

High-throughput identification of RNA nuclear enrichment sequences

Chinmay J Shukla, Alexandra L McCorkindale, Chiara Gerhardinger, Keegan D Korthauer, Moran N Cabili, David M Shechner, Rafael A Irizarry, Philipp G Maass & John L Rinn

Review timeline:

Submission date:	19 October 2017
Editorial Decision:	20 November 2017
Revision received:	3 December 2017
Editorial Decision:	14 December 2017
Revision received:	18 December 2017
Accepted:	20 December 2017

Editor: Anne Nielsen

Transaction Report:

(Note: With the exception of the correction of typographical or spelling errors that could be a source of ambiguity, letters and reports are not edited. The original formatting of letters and referee reports may not be reflected in this compilation.)

1st Editorial Decision

20 November 2017

Thank you for submitting your manuscript for consideration as a resource paper by The EMBO Journal. It has now been seen by two referees whose comments are shown below.

As you will see from the reports, both referees highlight the technical quality of the work and express interest in the findings reported in your manuscript. I want to add that I had an additional round of discussion with the referees (in light of the concerns from ref #2 about the level of mechanistic insight provided), emphasizing that we are considering this study for our methods/resource section. The outcome is that both referees find the method itself and the data provided to be important and valuable to the field and support its publication here.

Given the referees' positive recommendations, I would thus like to invite you to submit a revised version of the manuscript, addressing the comments of both reviewers. I should add that it is EMBO Journal policy to allow only a single round of revision, and acceptance of your manuscript will therefore depend on the completeness of your responses in this revised version.

For the revised manuscript I would particularly ask you to focus your efforts on the following points:

- Referee #1 asks for additional clarification of the retention effects for mRNAs vs lncRNAs. While this may not give a clear either-or answer, the analysis would be informative and can presumably be done with the current dataset.
- Referee #2 asks for an additional control to test if a difference in nuclear and cytoplasmic decay rates could contribute to the localization of the reporter RNA
- We noticed that the manuscript currently has 4 main figures while a lot of data is found in the 8 supplemental figures. In order to make the data more directly accessible to the readers I'd suggest that you move some of this supplemental data into the main figures.

Thank you for the opportunity to consider your work for publication. I look forward to your revision.

 REFEREE REPORTS

Referee #1:

The study uses barcoded oligonucleotides of 110nt to identify 109 nuclear-enriched RNA regions, termed "differential regions" (DRs). Presence of DRs in exons is shown to correlate with their increased nuclear retention. The identification and validation of DR motifs is convincing, and my main request is to corroborate their role in the retention of lncRNAs vs mRNAs.

The study could address more thoroughly the question why lncRNAs are more retained in the nucleus than mRNAs. The sequence motifs enriched in the DRs are used in Fig 3H-I to show that nuclear retention of DR-containing endogenous RNAs is increased. However, all classes of RNAs are grouped for this analysis. If DRs are more common in the classes that are generally more retained for other reasons, the result could be of correlative nature. The authors would need to analyse nuclear/cytoplasmic ratios separately for different classes of transcripts, especially separating lncRNAs from mRNAs in Fig 3H-I. The authors could then compare nuclear retention of DR-containing lncRNAs vs mRNA, and DR-lacking lncRNAs vs mRNA. How does nuclear retention of lncRNAs lacking DRs compare to mRNAs containing DRs?

Other comments:

- P-values are reported in the figures as 2.2×10^{-16} and for the same test in the text as $1/10^6$. Why are these values different?
- The statistics should be reported for the motifs in F3A,D: for a discriminative analysis, the number of motifs in the DR and in the non-DR sets, and the associated p-value for each motif should be reported.
- The Kolmogorov-Smirnov test might be more appropriate than the Mann-Whitney test in F3G-I, as it is usually used as a goodness-of-fit test to check if a set of continuous values follows a certain theoretical distribution, since this test relies on the statistical distance between the cumulative distribution probabilities as opposed to testing the actual values.
- The authors report that some of the DRs overlap with repC in Xist, which is known to be involved in the nuclear localisation. As I recall, repC is quite long, and I cannot see any DRs around this region. It would help to show the overlap with repC and repD in supplementary material.

