THE ROYAL SOCIETY PUBLISHING

PROCEEDINGS B

Comparative genomics reveals high rates of horizontal transfer and strong purifying selection on rhizobial symbiosis genes

Brendan Epstein and Peter Tiffin

Article citation details
Proc. R. Soc. B **288**: 20201804.
http://dx.doi.org/10.1098/rspb.2020.1804

Review timeline

Original submission:	27 July 2020
1st revised submission:	20 October 2020
2nd revised submission:	4 December 2020
Final acceptance:	8 December 2020

Note: Reports are unedited and appear as submitted by the referee. The review history appears in chronological order.

Review History

RSPB-2020-1804.R0 (Original submission)

Review form: Reviewer 1 (Ellen Simms)

Recommendation
Accept with minor revision (please list in comments)

Scientific importance: Is the manuscript an original and important contribution to its field?
Excellent

General interest: Is the paper of sufficient general interest?
Excellent

Quality of the paper: Is the overall quality of the paper suitable?
Good

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
Yes

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
No

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
Yes

Is it clear?
Yes

Is it adequate?
Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
See my comments in the attached file. (See Appendix A)

Review form: Reviewer 2

Recommendation
Accept with minor revision (please list in comments)

Scientific importance: Is the manuscript an original and important contribution to its field?
Excellent

General interest: Is the paper of sufficient general interest?
Excellent

Quality of the paper: Is the overall quality of the paper suitable?
Good

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
No

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
Yes

Is it clear?
Yes

Is it adequate?
Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
Thank you for the opportunity to review the manuscript "Comparative genomics reveals high rates of horizontal transfer and strong purifying selection on rhizobial symbiosis genes" by Brendan Epstein and Peter Tiffin. In this work, the authors compared the genome sequences of 27 strains of rhizobia to estimate the frequency of horizontal transfer of non-symbiotic and symbiotic genes falling into several functional classes. They then carry out statistical analyses to infer the forms of selection acting on symbiotic and non-symbiotic genes. I think this is important work that brings novel insights into the evolutionary dynamics of this keystone symbiosis. The work should be of interest to a wide readership of ecologists and evolutionary biologists.

The writing is lucid and appealing. I only have a few editorial suggestions, listed below.

I am not an expert in these sorts of comparative genomic analyses. As far as I could tell, the methods were sound. I note that they used very recent, presumably cutting-edge programs for many analyses, such as finding putatively horizontally transferred genes. I was a bit surprised that there was no mention of gene location, i.e. on plasmids versus the main chromosome. Surely plasmid genes would be expected to have higher rates of HGT?

My only major suggestions for improvement have to do with acknowledging and discussing some caveats in interpretation. In particular, I think some discussion of the definitions of the symbiosis gene categories is warranted. For example, some of the core fixation genes must also affect host benefit? In other words, I suspect some genes could be categorized in multiple ways, meaning the categories are not absolutely distinct and there is some subjectivity in the categorization and resulting comparisons. Furthermore, the "benefits derived" categories (host benefit and rhizobial fitness) seem particularly speculative to me, since they are based on GWAS studies in only a single rhizobium-host combination. The same genes may be unrelated to host or bacterial benefit in other genetic backgrounds and other genes that do impact benefits in other rhizobium-host combinations have likely been mis-characterized as "non-symbiosis" genes. I doubt these potential sources of error would overturn the major findings of this study, but I would like to see an acknowledgement of these issues.

I also wonder whether drawing conclusions about nodulation genes having reached selective optima rather than co-evolving (particularly emphasized in the Abstract) is justified, given that fitness landscapes were not directly measured in this study. The findings of purifying selection on most nodulation genes, even those that have been transferred, is interesting and suggestive of evolutionary stasis. But additional studies that directly measure rhizobial fitness are needed to fully test this conclusion.

Comments by line number:
22 "involved IN initiating"
37-40 Though logical, this statement would benefit from a supporting reference.
43 I'm uncomfortable with describing the symbiosis as facultative. This could be debated endlessly.
64-66 I think the "confer a fitness advantage" claim is an oversimplification. New symbiosis genes can exclude prior hosts, which could confer a large cost. I.e. not all transfers will expand the host range.

110 Really the major clades of alphaproteobacterial rhizobia, not ALL major clades
113 What does "MCL" stand for?
232 What does "similar numbers of strains" mean here? Non-symbiosis genes present in similar numbers of strains?
250 The subscripts here are a bit confusing.
371 change "the approximately" to "approximately the"
Figure 1 Maybe remind the reader what delta means in the legend

Review form: Reviewer 3

Recommendation
Major revision is needed (please make suggestions in comments)

Scientific importance: Is the manuscript an original and important contribution to its field?
Acceptable

General interest: Is the paper of sufficient general interest?
Acceptable

Quality of the paper: Is the overall quality of the paper suitable?
Marginal

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
No

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
N/A

Is it clear?
N/A

Is it adequate?
N/A

Do you have any ethical concerns with this paper?
No

Comments to the Author
The authors present a study on the comparative genomic analysis of 27 alpha-rhizobia bacteria, with the aims to assess the extent of HT of symbiosis genes, the effect of HT on evolution of the symbiosis and how selection has contributed to the divergence of symbiosis genes. They show that nodulation genes are less conserved across the genomes analysed than nitrogen fixation

genes, that symbiotic genes experience more HT than non-symbiotic and that symbiotic genes appear to harbor signatures of purifying selection. There are a number of issues in the way in which these data have been interpreted and discussed. First, it would have been informative for the authors to provide a list of nodulation, n-fixation, GWA host and symbiont benefit genes that were identified in the study. There are only a relatively small numbers of these genes (65-10 for each category) and it would have really improved the interpretation of the results to know their identity. Second, the finding that symbiosis genes show a high rate of HT should be discussed in the context of these genes being encoded on mobile genetic elements (plasmids or symbiosis islands) that have been frequently shown to transfer via bacterial conjugation, while the bulk of non-symbiotic genes are encoded on bacterial chromosomes and therefore are not subject to conjugal transfer (although transformation and transduction will play some role). Third, the logic behind the selection of the genomes that were analysed is not clear. Granted that the authors have focused on strains where completed closed genomes are available, (some have been missed), but there appears to be no indication in the paper that the authors have considered the vast host-range differences of their selected strains. For example, NGR234 is an incredibly broad-host range strain capable of nodulating 112 different legume genera, while Mesorhizobium australicum (species named misspelt in FigS1) WSM2073 appears only able to nodulate one. These vast differences would no doubt have an impact on the suite of nodulation genes harboured by these strains and hence the analysis of the data. Four, On line 279, the authors state that "a typical annotated nodulation gene is found in less than half of the rhizobia genomes we examined, and only 11 of the 65 nodulation genes were found in all genomes". Presumably these 11 genes would encode the core Nod Factor synthesis genes, transcriptional regulators and NF secretion systems, which have been shown in many previous studies to be widely conserved across rhizobial genera? While the remaining 54 genes would likely be involved in modifying core NF? Many non-core NF modifying genes have been characterised previously (See Rodopothong et al., 2009 MPMI; Perret et al., 2000 MMB Reviews as an example) but this was not discussed in this study. Given that the strains chosen for this study represent rhizobia spanning 7 different genera, and that many of the strains within these genera have vastly different host ranges to each other, it is not surprising that the set of "essential" nodulation genes is relatively small. Finally, I draw the authors' attention to the fact that WSM2073 is the result of a HT event, having acquired its symbiosis genes directly from WSM1271, as reported in Haskett et al. 2016 PNAS.

