UNIVERSITY OF CALIFORNIA SAN DIEGO

Detection and Characterization of Cell Surface RNA Signals using Lipid Bead Pull-down

A Thesis submitted in partial satisfaction of the requirements
for the degree Master of Science

in

Biology

by

Xuerui Huang

Committee in charge:

Sheng Zhong, Chair
Nan Hao, Co-Chair
Barry Grant

2020

The Thesis of Xuerui Huang is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

Co-chair

_____

Chair

University of California San Diego

2020

Table of Contents

List of Figures

List of Tables

List of Schema

Acknowledgements

I would like to acknowledge Professor Sheng Zhong for his support as the chair of my committee. Through multiple drafts and many long nights, his guidance has proved to be invaluable.

I would also like to acknowledge Doctor Lucie Hebert and Doctor Kathia Zaleta Rivera of Zhong lab, who spent their valuable time in the development of experimental protocols and the generation of experimental data. It is their support that helped me in an immeasurable way.

This thesis is coauthored with Lucie Hebert, Kathia Zaleta-Rivera, Xiaochen Fan, and Norman Huang. The thesis author, Xuerui Huang, was the primary author of this thesis.

ABSTRACT OF THE THESIS

Detection and Characterization of Cell Surface RNA Signals Using Lipid Bead Pull-down

by

Xuerui Huang

Master of Science in Biology

University of California San Diego, 2020

Professor Sheng Zhong, Chair
Professor Nan Hao, Co-Chair

RNA has been proved to interact with lipid bilayers and various proteins near the membrane, however; the existence of cell surface RNA was less explored. Using CLICK reaction and fluorescence visualization, preliminary research in Zhong Lab had suggested the existence of cell surface RNAs (csRNA) that attached firmly on the outer membrane of mouse and human cells. For the first step in this study, functionality assay of

cytotoxicity, which measured by LDH release of human immune Natural Killer cell line NK92, showed that the global csRNA perturbation could significantly impact NK92's cell killing activity. To further capture and characterize csRNAs using sequencing data, we have developed a method to pull-down lipid-associated RNAs using lipid-coated beads followed by RNA-sequencing. csRNAs were identified based on differential analysis of RNAs bound to 8 types of membrane-associated lipid against the control bead. Candidate RNAs discovered were validated using two sequencing techniques developed by the Zhong Lab, i.e. SurfaceClick and SurfaceSeq, to further remove background noise. Functional and structural analysis of validated csRNAs showed significant enrichment on immune and cancer-related functions as miRNAs. Furthermore, analysis result further indicated potential cellular functions such as cell-cell recognition, structural support, and signaling regulation. In conclusion, this study indicated the existence of RNA on cell's outer membrane as functional and structural groups for cellular functions such as anti-tumor cytotoxicity and immune response.

Introduction

RNA molecules are known to interact with proteins, DNAs, and other RNA molecules in every compartment of mammalian cell while the localization of RNAs was essential for understanding their functions. However, the possibility of RNAs presenting on the cell outer membrane is less explored, although several studies have proved RNAs could interact with lipids and proteins near plasma membranes and regulate cellular functions. A decade ago, Micheal Yarus Laboratory investigated RNA affinity for lipids on reconstituted membrane surfaces. They first stated that the chemical and molecular properties of RNAs and lipids enable affinity, complex stability, and even RNA protection from degradation[1]. They showed that RNA aggregate can bind to "patch regions" of reconstituted phospholipid bilayer through their secondary structure. They further found that RNA affinity for lipid was dependent on the organization of lipid structures by showing that RNAs have a high affinity for highly ordered lipid bilayer such as lipid rafts and cholesterol-based vesicles[2] . More recently, a specific interaction has been studied between the *LinkA* long noncoding RNA and PIP3 (Phosphatidylinositol (3,4,5)-trisphosphate) at the inner leaflet of the plasma membrane[3]. Their research result showed that the *LinkA*-PIP3 interaction was single-nucleotide specific and demonstrated for the first time that an RNA-lipid interaction had important biological and cellular consequences. These findings are extremely interesting and suggest a whole unexplored class of cellular signaling function for both coding and non-coding RNA through their interaction with lipids at the plasma membrane in mammalian cells. Therefore, an extensive comprehension of RNA affinity for lipids and their functional implications

would be of great significance to cell signaling and other regulated activities at the cell membrane.

In order to demonstrate the existence of endogenous RNA molecules located at the plasma membrane, a functionality assay based on cell's cytotoxicity and a lipid beads-based sequencing technique, LipidSeq, have been developed by Zhong Lab to discover csRNA signals. LipidSeq is an unbiased method that utilized lipid-coated beads followed by RNA-sequencing to identify RNA that directly bind to lipids. RNA starting material was isolated from the membrane fraction of EL4 cells, in order to enrich for RNA species having affinity for lipids. Eight different types of lipid-coated beads were selected based on chemical characteristics, presence in specific leaflets of the plasma membrane, and their known role in the bilayer structure. After pulling-down of the lipid-binding RNA species, cDNA libraries were generated and sequenced to identify the affinity of RNA molecules for each type of lipid beads compared to their affinity for control beads (non-coated with lipids). All affinities were virtually mapped onto the lipid bilayer structures to help decipher the structural functions of csRNAs. Ultimately, LipidSeq detected csRNA signals were compared to the csRNA candidates identified using two other Zhong Lab developed techniques, SurfaceClick and SurfaceSeq for further validation. SurfaceClick and SurfaceSeq were based on totally different mechanisms and both showed positive csRNA signals on human and mouse cells. With further validation using the two previously developed techniques, the legitimacy of csRNA signals would be further solidified.

Result

2.1 Functionality Examination: csRNAs could act as functional groups to affect natural killer cell's killing potential

Based on previous work in Zhong lab, two cancerous immune cell lines from mouse and human, EL4 and NK92, were identified to have positive signals for csRNAs using imaging technique (***Supplementary Table 1, Supplementary Figure 1)***. The next essential step is to evaluate whether csRNAs display relevant functions for the cells. Discovering any csRNA function will rule out the possibility that csRNA are present at the membrane surface in a nonpurposive way (cell trash release) or that these RNAs are captured-content of exosomes, or cell-free RNA sticking at the surface of cell-membrane. Therefore, we decided to perform the functional evaluation of cytotoxicity for csRNA on NK92 cell line. NK92 cell line is a cancerous natural killer cell line known to have retained the natural killer cell features and cytotoxic properties[4]. Cytotoxicity was defined as natural killer cell's killing potential, which was measured by the amount of lactate dehydrogenase (LDH) released into culture media. LDH is a stable cytosolic enzyme released upon cell lysis, and the amount of LDH released into culture media is proportional to the number of lysed cells. To assess the effect of csRNA, we removed csRNA of NK92 cells(effector cells) using RNase prior to the co-culture with target cells and then compared NK92 cytotoxicity ability with and without RNase treatment. The efficiency of RNase treatment has previously been evaluated using microscopy.

Functional assay of cytotoxicity showed that RNA perturbation using RNase on cell surface had a significant effect on NK92's cytotoxicity. Cytotoxicity was measured

by LDH release of the cells under two conditions, with RNase and without RNase, each with two different ratios of NK92 cells versus MDA-MB-231 cells (E:T ratio, a.k.a. dose) in the well. Variation of cytotoxicity was measured using two-way ANOVA to determine whether the change in condition or the change in dose could significantly affect NK92's cytotoxicity. In this case, the interaction effect between condition and dose was ignored. Based on the ANOVA result (***Figure 1A***), we could conclude that applying RNase on the cell surface of NK92 could significantly affect cytotoxicity while dose, different E:T ratio, was not statistically significant. By plotting the average cytotoxicity with and without RNase, a significant decrease of LDH release (cytotoxicity) was observed after 1:250 RNase treatment, both with the NK-92 stimulation by MDA-MB-231 target cells using E:T ratio of 0.31:1 and 0.16:1 (***Figure 1B***). These results lead to the conclusion that adding RNase to cell's outer membrane for the removal of cell surface RNAs would impact the cytotoxicity of NK92 cells significantly. Thus, by using the RNase treatment and the measurement of cytotoxicity using LDH release, csRNA as functional groups on cell membrane could be confirmed.

A

Value	Sum Sq	Df	F_value	Pr(>F)	Sig_level
Condition	0.3307	1	7.1	0.008868	**
Dose	0.0116	1	0.2481	0.619389	
Residuals	5.1242	110			

B

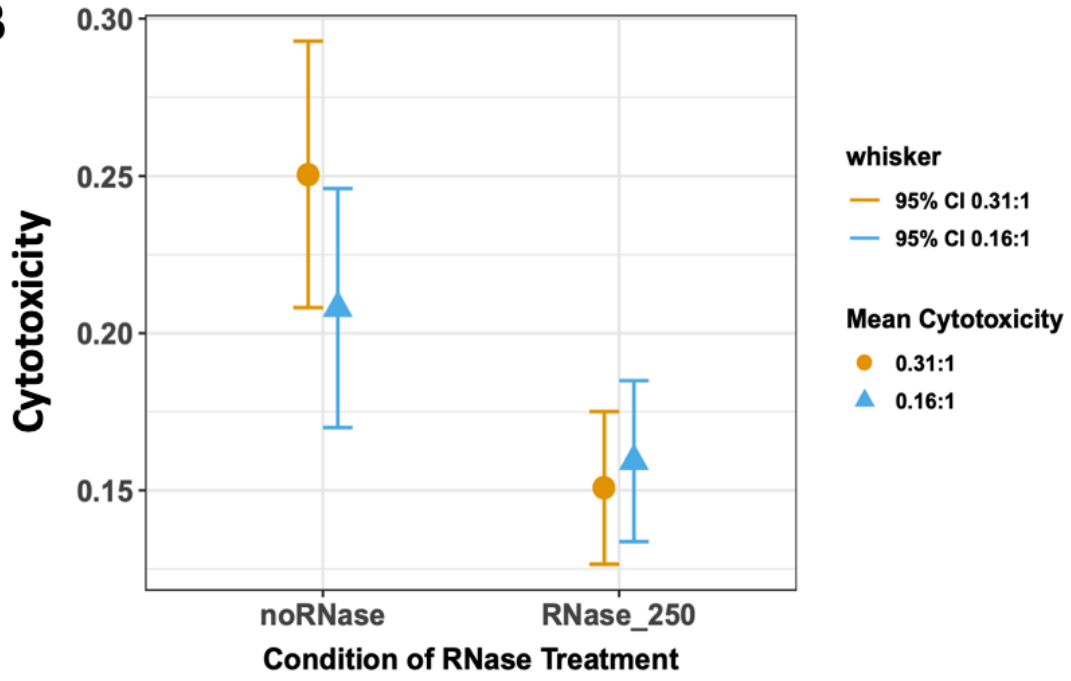


Figure 1: Result of LDH Release Cytotoxicity Assay and Two-way ANOVA: A. Result table of two-way ANOVA. Here, the condition is +/- RNase treatment, and Dose was the different E:T ratio to wells that applied. The F ratio is the ratio of two mean square values. If the null hypothesis is true, you expect F to have a value close to 1.0 most of the time. A large F ratio means that the variation among group means is more than you'd expect to see by chance. Pr(>F) is the p-value, which indicated the probability of null hypothesis is true. Here, the Pr(>F) for "Condition" is smaller than 0.01, indicated high possibility of the existence of cell surface RNAs. **B. The effect of RNase treatment on cytotoxicity**. Shapes in the middle is the average of all the NK cell's cytotoxicity value (4 data points in total) calculated under certain condition of RNase treatment. Color indicated different E:T ratio. E:T ratio is the ratio between number of effector cells, NK92, and number of target cells, MDA-MB-231, in the well. Larger the E, more K92 cells were put into the well, Yellow is the E:T ratio of 0.31:1, and blue is the E:T ratio of 0.16:1

2.2 LipidSeq Detected csRNAs: LipidSeq has the ability of detecting and grabbing csRNAs on cell outer membranes

RNA starting material was isolated from the EL4 membrane fractions. Nine different lipid-coated beads, with one control bead and 8 different types of beads that specifically bind to RNAs at outer leaflet, inner leaflet, and lipid rafts of the cell separately were selected (***Table 1, Supplementary Table 2***). Potential csRNA signals were defined as differential peaks with positive fold change between each type of bead with the control bead, which indicated significant enrichment from the control sample. Through pipeline, in total of 546,730 signals were detected with p-value threshold of 0.01, and 3,542 significant signals were selected with threshold of q-value smaller than 0.01 from all detected signals. Among 8 types of lipid beads, PC (*Phosphati-dycholines*) and PE (*Phosphati-dylethanolamines*) had a relative low number of signals while other beads all had around 90,000 signals detected. All beads had low percentage of significant signals, with PE beads had the least number of significant signals and Chol (*Cholesterol*) bead had the largest number of signals and significant signals detected **(*Table 1***).

To further check the pattern of distribution for these signals, read region distribution plot from transcription start site to transcription end site (***Figure 2A***) and chromosomal distribution (***Table 2, Figure2B***) were examined. The read region distribution plot presented that the reads pulled downed by beads PS (*Phosphati-dylethanolamines*), SM (*Sphingo-myelines*), Cer (*Ceramides*), and SS (*Sphingosines*) have strong enrichment on the transcription starting site and transcription ending site. The SM, Cer, SS, but not PS beads had specific affinity to bind RNAs on the outer leaflet, which could be the special property of reads displayed on the outer leaflet of cell

membrane. From the statistic and visualization of chromosomal distribution (***Table 2, Figure2B***), we observed a trend of high consistency, with more signals discovered on chromosome 2, chromosome 5, chromosome 11 and chromosome X. This could be an indication for specific enrichment of certain functions. Moreover, the number of signals discovered on each chromosome is on the same level except bead Chol (*Cholesterol*), which has comparably higher number of peaks discovered on each chromosome. Lipid bead Chol has affinity for both RNAs on lipid rafts and all the other types of RNAs. Therefore, it is very reasonable to discover more signals and significant signals in comparison to all other types of beads.

From the potential csRNA signals discovered by LipidSeq, as well as high similarity of read region distribution and chromosomal distribution, we could confirm that LipidSeq had the ability to capture cell surface specific RNA signals.