Referee #2:

In this manuscript Shukla et al. describe the use of a massive reporter assay to study the ability of RNA motifs to affect nuclear localization. In particular the authors use a tiling strategy to cover the regions of 38 well-characterized lncRNAs. They perform nuclear fractionation and identify RNA motifs associated to nuclear enrichment. Using MEME the authors identify a region commonly repeated in XIST, and a C-rich region present in multiple lncRNA. Finally the authors validate by smFISH that the identified XIST sequence, and other long differential regions associated to known lncRNAs (but not the C-rich motif) contribute to the nuclear localization. I found the proposed MPRNA approach very powerful and creative. However, I feel that the current study is still mainly descriptive and lacks the mechanistic insight that I would expect for EMBO Journal.

Major concerns:

1. Although the authors do not explain why they use a frame-shifted Sox1, I would imagine it is for the potential effect of NMD if an artificial stop codon is introduced in the library. Can the author determine if the observed change of localization is caused by variations in nuclear/cytoplasmic degradation? For example comparing the abundance of reporter DNA vs the RNA expression of the same reporters and their relative localization?

Minor:

1. Some figures are redundant. For example Fig 1D seems significant by definition (as it was based on the criteria to select the DR)

2. In Fig2F. What is the distribution of the tested regions (not only the differential) regarding the FISH classes? That information is necessary to know the background distribution and to be able to measure the enrichment.

1st Revision - authors' response

3 December 2017

We thank the reviewers for their positive and constructive comments of our manuscript entitled “High-throughput identification of RNA nuclear enrichment sequences.” We have taken this opportunity to significantly improve the quality of our manuscript based on the reviewers’ suggestions. To that end, we have revised the manuscript as described in the following point-by-point response to their comments.

Referee #1:

The study uses barcoded oligonucleotides of 110nt to identify 109 nuclear-enriched RNA regions, termed "differential regions" (DRs). Presence of DRs in exons is shown to correlate with their increased nuclear retention. The identification and validation of DR motifs is convincing, and my main request is to corroborate their role in the retention of lncRNAs vs mRNAs.

We appreciate the helpful comments of the reviewer and the careful revision of our manuscript. We have addressed the reviewer’s concerns by performing new analyses, and by adding new results to the revised manuscript. Please find below a point-by-point response to the specific comments.

The study could address more thoroughly the question why lncRNAs are more retained in the nucleus than mRNAs. The sequence motifs enriched in the DRs are used in Fig 3H-I to show that nuclear retention of DR-containing endogenous RNAs is increased. However, all classes of RNAs are grouped for this analysis. If DRs are more common in the classes that are generally more retained for other reasons, the result could be of correlative nature. The authors would need to analyze nuclear/cytoplasmic ratios separately for different classes of transcripts, especially separating lncRNAs from mRNAs in Fig 3H-I. The authors could then compare nuclear retention of DR-containing lncRNAs vs mRNA, and DR-lacking lncRNAs vs mRNA. How does nuclear retention of lncRNAs lacking DRs compare to mRNAs containing DRs?

We agree, and thank the reviewer for suggesting the neglected comparison of lncRNA vs mRNA nuclear retention. As suggested, we reanalyzed the nuclear enrichment of the identified motifs within the DRs by RNA classes. We found that the identified motifs have a stronger effect on nuclear retention of lncRNAs compared to mRNAs, but their presence still significantly increases the nuclear retention of both mRNAs and lncRNAs. We modified the results section accordingly, and replaced the CDF plots in figure 3 (panels H-I) with box plots (shown below), summarizing the results of the new analysis.

