Global analysis of regulatory divergence in the evolution of mouse alternative polyadenylation

Mei-Sheng Xiao, Bin Zhang, Yi-Sheng Li, Qingsong Gao, Wei Sun and Wei Chen

Corresponding author: Wei Chen, Southern University of Science and Technology

Review timeline:

Submission date:	07 October 2016
Editorial Decision:	28 October 2016
Revision received:	08 November 2016
Editorial Decision:	11 November 2016
Revision received:	13 November 2016
Accepted:	15 November 2016

Editor: Maria Polychronidou

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 28 October 2016

Thank you again for submitting your work to Molecular Systems Biology. We have now heard back from the two referees who agreed to evaluate your study. As you will see below, the reviewers acknowledge that the presented analyses generate interesting insights. However, they list several issues, which we would ask you to address in a revision. The reviewers' recommendations are rather clear so I think that there is no need to repeat the points listed below, but please let me know in case you would like to discuss any specific point.

REFEREE REPORTS

Reviewer #1:

The goal of the paper entitled "Global analysis of regulatory divergence in the evolution of mouse alternative polyadenylation" is to understand the degree of alternative polyadenylation (APA) divergence and the contributions of cis- and trans- regulatory elements to APA by applying F1 hybridization experiments into two evolutionarily distant mouse strains. Based on the sets of distinctly mapped reads in polyadenylation sites (pAs) from deep sequencing approaches in two strains, the authors provided corresponding experimental evidence for the following three observations. First, based on the frequency of divergent pAs in protein coding and non-coding regions, APA affecting their functions is more deleterious thus under strong negative selection. Secondly, through the comparison between two parental strains and their differences from the two alleles in the F1 hybrids, cis-effects are more dominant than trans-effects in APA. Lastly, both the stability of local RNA secondary structures and a poly(U) tract especially in the upstream region have considerable effects on gene regulation based on the measurement of the minimum free energy (MFE) of mRNA segments and sequence motifs analysis, respectively. Overall, most results are relatively clearly explained and their experimental results are independently supported by using human genome-scale data. The authors also introduced a recently published paper showing different patterns of positional stabilities of RNA secondary structures in ADA in Arabidopsis and provided three probable scenarios/hypotheses explaining the observational disparity. In 2015, using similar approach, Chen and his colleagues have already published a paper in the same journal for the regulatory divergence in the evolution of alternative splicing. I think this paper can additionally provide more complete pictures of evolutionary history for post-transcriptional regulation in mouse.

Thus, I recommend that this paper be accepted for publication after some minor points explained below are addressed.

1. In the subsection, "Construction of the pAs reference", authors demonstrated the quality of their data by saying that most representative cleavage sites of the pAs clusters were almost identical to the annotated 3' end. This sounds somewhat subjective. It would be better to show more objective evidence such as quantitative measurements of their agreement. In the last sentence from the same paragraph, the authors should cite a paper showing "previous" observation.

2. Related to Fig2B, authors used 20 genes for validating the accuracy of their allele specific APA analysis. Authors need to mention that high replicability can be seen regardless of the choice of the selected genes and the numbers chosen.

3. In the subsection, "RNA secondary structure in the upstream proximal region inhibits pAs usage", authors said "This trend became more evident if we restricted our analysis to the annotated most distal pAs, which were in general of higher strength than proximal ones". Is there any figure or table which we can see these trends? If so, it should be referenced here.

4. In figure 1D and F, please add actual numbers on top of the percentages.

Reviewer #2:

In this manuscript Xiao et al. perform a global analysis of alternative polyadenylation (APA) using fibroblasts from two divergent mouse lines as well as their F1 cross. They combine data from 3'READS and oligo-dT priming based 3' quantification to annotate and measure the relative expression of APA. They focus their work on cis-regulated APA events and investigate potential motives contributing to its regulation. The authors perform orthogonal confirmation of selected targets using a fluorescence based in vitro system and analyze the contribution of secondary structure and motives to APA usage.

General remarks:
The combined used of two different 3' quantification methods allow the authors to focus on median and high expressed APA events and remove from their analysis any APA event due to internal oligo-dT priming. The fact that alterations of core polyadenylation elements (eg. hexamer AAUAAA) impacts APA are not surprising. However the authors use an elegant experimental design that allows them to distinguish between cis- and trans-regulated APA.

Major points:
Due to the experimental designed used by the authors; I am surprised that they focus almost exclusively on the cis-regulated APA events. Adding a brief analysis of the trans-regulated APA events will significantly increase the interest of the paper and differentiate this work from other studies. For example, performing an hexamer analysis analogue to the one that the authors perform for the cis-regulated APA events. The authors could also study if different RNA Binding Proteins or miRNAs are putatively bound (or in proximity) to the alternative polyadenylated isoforms using available data (eg. PMID 23846655). And if so, analysis how is the expression of the putative RNA Binding Protein in the F1 cell line.

