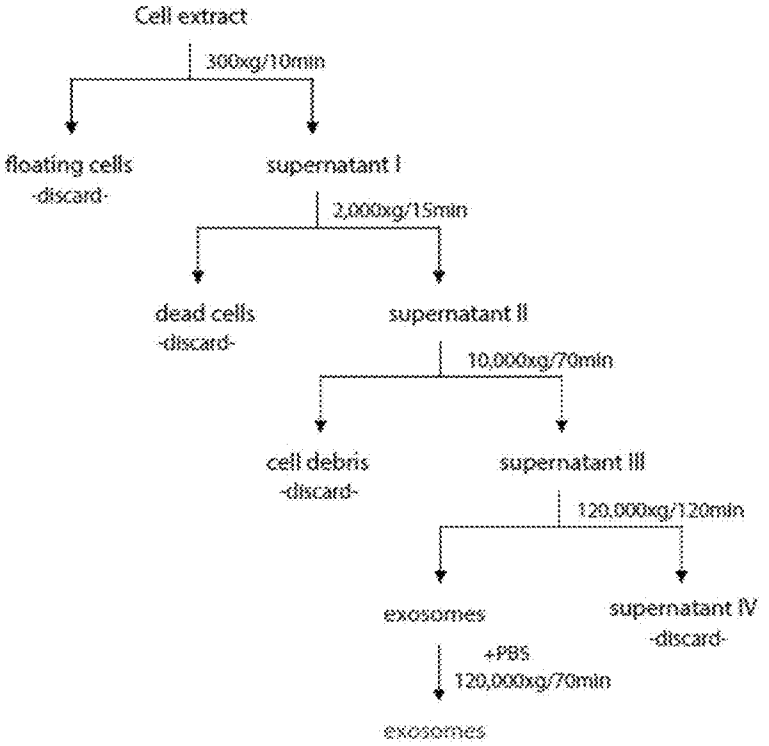

Figure 11a

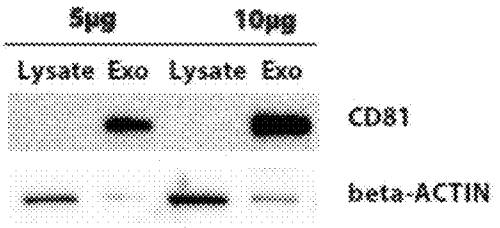

Figure 11b

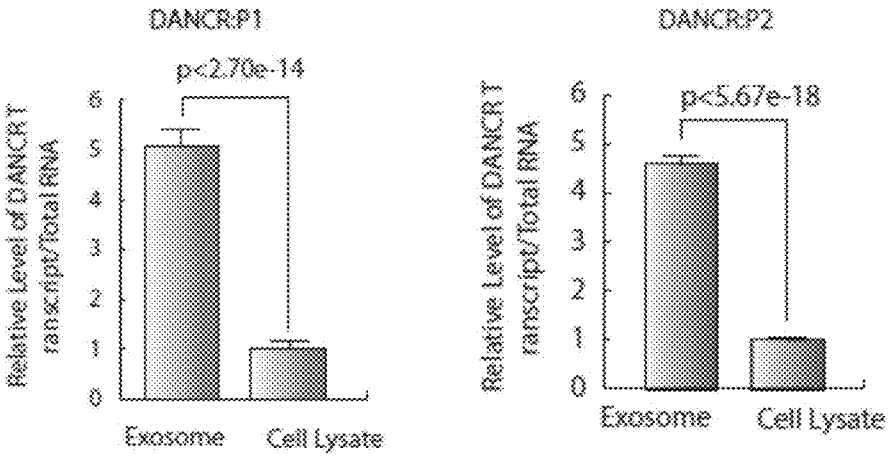

Figure 11c

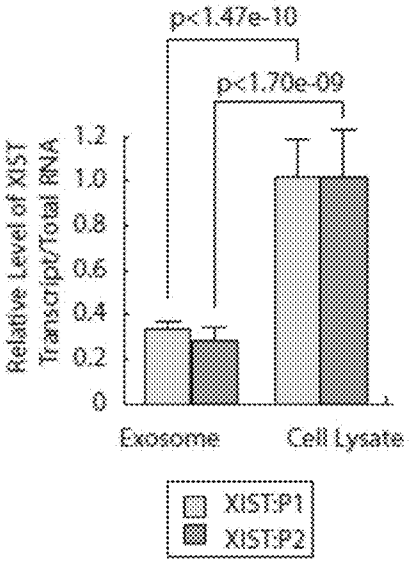

Figure 11d

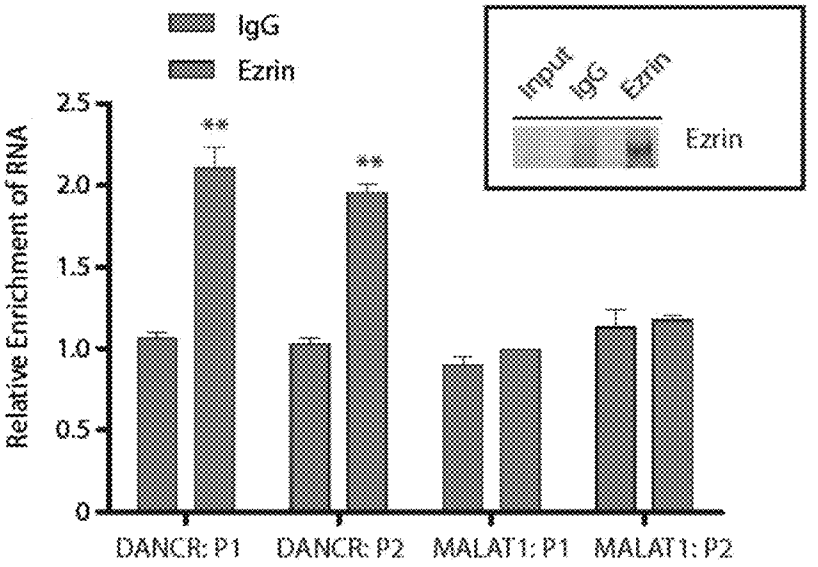

Figure 12a

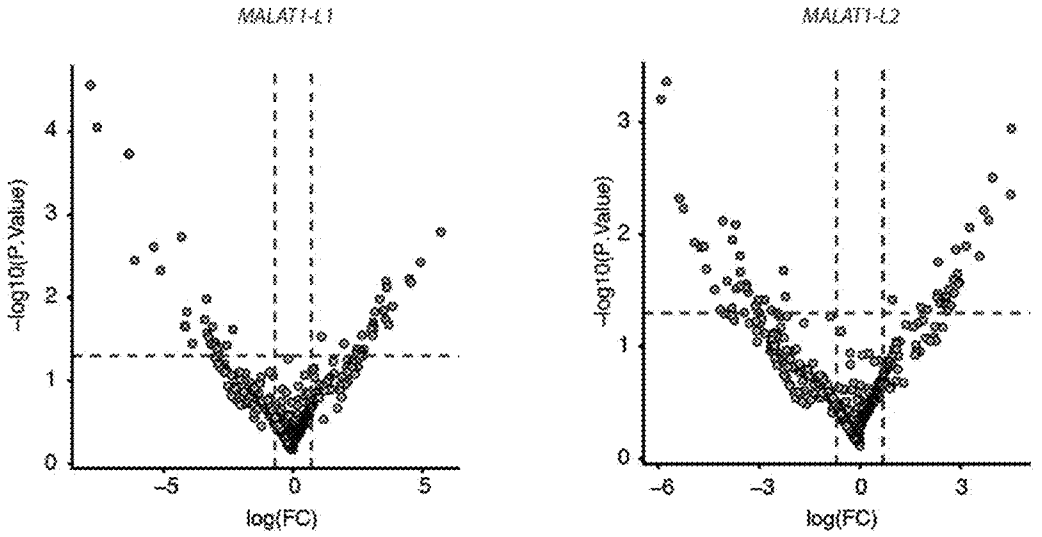

Figure 12b

METHOD FOR DETERMINING LONG NON-CODING RIBONUCLEIC ACID INTERACTION PROTEINS

SEQUENCE LISTING

[0001] The Sequence Listing file entitled "sequencelisting" having a size of 2,704 bytes and a creation date of Apr. 26, 2021, that was filed with the patent application is incorporated herein by reference in its entirety.