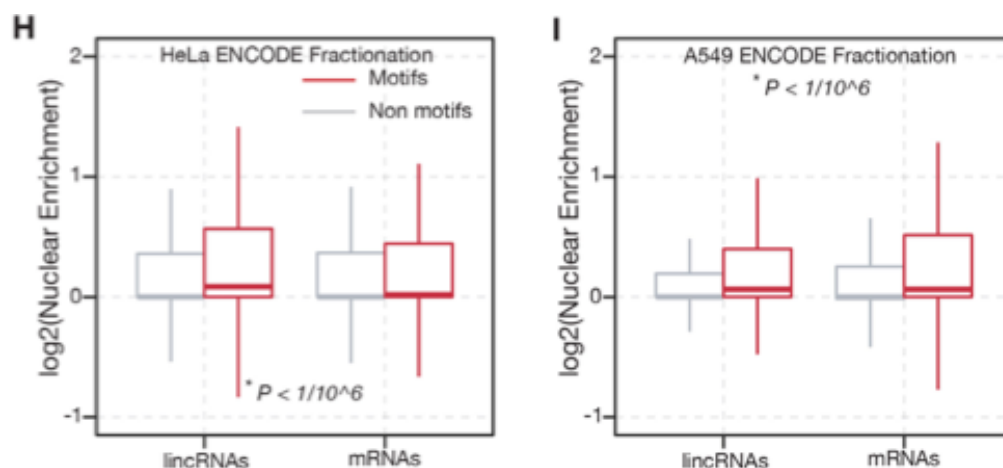


Figure 3: ... H–I. Novel nuclear enrichment motifs influence the localization of endogenous human transcripts. Comparison of the nuclear enrichment of both human mRNA and lncRNA transcripts with at least one occurrence of the discovered motifs, relative to all other transcripts, in HeLa and A549 cells³⁰. *P*-value: Mann Whitney Test.

Other comments:

- *P*-values are reported in the figures as 2.2×10^{-16} and for the same test in the text as $1/10^6$. Why are these values different?

We thank the reviewer for noting this discrepancy. The correct *p*-value is $< 1/10^6$ as found by the Wilcoxon test. We have now corrected this typo in the figures.

- The statistics should be reported for the motifs in F3A, D: for a discriminative analysis, the number of motifs in the DR and in the non-DR sets, and the associated *p*-value for each motif should be reported.

We thank the reviewer for these suggestions and have now included the exact *E* value for each motif reported by MEME.

- The Kolmogorov-Smirnov test might be more appropriate than the Mann-Whitney test in F3G-I, as it is usually used as a goodness-of-fit test to check if a set of continuous values follows a certain theoretical distribution, since this test relies on the statistical distance between the cumulative distribution probabilities as opposed to testing the actual values.

Based on the reviewer's comment we did perform the KS test that showed a statistically significant nuclear enrichment of the motifs similar to the Mann-Whitney test. The KS test was performed to test if there was a significant difference between nuclear enrichment of lncRNAs and mRNAs harboring the motifs compared to those without the motifs and gave a *p*-value $< 1/10^6$. Since both Mann-Whitney and KS tests are non-parametric and both show the significance of nuclear enrichment of motifs, we report the Mann-Whitney *p*-value, which, in our opinion, is a more appropriate test for evaluating significant shift of the medians. However, since we originally reported the shift in the median between all transcripts (lncRNAs and mRNAs) with and without the motif, we have now reanalyzed the enrichment data independently for the 2 class of transcripts, as suggested by the reviewer in point 1 and generated box plots that are more appropriate than the previous CDF plots (see above – Figure 3H and 3I).

- The authors report that some of the DRs overlap with repC in Xist, which is known to be involved in the nuclear localization. As I recall, repC is quite long, and I cannot see any DRs around this region. It would help to show the overlap with repC and repD in supplementary material.

We agree with the reviewer that showing the overlap would facilitate the interpretation of this finding. Thus, we have now included a schematic visualization of the location of the identified XIST DRs relative to all XIST repeats in the appendix (Appendix_Figure_S1, Panel A, see below).