Decision letter (RSPB-2020-1804.R0)

09-Sep-2020

Dear Dr Epstein:

Your manuscript has now been peer reviewed and the reviews have been assessed by an Associate Editor. The reviewers' comments (not including confidential comments to the Editor) and the comments from the Associate Editor are included at the end of this email for your reference. As you will see, the reviewers and the Editors have raised some concerns with your manuscript and we would like to invite you to revise your manuscript to address them.

We do not allow multiple rounds of revision so we urge you to make every effort to fully address all of the comments at this stage. If deemed necessary by the Associate Editor, your manuscript will be sent back to one or more of the original reviewers for assessment. If the original reviewers are not available we may invite new reviewers. Please note that we cannot guarantee eventual acceptance of your manuscript at this stage.

To submit your revision please log into http://mc.manuscriptcentral.com/prsb and enter your Author Centre, where you will find your manuscript title listed under "Manuscripts with

Decisions." Under "Actions", click on "Create a Revision". Your manuscript number has been appended to denote a revision.

When submitting your revision please upload a file under "Response to Referees" - in the "File Upload" section. This should document, point by point, how you have responded to the reviewers' and Editors' comments, and the adjustments you have made to the manuscript. We require a copy of the manuscript with revisions made since the previous version marked as 'tracked changes' to be included in the 'response to referees' document.

Your main manuscript should be submitted as a text file (doc, txt, rtf or tex), not a PDF. Your figures should be submitted as separate files and not included within the main manuscript file.

When revising your manuscript you should also ensure that it adheres to our editorial policies (https://royalsociety.org/journals/ethics-policies/). You should pay particular attention to the following:

Research ethics:
If your study contains research on humans please ensure that you detail in the methods section whether you obtained ethical approval from your local research ethics committee and gained informed consent to participate from each of the participants.

Use of animals and field studies:
If your study uses animals please include details in the methods section of any approval and licences given to carry out the study and include full details of how animal welfare standards were ensured. Field studies should be conducted in accordance with local legislation; please include details of the appropriate permission and licences that you obtained to carry out the field work.

Data accessibility and data citation:
It is a condition of publication that you make available the data and research materials supporting the results in the article. Please see our Data Sharing Policies (https://royalsociety.org/journals/authors/author-guidelines/#data). Datasets should be deposited in an appropriate publicly available repository and details of the associated accession number, link or DOI to the datasets must be included in the Data Accessibility section of the article (https://royalsociety.org/journals/ethics-policies/data-sharing-mining/). Reference(s) to datasets should also be included in the reference list of the article with DOIs (where available).

In order to ensure effective and robust dissemination and appropriate credit to authors the dataset(s) used should also be fully cited and listed in the references.

If you wish to submit your data to Dryad (http://datadryad.org/) and have not already done so you can submit your data via this link
http://datadryad.org/submit?journalID=RSPB&manu=(Document not available), which will take you to your unique entry in the Dryad repository.

If you have already submitted your data to dryad you can make any necessary revisions to your dataset by following the above link.

For more information please see our open data policy http://royalsocietypublishing.org/data-sharing.

Electronic supplementary material:
All supplementary materials accompanying an accepted article will be treated as in their final form. They will be published alongside the paper on the journal website and posted on the online figshare repository. Files on figshare will be made available approximately one week before the

accompanying article so that the supplementary material can be attributed a unique DOI. Please try to submit all supplementary material as a single file.

Online supplementary material will also carry the title and description provided during submission, so please ensure these are accurate and informative. Note that the Royal Society will not edit or typeset supplementary material and it will be hosted as provided. Please ensure that the supplementary material includes the paper details (authors, title, journal name, article DOI). Your article DOI will be 10.1098/rspb.[paper ID in form xxxx.xxxx e.g. 10.1098/rspb.2016.0049].

Please submit a copy of your revised paper within three weeks. If we do not hear from you within this time your manuscript will be rejected. If you are unable to meet this deadline please let us know as soon as possible, as we may be able to grant a short extension.

Thank you for submitting your manuscript to Proceedings B; we look forward to receiving your revision. If you have any questions at all, please do not hesitate to get in touch.

Best wishes,
Dr Locke Rowe
mailto: proceedingsb@royalsociety.org

Associate Editor
Comments to Author:
Your manuscript has been evaluated by three experts who find it has important and novel aspects. Two reviewers request several corrections and clarifications, and one reviewer asks for more substantial re-interpretation of results taking some additional information into account. Please note that Proceedings B does not allow multiple rounds of review, so you must make every effort to fully address all reviewers' concerns when preparing your revision.

Reviewer(s)' Comments to Author:
Referee: 1
Comments to the Author(s)
See my comments in the attached file.

Referee: 2
Comments to the Author(s)
Thank you for the opportunity to review the manuscript “Comparative genomics reveals high rates of horizontal transfer and strong purifying selection on rhizobial symbiosis genes" by Brendan Epstein and Peter Tiffin. In this work, the authors compared the genome sequences of 27 strains of rhizobia to estimate the frequency of horizontal transfer of non-symbiotic and symbiotic genes falling into several functional classes. They then carry out statistical analyses to infer the forms of selection acting on symbiotic and non-symbiotic genes. I think this is important work that brings novel insights into the evolutionary dynamics of this keystone symbiosis. The work should be of interest to a wide readership of ecologists and evolutionary biologists.

The writing is lucid and appealing. I only have a few editorial suggestions, listed below.

I am not an expert in these sorts of comparative genomic analyses. As far as I could tell, the methods were sound. I note that they used very recent, presumably cutting-edge programs for many analyses, such as finding putatively horizontally transferred genes. I was a bit surprised that there was no mention of gene location, i.e. on plasmids versus the main chromosome. Surely plasmid genes would be expected to have higher rates of HGT?

My only major suggestions for improvement have to do with acknowledging and discussing some caveats in interpretation. In particular, I think some discussion of the definitions of the symbiosis gene categories is warranted. For example, some of the core fixation genes must also affect host benefit? In other words, I suspect some genes could be categorized in multiple ways,

meaning the categories are not absolutely distinct and there is some subjectivity in the categorization and resulting comparisons. Furthermore, the "benefits derived" categories (host benefit and rhizobial fitness) seem particularly speculative to me, since they are based on GWAS studies in only a single rhizobium-host combination. The same genes may be unrelated to host or bacterial benefit in other genetic backgrounds and other genes that do impact benefits in other rhizobium-host combinations have likely been mis-characterized as "non-symbiosis" genes. I doubt these potential sources of error would overturn the major findings of this study, but I would like to see an acknowledgement of these issues.

I also wonder whether drawing conclusions about nodulation genes having reached selective optima rather than co-evolving (particularly emphasized in the Abstract) is justified, given that fitness landscapes were not directly measured in this study. The findings of purifying selection on most nodulation genes, even those that have been transferred, is interesting and suggestive of evolutionary stasis. But additional studies that directly measure rhizobial fitness are needed to fully test this conclusion.