TECHNICAL FIELD

[0002] The present invention relates to a novel method for determining long non-coding ribonucleic acid interaction proteins.

BACKGROUND ART

[0003] Although only 2% sequences of the human genome code protein, more than 70% of genomic DNA can be transcribed into RNA at different stages of development. For decades, these huge numbers of non-coding RNAs (ncRNAs) have been considered as "dark matter" and their functions remained to be developed. These ncRNAs, especially long non-coding ribonucleic acids (lncRNA, defined as ncRNAs with more than 200 nucleotides in length) have been of interest recently, which were generally considered to be an important part participating in epigenetic regulation. For example, various lncRNAs were involved in cell cycle regulation and proliferation, the dysregulation of which is associated with progression and metastasis of various cancers.

[0004] XIST (X-inactivation specific transcript) is one of the first widely studied lncRNAs, which directs X chromosome inactivation (XCI) in female mammals, thus balancing the dose of genes between XY males and XX females. Its ability to constrain chromatin binding proteins makes it possible to label allele and cis-regulate transcription. At present, XCI and XIST have become exemplary models for understanding the epigenetic regulation of lncRNA.

[0005] The interaction between lncRNA and RNA binding proteins (RBP) determines the function and fate of RNA molecules. Up to 8.5 percent of the entire human proteome was predicted to have RNA binding properties, further demonstrating the multiple functions of lncRNA in various biological events. Mutations in lncRNA transcripts or changes in the abundance of lncRNA may alter their associated proteome, leading to health problems. The identification of lncRNA-related proteins will reveal the molecular mechanisms of the cell function in complex human diseases.

[0006] Although there is a growing recognition of the functional importance of RBP, there are significant technical limitations in elucidating lncRNA-protein interactions in living cells at present. Current methods depend mainly on chemical or UV mediated cross-linking between RNA and protein molecules to achieve effective enrichment and separation for the complex. Such procedures may produce non-systematic biases in physiological situations and mask the interaction proteins.

[0007] Recently, Ramanathan, M. et al. developed a RaPID method that integrates the promiscuous but efficient biotin ligase BASU with a λN peptide navigation system that recognizes the stem-loop of RNA BoxB. Compared with other biotin ligase variants, BASU remains inactive until it is rapidly activated by the high concentration of exogenous biotin in the culture medium, thereby labeling proteins nearby in a smaller labeling radius (~10 nm) in a shorter reaction time. This feature greatly reduces non-specific background noise. However, the target RNA needs to be artificially fused with the BoxB stem-loop close to the RBP binding region, and it needs to be expressed ectopically in the cell. Therefore, RaPID must compromise on three important factors: firstly, the abundance of ectopically expressed target RNA greatly exceeds the endogenous level of the transcript, resulting in a non-physiological balance of the interaction between RNA and RBP; secondly, the incorporation of the hairpin structure BoxB into the target RNA may interfere with the natural structure of the transcript, thereby changing its binding protein; thirdly, BASU can only label the RBP adjacent to the stem loop of BoxB at the 5' or 3' end of the RNA, therefore, some important RBPs may be missed, especially for long transcripts such as XIST (~19 kb). Briefly, the potential shortcomings, including the loss of cellular background, extensive molecular engineering and possible destruction of the natural structure of RNA, greatly limit the wide application of this method.

SUMMARY OF INVENTION

[0008] An object of the present invention is to provide a novel method for determining long non-coding ribonucleic acid (lncRNA) interaction proteins.

[0009] The present invention provides a CRISPR-Assisted RNA-Protein Interaction Detection (CARPID) method, which integrates CRISPR/CasRx based RNA targeting and proximity markers to identify binding proteins of specific lncRNAs within cells at natural state.

[0010] The inventive technology for detecting CRISPR-assisted RNA-protein interaction can be used as a novel and powerful method to find the interaction proteins of lncRNA in living cells. The method uses the highly specific CRISPR/CasRx system fused to the promiscuous but efficient biotin ligase BASU. The interaction with various proteins plays a central role in the regulatory activity of lncRNA. The present invention relates to a fusion protein of BASU and dCasRx, which comprises the dCasRx of compact Type VI-D CRISPR single-effect system, which can find the target lncRNA by specific gRNA co-transfected into the target cell. Once binding to the target lncRNA, it can enable BASU to specifically biotinylate effector proteins nearby activated by a high concentration of biotin. The biotinylated protein is separated by streptavidin affinity-coupled magnetic beads, then eluted, trypsin-digested and quantitatively analyzed by label-free mass spectrometry. As a control group, cells transfected with BASU-dCasRx but without gRNA are used as background.

[0011] At the time of the results analysis, the proteins identified in the specific gRNA group are statistically compared with the control group (no gRNA) for enrichment or reduction. Rank product is a non-parametric statistical method used to calculate the false discovery rate (FDR) of enrichment. In order to generate a list of specific interaction proteins, the critical value is set as enrichment≥2 times and FDR≤0.05. The obtained protein can be used for gene ontology analysis or protein interaction network analysis. In addition, a comprehensive analysis is performed on the target protein by using gRNA targeting different regions, thereby obtaining a high-resolution spectrum of the target lncRNA interaction protein.

[0012] In an aspect, the present invention thus provides a fusion protein formed by BASU and dCasRx.

[0013] According to the embodiment of the present invention, the fusion protein can be BASU-dCasRx, or dCasRx-BASU.

[0014] In another aspect, the invention further provides an expression vector for expressing the fusion protein formed by BASU and dCasRx. Preferably, it is a mammalian expression vector.

[0015] In another aspect, the invention further provides a composition comprising: the fusion protein formed by BASU and dCasRx and/or the expression vector for expressing the fusion protein formed by BASU and dCasRx, as well as a gRNA targeting the target lncRNA.

[0016] In another aspect, the invention further provides a kit for determining lncRNA interaction proteins, comprising: the fusion protein formed by BASU and dCasRx and/or the expression vector for expressing the fusion protein formed by BASU and dCasRx, as well as a gRNA targeting the target lncRNA. Preferably, the kit further comprises a control reagent without gRNA (for example, co-transfected with the gRNA empty vector and the BASU-dCasRx fusion protein expression vector).

[0017] In another aspect, the present invention further provides a method for determining lncRNA interaction proteins, comprising:

[0018] co-transfecting the expression vector for expressing the fusion protein formed by BASU and dCasRx, and a gRNA that specifically targets the target lncRNA in a target cell, thereby BASU specifically biotin-labeling effector proteins nearby;

[0019] isolating the biotinylated proteins for analysis to determine the lncRNA interaction proteins.

[0020] According to a specific embodiment of the present invention, in the method for determining lncRNA-interaction proteins of the present invention, specifically, the biotinylated protein can be separated by streptavidin affinity-coupled magnetic beads, and then eluted and trypsin digested, and quantitatively analyzed by a label-free mass spectrometry.

[0021] According to a specific embodiment of the present invention, the method for determining the lncRNA interaction protein of the present invention is used to determine the lncRNA interaction protein in living cells.

[0022] According to a specific embodiment of the present invention, the method for determining lncRNA-interaction proteins of the present invention further comprises: statistically comparing the protein identified in the specific gRNA group with the control group without gRNA for enrichment or reduction.

