THE EMBO JOURNAL

Structural Basis for the Activation and Suppression of Transposition during Evolution of the RAG Recombinase

Yuhang Zhang, Elizabeth Corbett, Shenping Wu, and David Schatz
DOI: 10.15252/embj.2020105857

Corresponding author(s): David Schatz (david.schatz@yale.edu)

Review Timeline:

Submission Date:	5th Jun 20
Editorial Decision:	9th Jul 20
Revision Received:	17th Aug 20
Accepted:	20th Aug 20

Editor: Hartmut Vodermaier

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. Depending on transfer agreements, referee reports obtained elsewhere may or may not be included in this compilation. Referee reports are anonymous unless the Referee chooses to sign their reports.)

Thank you for submitting your manuscript on RAG transposase-to-recombinase evolution for our editorial consideration. We have now had it assessed by three expert referees, in light of whose positive comments we would be happy to offer publication of a revised version in The EMBO Journal.

As you will see from the reports copied below, most of the issues raised by the reviewers pertain to specific aspects of presentation/explanation, or are of curious/forward-looking nature, and I hope should therefore be straightforward to address within a limited revision period. Nevertheless, noting the current pandemic-related difficulties, I would of course be open for discussing any questions you may have regarding this revision or its timeline.

When preparing a revised manuscript, it would be great if you could already address various editorial aspects, as this should greatly facilitate our assessment at the time of resubmission.

REFEREE REPORTS

Referee #1:

The RAG recombinase (the RAG1 - RAG2 protein complex) initiates V(D)J recombination in the jawed-vertebrate immune system to establish a versatile response against a wide range of threats. Previous work has demonstrated that RAG has evolved from transposase enzymes and cuts DNA in a similar way. However, unlike transposases, RAG does not normally reintegrate excised DNA segments in the genome. Given the importance of V(D)J recombination and the potential harm of DNA reinsertion, understanding the evolutionary pathways and molecular principles of the suppression of RAG's integration activity is of high fundamental and medical relevance. Previous studies identified a specific mutation in RAG1, which reinstates transposition activity by switching a conserved arginine residue to the ancestral methionine found in RAG-like transposases. In this manuscript, the authors take advantage of this mutation (R848M) to shed light onto the mechanism of integration inhibition in RAG. They present two novel cryoEM structures of mouse Rag R848M bound to a DNA oligonucleotide that mimics the product of recombination signal sequence (RSS) insertion in a target DNA molecule (the strand-transfer complex, STC). By careful sub-classification of the cryoEM data they further report a previously undescribed structure with intact target DNA and disintegrated RSS (the target-capture complex, TCC). The structures reveal intriguing DNA architectures with remarkable bending and melting in the target DNA and implicate a key role for the 848 residue in promoting DNA deformations.
In agreement with previous reports, the authors suggest that formation of an integration-competent target DNA conformation presents an energetic barrier to integration and a methionine at position 848 in RAG-like transposases is critical to drive target deformation for efficient DNA insertion. Mutation of M848 to arginine during RAG evolution has likely contributed to a loss of integration activity so as to protect genome integrity.
The manuscript builds on significant previous findings by the authors and other groups in the field and greatly contributes to understanding RAG function and evolution.
The results are novel and well-presented, according to current standards in the field. The experimental designs (including protein and DNA variants used) are clever and the structures are very interesting. The manuscript is also well structured, clear and nicely supported with figures.
Together, the data provide important insights into RAG's molecular evolution and the mechanisms that prevent RSS reintegration during V(D)J recombination. Undoubtedly, this work will be of broad interest, and I only have a few suggestions, which the authors may wish to consider:

1. Introduction: It would be helpful to briefly describe the potential genomic impact of RSS integration and its relevance to disease to better highlight the significance of the work up front.
2. Page 9: Given that disintegration is enhanced in manganese at higher temperature, would it be possible to prepare a cleaner (more homogeneous) TCC structure using those conditions? This could help avoid "contamination" from STC in the TCC structure and result in higher resolution cryoEM maps. Such improved maps would be helpful to better support the observation of base flipping in the target DNA.
3. Page 10: For transposition, it seems counterproductive to flip out the C-t-2 base in the TCC such that it blocks the attack of the RSS 3'OH on the scissile phosphate, thereby preventing integration. Could the authors elaborate on if and how this could make sense for transposase function? In the manuscript, they argue that base flipping must be reverted prior to integration, but this would waste energy and potentially destabilize the bent target DNA structure.
4. Page 10, paragraph starting on line 195: Arginine has been shown to promote base flipping in other enzymes (including transposases), despite its positive charge. Why would it not be able to do the same in RAG? Can the authors explain this more clearly or test base flipping directly, for example, using a 2-aminopurine probe in the DNA?
5. Notably, mutation of R848 to alanine also considerably increased transposition in previous studies, whereas the R848 to leucine (a larger hydrophobic sidechain) mutation had less impact. Based on the authors arguments at the bottom of page 10, the opposite would be expected. Is it possible that in case of the R848A mutation, a neighboring conserved residue, M849 complements for M848's DNA bending function? Such complementation may be prohibited by the reduced flexibility of the protein backbone with bulkier residues such as leucine. It would be interesting to discuss the structural features of M849 in this context.
6. Also, can the sulfur atom of methionine provide a specific advantage for DNA bending (wedging and/or base flipping)? Are there other examples, where methionine plays similar roles in DNA deformations?
7. Page 12, line 239: The deletion in the RAG2 loop includes two highly conserved residues. Can this have any implications on RAG structure or function, besides stimulating integration? On a related note, it seems that R848 in RAG1 and the 333-342 loop in RAG2 have synergistic effect in inhibiting integration. Do these two protein regions interact?

Minor points and typos:
- The term "transposase-activated" sounds confusing. Perhaps "transposition-competent" or "integration-activated" would be better.
- Abstract, line 6: Please replace "donor" with "signal end" or similar. In the transposition literature, donor usually refers to the flanking DNA where the transposon is excised from (i.e. the DNA that 'donates' the transposon).
- Fig EV1A is a very useful figure. Would it be possible to move at least panels C and D to the main text, perhaps together with a simplified version of EV2?
- Page 5, line 77: The Mos1 STC structure should be cited here as well.
- Page 5, line 81: Please state here and in the Results section that the work was done with mouse RAG.
- Page 7, line 119: Please specify "very similar" by stating the RMSD. Are there any notable changes?
- Page 7, line 126: "near atomic-resolution" probably refers to the density maps here, not to the models. Obviously, all models are of atomic resolution.
- Page 8, line 133 - 135: This hypothesis was proposed previously, please cite Chen et al 2020b.
- Page 8, line 148 - 150: Please refer to the figure showing the cryoEM map of the Dynamic STC structure to better illustrate base flipping. Is it Fig EV4M?
- Page 11, line 212: citation to Chen et al 2020b is missing.

- Page 12, line 226: is the extrahelical position of the C-t-2 identical in the STC and TCC?
- Page 12, line 244: Probably figure references should be EV7J instead of S7J.
- Page 15, lines 307 - 309: I am a bit confused about this statement. Given the reduced number of hydrogen bonds, AT-rich sequences should be easier to melt.
- Page 18, line 370: "500mg of pTT5MP-RAG1", should this be 500 µg?
- Page 19, line 386: "Full length (FL) histidine-tagged human HMGB1" - the full length protein has not mentioned in any of the experiments. Perhaps it is worth stating also on page 6, line 101 that HMGB1ΔC was used.
- Page 20, line 410-411: "10% (v/v) PreScission Protease", please specify the protease concentration.

Referee #2:

The manuscript by Zhang et al. presents cryo-EM structures of RAG-mediated transposition intermediates, along with experiments to test some resulting mechanistic hypotheses, and speculation about the evolutionary implications of the data.

Overall, I think that this is an excellent manuscript. The structural and accompanying data provide new insights into aspects of the biochemistry of RAG1/RAG2, and should be of substantial interest for those in this field of research. The data are of high quality and as far as I can tell, the experimental work and analysis are sound and very competently done. Furthermore, the manuscript is beautifully presented - written in concise clear English with almost no errors, and with clear, informative figures.

I should note that I do not have specialist expertise in cryo-EM data analysis, so I hope that other reviewers will comment on that, but it all seems fine to me. My comments below are all of a minor nature.