Comments by line number:
22 "involved IN initiating"
37-40 Though logical, this statement would benefit from a supporting reference.
43 I'm uncomfortable with describing the symbiosis as facultative. This could be debated endlessly.
64-66 I think the "confer a fitness advantage" claim is an oversimplification. New symbiosis genes can exclude prior hosts, which could confer a large cost. I.e. not all transfers will expand the host range.
110 Really the major clades of alphaproteobacterial rhizobia, not ALL major clades
113 What does "MCL" stand for?
232 What does "similar numbers of strains" mean here? Non-symbiosis genes present in similar numbers of strains?
250 The subscripts here are a bit confusing.
371 change "the approximately" to "approximately the"
Figure 1 Maybe remind the reader what delta means in the legend

Referee: 3

Comments to the Author(s)
The authors present a study on the comparative genomic analysis of 27 alpha-rhizobia bacteria, with the aims to assess the extent of HT of symbiosis genes, the effect of HT on evolution of the symbiosis and how selection has contributed to the divergence of symbiosis genes. They show that nodulation genes are less conserved across the genomes analysed than nitrogen fixation genes, that symbiotic genes experience more HT than non-symbiotic and that symbiotic genes appear to harbor signatures of purifying selection. There are a number of issues in the way in which these data have been interpreted and discussed. First, it would have been informative for the authors to provide a list of nodulation, n-fixation, GWA host and symbiont benefit genes that were identified in the study. There are only a relatively small numbers of these genes (65-10 for each category) and it would have really improved the interpretation of the results to know their identity. Second, the finding that symbiosis genes show a high rate of HT should be discussed in the context of these genes being encoded on mobile genetic elements (plasmids or symbiosis islands) that have been frequently shown to transfer via bacterial conjugation, while the bulk of non-symbiotic genes are encoded on bacterial chromosomes and therefore are not subject to conjugal transfer (although transformation and transduction will play some role). Third, the logic behind the selection of the genomes that were analysed is not clear. Granted that the authors have focused on strains where completed closed genomes are available, (some have been missed), but there appears to be no indication in the paper that the authors have considered the vast host-range differences of their selected strains. For example, NGR234 is an incredibly broad-host range strain capable of nodulating 112 different legume genera, while Mesorhizobium australicum

(species named misspelt in FigS1) WSM2073 appears only able to nodulate one. These vast differences would no doubt have an impact on the suite of nodulation genes harboured by these strains and hence the analysis of the data. Four, On line 279, the authors state that "a typical annotated nodulation gene is found in less than half of the rhizobia genomes we examined, and only 11 of the 65 nodulation genes were found in all genomes". Presumably these 11 genes would encode the core Nod Factor synthesis genes, transcriptional regulators and NF secretion systems, which have been shown in many previous studies to be widely conserved across rhizobial genera? While the remaining 54 genes would likely be involved in modifying core NF? Many non-core NF modifying genes have been characterised previously (See Rodopothong et al., 2009 MPMI; Perret et al., 2000 MMB Reviews as an example) but this was not discussed in this study. Given that the strains chosen for this study represent rhizobia spanning 7 different genera, and that many of the strains within these genera have vastly different host ranges to each other, it is not surprising that the set of "essential" nodulation genes is relatively small. Finally, I draw the authors' attention to the fact that WSM2073 is the result of a HT event, having acquired its symbiosis genes directly from WSM1271, as reported in Haskett et al. 2016 PNAS.

Author's Response to Decision Letter for (RSPB-2020-1804.R0)

See Appendix B.

RSPB-2020-1804.R1 (Revision)

Review form: Reviewer 4 (Peter Young)

Recommendation
Major revision is needed (please make suggestions in comments)

Scientific importance: Is the manuscript an original and important contribution to its field?
Good

General interest: Is the paper of sufficient general interest?
Good

Quality of the paper: Is the overall quality of the paper suitable?
Good

Is the length of the paper justified?
Yes

Should the paper be seen by a specialist statistical reviewer?
No

Do you have any concerns about statistical analyses in this paper? If so, please specify them explicitly in your report.
No

It is a condition of publication that authors make their supporting data, code and materials available - either as supplementary material or hosted in an external repository. Please rate, if applicable, the supporting data on the following criteria.

Is it accessible?
Yes

Is it clear?
Yes

Is it adequate?
Yes

Do you have any ethical concerns with this paper?
No

Comments to the Author
This is the dullest article I have seen for a long time. A browsing reader will find just a single, small monochrome figure, from which they will deduce that the key point of the paper is the distribution of median delta in random genes. They will probably pass swiftly on in search of a more interesting paper. That would be a pity because, as earlier reviewers have said, this study actually addresses some interesting questions about the evolution of rhizobia, and of bacteria more generally.

I will not comment on the analyses, as previous reviewers have done this extensively and I think the authors have responded well to them, but the presentation really needs to be improved to help the reader to understand the work, especially if it is to made suitable for a journal with a wide readership like Proc Roy Soc B. Here are four points that I feel need attention.

1.
I found the manuscript frustrating to read because key information is missing and can only be found by poking around in the Supplementary Information. Importantly, the authors never tell us what organisms they are studying. They only used 27 genomes, so it would not be unreasonable to provide a list of them. In fact, Supplementary Figure S1 not only provides this information but also shows their degree of relatedness, so why not include it in the main manuscript? This is really important information because one would expect the probability of HT between two lineages to be highly dependent on their degree of relatedness, so the quantitative results of the study will be strongly influenced by the choice of genomes. It is immediately clear from Fig. S1 that a wide range of pairwise relatedness is covered, from strains that are virtually identical (apart from their symbiosis genes) to the largest divergences in the alpharhizobia.

2.
A major part of this study is concerned with the presence/absence of symbiosis-related genes. There are 143 such genes listed in Table S3, which is a perfectly feasible number to display in a figure. I suggest that the authors add, to the right of the phylogeny in Figure S1, a presence/absence matrix for these genes, colour coded by functional category and aligned with the relevant genome in the phylogeny. This would create an arresting figure that captured a central aspect of the study and provided immediate insight into the quantity and quality of the data that lie behind the dry statistics. An example of the sort of thing I have in mind would be Fig. 2 of Tian et al. (2012, https://doi.org/10.1073/pnas.1120436109).

3.
I am actually surprised that the authors do not cite this Tian et al. paper, because it has distinct parallels with their own study. Tian et al. sequenced 26 genomes of soybean symbionts and assessed the presence/absence of different symbiosis-related genes. They adopted a much wider definition of 'symbiosis-related', identifying 561 such genes, but these probably include most of those covered by Epstein and Tiffin. The focus of the analysis is different, so I think there is room for this new study, but some comparison of the two studies is needed.

4.
The phylogeny in Fig. S1 does not match the taxonomy. For example, neither Rhizobium etli nor the genus Rhizobium are monophyletic in the tree. This is not because the phylogeny is wrong, it is because the names are wrong. Strain IE4803 is not R. etli, it is R. sophoriradicis. IRBG74 is not Rhizobium sp., it is Agrobacterium pusense. Strain WSM2304 is not R. leguminosarum, it is R. acidisoli. The authors should check the strains they have used in the genome-based taxonomy databases at gtdb.ecogenomic.org and tygs.dsmz.de. It is never wise to trust the taxonomy assigned in GenBank.

Decision letter (RSPB-2020-1804.R1)

21-Nov-2020

Dear Dr Epstein:

Your manuscript has now been peer reviewed and a review has been assessed by an Associate Editor. The reviewers' comments (not including confidential comments to the Editor) and the comments from the Associate Editor are included at the end of this email for your reference. As you will see, the reviewer has raised some concerns with your manuscript and we would like to invite you to revise your manuscript to address them.

We do not allow multiple rounds of revision so we urge you to make every effort to fully address all of the comments at this stage. If deemed necessary by the Associate Editor, your manuscript will be sent back to one or more of the original reviewers for assessment. If the original reviewers are not available we may invite new reviewers. Please note that we cannot guarantee eventual acceptance of your manuscript at this stage.

To submit your revision please log into http://mc.manuscriptcentral.com/prsb and enter your Author Centre, where you will find your manuscript title listed under "Manuscripts with Decisions." Under "Actions", click on "Create a Revision". Your manuscript number has been appended to denote a revision.

When submitting your revision please upload a file under "Response to Referees" in the "File Upload" section. This should document, point by point, how you have responded to the reviewers' and Editors' comments, and the adjustments you have made to the manuscript. We require a copy of the manuscript with revisions made since the previous version marked as 'tracked changes' to be included in the 'response to referees' document.

Your main manuscript should be submitted as a text file (doc, txt, rtf or tex), not a PDF. Your figures should be submitted as separate files and not included within the main manuscript file.