[0023] According to a specific embodiment of the present invention, the method for determining a lncRNA-interaction protein of the present invention further comprises: calculating the false discovery rate of enrichment by Rank product; more preferably, the critical value is set as enrichment≥2 times and FDR≤0.05.

[0024] According to a specific embodiment of the present invention, the method for determining a lncRNA-interaction protein of the present invention further comprises:

[0025] performing a genetic ontological analysis or protein interaction network analysis on the obtained proteins; and/or performing a comprehensive analysis on the target protein by using gRNA targeting different regions to obtain a high-resolution spectrum of the target lncRNA interaction protein.

[0026] In another aspect, the invention further provides a method for analyzing enriched interaction proteins to specific regions of target lncRNA, comprising:

[0027] performing an enrichment analysis on the proteins with more than one peptide fragments detected; wherein the proteins preferably comprise human keratin;

[0028] normalizing and logarithmizing a LFQ abundance of each group;

[0029] replacing a missing value by a minimum value representing the detection limit of mass spectrometer;

[0030] determining the protein that is statistically enriched in samples of the gRNA transfection group compared to the control group transfected with gRNA empty vectors by rank product;

[0031] a protein with the adjusted p-value≤0.05 and the abundance change≥2 folds, is identified as RBP to the target lncRNA.

[0032] In another aspect, the present invention further provides a proteomics method for defining a high-resolution spectrum of the interaction protein for the target lncRNA. By applying specific gRNA to different regions of the target lncRNA, the interaction protein of the specific region can be obtained.

[0033] In another aspect, the present invention also provides an analysis system (device) for determining lncRNA interaction proteins, which includes a data analysis unit configured to enrich and analyze the protein with more than one peptide detected by the present invention, and further analyze to determine the lncRNA interaction protein. Specifically, the protein with more than one peptide includes human keratin. The specific analysis process comprises: normalizing and logarithmizing a LFQ abundance of each group; replacing a missing value by a minimum value representing the detection limits of mass spectrometer; determining the protein that is statistically enriched in samples of the gRNA transfection group compared to the control group transfected with gRNA empty vectors by rank product; a protein with the adjusted p-value≤0.05 and the abundance change≥2 folds, is identified as RBP to the target lncRNA.

[0034] In some specific embodiments of the present invention, CARPID is applied to three lncRNAs, namely XIST, DANCR and MALAT1, in the present invention, and reliably recognizes their known interaction proteins. It is worth noting that these three groups of interaction proteins have almost no overlap, showing the strong specificity of the method of the present invention.

[0035] In some specific embodiments of the present invention, the CARPID technology of the CRISPR auxiliary system of the present invention systematically detects the lncRNA XIST binding protein group in a non-crosslinking manner. Using CARPID, the present invention not only detects a number of previously reported XIST binding proteins, but also identifies many new factors, among which the present invention validates the TAF15 and SNF2L in this study through biochemical and functional verification. The data of the present invention supports the current consensus that XIST RNA regulates XCI by recruiting chromatin remodeling agents for chromosome condensation and isolating transcription mechanisms to further inhibit genes.

[0036] In order to maximize the credibility of the present invention and avoid false positive signals, the present invention controls the changes at both the experimental and statistical levels. Firstly, a self-cleavable GFP fusion is used to monitor the expression of BASU enzyme in cells and minimize the reaction time required for effective biotin labeling. Secondly, a multi-site targeting strategy is adopted to specifically target three different loci on XIST, and new proteins identified with at least two gRNA pairs are further verified in the present invention. Thirdly, for each group of gRNA, at least three repeated CARPID experiments are repeated. In addition, a triple simulation control is used to evaluate the statistical significance of the enrichment.

[0037] The present invention also proves that CARPID can be universally used to detect the bound proteome of lncRNA. The present invention specifically targets the other two lncRNA DANCR and MALAT1 with different length expression levels and subcellular localization. The dysregulations of DANCR and MALAT1 expression relate to a variety of malignant tumors, including liver cancer, breast cancer, glioma, colorectal cancer, gastric cancer and lung cancer. The research of the present invention shows that DANCR can interact with proteins largely enriched in extracellular exosomes. Interestingly, it has been reported in various studies that serum DANCR levels are elevated in cancer patients. In addition, the present invention also identifies the interaction between DANCR and Ezrin (an important structural protein in the cell cortex). Such findings reveal the new function of lncRNA in tumor development.

[0038] Label-free mass spectrometry is a direct and cost-effective method to apply CARPID. In addition, it has low technical requirements, thus ensuring wide applicability. The present invention has shown that CARPID is a powerful method for detecting RNA binding proteins, with high specificity and reproducibility. To further improve the resolution, quantitative mass spectrometry with different labeling strategies (such as TMT and other isobaric chemical labeling and SILAC labeling) can be incorporated into the CARPID channel.

[0039] In summary, the present invention combines CARPID, labeled quantitative mass spectrometry and non-parametric enrichment analysis, and can identify specific lncRNA interaction proteins in living cells with high confidence by using proteomics methods. The CARPID technology of the present invention can draw a high-resolution spectrum of various lncRNAs interaction proteins involved in human diseases. Such spectrum can provide guidance for therapies that interfere with the function of specific lncRNAs.

DESCRIPTION OF DRAWINGS

[0040] The patent or application file contains at least one drawing executed in color. Copies of this patent or patent application publication with color drawing(s) will be provided by the Office upon request and payment of the necessary fee.

[0041] FIGS. **1***a* to **1***e* show that CARPID is used to identify proteins associated with lncRNA XIST in living cells, wherein:

[0042] FIG. **1***a* is a schematic diagram of CARPID workflow. The target lncRNA is targeted by a group of gRNAs. The nuclease activity-free RNA nuclease CRISPR/CasRx (dCasRx) fused with engineered biotin ligase (BASU) is recruited to a specific site. After a treatment by biotin, the adjacent RNA binding protein (RBP) will be biotinylated by BASU. The red shading indicates the marking radius of BASU-dCasRx. Biotinylated proteins are enriched by streptavidin affinity-coupled magnetic beads (MyOne T1) for subsequent mass spectrometry (MS) identification and Western blot (WB) analysis.

[0043] FIG. **1***b* shows three groups of gRNAs (highlighted in different colors) against the human lncRNA XIST locus. Only one group of gRNA is expressed in each experiment.

[0044] FIG. **1***c* shows that CARPID is used to identify XIST-related proteomes. The volcano plot shows the enrichment of XIST-related proteins in HEK293T cells. The x-axis represents the log 2 converted value of the protein level fold change in the CARPID results of all three groups of XIST gRNAs compared with the control. The y-axis shows the negative logarithmic converted p-value (non-parametric rank product test). Significantly enriched proteins are marked as orange dots. Proteins previously verified to interact with XIST and recognized by CARPID are marked in orange font. The blue font represents SNF2L and TAF15, which are two newly identified XIST-related proteins.

[0045] FIG. **1***d* shows the XIST-protein interaction network identified by CARPID. White nodes indicate proteins recognized only in a group of gRNAs. The pink nodes represent the proteins recognized in two groups of gRNA. The red nodes show the proteins recognized in all three groups of gRNA. A line is connected between two proteins evaluated as interaction (STRING interaction score≥0.40). The width of the line is proportional to the STRING interaction score. The nodes with purple edges highlight the proteins involved in chromatin remodeling. XIST-related proteins are highlighted in bold orange (19 proteins). The blue font represents SNF2L and TAF15, which are two newly identified XIST-related proteins.

[0046] FIG. **1***e* shows the top five important gene ontology (GO) semantics of XIST-related proteins.