Specific comments
1. There are some nice introductory figures on general aspects of the system, but these are all relegated to supplementary material, so the first figure in the main section is the STC structure. I would suggest that something equivalent to the first two parts of Fig EV1 and all of Fig EV2 would be helpful for readers if included in the main body of the paper.

2. Lines 308 - 309. I wasn't very convinced by the explanation for the choice of a GC-rich target as being that 'GC basepairs have a particularly high breathing rate". One could make the opposite argument, that a 5-bp run of GC basepairs would be a very bad target, because of the higher stability/resistance to melting of the double helix.

3. Line 505, amino acids (not acid). Line 517, cells (not cell).

4. Figure 2. Should the legend state which integration site (i.e. 12RSS end?) is being shown here?

5. Figure 3C, D. The images show similarity of the target DNA in the STC and TCC complexes, but it might not be very clear to the reader what the point being made is. Would it be helpful to indicate somehow the places where the two structures differ significantly (e.g. with shading or boxes, and

refer to later figures showing the differences in more detail)? Also, it might help to show the rest of the DNA in the complexes, in a different colour.

6. Figure 4 (and text lines 183 - 185 etc.). It's not very obvious to me how the two conformations discussed here were deduced from the structural data. Some additional information or clarification of the analysis might be useful.

7. Figure 6A. The structural similarity between RAG2 and BbeRAG2L is not apparent in this figure. Is it possible to render the two images such that any structural similarity is more obvious?

8. Figure EV7I. I had to zoom in a long way to see the colonies on the plates clearly. It might be better to show representative areas rather than the whole plates.

9. Figure EV7H. I don't really understand this diagram. It would benefit from more explanation in the legend. Reading the relevant part of Materials and Methods didn't make it completely clear either. I'm presuming that the donor plasmid is Tet-resistance (from its name), and that the acceptor is kanamycin-resistance, so an integration product has both of these. What is the streptomycin for in the 'positive' (not the control) plates? (line 503)

Referee #3:

Schatz-EMBOJ-2020-105857

Assembly of functional antigen receptor genes in developing lymphocytes is the basis of adaptive immunity in vertebrates. In jawed vertebrates, this process is referred to as V(D)J recombination, because it brings together 2-3 distinct genomic elements into one unit that encodes the antigen binding site. Hence, the interest in the mechanism of V(D)J recombination is intense and a major focus of molecular immunology. Dr Schatz not only has discovered the genes responsible for this process, called RAG1 and RAG2, he has also provided critical insight into their evolution, and more recently into the precise molecular mechanism of programmed DNA rearrangement. Together, these studies tell a fascinating story of exaptation of genetic parasites, in the present context designated "molecular domestication", that forms the basis of a revolutionary reorganization of vertebrate immunity. From this perspective alone, any new information about the molecular details is welcome, not only to better understand how V(D)J recombination works, and how human RAG-gene mutations impact this process, but also because of the broader implications for our understanding of the structural requirements for DNA transposition.

In the present study, Schatz and colleagues elaborate on a previously unresolved yet important aspect of RAG function in vertebrates. In earlier work they made the peculiar observation that a single amino acid residue (R848) of RAG1 is an important suppressor of transposition activity, but the mechanism remained unexplained. The biochemical and structural features of this site in the protein are the focus of the present paper. Their elegant approach is evolutionarily informed: They replace R848 with M848, which is the equivalent in all RAG-like proteins of non-vertebrates, resulting in a transposase-activated form of RAG1. This allows them to examine the structures of target capture and strand transfer complexes using the cryo-EM technique and to relate the data to their previous analyses.

The results are as beautiful as they are revealing: The structures collectively show that it is the

methionine residue that facilitates the unstacking of DNA bases, and other rearrangements of the target sequence, ultimately leading to the formation of a sharp bend, facilitating the strand transfer reaction. These results leave no doubt that the presence of an arginine residue in RAG1 (as it is found in the vertebrate version of the protein) counters all of these activities. In a nutshell, evolutionary tinkering didn´t need to be extensive to achieve such a remarkable reversal of functional activities.