Table 1: Table of Peaks and Significant Peaks Detected. Peaks were detected using *MACS2*. Last column presented the percentage of significant peaks among all peaks detected. Different background color indicated specific binding position, which green background color indicated the bead is specific for outer leaflet. Blue background color indicated the bead is specific for RNAs on the inner leaflet, the red color indicated the bead is specific for RNAs on the lipid raft as well as all other types of RNAs, and the orange color indicated the bead is specific for RNAs on the outer leaflet and lipid raft.

Beads	Peaks (Pvalue<0.01)	Sig.peaks (Qvlaue<0.05)	Sig.Percentage
SM	74,443	585	0.79%
Cer	84,621	412	0.49%
PC	2,308	302	13.08%
PE	5,800	19	0.33%
PS	87,344	332	0.38%
PiP3	96,251	541	0.56%
Chol	100,543	981	0.98%
SS	95,420	370	0.39%
Total	546,730	3,542	0.65%

Table 2. Chromosomal Distribution of Peaks. This table showed the number of peaks detection on each chromosome for every type of beads. Different background color indicated specific binding position, which green background color indicated the bead is specific for outer leaflet. Blue background color indicated the bead is specific for RNAs on the inner leaflet, the red color indicated the bead is specific for RNAs on the lipid raft, and the orange color indicated the bead is specific for RNAs on the outer leaflet and lipid raft.

Name	chr1	chr2	chr3	chr4	chr5	chr6	chr7	chr8	chr9	chr10	chr11	chr12	chr13	chr14	chr15	chr16	chr17	chr18	chr19	chrM	chrX	chrY
SM	42	41	22	34	52	27	30	33	20	29	49	25	18	20	30	24	29	11	24	7	16	2
Cer	30	22	20	28	30	20	18	28	19	20	29	22	18	14	15	11	18	15	15	6	11	3
PC	24	25	19	15	19	15	22	7	16	14	22	15	12	8	11	11	8	12	11	0	15	1
PE	0	0	2	2	0	0	0	2	1	3	0	1	1	0	1	0	2	2	0	0	2	0
PS	19	27	14	14	21	16	18	22	21	18	26	13	15	11	11	11	16	10	8	3	13	4
PiP3	32	36	24	34	42	30	38	36	31	22	36	23	25	24	19	9	25	15	22	6	11	1
Chol	67	88	47	67	63	35	47	41	65	54	77	55	43	36	44	31	26	27	29	4	32	2
SS	23	30	16	22	31	15	17	19	26	18	20	17	21	13	16	7	15	20	15	2	7	0

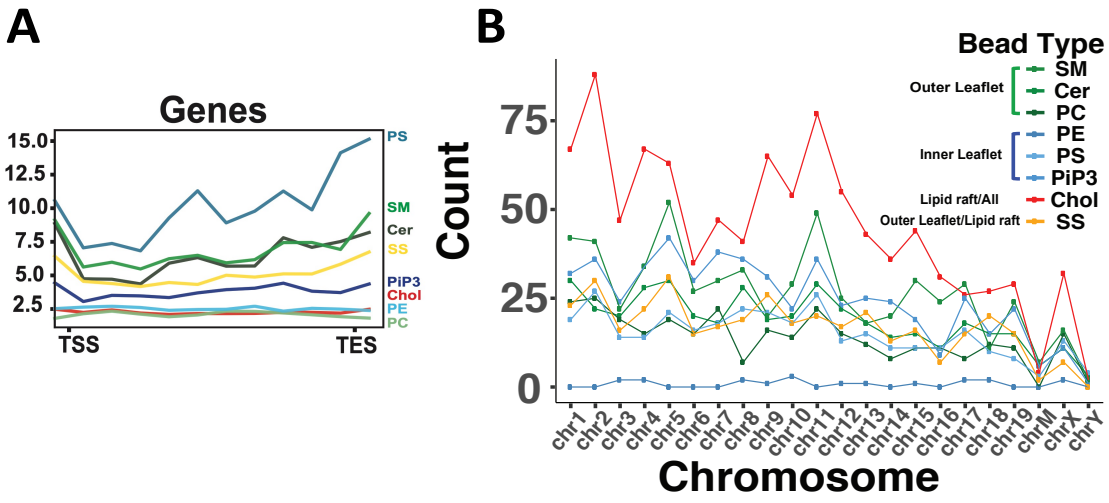


Figure 2: Peak Distribution: Different color indicated specific binding position, which green color indicated the bead is specific for outer leaflet. Blue background color indicated the bead is specific for RNAs on the inner leaflet, the red color indicated the bead is specific for RNAs on the lipid raft, and the orange color indicated the bead is specific for RNAs on the outer leaflet and lipid raft.
A. Average Chromosomal Distribution of the Peaks. This figure showed the average trend of peak distribution in the scale of 22 chromosomes of mouse genome from the transcription starting site to the transcription ending site. In total of 8 different beads were presented here, differentiated by different color.
B. **Visualization od Chromosomal Distribution of Peaks for Each Type of Beads**. This figure showed the trend for number of peaks detected on every chromosome of mouse genome for each type of bead. In total of 8 different beads were presented here, differentiated by different color.

2.3 Cross Validation between SurfaceClick and SurfaceSeq: SurfaceClick and SurfaceSeq are essential external source for LipidSeq signal validation

Since the accuracy of the detected signals using LipidSeq remained unknown, further validation was essential for extracting true positive signals. Here, validation was performed by signal overlapping with two other cell surface RNA sequencing techniques developed by Zhong Lab, SurfaceClick and SurfaceSeq. SurfaceSeq is a technique that based on drug delivery system that utilizes biodegradable polymeric nanoparticles (poly-lactic-co-glycolic acid)[5]. Nanoparticles would fuse with the cell membrane to produce membrane-coated nanoparticles for csRNA pull-down. SurfaceClick is a technique where cell surface RNAs are labeled on intact cells via CLICK reaction. Total RNA is further isolated, fragmented and purified streptavidin beads, where only the labeled csRNAs will be pulled down. Previous work showed that both techniques detected positive csRNA signals on EL4 cell line. Therefore, these two technologies could be essential source of further validation.

Nonetheless, the assessment of these two orthogonal technologies for the robustness was necessary before the validation. For that purpose, cross-validation was performed to compare the csRNA candidates obtained for EL4 cell line by the SurfaceClick technology and SurfaceSeq technology. The two techniques are technically and biologically drastically different and thus have different noise and background signal origin, which makes the comparison stronger to fulfill the overall goal.

Log2FoldChange (log2FC) for 46,191genes that measured by both technologies were collected and plotted as dot plot with log2FC of the gene detected in SurfaceClick on the X-axis and log2FC of the gene detected in SurfaceSeq on the Y-axis (***Figure 4A***).

All of these genes were used to fit the linear regression model and check for Pearson correlation. Result showed that genes detected by SurfaceClick is positively correlated with genes detected by SurfaceSeq with goodness-of-fit score 0.24 and Pearson correlation coefficient of 0.49 (P-value = 2.2e-16), which suggested high consistency between these two techniques. In each technique, significantly enriched genes were identified as genes with log2FC larger than 1 and adjusted p-value smaller than 0.05. In order to understand whether a gene identified as significantly enriched by one technique was also identified in the other technique, volcano plots with log2FC of the gene on the X-axis and padj on the Y-axis, were made, and hypergeometric test was performed on the intersection. To evaluate the distribution of detected genes, 680 significantly enriched surface RNA genes obtained with SurfaceClick (Dark Orange dots, ***Figure 3B***) were marked in the same color in the volcano plot with values obtained by SurfaceSeq **(*Figure 3D*)**. Vice versa, 1384 significantly enriched surface RNA genes obtained with SurfaceSeq (Dark blue dots, ***Figure 3C***) were marked in the same color in the volcano plot with values obtained by SurfaceClick (***Figure 3E***). Hypergeometric test was performed with 5 different cross-validation thresholds to test whether the probability of a gene identified as significantly enriched by SurfaceClick is equal to the probability of the same gene identified as significantly enriched by SurfaceSeq. A cross-validation threshold was applied to genes significantly enriched in the other technique (Dark Orange dots in Figure 3C or dark blue dots in Figure 3E). **Table 3** showed the result with cross-validation threshold of log2FC larger than 1 and padj < 0.05 and ***Supplementary Table 5*** showed the result with other 4 different cross-validation thresholds. Interestingly, 89/680 (14%) of the SurfaceClick detected genes and 89/1384 (6.4%) of the SurfaceSeq detected

genes were commonly enriched in surface samples in the two techniques. Since both p-values are smaller than 0.001 (***Table 3B, Table 3C***), the null hypothesis of the probability of a gene identified as significantly enriched by SurfaceClick is equal to the probability of the same gene identified as significantly enriched by SurfaceSeq could be rejected. Moreover, we can further conclude that genes identified by one technique has high probability to be identified as differentially expressed by the other technique.

These results indicated a large proportion of commonly enriched genes was detected by both techniques, confirmed the ability of these two technologies for identifying csRNAs, and legitimacy of using SurfaceClick and SurfaceSeq as outside source for validating LipidSeq detected signals.

Table 3*:* Hypergeometric Test in Detail and Test Result. Hypergeometric test for testing whether probability of a gene identified as significantly enriched by SurfaceClick is equal to the probability of the same gene identified as significantly enriched by SurfaceSeq. **A**. Gene stats with cross validation threshold of log2FC larger than 1 and padj < 0.05. **B**. The probability of selecting 89 significantly enriched genes from a sample of 680 genes taken from a SurfaceSeq gene pool containing 1384 significant enriched genes and 23037 non-significant enriched genes. **C**. The probability of selecting 89 significantly enriched genes from a sample of 1384 genes taken a SurfaceClick gene pool containing 680 significant enriched genes and 23740 non-significant enriched genes

A

	SurfaceClick	SurfaceSeq
Significantly Enriched Genes	680	1,384
insignificantly Enriched Genes	23,740	23,037
Cross-Validated Genes by the Other Tech	89	89

B

	SurfaceClick Detected Genes in SurfaceSeq	All SurfaceSeq Genes
Significant	89	1384
Insignificant	591	23,037
Total	680	24,421

dhyper.pValue= 8.075173e-14

C

	SurfaceSeq Detected Genes in SurfaceClick	All SurfaceClick Genes
Significant	89	680
Insignificant	1,295	23,740
Total	1,384	24,421

dhyper.pValue= 8.093368e-14

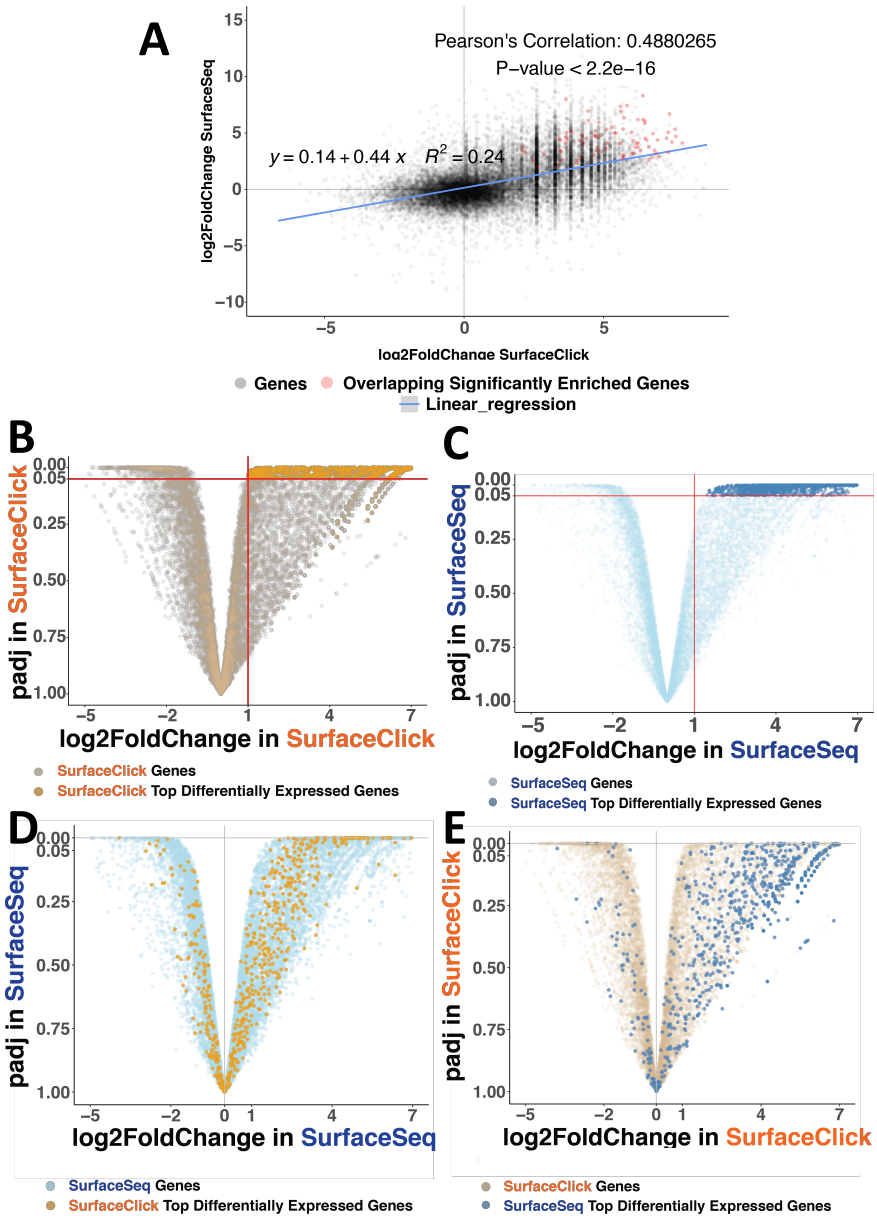


Figure 3: Cross Validation between SurfaceClick and SurfaceSeq: A. Global log2FC Scatter plot. Every dot represents a single gene, with x-value of log2FC in Surface-CLICK-seq and y-value of log2FC in SurfaceSeq technique. Red dots are genes that are detected by both techniques that passed the filter of log2FC larger than 1 and p value smaller 0.01. The linear regression is indicated in blue, showing a positive correlation. The Pearson correlation between the two techniques is indicated together with the p value **B.** Volcano plots for candidate genes detected in SurfaceSeq technique with log2FC and adjusted p-value from DEseq Call. Dark blue dots indicated significantly enriched SurfaceSeq genes **C**. Volcano plots for candidate genes detected in SurfaceSeq technique with log2FC and adjusted p-value from DEseq Call. Dark orange dots indicated significantly enriched SurfaceClick genes **D.** Volcano plots for candidate genes detected in SurfaceSeq technique with log2FC and adjusted p-value from DEseq Call. Dark blue dots indicated significantly enriched SurfaceSeq genes

2.4 LipidSeq Signal Validation using External Source: Legitimacy of LipidSeq detected signals could be ensured.

The cross-validation in the previous section showed positive correlation of genes detected by SurfaceClick and SurfaceSeq separately. Therefore, these two techniques were used as essential external source for further validation of LipidSeq detected signals. Significant technical signals from SurfaceClick and SurfaceSeq were detected and selected using the same data processing pipeline as in LipidSeq for consistency. In total of 20740 technical signals with 19,980 significant technical signals were detected using SurfaceSeq, and 32398 technical signals with 24,838 significant technical signals were detected using SurfaceClick (***Table 4A***). Although SurfaceSeq had 5 times more reads than SurfaceClick, more signals were detected from SurfaceClick technique (***Supplementary Table 3***). The validation was performed by overlapping LipidSeq detected signals with the two types of technical signals separately with two stringencies, general overlap and significant overlap (***Table 4A***). From validation, we could observe that the number of overlaps was round the same level, but signals had more general overlaps with SurfaceClick while having more significant overlaps with SurfaceSeq. In comparison to the number of signals that were validated by SurfaceClick or SurfaceSeq individually, less signals were validated by both techniques (***Table 4B*)**. Also, PC (*Phosphati-dycholines*) bead and PE (*Phosphati-dylethanolamines*) bead had relative low number of overlapping signals, which was consistent with the trend of signals detected in the previous section (section 2.2).