[0047] FIGS. **2***a* to **2***h* show the analysis results that verify the proteins associated with XIST, wherein:

[0048] FIG. **2***a* shows the verification of the XIST-TAF15 interaction using WB and immunoFISH. The upper panel shows that Western blotting is performed using anti-human TAF15 antibody after CARPID. From the lysate of HEK293T cells co-transfected with BASU-dCasRx and each of pre-gRNA (control), and the gRNAs specifically targeting locus 1 (XIST-L1), locus 2 (XIST-L2) and locus 3 (XIST-L3), the biotinylated protein is precipitated using streptavidin affinity-coupled magnetic beads. The experiment is carried out in three biological replicates and showed representative results. The lower panel shows the immunoFISH test results of TAF15 and XIST. HEK293T cells are fixed and incubated with anti-TAF15 antibody and corresponding secondary antibody with CF 488A (green). A specific oligonucleotide probe labeled as Cy3 (red) is used to detect XIST. The nucleus is counterstained with DAPI (blue). The box area in the left image is enlarged and displayed on the right. The scale is shown. This image is representative of three independent biological experiments.

[0049] FIG. **2***b* shows the TAF15 CLIP-seq data result of mouse brain tissue displayed by the genome browser view. This figure shows that TAF15 specifically binds to XIST RNA. The non-redundant readings of the two chains are displayed separately. The RefSeq gene also indicates chain information in different colors: red, forward strand; blue, reverse strand.

[0050] FIG. **2***c* shows a histogram showing the number of TAF15 binding peaks on the XIST transcript compared to the expected peak number (mean±SD) based on 10,000 random shuffle peak positions. The peak sites are directly extracted from the article by Kapeli et al. The p-value is based on the one-tailed Poisson test.

[0051] FIG. **2***d* shows TAF15 antibody is used in HEK293T cells with IgG as a control, and formaldehyde-assisted RIP is used to verify the XIST-TAF15 interaction. Quantitative enrichment is performed using RT-qPCR in three different regions of XIST and two different regions of MALAT1. GAPDH served as an internal control. Data is expressed as mean±SD, n=3; ** indicates $p<0.01$ using the unpaired Student's t test. The upper right figure shows the results of TAF15 antibody and IgG immunoprecipitation experiments.

[0052] FIG. **2***e* shows HTR-SELEX is used to verify the XIST-TAF15 interaction. The blue curve shows the predicted binding affinity of TAF15 along the XIST transcript. The x-axis represents the relative position of the human XIST transcript (~19 kb). The y-axis shows the gkm-SVM scores of 7 monomers starting from the corresponding position in XIST. Please note that the larger the gkm-SVM score, the higher the affinity of TAF15. The three colored vertical lines represent locus 1-3 of XIST. The blue curve shows the fitted value of the generalized additive model, while the gray area shows a confidence interval of 95%. As a genomic background, the orange curve shows the average gkm-SVM score of 1,000 sequences randomly sampled from the human genome (hg19).

[0053] FIG. **2***f* shows the verification of the XIST-SNF2L interaction using WB and immunoFISH. The upper panel shows that Western blotting is performed using anti-human SNF2L antibody after CARPID. From the lysate of HEK293T cells co-transfected with BASU-dCasRx and each of pre-gRNA (control), and the gRNAs specifically targeting locus 1 (XIST-L1), locus 2 (XIST-L2) and locus 3 (XIST-L3), the biotinylated protein is precipitated using streptavidin affinity-coupled magnetic beads. The experiment is carried out in three biological replicates and showed representative results. The lower panel shows the immunoFISH test results of SNF2L and XIST. HEK293T cells are fixed and incubated successively with anti-TAF15 antibody and corresponding secondary antibody with CF 488A (green). A specific oligonucleotide probe labeled as Cy3 (red) is used to detect XIST. The nucleus is counterstained with DAPI (blue). The box area in the left image is enlarged and displayed on the right. The experiment is performed in three biological replicates and showed representative results.

[0054] FIG. **2***g* shows SNF2L antibody is used in HEK293T cells with IgG as a control, and formaldehyde-assisted RIP is used to verify the XIST-TAF15 interaction. Quantitative enrichment is performed using RT-qPCR in three different regions of XIST and two different regions of MALAT1. GAPDH served as an internal control. Data is expressed as mean±SD, n=3; ** indicates $p<0.01$ using the unpaired Student's t test. The upper right figure shows the results of SNF2L antibody and IgG immunoprecipitation experiments.

[0055] FIG. **2***h* shows the effects of TAF15 and SNF2L in X-linked inhibition in mice. Female iMEF cells (E2C4) contain the GFP transgene on the inactivated X chromosome, and its expression is completely suppressed (NT). After 5-aza treatment (NT+5-aza), GFP is derepressed. Various shRNAs are used to knock out SmcHD1, TAF15 and SNF2L, and the GFP expression level is determined by RT-qPCR. Data is expressed as mean±SD, n=3, *$p<0.05$, **$p<0.01$, Student's t test.

[0056] FIGS. **3***a* to **3***e* show the identification of lncRNA DANCR and MALAT1 related proteins by CARPID in living cells, wherein:

[0057] FIG. **3***a* shows the identification of the DANCR-related proteome associated with CARPID. The volcano plot shows the enrichment of DANCR-related proteins in HEK293T cells. The x-axis represents the log 2 conversion of the protein level fold change in the CARPID combined with the two groups of DANCR gRNA in comparison with the control. The y-axis shows the negative logarithmic converted p-value (non-parametric rank product test). Significantly enriched proteins are marked as orange dots.

[0058] FIG. **3***b* shows the DANCR-protein interaction network identified by CARPID. White nodes indicate proteins recognized only in a group of gRNAs. The pink nodes represent the proteins recognized in two groups of gRNAs. A line is connected between two proteins evaluated as interaction (STRING interaction score≥0.40). The width of the line is proportional to the STRING interaction score.

[0059] FIG. **3***c* shows the proteome associated with MALAT1 identified by CARPID. The volcano plot shows the enrichment of MALAT1 related proteins in HEK293T cells. The x-axis represents the log 2 conversion of the protein level fold change in the CARPID combined with the two groups of MALAT1 gRNA in compared with the control. The y-axis shows the negative logarithmic converted p-value (non-parametric rank product test). Significantly enriched proteins are marked with orange dots.

[0060] FIG. **3***d* shows the MALAT1 protein interaction network identified by CARPID. White nodes indicate proteins recognized in a group of gRNAs. The pink nodes represent proteins recognized in two groups of gRNAs. A line is connected between two proteins evaluated as interaction (STRING interaction score≥0.40). The width of the line is proportional to the STRING interaction score.

[0061] FIG. **3***e* shows the comparison of CARPID results between different lncRNAs. The Venn diagram illustrates the unique and specific RBP between each two of the three lncRNAs (XIST, DANCR, and MALAT1).

[0062] FIGS. **4***a* to **4***e* are the optimized schematic diagrams of CARPID, wherein:

[0063] FIG. **4***a* shows the position of the three groups of gRNA on XIST. Different colors indicate different gRNA groups. Note that the interval between the two individual gRNAs in each group is approximately 15 nt.

[0064] FIG. **4***b* shows the scheme of the BASU-dCasRx construct. BASU is subcloned from BASU RaPID plasmid (Addgene #107250), and cloned into EF1a-dCasRx-2A-EGFP plasmid (Addgene #109050) in frame. LTR, long terminal repeat; T2A, self-cleaving peptide; eGFP, enhanced GFP.

[0065] FIG. **4***c* shows the transfection of HEK293T cells with BASU-dCasRx or dCasRx-BASU. After 48 hours, the cells are treated with 200 μM biotin for the specified time. Immunoblotting experiments are performed on whole cell lysates using streptavidin chelating HRP. The experiment is carried out in 3 biological replicates, and representative results are shown).

