Resubmission of NAR-02133-Met-N-2017

Dear Editor,

I am pleased to re-submit our original methods article entitled “riboSeed: leveraging prokaryotic genomic architecture to assemble across ribosomal regions”, by Nicholas R. Waters, Florence Abram, Ashleigh Holmes, Fiona Brennan, and Leighton Pritchard, for consideration for publication in NAR. This work describes a novel method and software tool for improving genome assemblies by reconstruction of repeat regions from short reads.

In this manuscript, we show that, frequently, the regions flanking equivalent rDNAs across taxa are conserved and that these flanking regions are uniquely distinguishable within a single isolate. These two findings allowed us to develop riboSeed, which uses the rDNA flanking regions from a reference genome as a template to resolve rDNA repeats. We show that assemblies produced by riboSeed are of very high quality when benchmarked against hybrid assembly and against other genome polishing methods, and that in some cases riboSeed can help close genome previously represented by multiple scaffolds.

This second resubmission incorporates comments from the reviewers largely focused on substantiating claims with examples from a wider variety of organisms. We have also clarified the intended use for the program in light of reviewer comments, and have substantially expanded the supplementary materials. We believe that the novelty of both the biological findings (characterization of rDNA flanking regions) and the method and open-source software implemented as riboSeed will be of interest to wide range of NAR readers: those interested in the computational leverage gained through our method, microbial ecologists, and those simply seeking to make the most out of their sequencing data. Further, the tools developed as part of riboSeed may prove useful for others hoping to explore rDNAs outside of their impact on genome assembly.

We would like to note that we believe that reviewer 3 did not appear to review the most recent version of the manuscript. Throughout their review, they appear to be referring to either a previous version of the manuscript or to the document showing the difference between the previous versions. This is supported by the fact that all the line numbers mentioned correspond not to the resubmitted manuscript but to the document showing the difference, referencing the lack of “violet” in figure 5 (changed from gray to violet for the previous resubmission), their question about broken links (broken in the document showing the difference, but not in either version of the actual manuscript).

This manuscript is unpublished and is not under consideration for publication elsewhere; it has been posted to bioRxiv to allow for early criticism of the work. We have no conflicts of interest to disclose. We trust that you will find our article of high quality and that you will find it suitable for publication in your journal. Please don’t hesitate to contact me if you require any further information. Thank you in advance for your consideration.

Best wishes and kind regards,

Note: We have attached the revised manuscript. Additionally, we have two files for the Editor: a highlighted version with changes made from the previous submission, and an itemized response to the reviewer’s comments along with a revision summary. We thank you for the thoroughness of the reviewing process and hope that we have now addressed all the reviewers’ comments to your satisfaction.

We submit the accompanying manuscript as a resubmission of the previously-presented manuscript with reference: NAR-02133-Met-N-2017.

# 

# 

# Summary of Changes to the Manuscript

## Changes to Text

We made changes to the text of the manuscript to address changes requested by the reviewers below.

* We add text to the Discussion outlining the limitations of riboSeed (e.g. the requirement for a appropriate reference), and note that for strains where rRNA coding regions are not organized into rDNA operons, riboSeed would be unnecessary.
* We link to the project homepage instead of the main GitHub page in response to ambiguity highlighted to reviewer 3.
* We have removed the defunct reference to *Klebsiella* from the legend of Figure 4.
* We made references in the main text to the new additions to the supplemental information.
* Minor corrections to grammar, spelling, and references were made where needed.

## Changes to Data

* n/a

## Changes to Analysis and Software

* We have expanded the performance section of the supplementary data to include performance examples covering all major taxonomic groups within bacteria for which riboSeed might be used (i.e. contain repeated rDNA regions), in addition to the archaeal data previously presented.

## Changes to Results

* In the supplemental information, we have added within-taxa entropy figures to show the relative conservation around a given rDNA region.
* In the supplementary information, we had neglected to

# 

# 

# Referee: 1

## Comments for the Author

The manuscript is considerably improved. However, the method needs wide validation to be accepted as novel and useful for the purpose

We thank the reviewer for their time and effort in addressing a new version of the manuscript , and the point is well taken that wide validation is required for the community to take up bioinformatics programs broadly or as part of common toolkits and workflows. However, we respectfully note that the referee does not describe a specific validation goal that can be clearly met by ourselves, but instead a hope for uptake by the community that would be assisted by publication.