Minor points:
In page 12-13 the authors briefly mention the method that they use for orthogonal confirmation (eg. Fig 3D and 4I). However, the description in the main text is too brief. I would recommend adding a couple of sentences describing the general principle of the approach and how the artificial constructs are assayed in the same cell lines.

Some small typos in the figures (eg. in FigEV1D " Cleavage Site").

1st Revision - authors' response 08 November 2016

Text continued on next page.

Reviewer #1:

The goal of the paper entitled "Global analysis of regulatory divergence in the evolution of mouse alternative polyadenylation" is to understand the degree of alternative polyadenylation (APA) divergence and the contributions of cis- and trans-regulatory elements to APA by applying F1 hybridization experiments into two evolutionarily distant mouse strains. Based on the sets of distinctly mapped reads in polyadenylation sites (pAs) from deep sequencing approaches in two strains, the authors provided corresponding experimental evidence for the following three observations. First, based on the frequency of divergent pAs in protein coding and non-coding regions, APA affecting their functions is more deleterious thus under strong negative selection. Secondly, through the comparison between two parental strains and their differences from the two alleles in the F1 hybrids, cis-effects are more dominant than trans-effects in APA. Lastly, both the stability of local RNA secondary structures and a poly(U) tract especially in the upstream region have considerable effects on gene regulation based on the measurement of the minimum free energy (MFE) of mRNA segments and sequence motifs analysis, respectively. Overall, most results are relatively clearly explained and their experimental results are independently supported by using human genome-scale data. The authors also introduced a recently published paper showing different patterns of positional stabilities of RNA secondary structures in ADA in Arabidopsis and provided three probable scenarios/hypotheses explaining the observational disparity. In 2015, using similar approach, Chen and his colleagues have already published a paper in the same journal for the regulatory divergence in the evolution of alternative splicing. I think this paper can additionally provide more complete pictures of evolutionary history for post-transcriptional regulation in mouse. Thus, I recommend that this paper be accepted for publication after some minor points explained below are addressed.

R: We thank the reviewer for her/his positive comments on our study.

1. In the subsection, "Construction of the pAs reference", authors demonstrated the quality of their data by saying that most representative cleavage sites of the pAs clusters were almost identical to the annotated 3' end. This sounds somewhat subjective. It would be better to show more objective evidence such as quantitative measurements of their agreement.

R: We would like to thank the review to point this out. To make our statement more quantitative, we calculated the number of pAs with the identified representative cleavage site exactly identical to the ENSEMBL annotated transcript ends and those locating within 5nt upstream or downstream of the annotated ends, respectively. In the revised Fig 1C, we added an inset, which shows that 39.5% and 41.1% of these pAs are identical to or within 5nt upstream of downstream of the annotated ends, respectively. We also added these numbers in the revised main text (Page 7) and Figure legend (Page 38).

2. In the last sentence from the same paragraph, the authors should cite a paper showing "previous" observation.

R: We thank the reviewer for the suggestion and in the revised manuscript, we added the citation for the corresponding paper (Page 7).

3. Related to Fig2B, authors used 20 genes for validating the accuracy of their allele specific APA analysis. Authors need to mention that high replicability can be seen regardless of the choice of the selected genes and the numbers chosen.

R: Thank the reviewer for the suggestion. In the revised manuscript, to further assess the reproducibility of our method on measuring the allelic difference in pAs usage, we compared the results from the two independent experimental replicates. As shown in the newly added Fig EV3, we observed the results from the two replicated correlated well (r = 0.90).

4. In the subsection, "RNA secondary structure in the upstream proximal region inhibits pAs usage", authors said "This trend became more evident if we restricted our analysis to the annotated most distal pAs, which were in general of higher strength than proximal ones". Is there any figure or table which we can see these trends? If so, it should be referenced here.

R: We are sorry for the confusion. Actually we have showed the observation in Fig EV5, but forgot to cite in the text. As shown in Fig EV5, the red curve represents the level of RNA secondary structure around the annotated most distal pAs. In the updated manuscript, we cited the Fig EV5 at the end of this sentence (Page 14) and made it more clear at the figure legend as well (Page 45).

5. In figure 1D and F, please add actual numbers on top of the percentages.

R: We thank the reviewer for the suggestion. In the revised manuscript, we have added the actual numbers in Fig 1D and F.