In addition to this remarkable insight, the authors provide interesting clues as to how the previously identified acidic region of RAG2 inhibits the transposition process. This is shown to be due to an evolutionary novelty, namely a vertebrate-specific loop. By removing the tip of the loop, that is, preventing its interference with the RAG/DNA complex, increases transposition activity. It is a striking example that the exaptation uses different means to achieve the same end: here it is a steric effect, whereas the M/R exchange in RAG1 specifically comes down to a different side-chain.

This study leaves little if anything to criticize. It is exemplary for its clear exposition of the biological problem, which emerges from previous work. The results are clearly described, and the conclusions well supported by the presented data. In my view, the study convincingly combines structural and functional analyses to provide a number of interesting insights. Finally the discussion is an insightful summary of the evolutionary sequence underlying the domestication process of a toxic enzyme to the advantage of the host.

I have one minor suggestion that may help the reader follow the flow of the experiments and their conclusions more easily: The cartoon presented as fig. EV2 should become a main figure.

Response to reviewers

We thank the reviewers for the time and attention they devoted to the review of our manuscript. We were gratified by the reviewers' uniformly enthusiastic evaluations that emphasized the high quality of the study and the broad significance of the findings.

As detailed below, we have addressed all of the reviewers' specific comments, and in so doing, have made changes to the manuscript and to the organization of the figures. Specifically, figures EV1 and EV2 have been combined to become the new main Figure 1 and what is now main Figure 4 has been altered to make the presentation of the findings clearer. Relatively minor changes to the text have been made and most are specified in the responses below. We are very grateful for the reviewers' insightful suggestions which have allowed us to substantially improve the manuscript. Reviewers' specific comments are in blue font. We have not quoted the initial paragraphs of the reviewers' comments but have quoted any comment that made a suggestion or raised a question.

Reviewer 1:

1. Introduction: It would be helpful to briefly describe the potential genomic impact of RSS integration and its relevance to disease to better highlight the significance of the work up front.

A：We have addressed this by adding a phrase on page 4 that describes/references the potential genomic impact. We have not addressed disease relevance as this would represent an additional level of speculation.

2. Page 9: Given that disintegration is enhanced in manganese at higher temperature, would it be possible to prepare a cleaner (more homogeneous) TCC structure using those conditions? This could help avoid "contamination" from STC in the TCC structure and result in higher resolution cryoEM maps. Such improved maps would be helpful to better support the observation of base flipping in the target DNA.

A: A good suggestion that we actually tried. However, we noticed that when disintegration was performed at higher temperature, the new target DNA generated by the reaction was released from the protein complex efficiently and that the TCC was not stable. Analysis of the cryo-EM maps indicated the absence of the TCC. We suspect that this is the reason why no one has been able to purify the TCC by assembling it directly from its constituents even after many years of effort.

3. Page 10: For transposition, it seems counterproductive to flip out the C-t-2 base in the TCC such that it blocks the attack of the RSS 3'OH on the scissile phosphate, thereby preventing integration. Could the authors elaborate on if and how this could make sense for transposase function? In the manuscript, they argue that base flipping must be

reverted prior to integration, but this would waste energy and potentially destabilize the bent target DNA structure.

A: A reasonable explanation for this observation is the temperature at which we purified and assembled the complex and froze the grid (4°C). We think that C-t-2 and Met848 adopt an energetically favorable configuration that we are able to observe at this low temperature, perhaps explaining why reaction efficiency is quite low at this temperature. At higher temperature, DNA in the local area would be more dynamic, particularly in the presence of Met848, allowing exchange between multiple configurations, some of which would be compatible with strand transfer. It's interesting to consider the possibility that base flipping of C-t-2 represents a distinct mechanism to suppress transposition that Met848 also helps to overcome. It seems premature to speculate on all this in the manuscript.

4. Page 10, paragraph starting on line 195: Arginine has been shown to promote base flipping in other enzymes (including transposases), despite its positive charge. Why would it not be able to do the same in RAG? Can the authors explain this more clearly or test base flipping directly, for example, using a 2-aminopurine probe in the DNA?