In order to make a clearer visualization for the relationship of inter and intra overlapping between lipid bead signals and two orthogonal technical signals, upset plot

was made, as a modified version of Venn diagram. In the upset plot, each column corresponds to a set, and each row corresponds to different conditions. Cells are either empty (light gray), indicating that this set is not part of that intersection, or filled by black dots, showing that the set is participating in the intersection. Preliminary data analysis showed that our data had an extensive amount of background; therefore, large number of consistent signals between LipidSeq signals and technical signals were hard to detect. Indeed, based on what we observed, most of the signals were unique to one type of bead specifically on the gene level, and a comparably small portion of signals were shared among four or more different beads or verified by both techniques, which would be highly consistent signals (***Figure 4***). Highly consistent signals were further validated by close-up genomic view (***Figure 5***). From the close-up signal visualization, we could observe a high consistency on the location of peak signals, which further solidified the legitimacy of detected signals.

Table 4: Number of Peaks Detected by Techniques and Overlap Counts. A. Total peaks (p-value < 0.01) and total significant peaks (q-value < 0.05) detected using MACS2. **B**. Number of Overlapping peaks with two different stringencies. Last column presented the number of common peaks that overlapped with both techniques with the stringency of p-value < 0.01. General overlap was performed by overlapping general LipidSeq signals (threshold of p-value < 0.01) and general technical signals (threshold of p-value < 0.01), while significant overlap (threshold of q-value < 0.05) was performed by overlapping significant LipidSeq signals with significant technical signals(threshold of q-value < 0.05).

A

Tech Type	Total Peak Num	Total Significant Peak Num
SurfaceSeq	20,740	19,908
SurfaceClick	32,298	24,838

B

	SurfaceSeq		SurfaceClick		
Beads	**Seq.Overlap**	**Seq.Sig.Overlap**	**Click.Overlap**	**Click.Sig.Overlap**	**Overlap with Both Techs**
SM	1806	92	2,655	59	107
Cer	1715	69	2,573	31	100
PC	151	37	162	33	6
PE	255	2	290	0	6
PS	1303	40	2,145	32	81
PiP3	1861	103	2,487	44	94
Chol	1525	111	2,208	124	74
SS	1671	61	2,292	26	84
Total	10,287	515	14,812	349	552

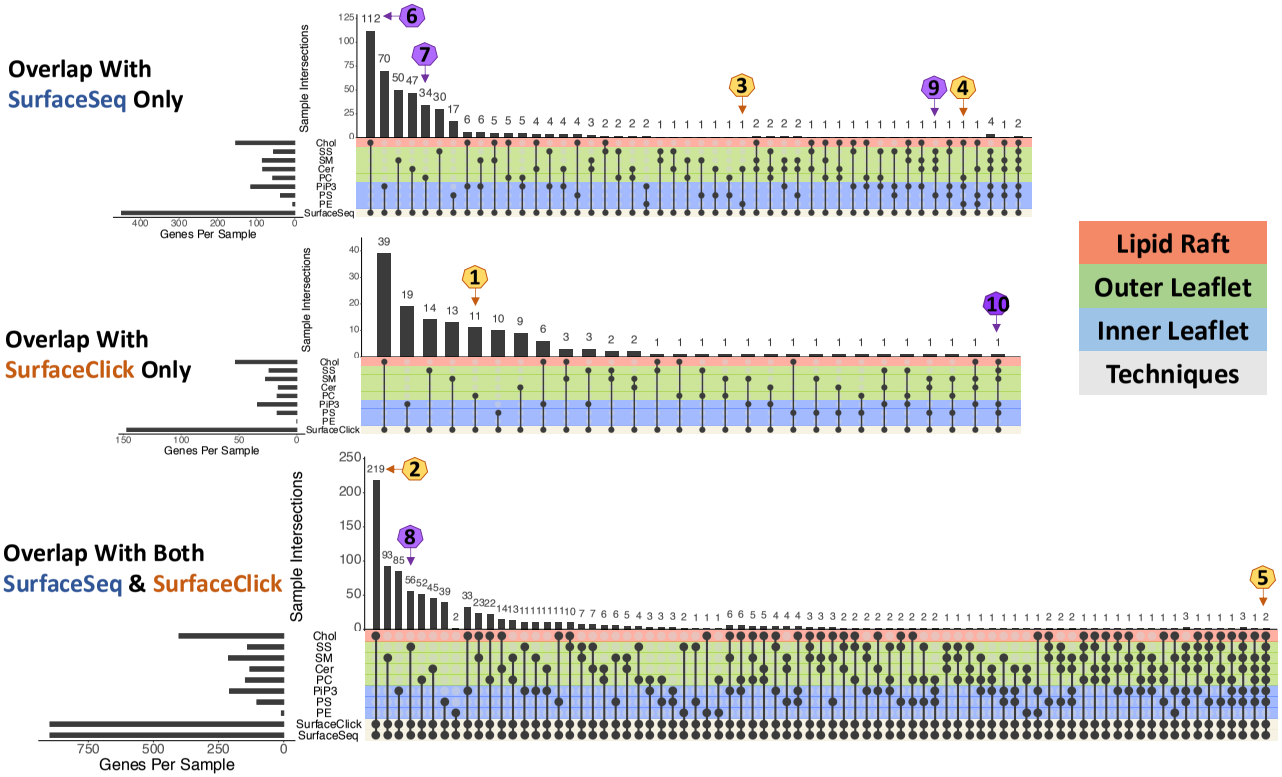


Figure 4: Upset plot for all 8 different types of beads and two types of cell surface sequencing technique, on the gene level. Green background color indicated the bead is specific for outer leaflet. Blue background color indicated the bead is specific for RNAs on the inner leaflet, the red color indicated the bead is specific for RNAs on the lipid raft. There were in total of three upset plots. The first plot showed the status of overlapping between signals and technique signals from **SurfaceSeq** only, the second plot showed the overlapping status of beads with technique **SurfaceClick** only, and the third one showed the overlapping status of beads with both **SurfaceSeq and SurfaceClick**. The upset plot showing the intersection relationships of significant signals, which filtered by applying the threshold of q-value smaller than 0.05, from 8 different types of beads and two orthogonal cell surface sequencing techniques on the gene level. On the plot, ten types of overlapping relationships were marked with tag numbers. 5 examples with yellow tags could be seen in ***Figure 6,*** and 5 more examples with purple tags were showed in ***Supplementary Figure 2.*** The occurrence of signals on different experimental trials in SurfaceClick and SurfaceSeq for these 10 close-up genomic regions could be found in ***Supplementary Figure 3***

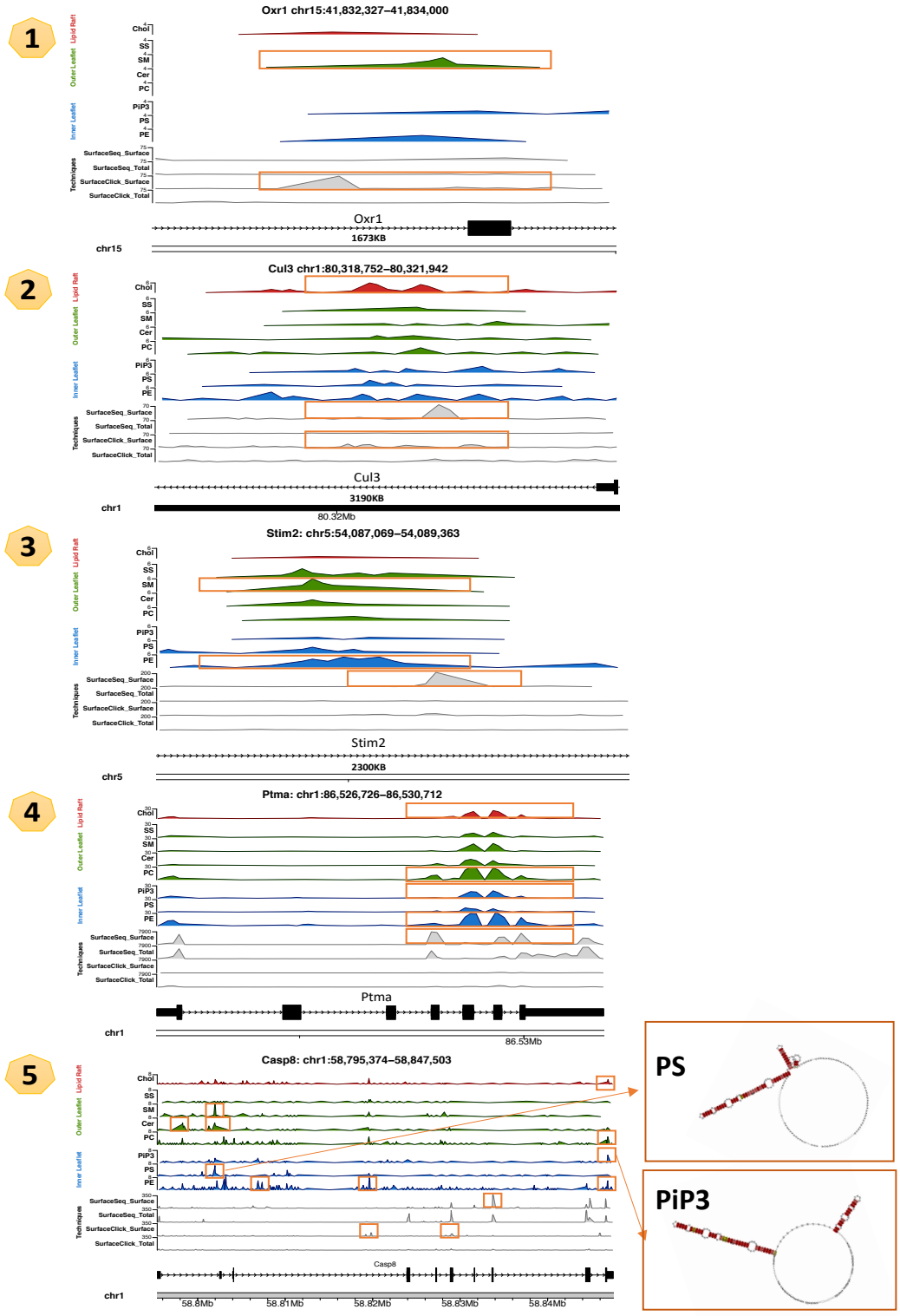


Figure 5: 5 Examples of Genomic View from Upset Plot for all 8 different types of beads and two types of cell surface sequencing technique, on the gene level. Same color scale as in **Figure 5**. Gene name and gene regions were annotated on top of each example. Peak regions were marked by orange rectangles. Example 5 had two errors pointed to matched motif secondary structures, which would be explained in the motif discovery section (**3.5**)

2.5 Motif and RNA Secondary Structure Analysis of csRNA Signals

Due to the fact the significant overlapped regions were generally short segments, motif analysis would be naturally applied on the current dataset to dig more information. By inputting significant signals with SurfaceClick background, SurfaceSeq background, and Control beads background into the program, in total of 489 motifs were discovered, and all of them were aligned to miRNAs/non-coding RNAs. Therefore, discovery of conservative motifs was significant to find similarities among motifs from different types of lipid bead. Similarity of motifs was measured in a pair-wise manner between all combinations of any two lipid beads. From the measurement, low number of similar motifs was observed for PC bead and PE bead, while other types of beads had roughly same levels of similar motifs discovered (***Figure 6B***). This observation could be caused by low amount of initial discovered signals of PC and PE, as we could observed in the previous section (Section 2.4).