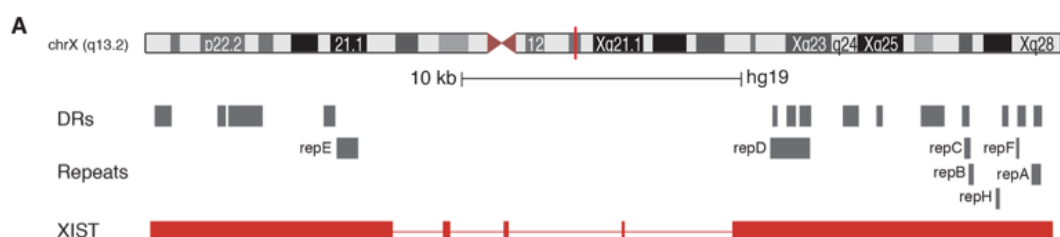


Figure S1. Sequence Features of Differential Regions A. UCSC genome browser tracks showing the overlap of XIST repeats and DRs.

Referee#2:

In this manuscript Shukla et al. describe the use of a massive reporter assay to study the ability of RNA motifs to affect nuclear localization. In particular the authors use a tiling strategy to cover the regions of 38 well-characterized lncRNAs. They perform nuclear fractionation and identify RNA motifs associated to nuclear enrichment. Using MEME the authors identify a region commonly repeated in XIST, and a C-rich region present in multiple lncRNA. Finally, the authors validate by smFISH that the identified XIST sequence, and other long differential regions associated to known lncRNAs (but not the C-rich motif) contribute to the nuclear localization. I found the proposed MPRNA approach very powerful and creative. However, I feel that the current study is still mainly descriptive and lacks the mechanistic insight that I would expect for EMBO Journal.

We thank the reviewer for acknowledging that “our MPRNA technique is powerful and the described results are novel and interesting for the fields of RNA biology” and for the helpful insights. We have addressed the reviewer’s concerns by performing new analysis, and by adding new results to the revised manuscript. Please find below a point-by-point response.

Major concerns:

1. Although the authors do not explain why they use a frame-shifted Sox1, I would imagine it is for the potential effect of NMD if an artificial stop codon is introduced in the library. Can the author determine if the observed change of localization is caused by variations in nuclear/cytoplasmic degradation? For example, comparing the abundance of reporter DNA vs the RNA expression of the same reporters and their relative localization?

We apologize for not describing the *fsSox2* reporter in detail. The *fsSox2* reporter was used and optimized in a previous study from our lab (Hacisuleyman et al. 2015). We previously determined that if the frameshift is not present, the transcript accumulates around the ER during translation which could influence cellular localization (moreover cell biology with induced Sox2 protein). Due to the frameshift in the *fsSox2* reporter, we were able to avoid translation-derived artefacts. While assessing nuclear/cytoplasmic degradation would not be feasible, our FISH analyses, both here and in Hacisuleyman et al. (2015), showed consistent cytosolic expression levels of *fsSox2* across different cell lines and various tested constructs, indicating uniform *fsSox2* processing. Altogether, our observations of uniform expression levels by smRNA-FISH show that (i) *fsSox2* is cytoplasmic and consistent in expression and (ii), this pattern changes with appended nuclear localization sequences (e.g. *MALATI*). This data suggests that *fsSox2* is a reasonable reporter for this proof-of principle. However, we agree with the reviewer that depending on the assay, other reporters may need to be optimized or characterized. We have now included a description why we used *fsSox2* in the text.

Minor:

1. Some figures are redundant. For example, Fig 1D seems significant by definition (as it was based on the criteria to select the DR)

We thank the reviewer for this suggestion. We now removed Fig 1D.

2. In Fig2F. What is the distribution of the tested regions (not only the differential) regarding the FISH classes? That information is necessary to know the background distribution and to be able to measure the enrichment.