We have attempted to validate riboSeed in the following ways:

* We have applied riboSeed to a synthetic dataset (artificial genome) based on rRNA regions from *E. coli*, and demonstrated that riboSeed performs better than *de novo* assembly for reconstruction of rDNA regions.
* We have demonstrated that riboSeed uses reference genome information to aid assembly through rRNA regions, but does not simply transpose the reference sequence into the final assembly, using both the synthetic dataset and published hybrid PacBio/Illumina data.
* We have established, using synthetic reads for completely sequenced genomes, that riboSeed assembles rRNA regions in their correct genomic context, for *E. coli* and *K. pneumoniae*.
* We have shown using published Illumina short reads that riboSeed is able to correctly assemble rRNA regions in their appropriate genomic context - as determined by hybrid assembly of both PacBio long reads and Illumina short reads derived from the same isolate - for *P. aeruginosa*.
* We have demonstrated that riboSeed is able to improve a published assembly of *S. aureus* by assembling and locating rRNA regions, bringing that assembly near to closure using only Illumina short reads.
* We have evaluated the performance of riboSeed against the *de facto* standard GAGE-B genomic datasets, including low GC content bacteria from the following genera: *Mycobacterium*, *Aeromonas*, *Bacillus*, *Rhodobacter*, *Staphylococcus*, *Vibrio*, *Xanthomonas* and *Bacteroides*. riboSeed was shown to improve assembly by correctly placing assembled rRNA regions in nearly all cases, establishing its broad applicability.
* We have established that riboSeed successfully assembles and locates rRNA regions with Illumina short read data for Archaeal genomes in the genera *Methanosarcina* and *Methanobacterium*, demonstrating the applicability of this tool across two microbial Kingdoms

Although we believe that our existing validation results exceed the usual efforts made in publishing comparable tools, we have aimed to meet the spirit of the reviewer’s criticism while not wishing to cause further delay in publication. We have in good faith performed further validation of riboSeed against publicly available genomes and short read datasets for *Corynebacterium diphtheriae, Chlamydia trachomatis, Clostridioides difficile, Burkholderia cepacia, Myxococcus xanthus, Helicobacter cinaedi,* and *Mycoplasma hominis*. With this new validation data we have demonstrated the utility of riboSeed against examples from all major bacterial clades for which rDNA regions are known to occur in operons and in several copies (results in supplementary information, Table S7). We hope that the editor will agree that we have met the reviewer’s requirements with these additional validation examples.

# 

# 

# Referee: 3

## Comments for the Author

### General comments

1. First, full disclosure: When I was asked to review this paper, I disclosed that I am not experienced in genomic assembly, nor in bioinformatics. However, my opinion was still welcome and actually strongly valued, so I accepted to be reviewer 3. I therefore might not understand all methods or visualization methods shown here, for which I apologize. I do however, have a strong background in microbial ecology, 16S rRNA based analyses, and a basic understanding of de Bruijn graphs so I am not a complete misfit either :-).

2. Second, the authors have addressed all comments raised by reviewers 1 and 2, and appear to have improved the manuscript a lot, so I am grateful to the work that went in the peer review process before I joined.

* We thank the reviewer for their insightful and helpful comments, and that they recognise the efforts we have made to improve the manuscript in accordance with the reviewers’ suggestions.

3. As one of the tests, the authors run E. coli fragments using Klebsiella pneumonia as the reference genome, and found that the reference is too divergent from the reads. These 2 species are very closely related; a BLAST comparison done by this reviewer comparing CP003200.1 (Klebsiella) and BA000007.2 (Escherichia) found that the 2 16S rRNA genes were 97% similar.