When revising your manuscript you should also ensure that it adheres to our editorial policies (https://royalsociety.org/journals/ethics-policies/). You should pay particular attention to the following:

Research ethics:
If your study contains research on humans please ensure that you detail in the methods section whether you obtained ethical approval from your local research ethics committee and gained informed consent to participate from each of the participants.

Use of animals and field studies:
If your study uses animals please include details in the methods section of any approval and licences given to carry out the study and include full details of how animal welfare standards

were ensured. Field studies should be conducted in accordance with local legislation; please include details of the appropriate permission and licences that you obtained to carry out the field work.

Data accessibility and data citation:
It is a condition of publication that you make available the data and research materials supporting the results in the article (https://royalsociety.org/journals/authors/author-guidelines/#data). Datasets should be deposited in an appropriate publicly available repository and details of the associated accession number, link or DOI to the datasets must be included in the Data Accessibility section of the article (https://royalsociety.org/journals/ethics-policies/data-sharing-mining/). Reference(s) to datasets should also be included in the reference list of the article with DOIs (where available).

In order to ensure effective and robust dissemination and appropriate credit to authors the dataset(s) used should also be fully cited and listed in the references.

If you wish to submit your data to Dryad (http://datadryad.org/) and have not already done so you can submit your data via this link
http://datadryad.org/submit?journalID=RSPB&manu=(Document not available), which will take you to your unique entry in the Dryad repository.

If you have already submitted your data to dryad you can make any necessary revisions to your dataset by following the above link.

For more information please see our open data policy http://royalsocietypublishing.org/data-sharing.

Electronic supplementary material:
All supplementary materials accompanying an accepted article will be treated as in their final form. They will be published alongside the paper on the journal website and posted on the online figshare repository. Files on figshare will be made available approximately one week before the accompanying article so that the supplementary material can be attributed a unique DOI. Please try to submit all supplementary material as a single file.

Online supplementary material will also carry the title and description provided during submission, so please ensure these are accurate and informative. Note that the Royal Society will not edit or typeset supplementary material and it will be hosted as provided. Please ensure that the supplementary material includes the paper details (authors, title, journal name, article DOI). Your article DOI will be 10.1098/rspb.[paper ID in form xxxx.xxxx e.g. 10.1098/rspb.2016.0049].

Please submit a copy of your revised paper within three weeks. If we do not hear from you within this time your manuscript will be rejected. If you are unable to meet this deadline please let us know as soon as possible, as we may be able to grant a short extension.

Thank you for submitting your manuscript to Proceedings B; we look forward to receiving your revision. If you have any questions at all, please do not hesitate to get in touch.

Best wishes,
Dr Locke Rowe
Editor, Proceedings B
mailto: proceedingsb@royalsociety.org

Associate Editor
Board Member: 1
Comments to Author:
Your revised manuscript has been assessed by an expert who finds significant problems with presentation that must be addressed before it can be considered for publication by Proceedings B.

Reviewer(s)' Comments to Author:

Referee: 4

Comments to the Author(s)
The first point that struck me about this manuscript was that the visual presentation is not very appealing. A browsing reader will find just a single, small monochrome figure, from which they will deduce that the key point of the paper is the distribution of median delta in random genes. They will probably pass swiftly on in search of a more interesting paper. That would be a pity because, as earlier reviewers have said, this study actually addresses some interesting questions about the evolution of rhizobia, and of bacteria more generally.

I will not comment on the analyses, as previous reviewers have done this extensively and I think the authors have responded well to them, but the presentation really needs to be improved to help the reader to understand the work, especially if it is to made suitable for a journal with a wide readership like Proc Roy Soc B. Here are four points that I feel need attention.

1.
I found the manuscript frustrating to read because key information is missing and can only be found by poking around in the Supplementary Information. Importantly, the authors never tell us what organisms they are studying. They only used 27 genomes, so it would not be unreasonable to provide a list of them. In fact, Supplementary Figure S1 not only provides this information but also shows their degree of relatedness, so why not include it in the main manuscript? This is really important information because one would expect the probability of HT between two lineages to be highly dependent on their degree of relatedness, so the quantitative results of the study will be strongly influenced by the choice of genomes. It is immediately clear from Fig. S1 that a wide range of pairwise relatedness is covered, from strains that are virtually identical (apart from their symbiosis genes) to the largest divergences in the alpharhizobia.

2.
A major part of this study is concerned with the presence/absence of symbiosis-related genes. There are 143 such genes listed in Table S3, which is a perfectly feasible number to display in a figure. I suggest that the authors add, to the right of the phylogeny in Figure S1, a presence/absence matrix for these genes, colour coded by functional category and aligned with the relevant genome in the phylogeny. This would create an arresting figure that captured a central aspect of the study and provided immediate insight into the quantity and quality of the data that lie behind the dry statistics. An example of the sort of thing I have in mind would be Fig. 2 of Tian et al. (2012, https://doi.org/10.1073/pnas.1120436109).

3.
I am actually surprised that the authors do not cite this Tian et al. paper, because it has distinct parallels with their own study. Tian et al. sequenced 26 genomes of soybean symbionts and assessed the presence/absence of different symbiosis-related genes. They adopted a much wider definition of 'symbiosis-related', identifying 561 such genes, but these probably include most of those covered by Epstein and Tiffin. The focus of the analysis is different, so I think there is room for this new study, but some comparison of the two studies is needed.

4.
The phylogeny in Fig. S1 does not match the taxonomy. For example, neither Rhizobium etli nor the genus Rhizobium are monophyletic in the tree. This is not because the phylogeny is wrong, it

is because the names are wrong. Strain IE4803 is not R. etli, it is R. sophoriradicis. IRBG74 is not Rhizobium sp., it is Agrobacterium pusense. Strain WSM2304 is not R. leguminosarum, it is R. acidisoli. The authors should check the strains they have used in the genome-based taxonomy databases at gtdb.ecogenomic.org and tygs.dsmz.de. It is never wise to trust the taxonomy assigned in GenBank.

Author's Response to Decision Letter for (RSPB-2020-1804.R1)

See Appendix C.

Decision letter (RSPB-2020-1804.R2)

08-Dec-2020

Dear Dr Epstein

I am pleased to inform you that your manuscript entitled "Comparative genomics reveals high rates of horizontal transfer and strong purifying selection on rhizobial symbiosis genes" has been accepted for publication in Proceedings B.

You can expect to receive a proof of your article from our Production office in due course, please check your spam filter if you do not receive it. PLEASE NOTE: you will be given the exact page length of your paper which may be different from the estimation from Editorial and you may be asked to reduce your paper if it goes over the 10 page limit.

If you are likely to be away from e-mail contact please let us know. Due to rapid publication and an extremely tight schedule, if comments are not received, we may publish the paper as it stands.

If you have any queries regarding the production of your final article or the publication date please contact procb_proofs@royalsociety.org

Open Access
You are invited to opt for Open Access, making your freely available to all as soon as it is ready for publication under a CCBY licence. Our article processing charge for Open Access is £1700. Corresponding authors from member institutions (http://royalsocietypublishing.org/site/librarians/allmembers.xhtml) receive a 25% discount to these charges. For more information please visit http://royalsocietypublishing.org/open-access.

Your article has been estimated as being 9 pages long. Our Production Office will be able to confirm the exact length at proof stage.

Paper charges
An e-mail request for payment of any related charges will be sent out after proof stage (within approximately 2-6 weeks). The preferred payment method is by credit card; however, other payment options are available

Electronic supplementary material:
All supplementary materials accompanying an accepted article will be treated as in their final form. They will be published alongside the paper on the journal website and posted on the online

figshare repository. Files on figshare will be made available approximately one week before the accompanying article so that the supplementary material can be attributed a unique DOI.