Reviewer #2:

In this manuscript Xiao et al. perform a global analysis of alternative polyadenylation (APA) using fibroblasts from two divergent mouse lines as well as their F1 cross. They combine data from 3'READS and oligo-dT priming based 3' quantification to annotate and measure the relative expression of APA. They focus their work on cis-regulated APA events and investigate potential motives contributing to its regulation. The authors perform orthogonal confirmation of selected targets using a fluorescence based in vitro system and analyze the contribution of secondary structure and motives to APA usage.

General remarks:
The combined used of two different 3' quantification methods allow the authors to focus on median and high expressed APA events and remove from their analysis any APA event due to internal oligo-dT priming. The fact that alterations of core polyadenylation elements (eg. hexamer AAUAAA) impacts APA are not surprising. However the authors use an elegant experimental design that allows them to distinguish between cis- and trans-regulated APA.

R: We thank the reviewer for her/his positive comments on our study.

Major points:
Due to the experimental designed used by the authors; I am surprised that they focus almost exclusively on the cis-regulated APA events. Adding a brief analysis of the trans-regulated APA events will significantly increase the interest of the paper and differentiate this work from other studies. For example, performing an hexamer analysis analogue to the one that the authors perform for the cis-regulated APA events. The authors could also study if different RNA Binding Proteins or miRNAs are putatively bound (or in proximity) to the alternative polyadenylated isoforms using available data (eg. PMID 23846655). And if so, analysis how is the expression of the putative RNA Binding Protein in the F1 cell line.

R: We thank the reviewer for the important suggestion. Following this suggestion, we applied a similar hexamer analysis to those trans-regulated pAs. In brief, we compared the frequency of all hexamers within 100nt upstream of the cleavage sites between trans-regulated pAs and controls. The control pAs were selected based on the following criteria: 1) pAs should have a minimum expression level, i.e. BL + SP > 10 reads; 2) pAs need to have a minimum pAs usage, i.e. BL + SP > 10%; 3) in the comparison between two parental strains, Benjamini-Hochberg-adjusted *P* value > 0.5 and delta percentage of pAs usage < 0.05.

As shown in Fig R1 A, no hexamers shows significantly biased frequency between the two groups. Moreover, as the reviewer suggested, we also downloaded both the RBP binding motifs (PMID 23856655) and predicted miRNA binding sites (TargetScan), then compared their frequencies between the control pAs and trans-regulated pAs. Again, we failed to observe any motifs showing significant bias.

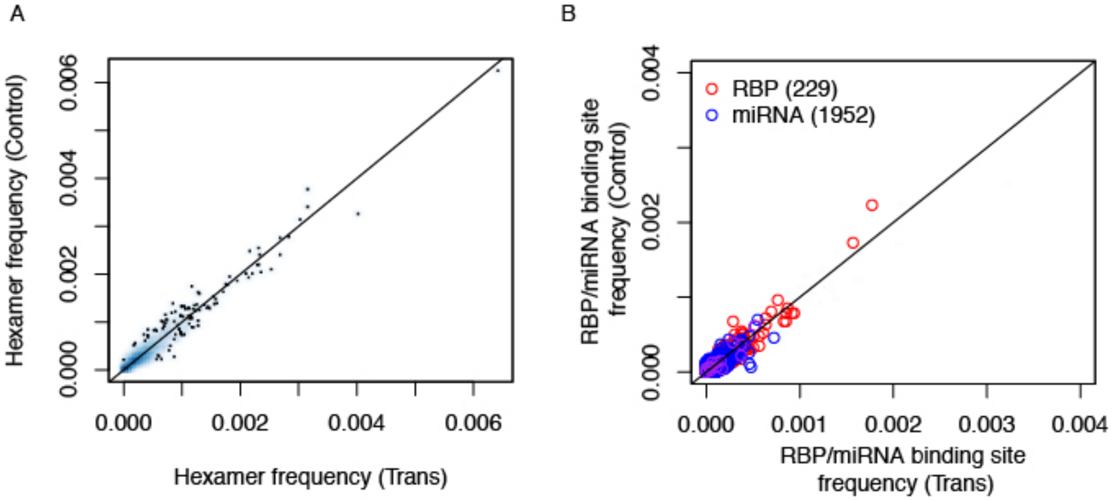

Fig R1: Scatterplot comparing the frequency of all hexamers (A) and RBP/miRNA binding sites (B) in the 100nt region upstream of cleavage sites between trans-regulatory pAs (X-axis) and control pAs (Y-axis).

Minor points:
In page 12-13 the authors briefly mention the method that they use for orthogonal confirmation (eg. Fig 3D and 4I). However, the description in the main text is too

brief. I would recommend adding a couple of sentences describing the general principle of the approach and how the artificial constructs are assayed in the same cell lines.