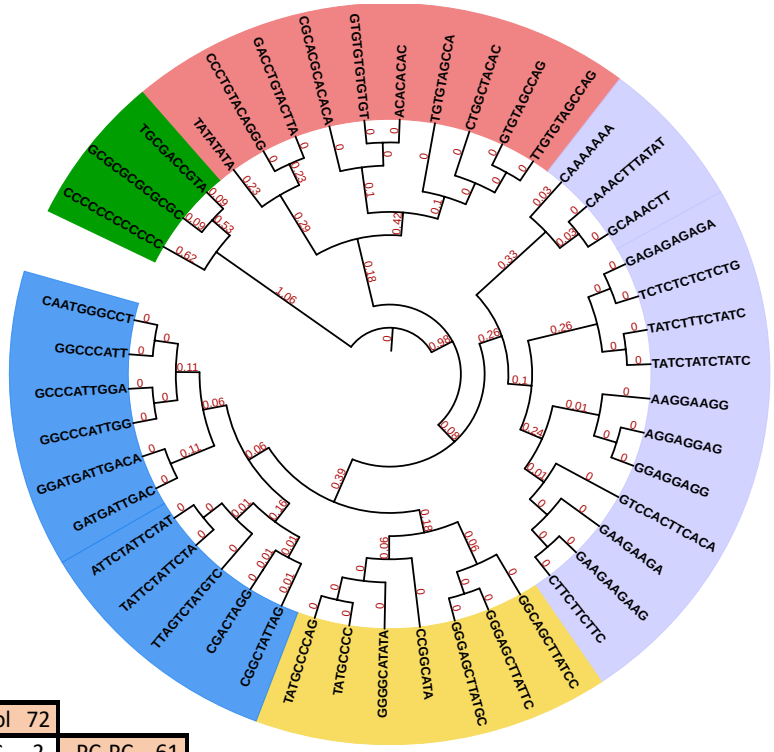
Within those similar motifs, we defined a motif as "conservative motif" if it showed up in more than 4 types of beads as similar motifs. Based on this rule, 45 conservative motifs were labeled. Clustering result of these highly conservative motifs showed they were closely clustered together with high similarity, which indicated high possibility of similar structure and cellular functions. To determine the possible functions of conservative motifs, top 10 conservative motifs were select based on the number of occurrences. Top conservative motifs were further aligned to miRBase[6] for functionality check (**Table 5**), and pre-identified functions related to cancer, inflammation, and immune system could be observed, which match the cellular functions of experimental cell line (EL4 – mouse cancerous immune cell line). This observation also matched the

result of cytotoxicity assay in section 2.1, which proved that csRNAs could act as functional groups to trigger cytotoxicity of natural killer cells as well as support cell immune response. Interestingly, other than directly related to cellular functions, some consensus motifs were annotated as promotor for known cancer related genes (*Malat1*) or structural basis for RNA-recognition site.

To further identify potential conserved structures formed by csRNAs, we performed secondary RNA structures prediction on motifs. Prediction was performed in a clustering-based manner, where similar signals (read segments) were grouped together for structural prediction. From the cluster evaluation, we could observe the pattern of high homogeneity and low completeness i.e. homogeneity score equals 1 and completeness score smaller than 1.17E-18 (***Table 6***). This pattern indicated the clustering itself finished perfectly but the clustering result was not accurate enough, and it might cause by short reads length and low number of input reads. To find the most representative structure, the structure predicted from the largest cluster with most data points of each type of lipid bead was selected. From the predicted secondary structure, we observed high consistency of structures for beads coated with lipids located at inner cell membrane, while the beads coated with lipids located at outer surface of cell membrane had low consistency in general. The next step is to find the co-occurrence of predicted motif secondary structures with previously LipidSeq detected significant signals. Sequences in these representative RNA secondary structures were used on searching RNA homologs in databases across mouse genome. Among the results, predicted structure of PS motif was matched to the differential peak region in exon region of Casp8 (***Figure 5, example 5***) with E-value of 5.8e-05, and predicted structure of PiP3 motif was

matched to the differential peak region at 3' UTR of Casp8 with E-value of 3e-36. This match between predicted secondary structure and lipid beads signals could lead to further exploration on the mechanism of csRNAs' modulation on cellular functions.

A



B

Cer-Cer 70								
Cer-Chol 24	Chol-Chol 72							
Cer-PC 0	Chol-PC 2	PC-PC 61						
Cer-PE 1	Chol-PE 0	PC-PE 0	PE-PE 53					
Cer-PiP3 15	Chol-PiP3 19	PC-PiP3 0	PE-PiP3 1	PiP3-PiP3 59				
Cer-PS 15	Chol-PS 27	PC-PS 1	PE-PS 0	PiP3-PS 19	PS-PS 55			
Cer-SS 9	Chol-SM 29	PC-SM 0	PE-SM 1	PiP3-SM 15	PS-SM 17	SM-SM 56		
Cer-SM 17	Chol-SS 17	PC-SS 1	PE-SS 1	PiP3-SS 11	PS-SS 18	SM-SS 2	SS-SS 63	

Figure 6: Result of Motif Discovery, Alignment, and Clustering. A. Clustering Tree of Conservative Motifs Discovered. Clustering tree was plotted in circular manner, and the length of tree branch is unproportionally to the distance marked in red. Different color block indicated different meta-branches that separated at the second deviation from the root. **B. Table for showing number of Conservative Motifs in Pair-wise Comparison.** This table showed the number of conservative motifs found in pair-wise comparison for every bead (order did not matter). The number on the diagonal was the number of motifs discovered for that type of be

Table 5: Motif Function Result. This table showed the result of pre-discovered function of the conservative motifs discovered in SurfaceLipid technique. Information included the original motif sequence, annotation from *miRBase*, confidence level from *miRBase*, and pre-discovered functions found in literatures.

Motif	Annotation	Confidence Level	Functions
GGCAGCTTATCC	hsa-miR-22	high confidence	Overexpression of miR-22 is neuroprotective via general anti-apoptotix effects
			May target specific Huntington's disease-related mechanisms
			Long non-coding RNA MALAT1 acts as a competing endogenous RNA to promote malignant melanoma growth and metastasis by sponging miR-22
			Inhibits tumor growth and metastasis in gastric cancer by directly targeting MMP14 and Snail
			Inhibits hepatocellular carcinoma cell migration and invasion via targeting CD147
GTCCACTTCACA	hsa-miR-3616-5p	high confidence	Identified as functional candidate microRNA seed sites in GWAS
GAAGAA	hsa-miR-4778	high confidence	Showed in GNRA tetraloop
			Structural basis for the dual RNA-recognition modes of human Tra2-β RRM
GAGAGA	hsa-miR-1281	not enough data	A positive regulator of global gene expression. Regulation of neutrophil senescence
			Showed in microRNA profile specific to cancer stem-like cells directly isolated from human larynx cancer specimens
CTTCTT	hsa-miR-1253	not enough data	Possible miiRNAs in lung cancer
			Potential antisense regulators in the archaron Sulfolobus solfataricus
			T-rich terminator-like sequence
ACACAC	hsa-miR-4455	not enough data	Potential Predictive Biomarker for Subclinical Hypothyroidism
			MicroRNA of CD8+ T Cells in Acute and Chronic Brucellosis
GTGTGT	hsa-miR-466	not enough data	Important miRNAs under Skin of Human and Mouse
			microRNAs in the Lymphatic Endothelium
AAGGAAGG	has-miR-4297	not enough data	microRNAs in human embryonic stem cells and neural precursors

Tale 6: Statistics pf Secondary Structure Discovery: This table showed the evaluation of clustering result and secondary structure prediction by using *GraphClust*. Table included information of Total Seq- the total number of input reads for each type of bead, Cluster Num-the number of final clusters from the clustering algorithm, completeness score and homogeneity score of final evaluation of the whole clustering, Largest Cluster-the tag number of the largest cluster for the specific type of beads, and cluster reads percent, which was the percentage of reads in comparison of all for the reads in the largest cluster.

Name	Total Seq	Average Seq Length	Total Seq Used	Cluster Num	completeness_score	homogeneity_score	Largest Cluster	Cluster Read	Cluster Reads Percent
SM	585	854	573	64	-1.17E-17	1	13	124	21.64%
Cer	412	849	395	47	1.48E-18	1	8	37	9.37%
PC	302	754	141	31	-1.27E-17	1	22	7	4.96%
PE	19	777	14	3	4.38E-17	1	3	7	50%
PS	331	866	328	43	7.94E-18	1	12	36	10.97%
PiP3	541	835	424	57	9.68E-18	1	2	65	15.33%
Chol	980	829	905	91	-3.89E-18	1	11	111	12.27%
SS	370	853	282	39	9.66E-18	1	5	44	15.60%

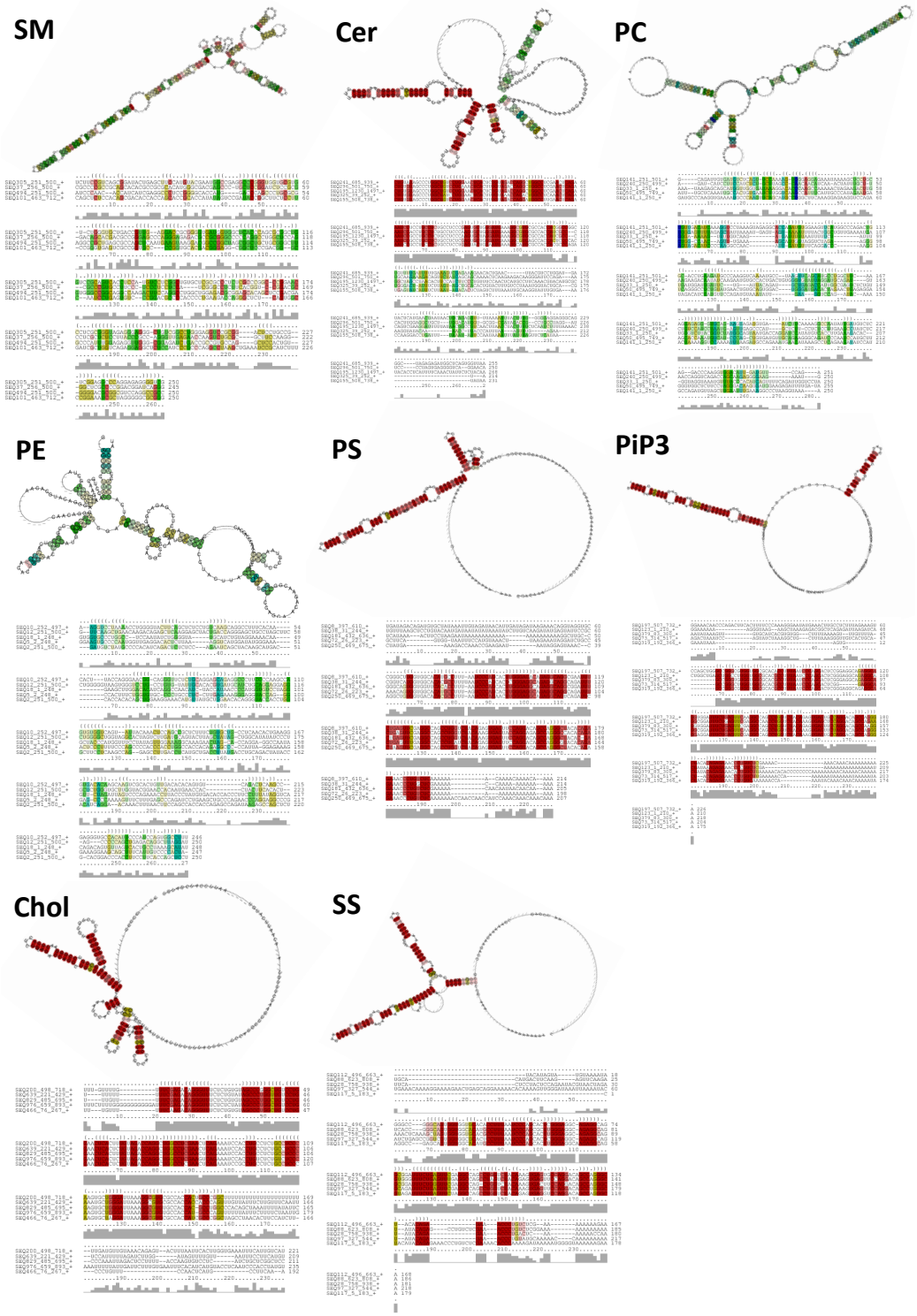


Figure 7: Representative Motif Structure and Alignment from the Largest Cluster for Each Type of Beads. Structures were on the top and alignments were at the bottom. The colored output was generated by aligning tRNA-aln4 from the *BRAliBase*[7] benchmark using *R-Coffee*[8] slow/accurate mode. Colors indicate the consistency of aligned residues with the primary library alignments and the predicted structures: blue to green means low consistency; yellow to red means good consistency. The dot bracket notation below the alignment indicates the consensus structure (predicted with *RNAalifold*[9]) and was added afterwards.

Discussion

3.1 Existence of csRNA and Success LipidSeq csRNA Pull-down

Previously, the existence of membrane-binding RNAs was proved *in vitro* on hydrophobic mica surfaces using fluorescence microscopy[1]. Also, bacteria bglG mRNA was shown to form pre-complex near membrane when co-transcribed with its membrane sensor[10], and mature human tRNAs were discovered to specifically retained in HeLa membrane as binding to liposomes[11]. More recently, interaction between the *LinkA* long noncoding RNA and PIP3 (Phosphatidylinositol (3,4,5)-trisphosphate) at the inner leaflet of the plasma membrane was proved[1–3]. However, specific functions and structures of those membrane-related RNAs were still remain unknown due to intrinsic and extrinsic noise of cell dynamics[12].

Preliminary work in Zhong Lab has successfully demonstrated the existence of csRNAs on the surface of intact cells using imaging technique. To further examine the potential csRNA signals, LipidSeq was developed based on pre-identified affinity between RNAs and lipid structures[1–3]. Potential csRNA candidates were successfully pulled down using LipidSeq with all 8 different types of lipid beads, followed by solid signal amplifications. Signals detected by LipidSeq showed high consistency on the distribution of signal amplifications, and signals were successfully validated using other two techniques.

Nevertheless, LipidSeq did have some limitations in the aspect of technical background noise. Due to the intrinsic sticky property of the beads, even control bead, which were not coated with lipid, could pull down certain amount of RNAs. These technical background noises were largely removed in the data processing steps by using

quality control, applying stringent thresholds, and performing signal validation. Thus, the legitimacy of detected signals could be ensured.

Although LipidSeq has some limitations, it could successfully detect and pull down csRNAs on cell membranes using lipid beads affinity. With validation of the detected signals, the existence of csRNA could be confirmed using LipidSeq.