Thank you for your fine contribution. On behalf of the Editors of the Proceedings B, we look forward to your continued contributions to the Journal.

Sincerely,

Dr Locke Rowe
Editor, Proceedings B
mailto: proceedingsb@royalsociety.org

Associate Editor:
Board Member
Comments to Author:
(There are no comments.)

Appendix A

Reviewer's comments for authors

Very interesting and important manuscript.

A hot topic in evolutionary biology is the drive to describe the population genetics of recombination via horizontal transfer (HT) in bacteria and to understand the circumstances under which evolution in these organisms will occur via HT and when it will occur via vertical transfer of mutations. Rhizobia provide excellent subjects for this research because their symbiosis genes exhibit strong evidence for such recombination.

The authors do a nice job introducing the questions and the study organisms. They provide good explanations of the observations supporting the HT theory for some kinds of genes. They also describe the complex life cycles of these symbionts and explain why genes important in the different life stages might exhibit different evolutionary patterns. They use comparative genomic analysis to test three different hypotheses that arise from the natural history of rhizobia. Two of these hypotheses regard the rates of HT across different kinds of genes. The third hypothesis uses more traditional genomic analysis to ask whether symbiosis genes have experienced positive selection, due to complex 3-party arms races, or purifying selection, due to stabilizing selection by host plants. These hypotheses are of great interest to a small group of biologists interested in the evolution of legume-rhizobium symbioses, but also important to a broad swath of evolutionary biologists seeking to understand evolutionary trajectories of all organisms experiencing particular kinds of selection pressures. Given the amount of genomic data now available on a wide array of rhizobia, and the importance of these questions, it's kind of remarkable that no one has done much of either of these kinds of analysis.

Specific comments:

Line 124: Calling these GWA genes was a bit disorienting. Why not call them "putative symbiosis benefit genes" of "candidate symbiosis benefit genes"?

Line 174: I'm curious to know what the implications are of excluding sequences with large Ks.

Lines 206 – 209: Is the number of nodulation gene copies in a genome was positively correlated with the genome size because symbiosis genes drive genome size? What causes this correlation?

Line 211: Should "GWAS" be "GWA"?

Line 231 – 233: Is gene duplication rate addressed in the discussion? It was not mentioned in the introduction.

Lines 245 – 247: "In fact, the difference between symbiotic and non-symbiotic gene Ka:Ks values is greater when we compare the symbiotic genes only to non-symbiotic genes with transfer rates that are greater than the genome-wide median." Is this discussed in the discussion?

Lines 248 – 251: Is a difference in Ka:Ks between 0.2 and 0.244 biologically important?

Lines 269 – 271: "each of the four classes of symbiosis genes are [is] less likely to be found in only a single genome, more likely to be [a] member[] of [a] larger gene famil[y], and more likely to be present in all genomes." Aside from the grammatical errors, why is this observation biologically interesting? It seems interesting, but isn't developed.

Lines 281 – 284: The idea that nodulation genes are involved in host specificity is not novel (it is well supported by observational data and elegant molecular genetic studies). Thus, rather than arguing that the sporadic appearance of nodulation genes only "supports the idea that the functional importance of at least some of these genes is host-specific," the authors could focus on the fact that these genes *are* thought to be host specific and *also* are not always present, which might tell us something interesting about the selection pressures they experience. For example, it might suggest that maintaining these genes in a genome is costly to bacterial fitness, such that these genes are quickly lost when they are not needed. Alternatively, such genes might disappear by genetic drift when they are not needed. Are there any analyses that the authors could do to distinguish between the fitness cost and fitness neutral hypotheses? Is there ever any evidence that these genes degrade (e.g. what happens to neutral genes in obligate symbioses) rather than just vanish entirely via some HT-associated process?

Lines 292 – 294: I'd like to see some discussion of why this pattern is evolutionarily interesting and what ecological situations might cause different patterns. For example, given that all the sequenced genomes came from rhizobia isolated from nodules, what patterns would you predict for rhizobia isolated directly from soil? I see that there is more true discussion (rather than reporting) of this result in lines 321 – 323.

Lines 297 – 299: As others have found, the authors found that nodulation genes experience strong purifying selection, suggestive of mutualistic stasis (i.e., legume host imposes stabilizing selection on nod genes by maintaining an evolutionarily stable receptor for detecting the bacterial symbiont). The key novelty of this manuscript's results is that the average nod gene Ka:Ks evolutionary rate has been compared with that of other symbiosis-related genes. Why are nodulation genes so conserved, given that they provide access to a presumably valuable resource (the host nodule)? Does this suggest that there are not "bad actors" (e.g. pathogens) seeking access to the host via this route. What characteristic makes this route so invulnerable? Why can't any other bacteria imitate rhizobial nodulation signaling? This is a fascinating evolutionary puzzle. Again, there is more *discussion* of this result in lines 328 – 330, which is good. However, I'd love to see the authors speculate a bit about why, in a kind of lock and key arms race, plants could experience positive selection on receptor genes yet bacteria experience purifying selection.

Lines 312 – 315: The authors also find that genes putatively involved in providing other symbiotic benefits to the host experience purifying selection. How can hosts impose purifying selection on rhizobium providing symbiotic benefits? Is this evidence that plants impose post-infection sanctions?

Lines 317 – 321: This paragraph is vague. What aspects of the previously described results are caused by the shifting identity of host species and which are caused by persistence of a host?

Appendix B

Associate Editor
Comments to Author:

Your manuscript has been evaluated by three experts who find it has important and novel aspects. Two reviewers request several corrections and clarifications, and one reviewer asks for more substantial re-interpretation of results taking some additional information into account. Please note that Proceedings B does not allow multiple rounds of review, so you must make every effort to fully address all reviewers' concerns when preparing your revision.

> Thank you for the opportunity to revise the paper and also for the three reviewers' generally overall positive assessments, constructive criticisms, and suggestions for improving the manuscript. We think the revised manuscript addresses the issues raised by the reviewers and hope that it is acceptable for publication in *Proc. B*. Below, we describe our response to the reviewers' comments.
>
> Sincerely,
>
> Brendan Epstein and Peter Tiffin

Referee: 1 (from the attached file)

Very interesting and important manuscript.

A hot topic in evolutionary biology is the drive to describe the population genetics of recombination via horizontal transfer (HT) in bacteria and to understand the circumstances under which evolution in these organisms will occur via HT and when it will occur via vertical transfer of mutations. Rhizobia provide excellent subjects for this research because their symbiosis genes exhibit strong evidence for such recombination.

The authors do a nice job introducing the questions and the study organisms. They provide good explanations of the observations supporting the HT theory for some kinds of genes. They also describe the complex life cycles of these symbionts and explain why genes important in the different life stages might exhibit different evolutionary patterns. They use comparative genomic analysis to test three different hypotheses that arise from the natural history of rhizobia. Two of these hypotheses regard the rates of HT across different kinds of genes. The third hypothesis uses more traditional genomic analysis to ask whether symbiosis genes have experienced positive selection, due to complex 3-party arms races, or purifying selection, due to stabilizing selection by host plants. These hypotheses are of great interest to a small group of biologists interested in the evolution of legume-rhizobium symbioses, but also important to a broad swath of evolutionary biologists seeking to understand evolutionary trajectories of all organisms experiencing particular kinds of selection pressures. Given the amount of genomic

data now available on a wide array of rhizobia, and the importance of these questions, it's kind of remarkable that no one has done much of either of these kinds of analysis.

We thank the reviewer for their generally positive assessment as well as their thoughtful and constructive suggestions for improving the manuscript.

Specific comments:

Line 124: Calling these GWA genes was a bit disorienting. Why not call them "putative symbiosis benefit genes" of "candidate symbiosis benefit genes"?