[0066] FIG. **4***d* is a scatter plot showing the comparison of gene expression levels in wild-type HEK293T cells or

HEK293T cells transfected with different gRNAs (XIST-L1/XIST-L2/XIST-L3). The x-axis in each figure represents the log 2 converted gene expression level in wild-type HEK293T cells. The y-axis represents the log 2 converted gene expression level of HEK293T cells transfected with gRNA in the CARPID experiment. Each figure indicates the expression level of XIST gene.

[0067] FIG. **4***e* shows the specificity of the CRISPR/CasRx system on XIST. HEK293T cells are co-transfected with CasRx and single gRNA (empty vector control, XIST-L1, XIST-L2 and XIST-L3). Total RNA is extracted from the treated cells and then subjected to reverse transcription and qPCR analysis to quantify the level of XIST specific sites. GADPH is used to normalize the level of XIST. Please note that the co-transfection with CasRx and gRNA specifically reduces the RNA transcription level at its target locus. Data is expressed as mean±SD, n=3, *** $p<0.001$, unpaired Student's t-test.

[0068] FIG. **5** shows the secondary structure location mapping of XIST gRNA. The XIST-L1/L2/L3 genome browser view on XIST hairpin structure information by Lu et al. uses black strings to represent regions with complementary pairs. The vertical lines in different colors highlight the location of the different gRNA groups of the targeted locus on the XIST lncRNA. The locations of known domains (A-H) are also indicated.

[0069] FIGS. **6***a* to **6***c* show the CARPID results of XIST, wherein:

[0070] FIG. **6***a* shows the use of the Venn diagram to illustrate the overlap/repeatability of the CARPID mass spectrometry identification results of three different XIST gRNAs (XIST-L1/XIST-L2/XIST-L3).

[0071] FIG. **6***b* shows the use of the Venn diagram to illustrate the overlap of the proteins significantly enriched in the CARPID mass spectrometry identification results of three different groups of XIST gRNA (XIST-L1/XIST-L2/XIST-L3).

[0072] FIG. **6***c* shows the identification of XIST binding proteins using CARPID. The volcano plot shows the enrichment of XIST-related proteins in HEK293T cells. The x-axis represents the log 2 converted value of the protein level fold change in the CARPID results of all three groups of XIST gRNAs compared with the control. The y-axis shows the negative logarithmic converted p-value (non-parametric rank product test). Significantly enriched proteins are marked with orange dots. Proteins previously known to interact with XIST and recognized by CARPID are marked in orange font. The blue font represents SNF2L and TAF15, which are two newly identified XIST-related proteins.

[0073] FIG. **7** shows the quantitative analysis of CARPID-WB. The results of CARPID are tested by western blotting, using TAF15 and SNF2L antibodies. The streptavidin affinity-coupled magnetic beads are added to lysate of the HEK293T cell transfected with BASU-dCasRx, as well as pre-gRNA (control), locus 1 (XIST-L1), locus 2 (XIST-L2) or locus 3 (XIST-L3), to precipitate the biotinylated protein. The experiment is carried out in three biological replicates and representative results are shown. The WB signal is quantified using ImageJ software (version 1.8.0_172).

[0074] FIGS. **8***a* to **8***d* show TAF15 HTR-SELEX, wherein:

[0075] FIG. **8***a* shows the design scheme for the oligonucleotide of HTR-SELEX. These oligonucleotides contain T7 promoter, Illumina adaptor (P5/P7) and 40-nt random sequence.

[0076] FIG. **8***b* is a schematic diagram of the HTR-SELEX experiment. First, the synthesized DNA template library is transcribed into RNA, and TAF15 protein is expressed in *Escherichia coli* Rosetta P3 DE LysS strain. After bound, washed and eluted, the remaining (bound) RNA is subjected to reverse transcription and PCR amplification to obtain an NGS sequencing library. The DNA library is sequenced by the Illumina Hiseq 4000 for molecular counting, and part of the library is used as input for the next round of HTR-SELEX (see methods for details).

[0077] FIG. **8***c* shows the RNA binding motif of TAF15 enriched from the HTR-SELEX analysis.

[0078] FIG. **8***d* shows the machine learning scheme. A machine learning algorithm (gkm-SVM) based on gap k-mer is adopted in the present invention to train a prediction model with HTR-SELEX data to evaluate the RNA sequence preference of TAF15. Due to computational power, both positive and negative sequences are randomly downsampled to 100,000 sequences. In order to find the best model, the present invention considers three key parameters of gkm-SVM: l, the length of the entire word includes spaces; k, the number of positions of information (i.e. no gaps) in each word; d, the maximum allowed Number of mismatches. The present invention uses 5-fold cross-validation for parameter combination search. When l=7, k=3, and d=4, the highest accuracy of cross-validation is 87.3%. Finally, the present invention uses the best model to score all 7-mers occurred in XIST, and draw a smooth gkm-SVM prediction score along the XIST transcript.

[0079] FIGS. **9***a* to **9***f* show the functional verification of TAF15 and SNF2L on XCI, wherein:

[0080] FIG. **9***a* shows that the cells are fixed with 3% PFA solution and then stained with DAPI. The GFP signal is observed on the fluorescence microscope under the FITC channel.

[0081] FIG. **9***b* shows verification of shRNA knockdown efficiency. The iMEF cells are infected with the lentivirus carrying the indicated shRNA. The gene expression level is detected by RT-qPCR. β-actin is used to normalize RNA expression under different conditions. Data is expressed as mean±SD, n=3, ** $p<0.01$, unpaired Student's t-test.

[0082] FIG. **9***c* shows the TAF15 knockdown recovery experiment. The packaged Taf15 specific shRNA (shTaf15-44) and anti-shRNA Taf15 virus are used to infect iMEF cells. The expression level of Taf15 is detected by RT-qPCR, and β-actin is used as an internal control. Data is expressed as mean±SD, n=3, unpaired Student's t-test.

[0083] FIG. **9***d* shows the replenishing X-linked GFP suppression phenotype. The GFP expression is determined by RT-qPCR under the same experimental conditions as in Figure c. Data is expressed as mean±SD, n=3, unpaired Student's t-test.

[0084] FIG. **9***e* shows the role of TAF15 in the transcription of mouse autosomal genes. Two different shRNA constructs (shTaf15-07 and shTaf15-44) are used to knock out TAF15 in female iMEF cells (E2C4), similarly as shown in FIG. **1**. Five autosomal genes are randomly selected from different chromosomes over 2 hours. The expression level is determined by RT-qPCR and normalized to β-actin. Data is expressed as mean±SD, n=3, *$p<0.05$, **$p<0.01$, unpaired

Student's t-test, ns=not significant. FIG. **9***e* further shows that allelic RNA-seq is used to demonstrate the effect of TAF15 on XCI. Genes that showed significant changes in allelic expression after 5-aza treatment (NT+5-aza vs. NT) are grouped into X chromosome (chrX) genes and autosomal genes. By defining the expression ratio between the minor and major alleles before processing as 0 (unbalanced), and setting the expression ratio of the minor and major alleles after processing as 1 (balance), the allele ratios of Taf15 gene knockout for X chromosome (blue) gene and autosomal gene (grey) are summarized respectively. Data is expressed as mean±SE, and data points for a single gene measured in two biological replicates are shown. P value is calculated using unpaired Student's t test.

[0085] FIG. **9***f* shows a working model of the dual role of XIST lncRNA in mediating XCI.

[0086] FIGS. **10***a* to **10***d* show the CARPID results of lncRNA DANCR, wherein:

[0087] FIG. **10***a* is a scatter diagram showing the comparison of gene expression levels in wild-type HEK293T cells or HEK293T cells transfected with pre-gRNA expression plasmids in CARPID experiments. The x-axis represents the gene expression level after log 2 conversion of HEK293T cells in the treatment group. The y-axis represents the gene expression after log 2 conversion in wild-type HEK293T cells. The expression levels of XIST, DANCR and MALAT1 are highlighted in red. Note that MALAT1 is abundant in HEAT293T cells, and the expression level of DANCR is much lower than XIST and MALAT1.