3.2 csRNAs Could Act as Recognition Motifs to Modulate Cell Functions

Based on the result of cytotoxicity functional assay, the global perturbation of csRNA resulted in a huge decrease of NK92 cytotoxicity functions, which pointed to the presence of csRNAs from another perspective and inferred their involvement in killer cells' major cellular function. Recently, RNAs have emerged as a target of pattern recognition receptors that drive activation of innate immunity[13]. Indeed, the preliminary analysis on SurfaceClick pull down from NK92 and EL4 both showed that the immune response appeared was one of the major pathways enriched in cell csRNA gene ontology analysis[14]. Different from the signals detected by SurfaceClick and SurfaceSeq, which were mostly mRNA signals, those detected by LipidSeq were comparable short segments with enrichment on 3' and 5' UTR regions. Based on this difference, motif-based analysis, instead of normal gene-based analysis, was performed on LipidSeq detected signals. Intermolecular RNA-RNA interactions are used by many RNAs, especially noncoding RNAs to achieve or modulate diverse cellular functions[15]. Interestingly, most of the LipidSeq detected motifs aligned to microRNAs, which have been validated to play vital roles in cancers development and self-immunity activation by targeting 3'UTC regions of downstream gene mRNAs[16]. Within those motifs, a highly conservative motif

"GGCAGCTTATCC" was aligned to miR22, which is a micro RNA that have been proved to fluctuate with cancer progression in body fluid to regulate cancer growth and trigger apoptosis as an immune response. Additionally, miR22 was discovered to have similar function with *Malat1*, which was one of the most famous lncRNA involved in the regulation of cancer growth[17]. The presence of *Malat1* was also detected by using SurfaceClick, SurfaceSeq, and Surface-FISH on EL4 cell lines[14,18].

From those observations, we could conclude that csRNAs has conserved motifs with critical functions in the innate immunity and cancer related functions. One step further, we can make the assumption that csRNAs can act as pattern recognition receptors that activates the cytotoxicity of natural killer cells, was well as can extend the possibility of csRNAs' vital role in regulating cellular functions for other types of cells.

3.3 csRNAs Could Act as Structural Group to Modulate Cell Functions

To identify potential structural units formed by csRNAs, motif and secondary RNA structure analysis were applied. Motif result showed that frequent poly-G containing motifs, which previously reported as may directly interact with lipid bilayers[19], were missing from current dataset. However, long G tracts are not essential for RNA's bilayer affinity, and might varied based on different cell types[20]. Thus, motifs discovered here could provide more possibilities of structures with affinity to cell membrane.

Several predicted secondary structures were able to be detected in significant differential peak regions. For example, most representative secondary structure from RNAs pulled by PiP3 was aligned to 3'UTR of Casp8. Casp8 was a gene that promotes cell migration, cell adhesion, and Rac activation by generating lipid products (PIP2 and PIP3) in normal and tumor cell lines[21]. Since the differential peaks of Casp8 were captured using PiP3 beads, the possibilities of RNA-RNA interactions as well as csRNA modulating cell functions as a structural group could be extended.

From the predicted secondary structures, the pattern of low consistency on the outer leaflet of the membrane while high consistency on the inner leaflet and lipid raft could be observed, though it was unclear what caused this pattern. One possibility was the variety of RNAs on the cell surface enabled more interactions. One thing worth noticing is that the topological consistency of predicted secondary structure is not strongly correlated to the base-pairing consistency. Besides difference on consistency, the potential functionality of many structures remained unknown. One possible function for those RNAs was forming signal recognition particle (SRP) complex on membrane as

binding site[10] or interact with liposomes as RNA structures[11]. Other possible functions for those RNAs structures that already been discovered including changing the permeability of cell membrane[20], stabilizing temporary pore formation[22], or even changing the physiological pH and the charge of cell membrane[23].

To sum up, the results provided possibilities of RNA secondary structures and provided a different point of view in future exploration of RNA as a structural functional group.

Conclusion

This study, using combination of fluorescence label, cytotoxicity functional assay, and second-generation sequencing data, suggests the existence of csRNAs as a class of functional and structural molecules on cell's outer plasma membrane. Preliminary analysis showed existence of cs RNA by using fluorescence dye on EL4 and NK92 cell lines. Due to the innate property of natural killer cells, cytotoxicity functionality examination was performed on NK92 by measuring LDH release. Result showed that the removal of csRNAs could largely impact natural killer cell's immune response, which firmly connected csRNA with cellular functions. Using 8 different types of lipid beads, which previously proved to have high affinity to RNA molecules, candidates csRNAs on outer, inner, and lipid raft were pull downed from intact cell surface of EL4 and successfully constructed sequenceable libraries. Besides self-selection for valid signals between lipid beads and control, this study utilized two newly developed csRNA sequencing techniques, SurfaceSeq and SurfaceClick, to perform inter and intra validation for further csRNA signals validation. Function alignment on highly conservative motifs showed enriched immune and cancer related functions. Combining these results to the preliminary imaging of csRNAs and cytotoxicity assay, we proved the existence of csRNAs on cell outer membrane as well as csRNAs' potential to act as regulatory factors for cellular functions. Furthermore, based on the result of secondary RNA structures prediction and overlapping between predicted structures with LipidSeq detected signals, we largely extended the possibility that csRNAs could have extensive functions such as acting as recognition site and providing structural support.

To sum up, investigating csRNAs on outer cell membrane would provide insights on regulation of cellular functions associated with RNA metabolism and structure, as well as its correlation with the membrane structure complexity. More importantly, these novel discoveries would help people to explore the field of RNA and understanding cell metabolism from different angels.

Material and Method

4.1 Cytotoxicity Assay: Functional examination of csRNAs

To test the effect of surface RNA to NK cells' cytotoxicity, NK cells (a.k.a. effector cells) were mixed with MDA-MB-231 cells (a.k.a. target cells), and NK cells' cytotoxicity was quantified by LDH assay[24]. In this assay, cytotoxicity is quantified as the ratio of the amount of NK-killed target cells to the amount of lysis-solution killed target cells[24]. Four independent experiments were carried as separated biological replicates (indexed by k, k = 1, 2, 3, 4). In each experiment, cytotoxicity of RNase treated NK cells (treatment group) were compared with cytotoxicity of untreated NK cells (control group). In the treatment group, NK cells were incubated with $4\mu l$ of RNase for 10 minutes, which partially removes surface RNA. In each experiment, and in both the treatment and control group, NK cells and MDA-MB-231 cells were mixed in two different E:T ratios (effector cell: target cell ratio), 0.16:1 and 0.31:1 (indexed by j, j = 1 (0.16:1), 2 (0.31:1)). Higher the previous number, more NK92 cells were put into the well. One individual 96-well-plate was prepared or each experiment. In each individual experiment, 12 technical replicates of LDH releases were measured under each RNase treatment and E:T ratios. Each technical replicate corresponds to one LDH scan, which gives readings from 3 or 4 wells. These 3 or 4 readings were averaged into 1 measurement as the reading of this technical replicate. Taken together, 192 ($12 \times i \times j \times k$) measurements of treatment group were obtained, which include 12 technical replicates for two conditions of RNase treatment with two different E:T ratios in 4 biological replicates. See below for the Equation for calculating cytotoxicity:

$$CytoToxicity$$

$$= \frac{[Death\ in\ the\ mixture] - [Effector\ spontaneous\ deaths] - [Target\ spontaneous\ deaths]}{[Target\ maximum\ deaths] - [Target\ spontaneous\ deaths]}$$

- Deaths in mixture: LDH concentration from both effector and target cells when mixed
- Effector spontaneous deaths: LDH concentration from effector cells only
- Target spontaneous deaths: LDH concentration from target cells only
- Target maximum deaths: Maximum amount of LDH concentration from target cells. 10µl of the Lysis Solution would be added to the well. This will result in complete lysis of target cells.

After calculated the cytotoxicity by using equation of $Y_{ijk} = \frac{Exp_{ijk} - E\_sp_{ijk} - T\_sp_k}{T\_max_k - T\_sp_k}$

with pre-selection of qualified data from the 756 measurements, two-way ANOVA test was used to measure either the change of RNase concentration (number of RNAs left on cell membrane) or change of E:T ratio (the number of NK92 and MDA-MB-231 cells put into wells) changed natural killer cell's cytotoxicity. Check ***Schema 1*** for specific explanation of the statistical model of two-way ANOVA.

4.2 LipidSeq Pull-down: Data processing and pre-selection of valid signals

To detected cell surface RNA signals, Zhong Lab developed an unbiased method using lipid-coated beads followed by RNA-sequencing to identify RNA that directly bind to lipids. Lipid-coated beads would be selected based on their chemical characteristics, their presence in specific leaflets of the plasma membrane, and their known role in the bilayer structure. In the experimental protocol, RNA starting material was isolated from the membrane fraction of EL4 cells. Nine different lipid-coated beads, with one control bead (not coated with any lipids) and 8 different types of beads that specifically bind to RNAs at outer leaflet, inner leaflet, and lipid rafts of the cell separately were selected. After pulling-down of the lipid-binding RNA species, cDNA libraries were generated and sequenced to identify the affinity of RNA molecules for each type of lipid beads in comparison with their affinity to control beads for the purpose of deciphering the structure-functions of membrane RNA (**Figure 8A**). Three technical replicates of each RNA type were generated simultaneously and sequenced, as well as water samples to evaluate specificity of amplification, since the starting quantities are very little (Did not showed up in the sample table). For every bead, in total of 6 libraries were generated from three biological replicates, and each biological replicate has two technical replicates (3*2=6) ***(Supplementary Table 2)***. In these three batches of experiments, batch three (biological replicate 3) were sequenced without equal-molar preparation. Thus, reads number of two technical replicated in batch 3 varied more than other two batches.

Raw sequencing files were processed by a self-build data processing pipeline (***Figure 8B***). Before alignment, adaptors were trimmed from raw reads using *Trimmomatic*[25], and quality control of reads was done by using *FastQC*[26]. Pre-processed

RNA sequencing data was aligned to mouse genome (mm10) using *STAR*[27], and two post-alignment data processing steps were be performed on the aligned bam files: selection of uniquely mapped reads and ribosomal RNA removal. To make sure the quality of data for downstream analysis, unique mapped reads were selected by using *STAR* flag of 255. Also, based on preliminary analysis, every sample had at least 20 percent of ribosomal RNAs (rRNA). Since experimental protocol was designed to capture csRNA signals, rRNAs were treated as contamination of signals from inside of cells. Thus, a step of rRNA removal was significant for getting clear signals for downstream analysis. Here, rRNAs were depleted bioinformatically by removing the overlapped regions between input reads and annotated repeat masker regions for ribosomal RNAs that downloaded from *UCSC genome browser*[28] using *BEDTools*[29]. Batch effect was the subsequent issue that need to be solved since three replicates that collected under different time stamps were involved in analysis, and we need to determine whether the differential signals were truly caused by experimental conditions. Therefore, batch effect was neutralized by merging all biological and technical replicates to one sample per bead. Final sample set included in total of 9 samples, which included 8 libraries for 8 different lipid beads and 1 library for control bead, each with around 1M reads (***Table 7***).

Table 7. Sample Information after Data Processing and Pre-selection: Table of reads information after processed by data pipeline and merging of batches. Table includes bead types, specific binding positions for the bead, total reads after mapping, total amount of ribosomal RNA each sample, and percentage of non-rRNA reads left after rRNA removal.

Binding Position	Beads	Total Reads	rRNA removed reads	% of non-rRNA reads
Lipid raft	Chol	4,670,130	1,025,751	21.96%
Outer leaflet	SS	3,991,879	768,822	19.26%
Outer leaflet	SM	4,604,513	1,060,729	23.04%
Outer leaflet	Cer	4,526,676	932,249	20.59%
Outer leaflet	PC	5,093,154	1,294,594	25.42%
Inner leaflet	PiP3	4,445,192	821,222	18.47%
Inner leaflet	PS	4,347,357	912,399	20.99%
Inner leaflet	PE	5,546,178	1,158,795	20.89%
Control	CT	5,084,689	1,131,073	22.24%

4.3 Signal Detection: Finding and selection of differential peaks

MACS2[30] was used to detect the significant differential peaks between every type of bead with control bead (CT) (***Figure 8B***). The peak searching process will be determined by tag size, min-length, max-gap, and fragment size. In order to keep the tag size flexible, the tag size was determined by the program using first 10 sequences from the input bead signal file, which were all within the range of 60bps-70bps. To fine-tune the peak calling behavior of *MACS2*, minimum length of a called peak and the maximum allowed a gap between two nearby regions to be merged were specified by using the program predicted fragment size d. In our cases, the fragment size for all samples were all predicted to be around 300bps. Moreover, *MACS2* will calculate p-value and q-value, which is adjusted p-value by FDR method, for each peak based on the whole distribution automatically. Based on the property of libraries that generated by using lipid beads pulldown, even with all previous steps of data processing, samples still contained extend amount of background noises. Thus, using the default cutoff of q-value smaller than 0.05 would be too stringent to get enough valid peaks from the program itself. To loosen the

filtering condition, p-value threshold of 0.01 was specified to override the default setting for valid peak selection. After got preliminary peaks from the program, an additional condition of q-value threshold of 0.05 would be applied on the peak summit for further narrowing down the range to eliminate false positive signal. To further check the pattern of distribution for these peaks, average read region distribution plot was calculated and plotted using *computMatrix* in package *deepTools*[31] with parameters of afterRegionStart length equals to 10, binSize equals to 10, and regionBodLength equals to 100.