The reviewer raises an interesting point about the distributions of transcripts in which DRs were called. We think the reviewer is considering a scenario where the length of transcripts in FISH class 1 were longer and would thus have more DRs by length and not by multiplicity. To address this, we now report the number of DR sequences/kb of sequence tiled, which would help us to remove any effect of different background sequences. We indeed see that this normalization removes the differences in DRs found in different FISH classes since all the FISH classes seem to harbor ~10 DRs per 10 kb of sequence tested (see Figure 2F).

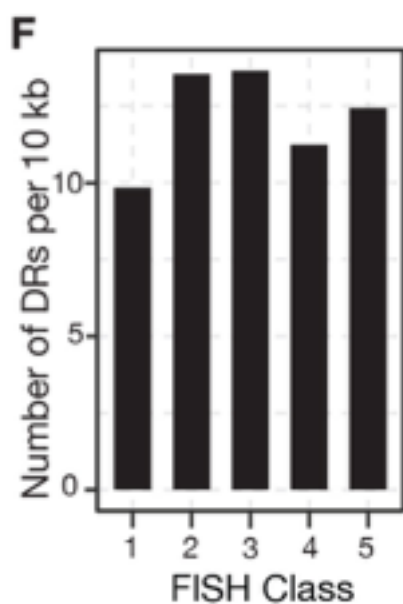


Figure 2: ... F. The number of DRs discovered per 10 kb of lncRNA sequence tiled, is highly similar for each FISH Class.

2nd Editorial Decision

14 December 2017

Thank you for submitting a revised version of your manuscript. It has now been seen by one of the original referees and this person's comments are shown below. As you will see the referee finds that all criticisms have been sufficiently addressed and recommend the manuscript for publication. However, before we can go on to officially accept the manuscript, I have to ask you to address the following editorial issues concerning text and figures [...].

Thank you again for giving us the chance to consider your manuscript for The EMBO Journal, I look forward to receiving your final revision.

REFeree REPORT

Referee #1:

The authors have adequately addressed all of my suggestions, and I find the manuscript ready for publication, and of great interest to the broad readership of the EMBO Journal.

YOU MUST COMPLETE ALL CELLS WITH A PINK BACKGROUND ↓
PLEASE NOTE THAT THIS CHECKLIST WILL BE PUBLISHED ALONGSIDE YOUR PAPER

Corresponding Author Name: John Rinn
Journal Submitted to: EMBO
Manuscript Number: EMB0J-2017-98452R

Reporting Checklist For Life Sciences Articles (Rev. June 2017)

This checklist is used to ensure good reporting standards and to improve the reproducibility of published results. These guidelines are consistent with the Principles and Guidelines for Reporting Preclinical Research issued by the NIH in 2014. Please follow the journal's authorship guidelines in preparing your manuscript.

A- Figures

1. Data

The data shown in figures should satisfy the following conditions:

- the data were obtained and processed according to the field's best practice and are presented to reflect the results of the experiments in an accurate and unbiased manner.
- figure panels include only data points, measurements or observations that can be compared to each other in a scientifically meaningful way.
- graphs include clearly labeled error bars for independent experiments and sample sizes. Unless justified, error bars should not be shown for technical replicates.
- if n< 5, the individual data points from each experiment should be plotted and any statistical test employed should be justified
- Source Data should be included to report the data underlying graphs. Please follow the guidelines set out in the author ship guidelines on Data Presentation.

2. Captions

Each figure caption should contain the following information, for each panel where they are relevant:

- a specification of the experimental system investigated (eg cell line, species name).
- the assay(s) and method(s) used to carry out the reported observations and measurements
- an explicit mention of the biological and chemical entity(ies) that are being measured.
- an explicit mention of the biological and chemical entity(ies) that are altered/varied/perturbed in a controlled manner.
- the exact sample size (n) for each experimental group/condition, given as a number, not a range;
- a description of the sample collection allowing the reader to understand whether the samples represent technical or biological replicates (including how many animals, litters, cultures, etc.).
- a statement of how many times the experiment shown was independently replicated in the laboratory.
- definitions of statistical methods and measures:
 - common tests, such as t-test (please specify whether paired vs. unpaired), simple χ^2 tests, Wilcoxon and Mann-Whitney tests, can be unambiguously identified by name only, but more complex techniques should be described in the methods section;
 - are tests one-sided or two-sided?
 - are there adjustments for multiple comparisons?
 - exact statistical test results, e.g., P values = x but not P values < x;
 - definition of 'center values' as median or average;
 - definition of error bars as s.d. or s.e.m.