* Although the 16S genes themselves are highly similar, riboSeed’s methodology for placement and separate assembly relies predominantly on the rDNA flanking regions, which are typically more less conserved in sequence than 16S sequence (Figure 3 and Figure S4), and so more representative of the genome at large. Indeed, using the program pyani ([https://widdowquinn.github.io/pyani/](https://widdowquinn.github.io/pyani/link)) to calculate average nucleotide identity between our two reference strains (*E. coli* BA000007.2; *K. pneumoniae* CP003200.1), we find that the genomes align with 85% identity, but with only 20% alignment coverage (i.e. no more than 20% of either organism’s genome aligns to the other). We acknowledge that riboSeed’s usefulness is maximised with genera where reasonably close and reliable references exist, which limits applications in community ecology dealing with novel genera, and now note this in the discussion.

3a. In many cases where researchers might want to use this tool, there might not be a reference genome that is as close as these 2 are together. It is great that the method can help assemble a genome when there is another complete genome of the same species available, but that might not always be the case. In many cases where microbial ecologists try to assemble microbial genomes, there is not even another genome in the same family available. If the method already fails if the closest relative should be in the same species - how useful is this tool going to be? It would be great if the authors could devote some sentences in the discussion acknowledging this limitation; this will clearly not be a tool for environmental microbiology on e.g. phylum SR1 genomes, but will have value for e.g. clinical microbiology where a lot of strains within a species have already been sequenced.

* The point that the reviewer raises is correct, and we note this in the Discussion: “The method of constructing pseudocontigs implemented by riboSeed relies on having a relevant reference sequence, where the rDNA regions act as bait”, where we also propose a future improvement to ameliorate this problem: “By using a database of sequence profiles (e.g. hidden Markov Models) from homologous rDNAs in a taxon, the step of choosing a single most appropriate reference might be circumvented”. The latter improvement has not yet been made.
* We have added another sentence to this paragraph to highlight fields in microbial genomics where we believe that riboSeed will be more or less useful and in particular that, although riboSeed is especially useful for revisiting extant short read datasets, and obtaining greater value from novel short read datasets in genera where at least one complete genome is available, applications to sequencing novel organisms from unsequenced genera (e.g in community ecology) are limited by the requirement for a reference genome.

4. P1,L25L and many other places. "rDNAs" is not a good term. Yes, they encode rRNAs, but there is no DNA that ends up in a ribosome. Could I recommend to use 'rRNA genes" or "rrn operons"?

* We consider “rDNA” to be a term in general use for DNA that encodes ribosomal RNA. It has been widely adopted since its first appearance in the literature in 1969 (<https://www.ncbi.nlm.nih.gov/pubmed/5345388>), so we would argue that it is well-understood in the community, and does not need to be changed. If the editor considers our intended meaning to be inappropriate in the context of NAR, we are happy to conform to house style.

5. P2,L31L and other places. "Prokaryotes" (also in L36) is a heavily criticized term (sorry, Reviewer 2!), since it suggests a common ancestor between bacteria and archaea. How about using "bacteria and archaea" here and "microbial" at some other places?

* We are unable to use “microbial” with this intent as the method is not intended for eukaryotic genomes, which includes microbial organisms such as fungi, oomycetes and algae. We recognise “prokaryote” as a term describing the common cellular structure of both Archaea and Bacteria, rather than making a strong implication of common ancestry. If the editor considers our intended meaning to be inappropriate in the context of NAR, we are happy to conform to house style.

6. This is very minor, but a space is missing in a lot of citations, for example P2L38L "regions(28)".

* We have corrected this throughout the manuscript.

### Specific comments:

#### 7. P1, L32L Abstract. "a method which constructs" --> "a method that constructs"

* We have changed this to “a method to construct” for clarity.

#### 8. P1, Figure 1. Dumb question: For bacterial genomes, what is the difference between "complete genome" and "chromosome"? Does this refer to bacteria with multiple chromosomes, such as Vibrio cholerae?

* We thank the reviewer for noting this; a summary of NCBI’s Genome Assembly Levels can be found at the following link: <https://support.ncbi.nlm.nih.gov/link/portal/28045/28049/Article/752/How-are-genome-assemblies-generated-and-what-are-assembly-levels>
* In brief, a “chromosome” assembly is one in which there is a record for each chromosome [the assemblies themselves may have gaps], whereas a “complete” assembly is one in which there are no gaps in any of the assembled chromosomes. “Complete” assemblies are therefore a well-defined subset of “chromosome” assemblies, but it is usual not to refer to “complete” assemblies as “chromosome” assemblies in order to highlight their improved assembly status.
* We have added a citation for Kitts et al’s 2016 paper about the NCBI Assembly database, which explains these levels and their definitions.