A: Yes, we agree that Arg can cause base flipping. We would raise three relevant points. First, in most of cases, the Arg side chain replaces the base inside the helix, with other residues nearby to stabilize the conformation of both the Arg and the DNA. However, in the RAG TCC, it is not obvious how such stabilization would be achieved, particularly of the flipped C-t-2. Secondly, RAG1 R848 (WT) possesses weak transposition activity in vitro, so even Arg is likely to accomplish base flipping at low efficiency. Third, as noted in the manuscript, we propose that due to its hydrophobicity, Met is particularly effective at destabilizing the target DNA, thereby accelerating the reaction. Putting 2-aminopurine in the target DNA is a reasonable suggestion and would presumably yield the same result (increased transposition efficiency) as seen by us and others using target DNA containing mismatch regions as small as one bp.

5. Notably, mutation of R848 to alanine also considerably increased transposition in previous studies, whereas the R848 to leucine (a larger hydrophobic sidechain) mutation had less impact. Based on the authors arguments at the bottom of page 10, the opposite would be expected. Is it possible that in case of the R848A mutation, a neighboring conserved residue, M849 complements for M848's DNA bending function? Such complementation may be prohibited by the reduced flexibility of the protein backbone with bulkier residues such as leucine. It would be interesting to discuss the structural features of M849 in this context.

A: A very good suggestion that we had not considered, thank you. We think that transposition activity is highest with M848 because Met has the longest side chain of the hydrophobic residues. But the reviewer is correct that when position 848 is Ala,

M847 has the possibility of compensating for the function of M848 to some extent. We have added text that addresses this possibility at the top of page 11.

6. Also, can the sulfur atom of methionine provide a specific advantage for DNA bending (wedging and/or base flipping)? Are there other examples, where methionine plays similar roles in DNA deformations?

A: The sulfur in M848 may interact with the aromatic ring of the base in signal end DNA using its unshared electron to interact with the π electron cloud, which might be predicted to stabilize the strand transfer product and inhibit disintegration. However, disintegration activity also very high with M848, so it is not clear that this interaction makes a major contribution to the reaction.

7. Page 12, line 239: The deletion in the RAG2 loop includes two highly conserved residues. Can this have any implications on RAG structure or function, besides stimulating integration? On a related note, it seems that R848 in RAG1 and the 333-342 loop in RAG2 have synergistic effect in inhibiting integration. Do these two protein regions interact?

A: In the sequence alignment, these two residues do show considerable conservation, but in the structures, they are shifted relative to one another between RAG2 and ProtoRAG2, so the functional significance of the conservation is unclear. We don't see an obvious interaction between RAG1 M848 and the RAG2 loop (closest approach is > 8 Å). Given the extremely strong inhibition of transposition due to R848, we think it is hard to conclude that the two modifications are actually synergistic in their effects.

Minor points and typos:

- The term "transposase-activated" sounds confusing. Perhaps "transposition-competent" or "integration-activated" would be better.

Changed to "integration-activated".

- Abstract, line 6: Please replace "donor" with "signal end" or similar. In the transposition literature, donor usually refers to the flanking DNA where the transposon is excised from (i.e. the DNA that 'donates' the transposon).

Replaced with "transposon ends"

- Fig EV1A is a very useful figure. Would it be possible to move at least panels C and D to the main text, perhaps together with a simplified version of EV2?

A: We agree. As a result of this suggestion and the need to reduce the number of EV figures to five, we have created a new Figure 1 which combines most of the elements of

the original EV1 and EV2. This has the added benefit of reducing the number of EV figures to five, as requested by the Editor.

- Page 5, line 77: The Mos1 STC structure should be cited here as well.

Thank you for pointing this out. We have added two references relating to Mos1.

- Page 5, line 81: Please state here and in the Results section that the work was done with mouse RAG.

Good suggestion. Done.

- Page 7, line 119: Please specify "very similar" by stating the RMSD. Are there any notable changes?

RMSD is 0.45Å, and this information has been added as suggested.

- Page 7, line 126: "near atomic-resolution" probably refers to the density maps here, not to the models. Obviously, all models are of atomic resolution.

Thank you for catching this; we have removed the phrase "near atomic resolution".

- Page 8, line 133 - 135: This hypothesis was proposed previously, please cite Chen et al 2020b.

Done.