Any descriptions too long for the figure legend should be included in the methods section and/or with the source data.

In the pink boxes below, please ensure that the answers to the following questions are reported in the manuscript itself. Every question should be answered. If the question is not relevant to your research, please write NA (non applicable). We encourage you to include a specific subsection in the methods section for statistics, reagents, animal models and human subjects.

B- Statistics and general methods

Please fill out these boxes ↓ (Do not worry if you cannot see all your text once you press return)

1.a. How was the sample size chosen to ensure adequate power to detect a pre-specified effect size?	We used 6 replicates for MPRNA to ensure we had enough power and in-sample variance estimation. For the genomic analyses, again multiple replicates were always used to ensure adequate power.
1.b. For animal studies, include a statement about sample size estimate even if no statistical methods were used.	NA
2. Describe inclusion/exclusion criteria if samples or animals were excluded from the analysis. Were the criteria pre-established?	NA
3. Were any steps taken to minimize the effects of subjective bias when allocating animals/samples to treatment (e.g. randomization procedure)? If yes, please describe.	Yes. For the FISH analyses, the images were double blinded before quantification
For animal studies, include a statement about randomization even if no randomization was used.	NA
4.a. Were any steps taken to minimize the effects of subjective bias during group allocation or/and when assessing results (e.g. blinding of the investigator)? If yes please describe.	Yes. For the FISH analyses, the images were double blinded before quantification
4.b. For animal studies, include a statement about blinding even if no blinding was done	NA
5. For every figure, are statistical tests justified as appropriate?	Yes
Do the data meet the assumptions of the tests (e.g., normal distribution)? Describe any methods used to assess it.	Non parametric tests were used to assess p-values. The sample sizes in both groups for Mann Whitney were more than 20
Is there an estimate of variation within each group of data?	Yes. We show boxplots for all our genomic analyses and SDs for the raw data to demonstrate variance

USEFUL LINKS FOR COMPLETING THIS FORM

- <http://www.antibodypedia.com>
- <http://1degreebio.org>
- <http://www.equator-network.org/reporting-guidelines/improving-bioscience-research-repo>
- <http://grants.nih.gov/grants/olaw/olaw.htm>
- <http://www.mrc.ac.uk/Ourresearch/Ethicsresearchguidance/Useofanimals/index.htm>
- <http://ClinicalTrials.gov>
- <http://www.consort-statement.org>
- <http://www.consort-statement.org/checklists/view/32-consort/66-title>
- <http://www.equator-network.org/reporting-guidelines/reporting-recommendations-for-tun>
- <http://datadryad.org>
- <http://figshare.com>
- <http://www.ncbi.nlm.nih.gov/gap>
- <http://www.ebi.ac.uk/ega>
- <http://biomodels.net/>
- <http://biomodels.net/miriam/>
- <http://jij.biochem.sun.ac.za>
- http://oba.od.nih.gov/biosecurity/biosecurity_documents.html
- <http://www.selectagents.gov/>

Is the variance similar between the groups that are being statistically compared?	Yes
---	-----

C- Reagents

6. To show that antibodies were profiled for use in the system under study (assay and species), provide a citation, catalog number and/or clone number, supplementary information or reference to an antibody validation profile. e.g., Antibodypedia (see link list at top right), 1DegreeBio (see link list at top right).	Hacisuleyman et al. (2015) Function and evolution of local repeats in Firre locus Nature Communications
7. Identify the source of cell lines and report if they were recently authenticated (e.g., by STR profiling) and tested for mycoplasma contamination.	HeLa cells are from ATCC (HeLa ATCC CCL-2). We did not authenticate them. We tested for mycoplasma (the stock and some of the passages, not all though) and they were negative.