#### 9. P2, L33L. Since we should be inclusive of archaea, maybe use "microbial" here?

* We do not validate the software for fungal or protist genomes, so we would prefer to avoid claims of applicability to all microbial genomes.

#### 10. P2, L42R. This is probably a very novice question, but where is the link to riboSeed? I only see links to other GitHub programs and documentation.

* We take the reviewer’s point about the GitHub page itself being potentially confusing for non-bioinformaticians, and have prepared a more user-friendly landing page that will direct interested readers to the pipeline and datasets referred to in the paper, as well as extensive documentation for the tool itself. The landing page also has links so that more proficient bioinformaticians can proceed directly to the source code on GitHub.

#### 11. P2, L60R: The link leads to the wrong page (has an "X" behind the html).

* We believe that this is referencing the document showing the differences between the two versions of the manuscript; the link is deactivated in that document, but fully functional in the actual manuscripts.

#### 12. P3, Figure 2. It was not easy to follow how the steps mentioned in the legend matched the steps shown in the figure. Maybe include the steps (e.g. "c)") behind the appropriate step in the description? In the 2nd sentence, does "Reads are mapped to a reference genome" refer to the de novo or to the de fere novo method? I assume it is the latter but it might be helpful to add. "those reads that align to rDNA and flanking regions are extracted." - is this done based on the annotation of the reference genome?

* We have clarified by adding “In riboSeed, reads are mapped to […]” to the second sentence, clarifying that the proceeding steps describe the *de fere novo* assembly.

#### 13. P3, L38L. Is this enough information for most to install the program? Maybe a link to [http://riboseed.readthedocs.io/en/latest/INSTALLATION.html](https://extranet.hutton.ac.uk/owa/,DanaInfo=duexc01.ad.hutton.ac.uk,SSL+redir.aspx?SURL=04oUKvfCbjSNCJ9vAfGKFD_MM8-GvYSFJlLuK3tnA9_2066wE0bVCGgAdAB0AHAAOgAvAC8AcgBpAGIAbwBzAGUAZQBkAC4AcgBlAGEAZAB0AGgAZQBkAG8AYwBzAC4AaQBvAC8AZQBuAC8AbABhAHQAZQBzAHQALwBJAE4AUwBUAEEATABMAEEAVABJAE8ATgAuAGgAdABtAGwA&URL=http%3a%2f%2friboseed.readthedocs.io%2fen%2flatest%2fINSTALLATION.html) would be helpful here.

* In our experience, explicit installation steps for software are not typically provided in a manuscript. By mentioning “Installation (via either conda, pip, or GitHub)”, we expect that most readers will know that they have the option of installing from source (GitHub), using Python’s package manager(PyPI), or using the widely-used conda ecosystem, and that instructions are likely available in the documentation (such as at the link the reviewer notes).

#### 14. P3, L41L: "riboseed pipeline" --> "riboSeed pipeline"

* This has been corrected in the text.

#### 15. P3, L56L. How does Barrnap work? Is this a sequence homology search, or based on user annotation within the reference genome?

* Briefly, Barnap has a library of hidden Markov models (HMMs) for rRNA families, and uses NHMMER to query with these against input sequences. For further information we refer the reviewer to the project page at https://github.com/tseemann/barrnap.

#### 16. P3, L41R. Does the "Jenks Natural Breaks algorithm" need a citation?

* The Jenks Natural Breaks algorithm (aka the Jenks Optimisation Method) is a standard statistical method that is typically used without citation in the literature. The implementation used in riboSeed is provided by a third-party Python package that may be obtained using PyPI at https://pypi.python.org/pypi/jenkspy.

#### 17. P3, L55R: What does "BWA" stand for?

* BWA may stand for Burrows-Wheeler Algorithm, or for “BWA” - a program implementing the Burrows-Wheeler Algorithm for sequence mapping. We recognise that this is sometimes confusing (in the same way that R may refer to the programming language R, or R - the program that interprets other programs written in the language called R). In this manuscript we are referring to the software package, and we believe this is made clear in the sentence context, and by the associated reference.