- Page 8, line 148 - 150: Please refer to the figure showing the cryoEM map of the Dynamic STC structure to better illustrate base flipping. Is it Fig EV4M?

Done.

- Page 11, line 212: citation to Chen et al 2020b is missing.

Reference has been added.

- Page 12, line 226: is the extrahelical position of the C-t-2 identical in the STC and TCC?

The C-t-2 base adopts different positions in the STC and the TCC. This is not relevant to the sentence mentioned by the reviewer.

- Page 12, line 244: Probably figure references should be EV7J instead of S7J.

This has been corrected. Note that all EV figures have been renumbered in the revised manuscript.

- Page 15, lines 307 - 309: I am a bit confused about this statement. Given the reduced number of hydrogen bonds, AT-rich sequences should be easier to melt.

We agree, this seems paradoxical. The Dornberger et al study that we cite notes that while isolated GC base pairs have slower breathing rates than AT base pairs (as expected), their findings show that GC base pairs within GC tracts have "unusually rapid base pair dynamics". The GC tracts examined by Dornberger et al were 4-6 bp in length, spanning the RAG transposition target site of 5 bp. We had failed to specify "GC tracts" in our original statement of the finding, but now do so. To our knowledge, the underlying physical explanation for this rapid breathing of GC bp in GC tracts has not been established; regardless, we feel it could be relevant to the preference of RAG for GC-rich target sites.

- Page 18, line 370: "500mg of pTT5MP-RAG1", should this be 500 μg?

Yes, it is 500 ug. Thank you for catching this. This has been corrected.

- Page 19, line 386: "Full length (FL) histidine-tagged human HMGB1" - the full length protein has not mentioned in any of the experiments. Perhaps it is worth stating also on page 6, line 101 that HMGB1ΔC was used.

The disintegration reactions in Figure 3 use full length HMGB1. We have specified this in the figure legend and have specified on page 6 that HMGB1ΔC was used for the structural analysis.

- Page 20, line 410-411: "10% (v/v) PreScission Protease", please specify the protease concentration.

Done.

Reviewer 2

Specific comments:

1. There are some nice introductory figures on general aspects of the system, but these are all relegated to supplementary material, so the first figure in the main section is the STC structure. I would suggest that something equivalent to the first two parts of Fig EV1 and all of Fig EV2 would be helpful for readers if included in the main body of the paper.

We agree, and as noted in our response to Reviewer 1, have moved almost all of EV1 and 2 to create a new main Figure 1.

2. Lines 308 - 309. I wasn't very convinced by the explanation for the choice of a GC-rich target as being that 'GC basepairs have a particularly high breathing rate". One could make the opposite argument, that a 5-bp run of GC basepairs would be a very bad target, because of the higher stability/resistance to melting of the double helix.

This has been addressed in our response to Reviewer 1. The high breathing rate is seen with tracts of GC bp and that is now explained in the text.

3. Line 505, amino acids (not acid). Line 517, cells (not cell).

We have corrected these errors; thank you for catching them.

4. Figure 2. Should the legend state which integration site (i.e. 12RSS end?) is being shown here?

Yes, thank you. We now specify that this is the 12RSS integration site. The 23RSS integration site exhibits similar structural features.

5. Figure 3C, D. The images show similarity of the target DNA in the STC and TCC complexes, but it might not be very clear to the reader what the point being made is. Would it be helpful to indicate somehow the places where the two structures differ significantly (e.g. with shading or boxes, and refer to later figures showing the differences in more detail)? Also, it might help to show the rest of the DNA in the complexes, in a different colour.

An excellent suggestion. We have created a new Figure (now Figure 4) that shows a zoom in on the DNA and depicts the important differences between the STC and TCC. This will indeed be helpful for readers.

6. Figure 4 (and text lines 183 - 185 etc.). It's not very obvious to me how the two conformations discussed here were deduced from the structural data. Some additional information or clarification of the analysis might be useful.

We see density that is consistent with two different positions for the M848 side chain, and that is stated in the text. We think that the figure (particularly what is now Fig. 5B) makes this clear, or at least as clear as we can make it. We have not made changes to the text or figure.

7. Figure 6A. The structural similarity between RAG2 and BbeRAG2L is not apparent in this figure. Is it possible to render the two images such that any structural similarity is more obvious?