4.4 Pre-validation step: Cross-validation of two Newly Developed Techniques, SurfaceClick and SurfaceSeq

Since the lipid bead samples contained certain amount of background noise due to its natural affinity to RNA molecules, it was hard to completely eliminate these noises from experiment protocol or bioinformatically. Thus, further validation from outside source was significant for getting true signals from the chaos. To detect and capture csRNAs signals on the outer membrane of cells, two orthogonal technologies were developed and applicated by Zhong lab: SurfaceSeq and SurfaceClick. SurfaceSeq is a drug delivery system that utilizes biodegradable polymeric nanoparticles (poly-lactic-co-glycolic acid)[5] that are fused with the cell membrane producing a membrane-coated nanoparticle to avoid the immune system. This technology is crucial because not only isolate the cell membrane, but it also minimizes contamination of intracellular and extracellular RNAs producing high yield cs RNAs knowing that RNAs are prone to degradation. SurfaceClick is a technique that csRNAs are labeled on intact cells and a

subset of cells are imaged to control for the CLICK reaction. Total RNA is further isolated, fragmented and subjected to streptavidin beads, where only the biotinylated csRNAs will be pulled down. Stringent urea and high salt-based washes were used to remove non-specific binding of non-biotinylated RNA[15]. These two lab-developed technologies could be essential source of further validation. Nevertheless, before utilizing information gathered from SurfaceSeq and SurfaceClick, an internal validation for csRNA signals captured by these two new techniques is necessary.

csRNAs pull down by using these two techniques with correspond background control were performed on the mouse EL4 cell line (**Supplementary Table 3**). In total of 5 SurfaceSeq libraries, includes 3 surface RNA samples and 2 total RNA samples, and 6 SurfaceClick libraries, include 3 surface RNA samples and 3 total RNA samples were used in this cross-tech validation. Raw data were processed using the same pipeline in pre-proccing and mapping step **(*Figure 8B*)**. After that, *featureCounts*[32] was used to get the count of RNA on gene level for each sample individually, and differential expression analysis was performed using *DESeq2*[33] to compare gene counts between surface samples and total samples. From the result of differential expression analysis, log2FC of the gene expression level changed from total RNAs to csRNAs and its corresponding q-value, which was an adjusted p-value by using FDR method, were collected for 46,191 annotated genes in mouse genome. Some of the genes were captured only in the surface RNA sample but not in total RNA sample, therefore; it was not possible to calculate the log2FC for those genes. After filtration of those genes, we got 24420 genes for SurfaceClick and 24421 genes for SurfaceSeq with valid log2FC value. Top differentially expressed gene, which were genes with positive log2FC and small q-value, were selected

as candidate cell surface signal for the method. To further valid the signal, cross-validation between these two orthogonal methods was performed to test whether top differentially expressed genes from one assay is differentially expressed in the other assay. Top differentially expressed genes with threshold of adjusted p-value smaller than 0.05 and log2FC larger than 2 from both techniques would be select. Distribution as well as the size of intersection with p-value from hyper-geometric distribution were used to measure the level of similarity between two candidate gene set.

4.5 Signal validation: Signal overlapping among LipidSeq, SurfaceClick, and SurfaceSeq

After performed cross-validation to justify csRNA signals that detected by two newly developed csRNA sequencing techniques by Zhong lab, SurfaceSeq and SurfaceClick, these two techniques were used to validated potential surface RNAs that pulled by using lipid beads. In order to make the comparison on the same row, differential peaks between surface RNA samples and total RNA samples need to be detected using the same method, MACS2[30], with the same parameters. For finding the differential peak region between surface samples and total samples of SurfaceSeq and SurfaceClick, technical replicates need to be merged into one. After merging, four samples were used in downstream analysis: SurfaceSeq-surface sample, SurfaceSeq-Total sample, SurfaceClick-surface sample, and SurfaceClick-total sample ***(Supplementary Table 3)***.

Ultimately, the RNA-binding lipids candidates will be compared to the cell-surface RNA candidates identified using SurfaceSeq and SurfaceClick in EL4 cells in order to narrow down the localization and structure of csRNA molecules at the surface of plasma membrane. Peak overlapping between surface lipid peaks (peak II) and two orthogonal methods (Peak I for surface-CLICK-seq and Peak III for nanoparticle-seq) was done by using *BEDTools interest*[29] for finding the complete intersect regions longer than 50bps. Calculated overlapped regions would be further verify by genomic region visualization. All the valid regions would be used in the motif analysis for providing more specific information.

4.6 Motif Analysis Process: Functional analysis and secondary structure prediction

Due to the fact the significant overlapped regions were generally short segments, motif analysis would be naturally the next step. Motif analysis was performed by using *Homer "findMotifGenome"*[34] with parameter of *-rna* and background of control beads to find motif with length of 8bps, 10bps, and 12 bps for each type of lipid bead separately. Each Homer motif finding returned top 50 motifs with highest confidence. Program itself would mark the motifs with high possibility being false positive signal, and those motifs were filtered to raise accuracy.

Next step was to find conservative motifs across different types of beads. By using *biopython-motif alignment*[35], distance and offset were measured for each pair of motifs. Here, distance was measured by 1-Pearson correlation of count matrix, where smaller the distance, and offset is the shift distance between two motifs in unit of base pair, smaller the number, longer overlapping segments between two motifs. Conservative motifs across different types of beads were selected by using the threshold of distance smaller than 0.2 and offset smaller or equal to 2. To further discuss the properties of discovered conservative motifs, candidates were aligned to RNABase[36] for function check. Besides function, secondary structure of motif could be interesting to discover too. Pre-build data pipeline *GraphClust*[37] was used to explore possible secondary structure of motifs (***Supplementary Figure 4)***. Since there number of signals that had been verified by both techniques was limited, significant signals were used as input of the pipeline. Input for *Graphclust* were differential peaks regions in fasta format and secondary structures were predicted based on clusters, the largest cluster with most data points from each bead was selected to represent the most possible secondary structure. The RNA

secondary model was predicted and calibrated using CMfinder[5] and CMsearch[6]. Due to the fact there were limited number of differential peaks that were validated by both techniques, significant differential peaks between every type of bead and control bead were used as input to avoid the inaccuracy that caused by small input.

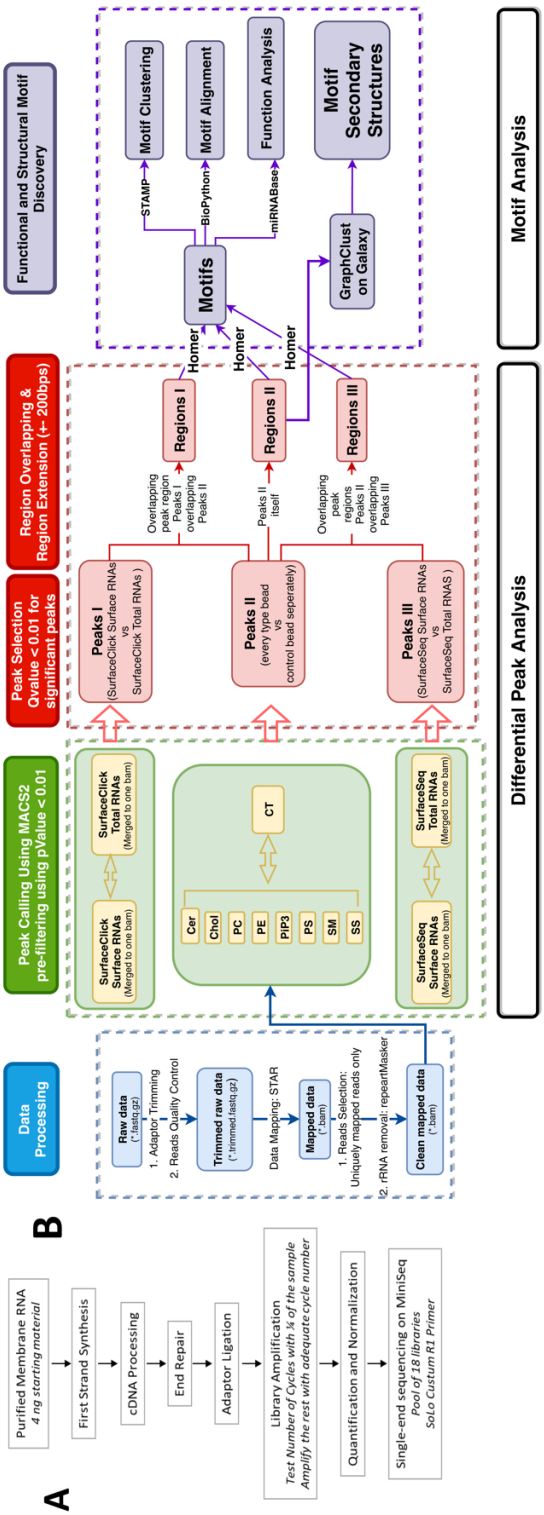


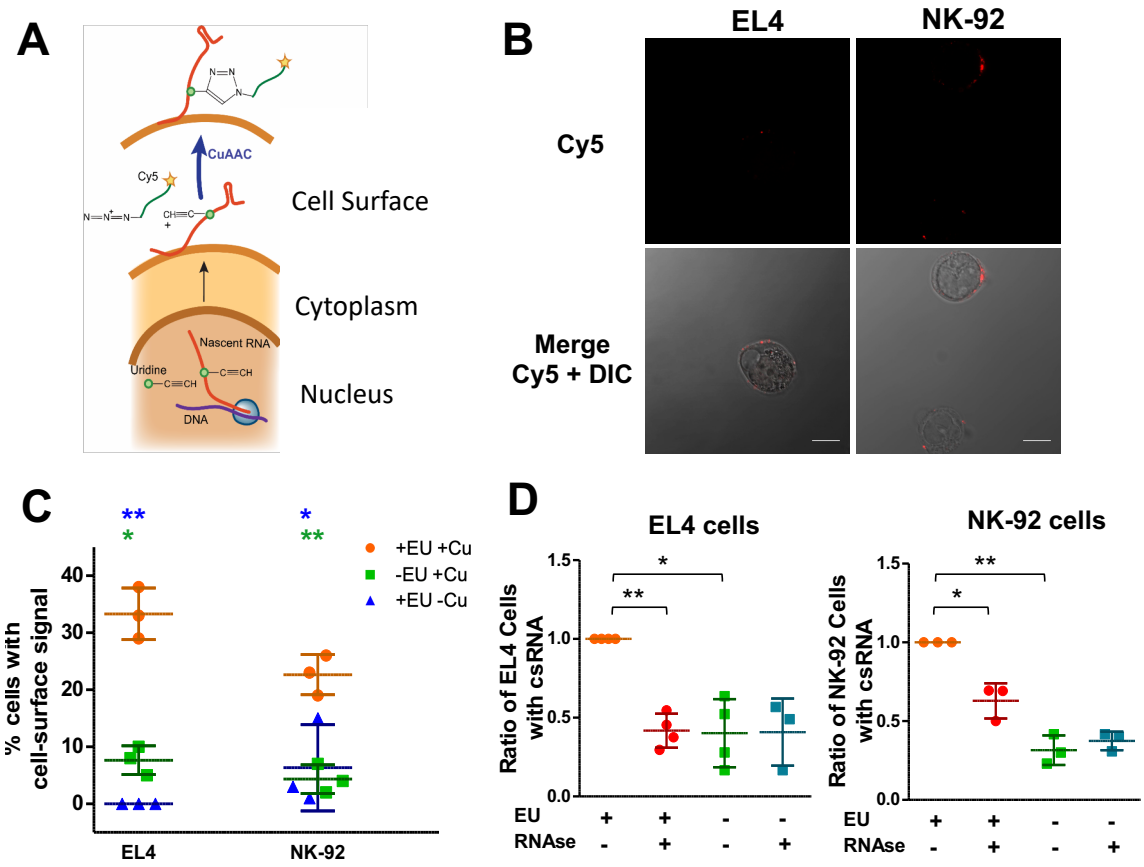
Figure 8: Experiment Protocal and Data Processing Pipeline: A. Workflow of generating surface lipid libraries using the Ovation® SoLo RNA-Seq kit from Nugen. **B**. Flow Chart for Data Processing and Peak Analysis Pipeline, which includes the standard data processing steps, peak detection and selection processing steps, peak region overlapping processing steps, and motif analysis processing steps based on differential peak regions.

This thesis is coauthored with Lucie Hebert, Kathia Zaleta Rivera, Xiaochen Fan, and Norman Huang. The thesis author, Xuerui Huang, was the primary author of this thesis.

Supplementary Materials

Supplementary Table 1: Characteristics of cell-surface RNA positive cell lines

Cell line Name	Cell type/Disease	Organism	Tissue of Origin	Cell morphology	Age of donor	Gender	% cells with csRNA
EL4	Lymphoma	Mouse	T lymphocyte	lymphoblast	Unknown	Unknown	29%-38%
NK-92	Non-Hodgkin's lymphoma	Human	Natural killer cell	lymphoblast	50y	Male	19%-26%



Supplementary Figure 1: Previous Result of Identification of mammalian cell lines displaying csRNA using Surface-CLICK technology: A. Workflow of the cell-surface CLICK technology. **B.** Representative images of the 2 cell lines showing cell-surface signal after CLICK reaction. Upper panels show signal form cy5 channel, lower panels show a merge image of DIC channel and cy5 channel. Scale bars: 10um. **C.** Graph representing the percentage of cells exhibiting cell-surface signal. Blue stars show significance difference between (+EU +Cu) and (+EU –Cu) for each cell line. Green stars show significance difference between (+EU +Cu) and (-EU +Cu) for each cell line. **D.** Graphs representing the percentage of cells presenting with cell-surface signal after RNase treatment in EL4 and NK-92 cell lines, normalized on the (-EU +Cu) condition. Error bars represent the standard deviation of three independent experiments. Statistic indicators: *p<0.05, **p<0.01, ***p<0.001.

Explanation: Mathematical Model of Toxicity Formula

Y: Toxicity, a measurement of the killing potential of NK-92 cells, calculated by

$$Y_{ijk} = \frac{Exp_{ijk} - E\_sp_{ijk} - T\_sp_k}{T\_max_k - T\_sp_k}$$

i: RNase treatment of NK-92 cells. In total of two different treatments

i	*RNA treatment*	*RNase μl (dilution)*	*Val PBS (μl)*
0	No RNase	0	1000
1	RNase A+T1 1:250	4 ul (1:250)	996

j: E:T ratio, which measures the ration between number of effector cells (NK-92) and number of target cells (MDA-MB-231). In total of four different ratios

j	*E:T ratio*	*Condition(E:T/E)*
1	0.31:1	$1.6 \times 10^3 : 5.1 \times 10^3$
2	0.16:1	$8.0 \times 10^2 : 5.1 \times 10^3$
3	0.31:0	1.6×10^3
4	0.16:0	8.0×10^2

k: Different trials, each trial is an individual 96-well plate. In total of four different trials

k	Plate Number	Experiment Date(D/M/Y)
1	7	8/21/2018
2	8	8/21/2018
3	10	8/28/2018
4	11	9/1/2018

$Exp_{ijk,j=1,2}$: The collection of data that contains all the measurements of absorbance at 490nm of plate k in the i, j condition by using platereader. The absorbance measures concentration of LDH (lactate dehydrogenase), which is proportional to the number of lysed cells.