* for all hyperlinks, please see the table at the top right of the document

D- Animal Models

8. Report species, strain, gender, age of animals and genetic modification status where applicable. Please detail housing and husbandry conditions and the source of animals.	NA
9. For experiments involving live vertebrates, include a statement of compliance with ethical regulations and identify the committee(s) approving the experiments.	NA
10. We recommend consulting the ARRIVE guidelines (see link list at top right) (PLoS Biol. 8(6), e1000412, 2010) to ensure that other relevant aspects of animal studies are adequately reported. See author guidelines, under 'Reporting Guidelines'. See also: NIH (see link list at top right) and MRC (see link list at top right) recommendations. Please confirm compliance.	NA

E- Human Subjects

11. Identify the committee(s) approving the study protocol.	NA
12. Include a statement confirming that informed consent was obtained from all subjects and that the experiments conformed to the principles set out in the WMA Declaration of Helsinki and the Department of Health and Human Services Belmont Report.	NA
13. For publication of patient photos, include a statement confirming that consent to publish was obtained.	NA
14. Report any restrictions on the availability (and/or on the use) of human data or samples.	NA
15. Report the clinical trial registration number (at ClinicalTrials.gov or equivalent), where applicable.	NA
16. For phase II and III randomized controlled trials, please refer to the CONSORT flow diagram (see link list at top right) and submit the CONSORT checklist (see link list at top right) with your submission. See author guidelines, under 'Reporting Guidelines'. Please confirm you have submitted this list.	NA
17. For tumor marker prognostic studies, we recommend that you follow the REMARK reporting guidelines (see link list at top right). See author guidelines, under 'Reporting Guidelines'. Please confirm you have followed these guidelines.	NA

F- Data Accessibility

18: Provide a "Data Availability" section at the end of the Materials & Methods, listing the accession codes for data generated in this study and deposited in a public database (e.g. RNA-Seq data: Gene Expression Omnibus GSE39462, Proteomics data: PRIDE PXD000208 etc.) Please refer to our author guidelines for 'Data Deposition'. Data deposition in a public repository is mandatory for: a. Protein, DNA and RNA sequences b. Macromolecular structures c. Crystallographic data for small molecules d. Functional genomics data e. Proteomics and molecular interactions	Data Availability section included in the manuscript
19. Deposition is strongly recommended for any datasets that are central and integral to the study; please consider the journal's data policy. If no structured public repository exists for a given data type, we encourage the provision of datasets in the manuscript as a Supplementary Document (see author guidelines under 'Expanded View' or in unstructured repositories such as Dryad (see link list at top right) or Figshare (see link list at top right)).	Data submitted to GEO - number in the manuscript. Other raw data in Supplementary Table 1 and 2
20. Access to human clinical and genomic datasets should be provided with as few restrictions as possible while respecting ethical obligations to the patients and relevant medical and legal issues. If practically possible and compatible with the individual consent agreement used in the study, such data should be deposited in one of the major public access-controlled repositories such as dbGAP (see link list at top right) or EGA (see link list at top right).	NA
21. Computational models that are central and integral to a study should be shared without restrictions and provided in a machine-readable form. The relevant accession numbers or links should be provided. When possible, standardized format (SBML, CellML) should be used instead of scripts (e.g. MATLAB). Authors are strongly encouraged to follow the MIRIAM guidelines (see link list at top right) and deposit their model in a public database such as Biocompare (see link list at top right) or JWS Online (see link list at top right). If computer source code is provided with the paper, it should be deposited in a public repository or included in supplementary information.	Computer code deposited in publicly accessible GitHub Repo oligoGames. Please see github.com/cshukla

G- Dual use research of concern

22. Could your study fall under dual use research restrictions? Please check biosecurity documents (see link list at top right) and list of select agents and toxins (APHIS/CDC) (see link list at top right). According to our biosecurity guidelines, provide a statement only if it could.	NA
---	----