We agree with the reviewer that the structural similarities are not as clear as one would like. Unfortunately, BbeRAG2L was not resolved at high resolution throughout, with the consequence that some of the beta strands can't be depicted clearly in the model. The two structures are shown from the same perspective and we don't think we can make the similarity clearer given the limitations in the BbeRAG2L structural data. We have added a sentence to the figure legend that explains this.

8. Figure EV7I. I had to zoom in a long way to see the colonies on the plates clearly. It might be better to show representative areas rather than the whole plates.

We have deleted Fig. EV7H, the schematic diagram of the plasmid-to-plasmid transposition assay, from this EV figure to make room for an enlarged version of the panel in question in which the colonies can be seen more clearly. See new Fig. EV5H.

9. Figure EV7H. I don't really understand this diagram. It would benefit from more explanation in the legend. Reading the relevant part of Materials and Methods didn't make it completely clear either. I'm presuming that the donor plasmid is Tet-resistance (from its name), and that the acceptor is kanamycin-resistance, so an integration product has both of these. What is the streptomycin for in the 'positive' (not the control) plates? (line 503)

We agree that this wasn't very clear and have deleted this figure panel. The details of the assay are not vital for this study. Readers who are interested in the details of the plasmid-to-plasmid transposition assay can refer back to our previous paper (Zhang et al., 2019) where a diagram is provided. This reference is now provided both in the relevant figure legend (new Figure 7D) and in Methods.

Reviewer 3

I have one minor suggestion that may help the reader follow the flow of the experiments and their conclusions more easily: The cartoon presented as fig. EV2 should become a main figure.

As described above, EV2 has been incorporated into main Figure 1, as suggested.

Accepted **20th Aug 2020**

Thank you for submitting your final revised manuscript for our consideration. I have now had a chance to look through it and to assess your responses to the comments raised by the original reviewers, and I am happy to inform you that we have now accepted it for publication in The EMBO Journal.

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: David Schatz
Journal Submitted to: EMBO Journal
Manuscript Number: EMBOJ-2020-105857

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 iustified
- 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.

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.

USEFUL LINKS FOR COMPLETING THIS FORM

http://www.antibodypedia.com
http://1degreebio.org
http://www.equator-network.org/reporting-guidelines/improving-bioscience-research-repor

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-tumo

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?	For experiments where statistics were applied, at least three independent experiments (biological replicates) were performed.
1.b. For animal studies, include a statement about sample size estimate even if no statistical methods were used.	N/A
2. Describe inclusion/exclusion criteria if samples or animals were excluded from the analysis. Were the criteria pre-established?	N/A
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.	N/A
For animal studies, include a statement about randomization even if no randomization was used.	N/A
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.	N/A
4.b. For animal studies, include a statement about blinding even if no blinding was done	N/A
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.	Where sufficient data points were available for testing of normal distribution (as in Fig. 7D), the Kolmogorov-Smirnov Test of Normality was used to ensure that the data did not differ significantly from that which was normally distributed. For other relevant figures, the data points are well clustered and no test to determine normal distribution was run.
Is there an estimate of variation within each group of data?	Yes

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).	N/A
7. Identify the source of cell lines and report if they were recently authenticated (e.g., by STR profiling) and tested for mycoplasma contamination.	HEK293T cells were obtained from ATCC and were not tested for mycoplasma contamination.

* 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.	N/A
9. For experiments involving live vertebrates, include a statement of compliance with ethical regulations and identify the committee(s) approving the experiments.	N/A
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.	N/A

E- Human Subjects

11. Identify the committee(s) approving the study protocol.	N/A
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.	N/A
13. For publication of patient photos, include a statement confirming that consent to publish was obtained.	N/A
14. Report any restrictions on the availability (and/or on the use) of human data or samples.	N/A
15. Report the clinical trial registration number (at ClinicalTrials.gov or equivalent), where applicable.	N/A
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.	N/A
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.	N/A

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	The structural coordinates data have been deposited to the Protein Data Bank. The associated density maps were deposited in the Electron Microscopy DataBank.
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).	N/A
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).	N/A
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 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.	N/A

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.	No.