$E\_sp_{ijk,j=3,4}$: The collection of data that contains all the measurements of absorbance at 490nm of plate of plate k in the i, j condition by using platereader. The absorbance measures concentration of LDH, which is proportional to the number of naturally lysed NK-92 cells.

- There is a relationship between the j in Exp and j in E_sp. Correlation is presented in the table below

Exp	E_sp
1	3
2	4

T_sp_k: The mean value of all the measurements of absorbance at 490nm with 5.1×10^3 target cells (MDA-MB-231) of plate k by using platereader. The absorbance measures concentration of LDH, which is proportional to the number of naturally lysed MDA-MB-231 cells.

T_max_k: The mean value of all the measurements of absorbance at 490nm of plate k by using platereader with 5.1×10^3 target cells and $10\mu l$ lysis solution. The adding of lysis solution will result in complete lysis of target cells.

ANOVA Model Explanation:

- $Y_{ijk} = \mu + bT_i + cB_j + \varepsilon_{ijk}$
- In this model, 1 trial of Y_{ijk} will be calculated

Y_{ijk}: The observation for which $X_1 = i$ and $X_2 = j$ in trial k

μ: The general location factor

X_1: The primary factor in the linear model

X_2: The blocking factor in the linear model

T_i: The effect for being in treatment i (of factor X_1)

B_j: The effect for being in block j (of factor X_2)

ε_{ijk}: random error under condition ij in trial k

H_0: The means of the measurement variable for each group are equal, $b = 0$. Y_{ij} is independent of T_i (Cyto-toxicity independent from different RNase treatment)

H_1: The means of the measurement variable for each group are unequal, $b \neq 0$. Y_{ij} is dependent of T_i (Cyto-toxicity related to different RNase treatment)

Schema 1. Equation for the Measurement of Cytotoxicity and ANOVA Model Explanation

Supplementary Table 2: Library and mapping information. This table showed the sequencing and mapping information for each technical and biological replicate for every bead. There are in total of 9 types of beads, includes one control bead (no lipid affinity) and 8 beads with different affinity. Different color indicated specific binding position, which green background color indicated the bead is specific for outer leaflet. Blue background color indicated the bead is specific for RNAs on the inner leaflet, the red color indicated the bead is specific for RNAs on the lipid raft, and the orange color indicated the bead is specific for RNAs on the outer leaflet and lipid raft.

Sample Name	Lipid Name	Lipid Localization	Biological Rep	Tech Rep	M Seqs	M Aligned	% Aligned
CT-1_1	Control No lipid	NA	1	1	2.2	1.1	52.40%
CT-1_2			1	2	1.7	0.9	53.80%
CT-2_1			2	1	2	1.2	61.80%
CT-2_2			2	2	2.1	1.3	61.30%
CT-3_1			3	1	1.4	0.8	55.40%
CT_3_2			3	2	2.9	1.6	54.50%
SM-1_1	Sphingo-myelins	Outer leaflet	1	1	1.8	0.9	53.40%
SM-1_2			1	2	1.8	1	53.30%
SM-2_1			2	1	2.2	1.4	60.90%
SM-2_2			2	2	2.1	1.3	61.20%
SM-3_1			3	1	1.3	0.7	51.30%
SM-3_2			3	2	1.2	0.6	49.60%
Cer-1_1	Ceramides	Outer leaflet	1	1	1.6	0.9	54.50%
Cer-1_2			1	2	1.9	1	54.90%
Cer-2_1			2	1	2.1	1.3	61.50%
Cer-2_2			2	2	2.2	1.3	61.60%
Cer-3_1			3	1	1.3	0.7	52.80%
Cer-3_2			3	2	0.8	0.4	52.90%
PC-1_1	Phosphati-dylcholines	Outer leaflet	1	1	1.8	1	54.50%
PC-1_2			1	2	1.9	1	54.00%
PC-2_1			2	1	1.8	1.1	61.40%
PC-2_2			2	2	2.2	1.4	60.80%
PC-3_1			3	1	3.2	1.7	54.00%
PC-3_2			3	2	2.6	1.4	53.50%
PE-1_1	Phosphati-dylethanolamines	Inner leaflet	1	1	1.8	1	54.00%
PE-1_2			1	2	1.5	0.8	53.40%
PE-2_1			2	1	2.3	1.4	58.20%
PE-2_2			2	2	2.1	1.2	59.50%
PE-3_1			3	1	3	1.6	52.70%
PE-3_2			3	2	6.4	3.5	54.40%
PS-1_1	Phosphati-dylserines	Inner leaflet	1	1	1.6	0.9	53.70%
PS-1_2			1	2	2.1	1.1	54.50%
PS-2_1			2	1	1.9	1.2	61.30%
PS-2_2			2	2	1.8	1.1	61.20%
PS-3_1			3	1	0.7	0.4	50.20%
PS-3_2			3	2	0.8	0.4	56.50%
PiP3-1_1	Phosphatidylinositol (3,4,5)-trisphosphate	Inner leaflet	1	1	1.9	1	54.20%
PiP3-1_2			1	2	1.5	0.8	53.90%
PiP3-2_1			2	1	2	1.2	60.70%
PiP3-2_2			2	2	2.1	1.3	62.00%
PiP3-3_1			3	1	1.2	0.6	51.70%
PiP3-3_2			3	2	1.3	0.7	52.90%
Chol-1_1	Cholesterol	Lipid raft / all	1	1	1.8	1	55.50%
Chol-1_2			1	2	1.6	0.9	54.30%
Chol-2_1			2	1	1.9	1.2	60.90%
Chol-2_2			2	2	2.2	1.4	61.00%
Chol-3_1			3	1	1.1	0.6	52.70%
Chol-3_2			3	2	2.5	1.3	53.00%
SS-1_1	Sphingosines	Outer leaflet/ lipid rafts	1	1	1.5	0.8	53.50%
SS-1_2			1	2	1.8	1	54.30%
SS-2_1			2	1	1.6	1	60.60%
SS-2_2			2	2	1.9	1.1	61.50%
SS-3_1			3	1	0.3	0.2	53.50%
SS-3_2			3	2	1.9	1	52.00%

Supplementary Table 3: Library information for SurfaceSeq and SurfaceClick

SurfaceSeq				SurfaceClick			
SampleID	**RNA_type**	**Mapping**	**Non-rRNA Uniq Reads**	**SampleID**	**RNA_type**	**Mapping**	**Non-rRNA Uniq Reads**
EL4-Seq-1	Surface	21,981,145 (28.4%)	3,406,436	EL4-Click-1	Surface	2,600,269 (64.7%)	2,220,649
EL4-Seq-2	Surface	96,871,019 (80.8%)	25,317,441	EL4-Click-2	Surface	5,027,700 (81.8%)	4,162,170
EL4-Seq-3	Surface	81,245,588 (76.8%)	4,796,347	EL4-Click-3	Surface	3,108,187 (75.9.%)	3,745,353
EL4-Tot-Seq-1	Total	49,807,704 (72.8%)	35,212,630	EL4-Tot-Click-1	Total	4,418,618 (57.1%)	583,389
EL4-Tot-Seq-2	Total	57,035,602 (90.5%)	51,679,699	EL4-Tot-Click-2	Total	17,835,770 (74.7%)	6,966,723
				EL4-Tot-Click-3	Total	5,158,649 (78.5%)	2,339,436
SurfaceSeq Surface		Reads Num	33,520,224	**SurfaceClick Surface**		Reads Num	10,128,172
SurfaceSeq Total		Reads Num	86,892,329	**SurfaceClick Total**		Reads Num	9,889,548

Supplementary Table 4: Total number of reads for combined SurfaceClick surface samples, combined SurfaceClick total samples, SurfaceSeq surface samples, and SurfaceSeq total samples

Lib Type	Total Reads
SurfaceClick Surface	10,128,172
SurfaceClick Total	11,889,548
SurfaceSeq Surface	53,520,224
SurfaceSeq Total	86,892,329

Supplementary Table 5: P-value of Hypergeometric Test using 4 Different Threshold. This table showed the result of hypergeometric test using 4 different thresholds. The P-value indicated the probability of a gene identified as significantly enriched by SurfaceClick is equal to the probability of the same gene identified as significantly enriched by SurfaceSeq. Results lead to the conclusion of the probability of a gene identified as significantly enriched by SurfaceClick is not equal to the probability of the same gene identified as significantly enriched by SurfaceSeq. Therefore, the further conclusion of genes identified by one technique were not evenly distributed in the other could be made.

padj < 0.1 and log2FoldChange > 1

	SurfaceClick Detected Genes in SurfaceSeq	All SurfaceSeq Genes
Significant	110	1384
Insignificant	570	23,037
Total	680	24,421

dhyper.pValue = 7.578951e-24

	SurfaceSeq Detected Genes in SurfaceClick	All SurfaceClick Genes
Significant	158	680
Insignificant	1,226	23,740
Total	1,384	24,421

dhyper.pValue = 9.427191e-55

padj < 0.2 and log2FoldChange > 1

	SurfaceClick Detected Genes in SurfaceSeq	All SurfaceSeq Genes
Significant	161	1384
Insignificant	519	23,037
Total	680	24,421

dhyper.pValue = 5.213349e-57

	SurfaceSeq Detected Genes in SurfaceClick	All SurfaceClick Genes
Significant	296	680
Insignificant	1,096	23,740
Total	1,384	24,421

dhyper.pValue = 8.677003e-191

Cross validation threshold of padj < 0.05 and log2FoldChange > 2

	SurfaceClick Detected Genes in SurfaceSeq	All SurfaceSeq Genes
Significant	88	1384
Insignificant	592	23,037
Total	680	24,421

dhyper.pValue = 2.102588e-13

	SurfaceSeq Detected Genes in SurfaceClick	All SurfaceClick Genes
Significant	85	680
Insignificant	1,299	23,740
Total	1,384	24,421

dhyper.pValue = 3.421948e-12

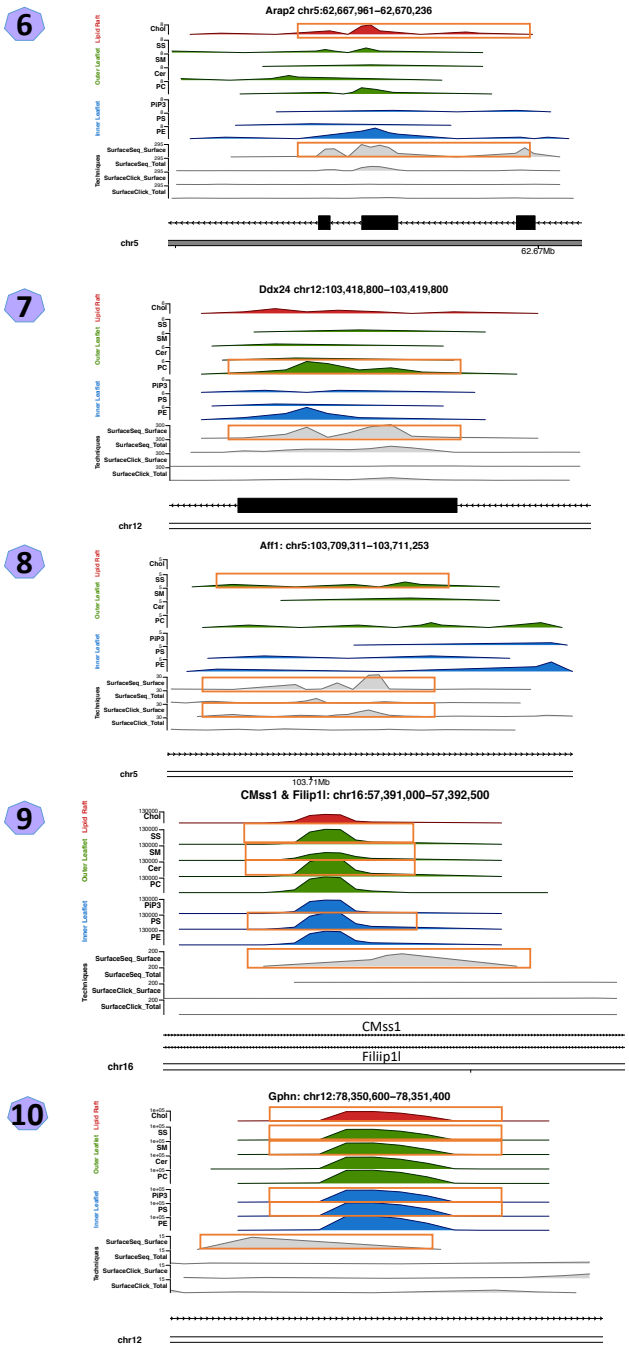
Cross validation threshold of padj < 0.01 and log2FoldChange > 2

	SurfaceClick Detected Genes in SurfaceSeq	All SurfaceSeq Genes
Significant	59	1384
Insignificant	621	23,037
Total	680	24,421