We appreciate the suggestion and have adopted the phrases "host benefit genes" and "rhizobia fitness genes".

Line 174: I'm curious to know what the implications are of excluding sequences with large Ks.

The Reviewer refers to our exclusion of ~ 20% of sequence pairs that differed at too many sites for Ks to be estimated reliably due to multiple evolutionary changes (i.e. multiple mutations / site). We have no reason to think that these sequence pairs are of particular biological interest and think they sequence differences simply reflect time since divergence. In response to the Reviewer's comment, we now note that these genes also tended to have very high Ka values (median Ka = 0.71 vs. 0.35 for genes with more reliable Ks estimates, line 184). In addition, Ks is only weakly correlated with Ka:Ks (r=0.22).

Lines 206 – 209: Is the number of nodulation gene copies in a genome was positively correlated with the genome size because symbiosis genes drive genome size? What causes this correlation?

We don't think the number of symbiosis genes drives genome size, and we certainly don't have any data to support that. Rather, as we mention in the revised manuscript (lines 214-216), we think that the correlation might result from among-lineage variation in the efficiency of removing non-essential / non-advantageous genes from their genomes (i.e. HT genes that do not confer a fitness advantage).

Line 211: Should "GWAS" be "GWA"?

We made this change, thanks.

Line 231 – 233: Is gene duplication rate addressed in the discussion? It was not mentioned in the introduction.

An earlier version of the manuscript included discussion of duplication rates in the Introduction and Discussion. However, we removed that material due to both space limitations and because we think it distracted from the major objectives of the manuscript. At the same time, we think it is important to report this result both as a basic description of the data and because it helps the reader evaluate the sources of increased gene family size—i.e. the relative contribution of HT vs. duplication vs. ancestral gene copies.

Lines 245 – 247: "In fact, the difference between symbiotic and non-symbiotic gene Ka:Ks values is greater when we compare the symbiotic genes only to non-symbiotic genes with transfer rates that are greater than the genome-wide median." Is this discussed in the discussion?

We included this information in the Results to show that the lower Ka:Ks values of symbiosis genes is not due to a bias associated with the higher HT rates of symbiosis genes. We think it is valuable to include this in case a reader is concerned about such potential bias. We don't have much more to say about the relationship. We do agree that the relationship between HT and evolutionary rates would be an interesting topic to pursue, but also think it is beyond the scope of this manuscript. Moreover, results from some preliminary analyses we conducted indicate our data are not extensive enough to come to any robust conclusions about the relationship between HT and Ka:Ks.

Lines 248 – 251: Is a difference in Ka:Ks between 0.2 and 0.244 biologically important?

That is an interesting question, and to be honest we're not sure. However, prompted in part by this comment, in the revised manuscript we put less emphasis on the purifying selection and more emphasis on the result that the symbiosis genes are clearly not evolving in response to strong positive selection (at least relative to the rest of the genome).

Lines 269 – 271: "each of the four classes of symbiosis genes are [is] less likely to be found in only a single genome, more likely to be [a] member[] of [a] larger gene famil[y], and more likely to be present in all genomes." Aside from the grammatical errors, why is this observation biologically interesting? It seems interesting, but isn't developed.

We think that this, and several comments below, resulted from us unnecessarily repeating results in the Discussion section. We think these basic descriptors of the data are valuable to the reader, but as the Reviewer notes, we don't have lot to say about them. While revising the manuscript we realized we did not need to repeat the description of the data in the Discussion and that doing so was distracting. As such, we

have deleted that material and focus on the interpretation of our results and their relevance for understanding the evolution of symbiosis.

*Lines 281 – 284: The idea that nodulation genes are involved in host specificity is not novel (it is well supported by observational data and elegant molecular genetic studies). Thus, rather than arguing that the sporadic appearance of nodulation genes only "supports the idea that the functional importance of at least some of these genes is host-specific," the authors could focus on the fact that these genes *are* thought to be host specific and *also* are not always present, which might tell us something interesting about the selection pressures they experience. For example, it might suggest that maintaining these genes in a genome is costly to bacterial fitness, such that these genes are quickly lost when they are not needed. Alternatively, such genes might disappear by genetic drift when they are not needed. Are there any analyses that the authors could do to distinguish between the fitness cost and fitness neutral hypotheses? Is there ever any evidence that these genes degrade (e.g. what happens to neutral genes in obligate symbioses) rather than just vanish entirely via some HT-associated process?*

We appreciate the reviewer's perspective and insight. Following their suggestion, we have revised this section to discuss the possible roles of drift and selection in the loss (and gain) or symbiosis genes (lines 283-301). We do not think that the data we analyzed allow us to distinguish between the roles of drift and selection. However, in the revised manuscript we highlight results from two studies (Klinger et al. 2016 and Cavassim et al. 2020) that show evidence of positive selection on symbiosis genes, one involving local adaptation to a high-N environment, and one showing the spread of a recently introgressed gene.

Lines 292 – 294: I'd like to see some discussion of why this pattern is evolutionarily interesting and what ecological situations might cause different patterns. For example, given that all the sequenced genomes came from rhizobia isolated from nodules, what patterns would you predict for rhizobia isolated directly from soil? I see that there is more true discussion (rather than reporting) of this result in lines 321 – 323.

To address this issue we have revised and reorganized the Discussion so that the interpretation of data (e.g. lines 321 – 323) is not displaced from the presentation of the data being interpreted (e.g. 292 – 294). Both drift as well as environmentally-variable selection might be responsible for high rates of presence-absence variation of symbiosis genes. The particularly high rates of such variation in nodulation genes might be related to costs (Triplett and Sadowsky 1992) or local changes in host identity. Although genomic data from soil-isolated rhizobia are limited, the available data (e.g. Hollowell et al. 2016, Gano-Cohen et al. 2020, which we now cite) indicate that soil isolated rhizobia do, in fact, vary in the presence of symbiosis genes. The revised discussion of this topic is found on lines 283-301.

*Lines 297 – 299: As others have found, the authors found that nodulation genes experience strong purifying selection, suggestive of mutualistic stasis (i.e., legume host imposes stabilizing selection on nod genes by maintaining an evolutionarily stable receptor for detecting the bacterial symbiont). The key novelty of this manuscript's results is that the average nod gene Ka:Ks evolutionary rate has been compared with that of other symbiosis-related genes. Why are nodulation genes so conserved, given that they provide access to a presumably valuable resource (the host nodule)? Does this suggest that there are not "bad actors" (e.g. pathogens) seeking access to the host via this route. What characteristic makes this route so invulnerable? Why can't any other bacteria imitate rhizobial nodulation signaling? This is a fascinating evolutionary puzzle. Again, there is more *discussion* of this result in lines 328 – 330, which is good. However, I'd love to see the authors speculate a bit about why, in a kind of lock and key arms race, plants could experience positive selection on receptor genes yet bacteria experience purifying selection.*

This is another interesting and challenging question. To address it we have expanded our discussion (lines 330-343). In particular, we note that although we detect an overall signature of purifying selection, four of the symbiosis genes have high Ka:Ks relative to genome-wide values, suggesting a possible role of positive selection (although these values may also reflect lack of selective constraint). Alternatively, it is possibility that arms-race type coevolution is happening, but not at the genes we have looked at, but rather at either symbiosis genes that are not among those we studied (many of which were originally identified through forward genetic screens and may thus form some of the core machinery for symbiosis. There are likely less well characterized modifier genes and these may be subject to different selective pressures) or by targeting proteins that don't play direct roles in symbiosis formation of function.

Lines 312 – 315: The authors also find that genes putatively involved in providing other symbiotic benefits to the host experience purifying selection. How can hosts impose purifying selection on rhizobium providing symbiotic benefits? Is this evidence that plants impose post-infection sanctions?