R: We thank the reviewer for the suggestion. In the revised manuscript, we described the general principle of the approach and how the artificial constructs are assayed in cell line (Page 13).

Some small typos in the figures (eg. in FigEV1D "Cleavage Site").

R: Thank the reviewer for helping us find the mistake. We have already checked and corrected the typos in Fig EV1D and other Figs as well.

2nd Editorial Decision 11 November 2016

Thank you again for submitting your work to Molecular Systems Biology. We have now evaluated the revised study and we think that the issues raised by the reviewers have been satisfactorily addressed. We would only ask you to include a couple of sentences in the main text referring to the analysis of trans-regulated pAs that was performed after the recommendation of reviewer #2.

2nd Revision - authors' response 13 November 2016

We thank you for your suggestion. Now, we added the analysis on trans-regulated pAs to the section "*Sequence motifs associated with pAs strength"* (Page 16) :

"Encouraged by the success of this motif analysis, we applied a similar hexamer analysis also to the trans-regulated pAs. Here, we compared the frequency of all hexamers within 100nt upstream of the cleavage sites between trans-regulated and control pAs without parental divergence. However, no hexamers showed significantly biased frequency between the two groups (Fig EV6F)."

In addition, the legend for Fig EV6F is also added on Page 46.

We hope that you find our revised manuscript now suitable for publication in Molecular Systems Biology.

3rd Editorial Decision 15 November 2016

Thank you again for sending us your revised manuscript. We are now satisfied with the modifications made and I am pleased to inform you that your paper has been accepted for publication.

EMBO PRESS

YOU MUST COMPLETE ALL CELLS WITH A PINK BACKGROUND ↓

PLEASE NOTE THAT THIS CHECKLIST WILL BE PUBLISHED ALONGSIDE YOUR PAPER

Corresponding Author Name: Wei Chen
Journal Submitted to: Molecular Systems Biology
Manuscript Number: MSB-16-7375

Reporting Checklist For Life Sciences Articles (Rev. July 2015)

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 χ2 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.

Please ensure that the answers to the following questions are reported in the manuscript itself. We encourage you to include a specific subsection in the methods section for statistics, reagents, animal models and human subjects.

In the pink boxes below, provide the page number(s) of the manuscript draft or figure legend(s) where the information can be located. Every question should be answered. If the question is not relevant to your research, please write NA (non applicable).

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://jjj.biochem.sun.ac.za
http://oba.od.nih.gov/biosecurity/biosecurity_documents.html
http://www.selectagents.gov/

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?	NA
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.	NA
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.	NA
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.	Fig 2C, DEXSeq, negative binomial distribution assumption was not assessed. Fig 3A, 3C, 3E and Fig4 E-G, Mann-Whitney U test,no distribution assumption. Fig 3D, t-test, the normal distribution assumption was not assessed.
Is there an estimate of variation within each group of data?	NA
Is the variance similar between the groups that are being statistically compared?	NA

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).	NA
7. Identify the source of cell lines and report if they were recently authenticated (e.g., by STR profiling) and tested for mycoplasma contamination.	No

* 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 accession codes for deposited data. See author guidelines, under '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	All the sequencing data generated from this study has been submitted to the European Nucleotide Archive (http://www.ebi.ac.uk/ena) under the accession number PRJEB15336.
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).	NA
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. As far as possible, primary and referenced data should be formally cited in a Data Availability section. Please state whether you have included this section. Examples: **Primary Data** Wetmore KM, Deutschbauer AM, Price MN, Arkin AP (2012). Comparison of gene expression and mutant fitness in Shewanella oneidensis MR-1. Gene Expression Omnibus GSE39462 **Referenced Data** Huang J, Brown AF, Lei M (2012). Crystal structure of the TRBD domain of TERT and the CR4/5 of TR. Protein Data Bank 4O26 AP-MS analysis of human histone deacetylase interactions in CEM-T cells (2013). PRIDE PXD000208	Primary Data Xiao MS, Zhang B, et al. (2016) Global analysis of regulatory divergence in the evolution of mouse alternative polyadenylation; PRJEB15336 Referenced Data Gao Q, Sun W, Ballegeer M, Libert C, Chen W (2015) Predominant contribution of cis-regulatory divergence in the evolution of mouse alternative splicing. Molecular systems biology 11: 816; ERP006913 Wan Y, Qu K, Zhang QC, Flynn RA, Manor O, Ouyang Z, Zhang J, Spitale RC, Snyder MP, Segal E, Chang HY (2014) Landscape and variation of RNA secondary structure across the human transcriptome. Nature 505: 706-709; SRA100457
22. 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 Biomodels (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.	NA

G- Dual use research of concern

23. 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