[0088] FIG. **10***b* shows the location of the gRNA group of DANCR used in CARPID (upper: DANCR-L1/2; lower: MALAT1-L1/L2) and qPCR primers (upper: DANCR: P1/P2; lower: MALAT1: P1/P2). F, forward primer; R, reverse strand primer. The number in parentheses indicates the position of the gRNA set starting from 1 nt in the corresponding RNA transcript.

[0089] FIG. **10***c* shows that LncRNA DANCR combines proteomic identification with CARPID and then MS. The volcano map shows that in HEK293T cells expressing BASU-dCasRx, the enrichment of lncRNA DANCR-related proteins in each group of gRNA exceeds that of the control (empty gRNA expression vector). The x-axis represents the logarithmic change fold of the protein level in CARPID of each group of DANCR gRNA relative to the control. The y-axis shows the negative log 10 converted p-value after (non-parametric rank product test). Significantly enriched proteins are marked with orange dots.

[0090] FIG. **10***d* shows the top six important gene ontology (GO) term of DANCR-related proteins.

[0091] FIGS. **11***a* to **11***d* show verification of the presence of DANCR in exosomes, wherein:

[0092] FIG. **11***a* is a schematic diagram of the isolation of exosomes from cultured human cells. The exosomes used for DANCR detection are highlighted in red font.

[0093] FIG. **11***b* shows the purification of exosomes is examined by immunoblotting. 5 μg or 10 μg cell lysates and exosomal fractions are separated on SDS-PAGE gels, and subjected to Western blot analysis for the indicated proteins. Note that CD81 is highly enriched in purified exosomes.

[0094] FIG. **11***c* shows the comparison of DANCR levels in cell lysates and exosomes. Total RNA is extracted from whole cell lysates and exosomes of HEK293T cells, and DANCR transcription levels are quantified by reverse transcription followed by qPCR. Two sets of qPCR primers are used. The y-axis represents the relative content of DANCR RNA in isolated exosomes (Exosome) or whole cell lysate (Cell Lysate), expressed as an equivalent amount of total RNA. The p-value is calculated using the unpaired student's t-test. Note that DANCR is enriched in exosomes relative to whole cell lysates.

[0095] FIG. **11***d* shows the comparison of XIST levels in HEK293T whole cell lysate and exosomes. Total RNA is extracted from whole cell lysates and exosomes. Then, the same amount of RNA is reverse transcribed into cDNA for the following qPCR analysis. XIST: P1/P2 represents two different sets of qPCR primers used for XIST detection. The p value is calculated using the unpaired Student's t test.

[0096] FIGS. **12***a* to **12***b* show the CARPID results of MALAT1, wherein:

[0097] FIG. **12***a* shows that formaldehyde-assisted RIP assay is used to verify the XIST-Ezrin interaction in HEK293T cells with Ezrin antibody and IgG as controls. RT-qPCR quantitative enrichment is performed by using DANCR in two different regions and MALAT1 in two different regions. GAPDH is used as an internal control in the experiment. Data is expressed as mean±SD, n=3; ** indicates $p<0.01$ using the unpaired Student's t test. The upper right panel shows the abundance of Ezrin in the input sample and in the immunoprecipitation, followed by Western blotting using Ezrin antibody or IgG.

[0098] FIG. **12***b* shows the identification of the MALAT1-related proteome associated with CARPID. The volcano plot shows the enrichment of MALAT1 related proteins in HEK293T cells. The x-axis represents the log 2 transformation value of the protein level fold change in the CARPID results of all three groups of XIST gRNAs in compared with the control. The y-axis shows the negative logarithmic converted p-value (non-parametric rank product test). Significantly enriched proteins are marked with orange dots.

DETAILED DESCRIPTION OF THE INVENTION

[0099] The technical solutions of the present invention are now described in detail below for a better understanding of the technical features, objectives, and advantageous effects of the present invention, but they should not be interpreted as limiting the scope of the present invention. The experimental methods without specifying specific conditions in the examples are conventional means and conventional conditions well known in the art, or according to the conditions recommended by the manufacturer.

Example 1. Establishment of CARPID Technology

[0100] Referring to FIG. **1***a*, the present invention provides a method called CRISPR-assisted RNA-protein interaction detection (CARPID), which can be used to detect RBP bound to endogenous lncRNA transcripts in living cells. The present invention designs a guide RNA (gRNA) array, which is composed of two gRNA sequences separated by a 30 nt direct repeat (DR) to target two adjacent loci on same lncRNA transcript (FIG. **4***a*). In theory, this can improve targeting specificity, thereby reducing background noise.

[0101] In order to identify the RBP that binds to the target lncRNA, dCasRx is fused with the engineered biotin ligase BASU in the present invention. In order to monitor and minimize the changes caused by the heterogeneous expres-

sion of BASU enzymes between cells, BASU-dCasRx is cloned in reading frame with self-cleaving T2A peptide and eGFP cDNA (FIG. **4***b*) in the present invention. In order to optimize the reaction conditions, various induction times are tried (FIG. **4***c*). The present invention also compares the enzymatic activity by reversing the order of BASU and dCasRx in the fusion protein, and no obvious difference is observed (FIG. **4***c*). The present invention selects BASU-dCasRx, and the cells are treated with 200 μM biotin for 15 minutes in the subsequent analysis, as the shortest but sufficient reaction time. By co-expressing BASU-dCasRx fusion protein with gRNA targeting specific regions of lncRNA (XIST), biotinylation is induced and then biotinylated proteins are enriched for mass spectrometry-mediated protein identification and quantification (FIG. **1***a*), so as to carry out the CARPID. No significant changes in gene expression are observed in cells overexpressing BASU-dCasRx and gRNA, confirming that CARPID does not change the physiological functions of transfected cells (FIG. **4***d*).

Example 2 Performance Evaluation of CARPID

[0102] XIST is one of the most interesting and intensively researched mammalian lncRNA genes. It is located on the long (q) arm of the X chromosome in the human genome and is only expressed in the inactive X chromosome (Xi) to regulate cis-XCI in differentiated female cells. Technical research has revealed a variety of XIST binding proteins in the art and gradually revealed potential molecular pathways. This example focuses on XIST to evaluate the performance of CARPID.

[0103] HEK293T cells are transfected with a vector expressing BASU-dCasRx and three different gRNAs targeting different regions of XIST in the present invention (FIG. **1***b*; see Table 1 for the three groups of gRNA).

TABLE 1

XIST L1-1	TGAAAAGACCTTGAAAACACCTGGTGTACC (SEQ ID No. 1)
XIST L1-2	AGGAGGGGACAAATAAGAGGGGACAGAGGT (SEQ ID No. 2)
XIST L2-1	TATGTGGAGAGGACCCTCCTTTTCTAGTGC (SEQ ID No. 3)
XIST L2-2	AGTCTTATGGAGTGGGCACTCCCTGCTGGA (SEQ ID No. 4)
XIST L3-1	AGTAGAGGGGTTCATGTATAATGGGTGGGA (SEQ ID No. 5)
XIST L3-2	AGAAGGGGCTTTGGGTAGTCAGCATACTCA (SEQ ID No. 6)
DANCR L1-1	TAAGAGACGAACTCCTGGAGCTCAAGGTCG (SEQ ID No. 7)
DANCR L1-2	GCTGCCTCAGTTCTTAGCGCAGGTTGACAA (SEQ ID No. 8)
DANCR L2-1	TTCCTATTGTAACTGAAGGGATAGTTGGCT (SEQ ID No. 9)
DANCR L2-2	CCAAATATGCGTACTAACTTGTAGCAACCA (SEQ ID No. 10)

TABLE 1-continued

MALAT1 L1-1	AGTTGCGGGGCCCCAGTCCTTTACAGAAGT (SEQ ID No. 11)
MALAT1 L1-2	TTCTGCGTTGCTAAAATGGCGCTGCGCTTA (SEQ ID No. 12)
MALAT1 L2-1	AATCTTAGAAACGTGAAAACCCACTCTTGG (SEQ ID No. 13)
MALAT1 L2-2	TTGCTTTTTTGTTCGAGAAATCGGAGCAGC (SEQ ID No. 14)

[0104] The specificity of these gRNA groups is confirmed by co-transfection with active CasRx, the CasRx co-transfection shows that the target area is specifically digested without affecting other areas (FIG. **4***e*). Due to the highly ordered structure of XIST, the present invention also avoids targeting the expected XRNA hairpin structure (FIG. **5**).