We do think that purifying selection is consistent with sanctions in that it suggests that mutations that alter the protein tend to reduce the fitness of individuals that express those alleles. In the revised Discussion we present this perspective in lines 311-314.

Lines 317 – 321: This paragraph is vague. What aspects of the previously described results are caused by the shifting identity of host species and which are caused by persistence of a host?

We have expanded and clarified this aspect of the Discussion. To improve the flow of the Discussion we also have moved this section to earlier in the paper (now lines 283-301).

Referee: 2

Comments to the Author(s)
Thank you for the opportunity to review the manuscript "Comparative genomics reveals high rates of horizontal transfer and strong purifying selection on rhizobial symbiosis genes" by Brendan Epstein and Peter Tiffin. In this work, the authors compared the genome sequences of 27 strains of rhizobia to estimate the frequency of horizontal transfer of non-symbiotic and symbiotic genes falling into several functional classes. They then carry out statistical analyses to infer the forms of selection acting on symbiotic and non-symbiotic genes. I think this is important work that brings novel insights into the evolutionary dynamics of this keystone symbiosis. The work should be of interest to a wide readership of ecologists and evolutionary biologists.

The writing is lucid and appealing. I only have a few editorial suggestions, listed below.

We appreciate both the positive evaluation and the helpful suggestions -- Thanks.

I am not an expert in these sorts of comparative genomic analyses. As far as I could tell, the methods were sound. I note that they used very recent, presumably cutting-edge programs for many analyses, such as finding putatively horizontally transferred genes. I was a bit surprised that there was no mention of gene location, i.e. on plasmids versus the main chromosome. Surely plasmid genes would be expected to have higher rates of HGT?

We agree with the reviewer that genes located on plasmids and other mobile genetic elements, as is often the case for symbiosis genes, would be expected to have a higher rate of HT. We were remiss in not making this clear, and now mention that symbiosis genes are often found on plasmids in the Introduction (line 65) and Discussion (lines 269-270).

My only major suggestions for improvement have to do with acknowledging and discussing some caveats in interpretation. In particular, I think some discussion of the definitions of the symbiosis gene categories is warranted. For example, some of the core fixation genes must also affect host benefit? In other words, I suspect some genes could be categorized in multiple ways, meaning the categories are not absolutely distinct and there is some subjectivity in the categorization and resulting comparisons. Furthermore, the "benefits derived" categories (host benefit and rhizobial fitness) seem particularly speculative to me, since they are based on GWAS studies in only a single rhizobium-host combination. The same genes may be unrelated to host

or bacterial benefit in other genetic backgrounds and other genes that do impact benefits in other rhizobium-host combinations have likely been mis-characterized as "non-symbiosis" genes. I doubt these potential sources of error would overturn the major findings of this study, but I would like to see an acknowledgement of these issues.

There is limited overlap among gene categories—particularly the putative host benefit genes and the annotated fixation genes which have 4 genes in common. There is, however, no overlap between annotated fixation and annotated nodulation genes or between host benefit and symbiont benefit genes. The overlap between categories is presented in Table 1, and in the revised methods we make a clear statement about overlap between categories (lines 146-148, "The rhizobia fitness genes and host benefit genes partially overlapped with the nodulation and fixation genes, but did not overlap with each other (Table 1)").

We agree with the reviewer's related points that the genes identified by association analyses may have different functions in different lineages. We now clearly present this issue in the Methods (lines 145-146, "We note that the function of the genes identified by GWA may differ among lineages and may not be related to symbiosis outside of *Ensifer.*"). We also acknowledge that our analyses was primarily focused on well characterized genes, and that less well characterized genes might have fundamentally different histories of selection than the genes we included in our analyses (lines 333-338).

I also wonder whether drawing conclusions about nodulation genes having reached selective optima rather than co-evolving (particularly emphasized in the Abstract) is justified, given that fitness landscapes were not directly measured in this study. The findings of purifying selection on most nodulation genes, even those that have been transferred, is interesting and suggestive of evolutionary stasis. But additional studies that directly measure rhizobial fitness are needed to fully test this conclusion.

We have rephrased this section of the abstract and now write: "These patterns are consistent with rhizobia adapting to the host environment through the loss and gain of symbiosis genes, but not with host-imposed positive selection driving divergence of symbiosis genes through recurring bouts of positive selection." We also have revised the Discussion to eliminate claims or suggestions that symbiosis genes are at selective optima. We still use the phrase selective optima on lines 80 and 317, in saying that one of perspectives in the discussion of the nature of selection on mutualistic partners is that they are at selective optima.

Comments by line number:

22 "involved IN initiating"

Fixed.

37-40 Though logical, this statement would benefit from a supporting reference.

We have added a reference to Vos (2009) (https://doi.org/10.1016/j.tim.2009.03.001).

43 I'm uncomfortable with describing the symbiosis as facultative. This could be debated endlessly.

By facultative we mean that it is not an obligate relationship (both partners can survive and reproduce in the absence of the other), and have clarified our meaning (line 40). We'd be interested to know more about the reviewer's perspective here as we, and others, have used this phrase for years and not had anyone tell us they disagree.

64-66 I think the "confer a fitness advantage" claim is an oversimplification. New symbiosis genes can exclude prior hosts, which could confer a large cost. I.e. not all transfers will expand the host range.

We agree that not all acquisitions of symbiosis genes will expand the host range or confer a fitness advantage, and have replaced "would offer an opportunity" to "potentially offers an opportunity" (line 63). However, "confer a fitness advantage" refers to associating with a plant host in general, which we think is a reasonable claim.

110 Really the major clades of alphaproteobacterial rhizobia, not ALL major clades

Fixed.

113 What does "MCL" stand for?

We now provide the meaning of this acronym (clearly we should have done so before). MCL stands for Markov CLuster algorithm, a method for identifying clusters in graphs that is very commonly used to group gene sequences into orthologous sets.

232 What does "similar numbers of strains" mean here? Non-symbiosis genes present in similar numbers of strains?

Yes. We have altered the wording to make this more clear (now lines 240-241).

250 The subscripts here are a bit confusing.

We have replaced the subscripts with a more verbose description (now line 262).

371 change "the approximately" to "approximately the"

Fixed.

Figure 1 Maybe remind the reader what delta means in the legend

Done.

Referee: 3 Comments to the Author(s)

The authors present a study on the comparative genomic analysis of 27 alpha-rhizobia bacteria, with the aims to assess the extent of HT of symbiosis genes, the effect of HT on evolution of the symbiosis and how selection has contributed to the divergence of symbiosis genes. They show that nodulation genes are less conserved across the genomes analysed than nitrogen fixation genes, that symbiotic genes experience more HT than non-symbiotic and that symbiotic genes appear to harbor signatures of purifying selection. There are a number of issues in the way in which these data have been interpreted and discussed.

We appreciate the reviewer's constructive criticism and deeply knowledgeable review.

First, it would have been informative for the authors to provide a list of nodulation, n-fixation, GWA host and symbiont benefit genes that were identified in the study. There are only a relatively small numbers of these genes (65-10 for each category) and it would have really improved the interpretation of the results to know their identity.

We provided a list of genes in Table S3. We believe this list is too large to fit as a main text table, but we have added an additional reference to this table in the Results section (line 202) to make it easier to find.

Second, the finding that symbiosis genes show a high rate of HT should be discussed in the context of these genes being encoded on mobile genetic elements (plasmids or symbiosis islands) that have been frequently shown to transfer via bacterial conjugation, while the bulk of non-symbiotic genes are encoded on bacterial chromosomes and therefore are not subject to conjugal transfer (although transformation and transduction will play some role).