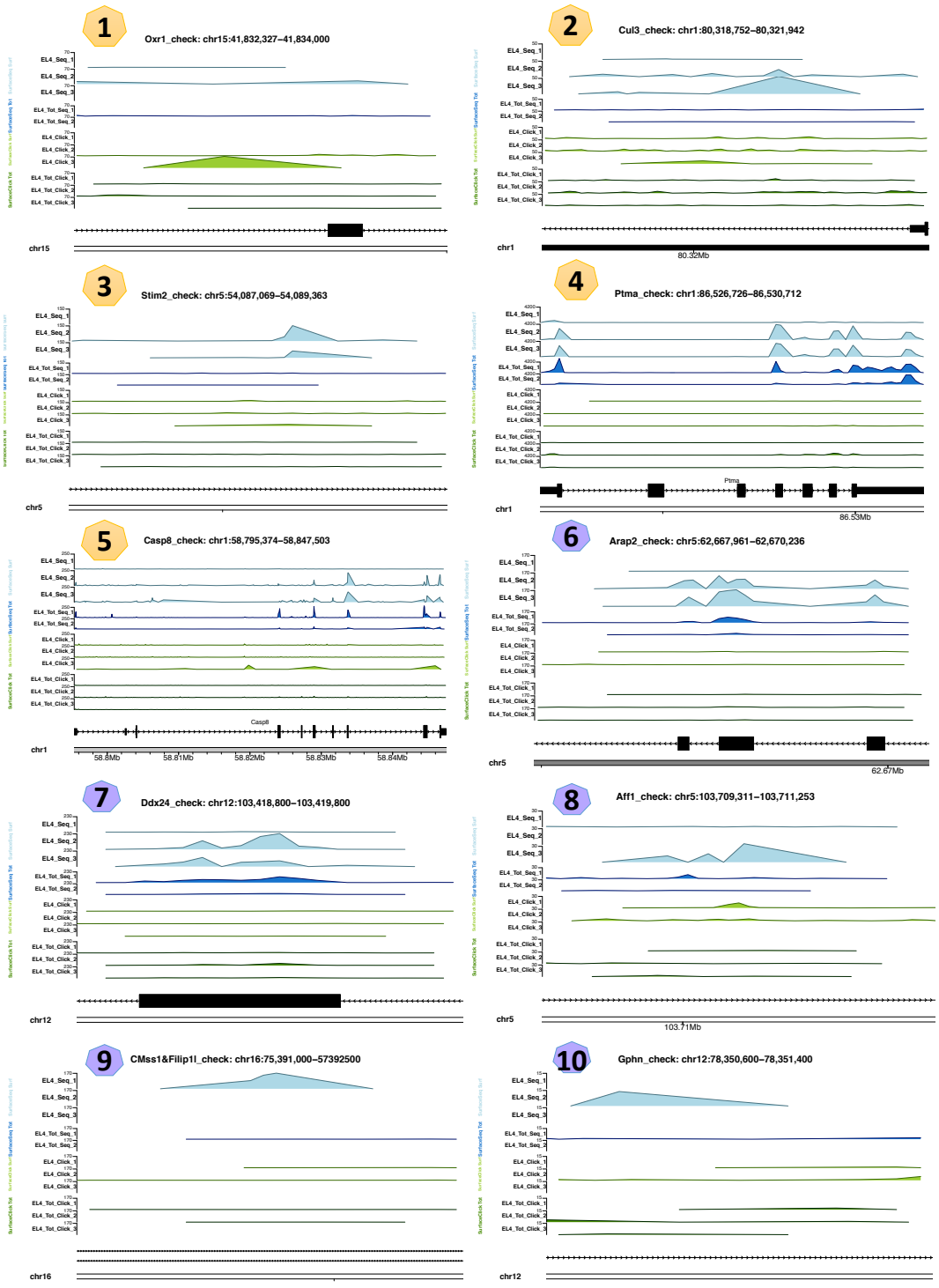
dhyper.pValue = 0.0003177944

	SurfaceSeq Detected Genes in SurfaceClick	All SurfaceClick Genes
Significant	35	680
Insignificant	1,349	23,740
Total	1,384	24,421

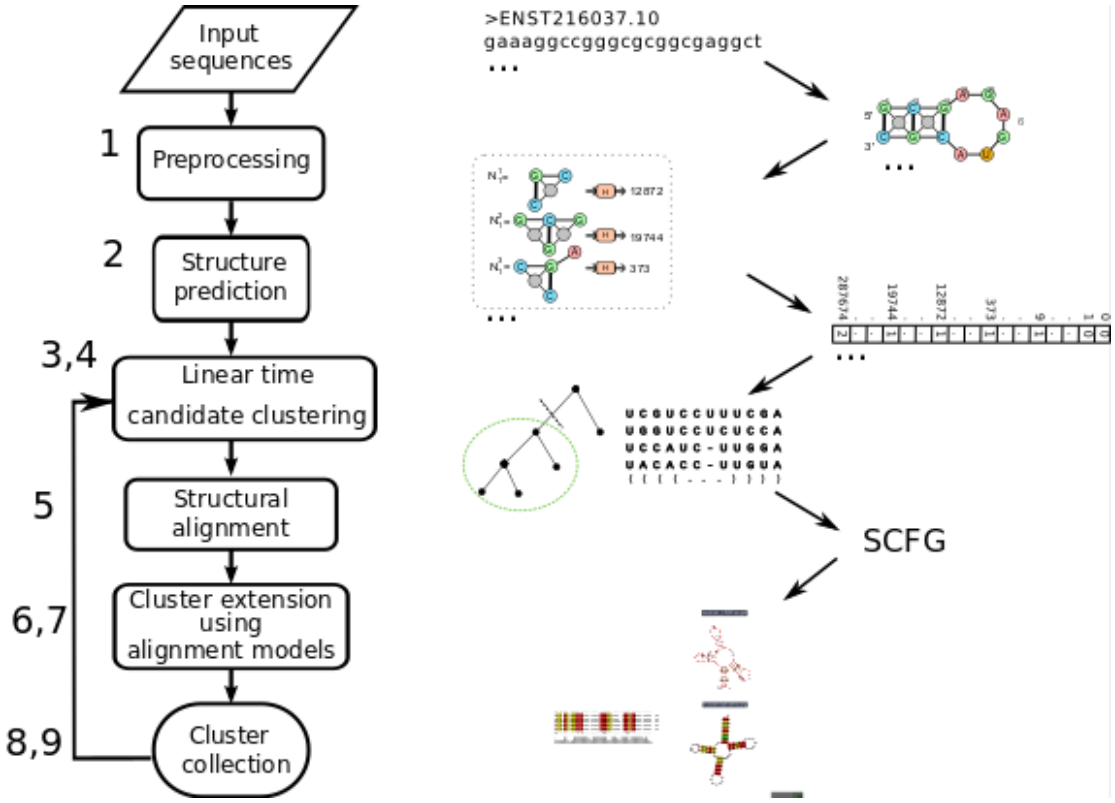
dhyper.pValue = 0.05830507



Supplementary Figure 2: Additional 5 Examples of Upset plot for all 8 different types of beads and two types of cell surface sequencing technique, on the gene level. Green background color indicated the bead is specific for outer leaflet. Blue background color indicated the bead is specific for RNAs on the inner leaflet, the red color indicated the bead is specific for RNAs on the lipid raft, and the orange color indicated the bead is specific for RNAs on the outer leaflet and lipid raft. The upset plot showing the intersection relationships of significant peaks, which filtered by applying the threshold of q-value smaller than 0.05, from 8 different types of beads and two orthogonal cell surface sequencing techniques on the gene level. On the plot, ten types of overlapping relationships were marked by blue arrow with tag number.



Supplementary Figure 3. **Situation of detected peaks shown in genomic view 1-10 figures shared among different experimental replicates within SurfaceClick or SurfaceSeq**: Blue trials on top were replicated from SurfaceSeq and Green trials at the bottom were replicated from SurfaceClick. Lighter color repressed the surface samples and darker color represented the total samples. Number on top of each figure is correspond to 10 examples on Upset plot in ***Figure 4*** and ***Supplementary Figure 2***



Supplementary Figure 4: **GraphClust Pipeline on Galaxy**. The pipeline for clustering RNA sequences and structured motif discovery is a multi-step pipeline. Overall it consists of three major phases: a) sequence-based pre-clustering b) encoding predicted RNA structures as graph features c) iterative fast candidate clustering then refinement

References

1. Janas, T. & Yarus, M. Visualization of membrane RNAs. *Rna* **9**, 1353–1361 (2003).

2. Janas, T., Janas, T. & Yarus, M. Specific RNA binding to ordered phospholipid bilayers. *Nucleic Acids Res.* **34**, 2128–2136 (2006).

3. Lin A, Hu Q, L. C. The LINK-A lncRNA interacts with PtdIns(3,4,5)P3 to hyperactivate AKT and confer resistance to AKT inhibitors. *Physiol. Behav.* **176**, 139–148 (2017).

4. Klingemann, H., Boissel, L. & Toneguzzo, F. Natural killer cells for immunotherapy - Advantages of the NK-92 cell line over blood NK cells. *Front. Immunol.* **7**, 1–7 (2016).

5. Hirenkumar, M., Steven, S. Poly Lactic-co-Glycolic Acid (PLGA) as Biodegradable Controlled Drug Delivery Carrier. *Polymers (Basel).***3**, 1–19 (2012).

6. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. MiRBase: From microRNA sequences to function. *Nucleic Acids Res.* **47**, D155–D162 (2019).

7. Löwes, B., Chauve, C., Ponty, Y. & Giegerich, R. The BRaliBase dent-a tale of benchmark design and interpretation. *Brief. Bioinform.* **18**, 306–311 (2017).

8. Wilm, A., Higgins, D. G. & Notredame, C. R-Coffee: A method for multiple alignment of non-coding RNA. *Nucleic Acids Res.* **36**, (2008).

9. Bernhart, S. H., Hofacker, I. L., Will, S., Gruber, A. R. & Stadler, P. F. RNAalifold: Improved consensus structure prediction for RNA alignments. *BMC Bioinformatics* **9**, 1–13 (2008).

10. Buskilay, A. A. A., Kannaiahy, S. & Amster-Choder, O. RNA localization in bacteria. *RNA Biol.* **11**, 1051–1060 (2014).

11. Janas, T., Janas, T. & Yarus, M. Human tRNA associates with HeLa membranes, cell lipid liposomes, and synthetic lipid bilayers. *Portafolio* n/a (2005) doi:10.1261/rna.035352.112.proteins.

12. Swain, P. S., Elowitz, M. B. & Siggia, E. D. Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc. Natl. Acad. Sci. U. S. A.* **99**, 12795–12800 (2002).

13. Freund, I., Eigenbrod, T., Helm, M. & Dalpke, A. H. RNA modifications modulate activation of innate toll-like receptors. *Genes (Basel).* **10**, (2019).

14. Herbet, L., Huang, X., Identification and Functional characterization of RNA at the surface of mammalian cells. *Unpubl. Manuscr.* (2020).

15. Mcmaster, M. L., Kristinsson, S. Y., Turesson, I., Bjorkholm, M. & Landgren, O. RNA-RNA Interactions Enable Specific Targeting of Noncoding RNAs to Nascent Pre-mRNAs and Chromatin Sites Jesse. *Clin. Lymphoma* **9**, 19–22 (2010).

16. Wang, J., Li, Y., Ding, M., Zhang, H., Xu, X., Tang, J. Molecular mechanisms and clinical applications of MIR-22 in regulating malignant progression in human cancer (Review). *Int. J. Oncol.* **50**, 345–355 (2017).

17. Amodio, N., Raimondi, L., Juli, G., Stamato, M., Caracciolo, D., Tagliaferri, P., Tassone, P. MALAT1: A druggable long non-coding RNA for targeted anti-cancer approaches. *J. Hematol. Oncol.* **11**, 1–19 (2018).

18. Huang, N., Fan, X., Rivera, K., Nguyen, T., Fang, R., Luo. Y, Gao, J., Chen, Z., Zhang, L., Zhong, S. Naturally occurring cell surface-display of genome encoded RNAs and their impacts on cell-cell interactions. *Unpubl. Manuscr.* (2020).

19. Yarus, M. Amino Acids as RNA Ligands: A Direct-RNA-Template Theory for the Code's Origin. *Mol. Evol.* **47**, 109–117 (1998).

20. Khvorova, A., Kwak, Y. G., Tamkun, M., Majerfeld, I. & Yarus, M. RNAs that bind and change the permeability of phospholipid membranes. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 10649–10654 (1999).

21. Helfer, B., Boswell, B., Finlay, D., Cipres, A., Vuori, K., Kang, T., Wallach, D., Dorfleutner, A., Lahti, J., Flynn, D., Frisch, M. Caspase-8 promotes cell motility and calpain activity under nonapoptotic conditions. *Cancer Res.* **66**, 4273–4278 (2006).

22. Janas, T., Janas, T. & Yarus, M. A membrane transporter for tryptophan composed of RNA. *Rna* **10**, 1541–1549 (2004).

23. Broude, N. E. Analysis of RNA localization and metabolism in single live bacterial cells: Achievements and challenges. *Mol. Microbiol.* **80**, 1137–1147 (2011).

24. Konjević, G., Jurišić, V. & Spužić, I. Corrections to the original lactate dehydrogenase (LDH) release assay for the evaluation of NK cell cytotoxicity. *J. Immunol. Methods* **200**, 199–201 (1997).

25. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

26. S, A. FastQC: a quality control tool for high throughput sequence datale. http://www.bioinformatics.babraham.ac.uk/projects/ (2010).

27. Dobin, A., Davia, C., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., Gingeras, T. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

28. Kuhn, R. M., Haussler, D. & James Kent, W. The UCSC genome browser and associated tools. *Brief. Bioinform.* **14**, 144–161 (2013).

29. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

30. Zhang, Y., Liu, T., Meyer, C., Eeckhoute, J., Jhonson, D., Bernstein, B., Nusbaum, C., Myers, R., Brown, M., Li, W., Liu, S. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, (2008).

31. Ramírez, F., Ryan, D., Gruning, B., Bhardwaj, V., Kilpert, F., Richter, A., Heyne, S., Dunder, F., Manke, T. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).

32. Liao, Y., Smyth, G. K. & Shi, W. FeatureCounts: An efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

33. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 1–21 (2014).

34. Sven, H., Benner, C., Spann, N., Bertolino, E., Lin, Y., Laslo, P., Cheng, Jason., Murre, C., Singh, C., Glass, C. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).

35. Cock, P. J. A., Antao, T., Chang, J., Chapman, B., Cox, C., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., Hoon, M. Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).

36. Venkatesh L. Murthy and George D.Rose. RNABase: an annotated database of RNA structures. *Oxford Univ. Press* (2013).

37. Heyne, S., Costa, F., Rose, D. & Backofen, R. Graphclust: Alignment-free structural clustering of local RNA secondary structures. *Bioinformatics* **28**, 224–232 (2012).