[0105] Three biological repetitions for each group of gRNA are performed in the present invention to further "dilute" the non-specific noise generated by random binding. To determine the baseline of background biotinylation, an empty gRNA vector is used in the present invention to perform a control CARPID. The protein identification based on mass spectrometry (MS) shows that most of the detected proteins with at least two peptides (447 proteins) are shared between different gRNA groups and in triplicate of each group, which indicates that CARPID has strong repeatability (FIG. **6***a*, FIG. **6***b*).

[0106] For the enrichment analysis, label-free MS quantitation and non-parametric rank product test are applied in the present invention, the enrichment cut-off value is >2 times, and the adjusted p value is <0.05. The results show that at least one group of gRNA significantly enriched in 73 XIST interaction proteins compared with the vector control group (FIG. **1***c*). In addition, 23 of the 73 proteins are found to have at least two different groups of gRNAs, and 13 of which are shared by all three gRNA pairs (FIG. **1***d*; FIG. **6***c*). Previous studies have reported more than a quarter (19/73) of these strong XIST interaction proteins (FIG. **1***d*), including a variety of functionally verified conjugates: Cohesin subunits RAD21 and SMC1A, an ATP-dependent helicase ATRX, SWI/SNF chromatin remodeling agent BRG1. It is also noted in the present invention that some known XIST interactive RBPs, such as SPEN and RBM15, are not in this list. The inventors believe that their binding to XIST may be weak or dynamic in living cells, and are difficult to enrich.

[0107] Gene ontology (GO) analysis of significantly enriched candidate proteins shows that the proteins interacting with XIST are largely involved in covalent chromatin modification and chromatin remodeling (FIG. **1***e*). The ATP-dependent helicase ATRX found in CARPID belongs to these categories. This gene is also reported in an independent study, asserting that it played a role in guiding the polycomb complex PRC2 to the X chromosome for inactivation and gene silencing. These findings of the present invention strongly indicate that CARPID is a highly reliable method for determining RBP.

[0108] In addition to the known XIST interactors, CARPID has also identified a variety of new factors, including the transcription initiation factor TFIID subunit TAF15 (FIG. **1***c*, FIG. **1***d*). TAF15 is known to interact with TATA-box binding protein (TBP) and RNA polymerase II, and act as a co-activator that recognizes the core promoter and promotes transcription initiation. The present invention

firstly confirms the association between TAF15 and XIST lncRNA by Western blotting (WB) and immune FISH (FIG. **2***a*, FIG. **7**). It is reported that TAF15 is an RNA binding protein in mouse tissues. Therefore, the present invention re-studies the TAF15 CLIP-seq data in the mouse brain, and found that TAF15 does significantly bind to XIST lncRNA, and the binding cluster enrichment degree exceeded the expected 9 times (FIG. **2***b*, FIG. **2***c*). In order to study whether the binding of TAF15 depends on its biochemical binding affinity to XIST lncRNA sequence characteristics, a library containing 40-nt RNA transcripts and random sequences is used in the present invention to perform HTR-SELEX experiments on TAF15 (FIG. **8***a*, FIG. **8***b*). HTR-SELEX identified significant enrichment of RNA sequence motifs of TAF15 (FIG. **8***c*), similar to previous reports. Given that the abundant hairpin structure in XIST lncRNA may lead to dinucleotide interdependence, the effect of RNA sequence on TAF15 binding cannot be fully described by a simple position weight matrix model. Therefore, a k-mer-based machine learning algorithm (gkm-SVM) is used in the present invention, and HTR-SELEX data is used to model the RNA binding specificity of the human TAF15 protein (FIG. **8***d*). Consistent with the WB and MS results, the HTR-SELEX results further support that the affinity of locus 1 and 2 where TAF15 binds to XIST lncRNA is higher than locus 3 (FIG. **2***e*).

[0109] The well-studied chromatin remodeling agent SNF2L is also found in CARPID (FIG. **1***c*), and confirmed by WB and immune FISH (FIG. **20**. Consistently, the RIP-qPCR results shows that XIST lncRNA is significantly correlated with SNF2L (FIG. **2***g*). SNF2L and SNF2H are two collateral ATP-dependent chromatin remodeling enzymes belonging to the ISWI (Imitation Switch) family, which can move nucleosomes along DNA. For more than ten years, IWSI has been known to be associated with cohesin complexes in human cells. Consistent with this, the present invention identified two cohesin subunits SMC1A and RAD21 that interact with XIST (FIG. **1***c*).

[0110] In order to verify the role of these two new RBPs in mammalian XCI, a transgene female mouse embryonic fibroblast cell line with Xi-linked GFP reporter gene is used in the present invention. Xi-linked GFP is silenced by multiple epigenetic mechanisms. Therefore, no GFP transcript is detected (FIG. **2***h*) and no fluorescent signal is observed (FIG. **9***a*). In contrast, when 5'-azacytosine (5-aza) is used to inhibit DNA methylation, both GFP mRNA and fluorescence signal increased significantly. In order to clarify their functional importance in XCI, TAF15 and SNF2L are depleted in the presence of 5-aza processing in the present invention (FIG. **9***b*). Surprisingly, after silencing TAF15, the level of 5-aza-enhanced GFP is significantly reduced, which can be partially rescued by ectopic expression of RNAi-resistant TAF15 cDNA (FIG. **9***c*, FIG. **9***d*). In order to rule out the possibility that TAF15-related XCI is specific to the knocked-in GFP locus, RNA sequencing is performed in the present invention after the exhaustion of TAF15 in the female mouse embryonic fibroblast cell line under the genetic background of MAF and Cast hybrid. The allelic expression of a gene can be determined by the availability of SNPs between two different genetic backgrounds. As expected, the genes on the X chromosome show greater allelic depletion than autosomal genes (FIG. **9***e*), which proves the role of TAF15 in antagonizing XCI. On the other hand, in 5-aza-treated cells, knockdown of SNF2L resulted in further suppression of GFP (FIG. **2***h*, FIG. **9***a*), indicating that SNF2L and XIST RNA act synergistically to promote XCI.

[0111] Functionally, SNF2L and TAF15 belong to transcription repressor and activator, respectively. The present invention found that SNF2L is used as XIST RBP, indicating that ISWI family proteins may promote XCI through their known function of regulating the higher-order structure of chromatin. The binding of TAF15 and XIST indicates that XIST RNA can repel TAF15 and possibly other transcriptional activators from binding to the promoter of Xi-linked genes, thereby preventing the expression of target genes. This phenomenon supports a multitasking model that recruits inhibitors (such as SNF2L) and expulsion transcription activators (such as TAF15), and they may be the basis of XIST-mediated X chromosome inactivation (FIG. **9***f*).