As we discuss above we should have more clearly acknowledged this in the original submission. In the revised manuscript we make it clear that symbiosis genes are often found on plasmids and other mobile elements (Introduction, line 65) and (Discussion, lines 269-270). Because genome structures and the genomic locations of symbiosis genes vary widely among rhizobial lineages (e.g. megaplasmids in *Ensifer*, smaller plasmids in *Rhizobium*, often a single chromosome in *Bradyrhizobium*), we did not analyze the effects of genome location on either HT or selective history.

Third, the logic behind the selection of the genomes that were analysed is not clear. Granted that the authors have focused on strains where completed closed genomes are available, (some

have been missed), but there appears to be no indication in the paper that the authors have considered the vast host-range differences of their selected strains. For example, NGR234 is an incredibly broad-host range strain capable of nodulating 112 different legume genera, while Mesorhizobium australicum (species named misspelt in FigS1) WSM2073 appears only able to nodulate one. These vast differences would no doubt have an impact on the suite of nodulation genes harboured by these strains and hence the analysis of the data.

Certainly the reviewer is correct that we did not include all complete rhizobial genomes. Our goal was to include a representative from each of the major clades and named groups of Alphaproteobacterial rhizobia. We have revised the Methods (lines 96-115) and the Discussion (lines 282-289) to clarify this point and to explicitly acknowledge the role of host range in generating the patterns we observe. We also agree that the range and identity of hosts nodulated by these strains could affect the results of our analysis—had we analyzed a restricted set of strains that were closely related and had the same host range, we might not have found as much evidence of HT. That said, variation in host range is a commonly noted characteristic of rhizobia, and so we think that the strains we included, and the processes we inferred, are representative of rhizobia in general and that the patterns we find reflect long-term evolutionary processes that contribute to host-range and other diversity we see today. Perhaps our analyses will inspire others to explore some of the interesting issues the reviewer raises (we hope so), but doing so seems beyond the scope of the current paper.

Also, thank you for catching the misspelling, which we have fixed.

Four, On line 279, the authors state that "a typical annotated nodulation gene is found in less than half of the rhizobia genomes we examined, and only 11 of the 65 nodulation genes were found in all genomes". Presumably these 11 genes would encode the core Nod Factor synthesis genes, transcriptional regulators and NF secretion systems, which have been shown in many previous studies to be widely conserved across rhizobial genera? While the remaining 54 genes would likely be involved in modifying core NF? Many non-core NF modifying genes have been characterised previously (See Rodopothong et al., 2009 MPMI; Perret et al., 2000 MMB Reviews as an example) but this was not discussed in this study. Given that the strains chosen for this study represent rhizobia spanning 7 different genera, and that many of the strains within these genera have vastly different host ranges to each other, it is not surprising that the set of "essential" nodulation genes is relatively small.

We agree with the reviewer that the magnitude of the variation among genomes in nodulation gene repertoire is affected by, or reflective of, the diversity of hosts nodulated and the phylogenetic range of our sample. As stated above, we think that this sample is representative of rhizobia in general and is appropriate for the questions we were interested in, and we have added some language to the manuscript to clarify this point (lines 96-115 & lines 282-289).

Finally, I draw the authors' attention to the fact that WSM2073 is the result of a HT event, having acquired its symbiosis genes directly from WSM1271, as reported in Haskett et al. 2016 PNAS.

Thank you, that is interesting, and we now mention it in the Discussion (lines 288-289) and Methods (lines 114-115).

Appendix C

Associate Editor
Board Member: 1
Comments to Author:
Your revised manuscript has been assessed by an expert who finds significant problems with presentation that must be addressed before it can be considered for publication by Proceedings B.

> *Thank you for the opportunity to revise our manuscript. We also thank the reviewer for their overall favorable assessment as well as their helpful corrections and suggestions for improving the manuscript. Following the reviewer's suggestions, we have moved a figure from the Supplemental materials to the main text, added color to both figures, updated the strain names, and added a reference. We think these changes both broaden the visual appeal of the manuscript and make it more informative.*

Reviewer(s)' Comments to Author:

Referee: 4

Comments to the Author(s)
The first point that struck me about this manuscript was that the visual presentation is not very appealing. A browsing reader will find just a single, small monochrome figure, from which they will deduce that the key point of the paper is the distribution of median delta in random genes. They will probably pass swiftly on in search of a more interesting paper. That would be a pity because, as earlier reviewers have said, this study actually addresses some interesting questions about the evolution of rhizobia, and of bacteria more generally.

I will not comment on the analyses, as previous reviewers have done this extensively and I think the authors have responded well to them, but the presentation really needs to be improved to help the reader to understand the work, especially if it is to made suitable for a journal with a wide readership like Proc Roy Soc B. Here are four points that I feel need attention.

1. I found the manuscript frustrating to read because key information is missing and can only be found by poking around in the Supplementary Information. Importantly, the authors never tell us what organisms they are studying. They only used 27 genomes, so it would not be unreasonable to provide a list of them. In fact, Supplementary Figure S1 not only provides this information but also shows their degree of relatedness, so why not include it in the main manuscript? This is really important information because one would expect the probability of HT between two lineages to be highly dependent on their degree of relatedness, so the quantitative results of the study will be strongly influenced by the choice of genomes. It is immediately clear from Fig. S1 that a wide range of pairwise relatedness is covered, from strains that are virtually identical (apart from their symbiosis genes) to the largest divergences in the alpharhizobia.

> *Thank you prompting us to make this change. We have followed the suggestion and moved this information from the Supplemental material to the main text (now Figure 1). This figure provides the phylogeny showing relatedness among strains and adds to the visual appeal.*

2. A major part of this study is concerned with the presence/absence of symbiosis-related genes. There are 143 such genes listed in Table S3, which is a perfectly feasible number to display in a figure. I suggest that the authors add, to the right of the phylogeny in Figure S1, a presence/absence matrix for these genes, colour coded by functional category and aligned with the relevant genome in

the phylogeny. This would create an arresting figure that captured a central aspect of the study and provided immediate insight into the quantity and quality of the data that lie behind the dry statistics. An example of the sort of thing I have in mind would be Fig. 2 of Tian et al. (2012, https://doi.org/10.1073/pnas.).

As indicated above, we now present this information in Figure 1 of the main text. We also follow their suggestion to present a color-coded heatmap showing the presence or absence of symbiosis genes in each strain. This figure is similar to that in Tian et al. 2012, and as the reviewer suggested this is a visually appealing and effective way to communicate results.

3. I am actually surprised that the authors do not cite this Tian et al. paper, because it has distinct parallels with their own study. Tian et al. sequenced 26 genomes of soybean symbionts and assessed the presence/absence of different symbiosis-related genes. They adopted a much wider definition of 'symbiosis-related', identifying 561 such genes, but these probably include most of those covered by Epstein and Tiffin. The focus of the analysis is different, so I think there is room for this new study, but some comparison of the two studies is needed.

We thank the reviewer for pointing us to Tian et al.. We have added a comparison of Tian et al.'s main results that have to do symbiosis gene evolution to the Discussion (lines 343-346).

4. The phylogeny in Fig. S1 does not match the taxonomy. For example, neither Rhizobium etli nor the genus Rhizobium are monophyletic in the tree. This is not because the phylogeny is wrong, it is because the names are wrong. Strain IE4803 is not R. etli, it is R. sophoriradicis. IRBG74 is not Rhizobium sp., it is Agrobacterium pusense. Strain WSM2304 is not R. leguminosarum, it is R. acidisoli. The authors should check the strains they have used in the genome-based taxonomy databases at gtdb.ecogenomic.org and tygs.dsmz.de. It is never wise to trust the taxonomy assigned in GenBank.

We have updated the species names throughout the manuscript, using names provided in gtdb.ecogenomic.org. In the supplemental material (Table S1) we also provide the strain name that was assigned to the genomes when deposited in GenBank.