Example 3. Application of CARPID

[0112] In order to generalize the application of CARPID and extend its application to non-nuclear lncRNAs, the present invention designs a gRNA set with respect to two other lncRNAs DANCR (differentiation antagonistic non-protein coding RNA) and MALAT1 (lung adenocarcinoma transcript 1 associated with metastasis). It is reported that DANCR transcripts mainly exist in the cytoplasm, and their overexpression is significantly related to the poor prognosis of a variety of cancers, including breast cancer, liver cancer, colorectal cancer and osteosarcoma. However, the molecular mechanism has not yet been elucidated. It is important to note that the length of DANCR is 1000 nucleotides, which is much shorter than XIST. This and its low abundance in cells (FIG. **10***a*) make it technically challenging to conduct research using currently available methods (such as ChIRP-MS), which requires spanning dozens of different RNA probes in order to sufficiently capture.

[0113] CARPID is used together with two sets of gRNA in HEK293T cells (FIG. **10***b*), and the present invention detects 640 DANCR lncRNA-related proteins (≥2 peptides), of which 35 and 26 proteins are significantly enriched in the locus 1 and locus 2 (FIG. **3***a*, FIG. **3***b*, FIG. **10***c*).

[0114] It is worth noting that GO-term analysis shows that most DANCR-related proteins are rich in extracellular vesicles, which indicates that DANCR is located in this specialized cell compartment (FIG. **10***d*). To verify this, the present invention purifies vesicles from HEK293T cells and checks the total RNA of exosomes and whole cell lysates (FIG. **11***a*, FIG. **11***b*). Indeed, quantitative RT-PCR analysis shows that the degree of enrichment of DANCR in exosomes is 5 times that of cell lysates (FIG. **11***c*). In contrast, XIST is largely consumed in exosomes (FIG. **11***d*). The present invention also notes an interesting DANCR binding protein Ezrin (EZR) (FIG. **10***d*), a membrane-bound cytoskeleton junction protein, which is associated with the poor prognosis of many cancers. RIP-qPCR verified the binding of Ezrin to DANCR lncRNA. Compared with the IgG control, DANCR lncRNA is approximately 2 times enriched in the Ezrin drop-down list, but it is not used for MALAT1 lncRNA (FIG. **12***a*).

[0115] CARPID is also performed in the present invention on another lncRNA MALAT1 that is known to be abundant in the nucleus but also present in the cytoplasm. Two groups of different gRNAs used for human MALAT1 can capture 484 proteins (≥2 peptides), of which 43 proteins are significantly enriched (FIG. **3***c*, FIG. **3***d*, FIG. **12***b*).

[0116] In this study, among the three tested lncRNAs, the comparison results of CARPID that partially share the subcellular distribution results in the candidates with almost no overlap, demonstrating the high specificity of the CARPID method (FIG. **3***e*). In summary, these data support that CARPID is a powerful tool for studying lncRNAs of various lengths and expression levels in various subcellular locations.

SEQUENCE LISTING

<160> NUMBER OF SEQ ID NOS: 14

<210> SEQ ID NO 1
<211> LENGTH: 30
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: gRNA

<400> SEQUENCE: 1

tgaaaagacc ttgaaaacac ctggtgtacc 30

<210> SEQ ID NO 2
<211> LENGTH: 30
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: gRNA

<400> SEQUENCE: 2

aggaggggac aaataagagg ggacagaggt 30

<210> SEQ ID NO 3
<211> LENGTH: 30
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: gRNA

<400> SEQUENCE: 3

tatgtggaga ggaccctcct tttctagtgc 30

<210> SEQ ID NO 4
<211> LENGTH: 30
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: gRNA

<400> SEQUENCE: 4

agtcttatgg agtgggcact ccctgctgga 30

<210> SEQ ID NO 5
<211> LENGTH: 30
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: gRNA

<400> SEQUENCE: 5

agtagagggg ttcatgtata atgggtggga 30

<210> SEQ ID NO 6
<211> LENGTH: 30
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: gRNA

-continued

<400> SEQUENCE: 6

agaaggggct ttgggtagtc agcatactca 30

<210> SEQ ID NO 7
<211> LENGTH: 30
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: gRNA

<400> SEQUENCE: 7

taagagacga actcctggag ctcaaggtcg 30

<210> SEQ ID NO 8
<211> LENGTH: 30
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: gRNA

<400> SEQUENCE: 8

gctgcctcag ttcttagcgc aggttgacaa 30

<210> SEQ ID NO 9
<211> LENGTH: 30
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: gRNA

<400> SEQUENCE: 9

ttcctattgt aactgaaggg atagttggct 30

<210> SEQ ID NO 10
<211> LENGTH: 30
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: gRAN

<400> SEQUENCE: 10

ccaaatatgc gtactaactt gtagcaacca 30

<210> SEQ ID NO 11
<211> LENGTH: 30
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: gRNA

<400> SEQUENCE: 11

agttgcgggg ccccagtcct ttacagaagt 30

<210> SEQ ID NO 12
<211> LENGTH: 30
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: gRNA

<400> SEQUENCE: 12

ttctgcgttg ctaaaatggc gctgcgctta 30

-continued

<210> SEQ ID NO 13
<211> LENGTH: 30
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: gRNA

<400> SEQUENCE: 13

aatcttagaa acgtgaaaac ccactcttgg 30

<210> SEQ ID NO 14
<211> LENGTH: 30
<212> TYPE: DNA
<213> ORGANISM: Artificial Sequence
<220> FEATURE:
<223> OTHER INFORMATION: gRNA

<400> SEQUENCE: 14

ttgctttttt gttcgagaaa tcggagcagc 30

1. A fusion protein formed by BASU and dCasRx.

2. The fusion protein according to claim **1**, which is BASU-dCasRx or dCasRx-BASU.

3. An expression vector for expressing the fusion protein according to claim **1**.

4. A composition comprising: the fusion protein according to claim **1** and/or the expression vector according to claim **3**, and a gRNA targeting the target lncRNA;

preferably, the target lncRNA is XIST, DANCR, or MALAT1.

5. A kit for determining lncRNA interaction proteins, comprising: the fusion protein according to claim **1** and/or the expression vector according to claim **3**, and a gRNA targeting the target lncRNA;

preferably, the kit further comprises a control reagent without gRNA.

6. A method for determining lncRNA interaction proteins, comprising:

co-transfecting the expression vector according to claim **3** and a gRNA that specifically targets the target lncRNA in a target cell, thereby BASU specifically biotin-labeling effector proteins nearby;

isolating the biotinylated proteins for analysis to determine the lncRNA interaction proteins.

7. The method according to claim **6**, which is for determining a lncRNA interaction protein in living cells.

8. The method according to claim **6**, further comprising:

statistically comparing the protein identified in the specific gRNA group with the control group without gRNA for enrichment or reduction; preferably, the method further comprises:

calculating the false discovery rate of enrichment by using Rank product; more preferably, the critical value is set as enrichment≥2 folds and FDR≤0.05;

preferably, the method further comprises:

applying the obtained proteins for genetic ontological analysis or protein interaction network analysis; and/or

analyzing the target protein comprehensively by using gRNA targeting different regions, thereby obtaining a high-resolution spectrum of the target lncRNA interaction protein.

9. A method for analyzing enriched interaction proteins to specific regions of the target lncRNA, comprising:

performing an enrichment analysis on the proteins with more than one peptide fragments detected by the method according to the claim **6**;

normalizing and logarithmizing a LFQ abundance of each group;

replacing a missing value by a minimum value representing the detection limits of mass spectrometer

determining the protein that is statistically enriched in samples of the gRNA transfection group compared to the control group transfected with gRNA empty vectors by using rank product;

a protein with an adjusted p-value 50.05 and the abundance change 2 folds, is identified as RBP to the target lncRNA.

10. The method according to claim **9**, wherein the protein subjected to the enrichment analysis including human keratin.

* * * * *