#### 18. P4, L15L. If the pseudocontigs are linked together with a random spacer, can their order change based on the iterative mapping of the reads? I assume so, since the purpose is to close the genome but not sure. Is the swap command the one that does that?

* As mentioned in the previous response to reviewers, the order of the pseudocontigs does not influence genome order in the final assembly. Their order may change between iterations, but because there are no reads that span the spacer, this cannot affect the final incorporation into the assembly..

#### 19. P4L52L. Should "a" have been deleted? It sounds better when it is left in.

* Corrected.

#### 20. P4, L57R. "E. coli MG1655" needs more clarification. Is this the E. coli Sakai strain mentioned above or a different strain? Based on the later text, it appears to be a different one, but this is not clear.

* We have added the strain accession number to further highlight that that two different strains are used in this analysis.

#### 21. P5, L35. Figure 3. "moving away from the the 16S" - please remove duplicated word.

* Corrected

#### 22. P5, L41R. Where are the results shown? The authors describe that each rRNA operon was categorized, but no results are shown here in this section. Should this text be part of the Methods? And if so, with which part of the Methods should it be integrated? There is a section called "3. Assessment and Visualization" - this seems to be the best place to integrate it. This text should either be moved to the Methods. Alternatively, the authors could combine this section with the next one ("Simulated reads"), since that one shows the results.

* We agree that this could be equally appropriate to have at the end of the Methods section, and have moved this section accordingly.

#### 23. P5, L59R. Typo: "missasembly"

* Corrected.

#### 24. P6, L35L. The reference to Figure 4 has now been deleted - this should be restored.

* We have now restored the reference to Figure 4.

#### 25. P6, Figure 4. I do not see how the de fere novo method has only 4 out of the 7 bridges correct. The de novo assembly appears to have all of them correct (I probably misinterpret this but they are all in the correct order). In addition, the figure appears to show that the Klebsiella reference results in 2 + 3 correct spacers, instead of a complete failure. I am also having trouble seeing the connections between the top 2 graphs, since they are well connected. Would the figure be more clear if the rrn operons would also be shown? What do the black vertical lines mean?

* We have clarified the caption for those unfamiliar with Mauve by inserting the following sentence: “Red regions represent rRNA coding sequences, vertical black lines indicate boundaries between assembled contigs, and shading represents synteny.”
* The pink regions connecting the “Artificial Chromosome” and “De fere novo” sequences indicate regions of sequence identity. In sequence from left to right, the first bar spans four of the rDNA regions indicated on the “Artificial Chromosome”, so these four rDNA regions are correctly assembled and ordered in this matching region. The remaining pink matching regions indicating sequence matches do not include any of the other rDNAs.
* We believe that the reviewer was not referring to the most up to date submission, in which the figure was simplified and made clearer.

#### 26. P6, Figure 5. This figure is very low quality which is probably why I am not sure what the little wiggly orange and blue things are. Shouldn't there be one value (i.e. the correct number of operons) for each substitution frequency? How can there be 7 values? Also, I do not see a "lilac" area at all, just a gray one.

* The colours of this figure were modified in the most recent submission, and should be clearer. Again, given that the gray color the reviewer refers to was only present in the previous version, we believe that the reviewer was not referring to the most up to date submission, in which the figure was improved.

#### 27. P7, Figure 6. What is shown on the X-axis? Length of the fragments tested? How do the values here (correct, incorrect, ambiguous) correspond to the earlier defined values of correct, unassembled, and incorrect? And assuming that these 3 categories always need to add up to 7 or 8, would it easier to interpret if the data was shown as a stacked column, with 3 colors? Also, what does the n=9 mean? That it was performed nine times? On 9 genomes? 9 randomizations? A bit more info would be welcome here.

* We realize that while under the Assessment and Visualization section we define the assembly categories, we neglected to define how the “ribo score” command classes an assembly as ambiguous. We have added that explanation to help clarify that section.
* A stacked bar plot, unlike the boxplots, would not permit us to easily represent the variation in assignment to “correct”, “incorrect” and “ambiguous” classes over a number of genomes. We believe that representing the uncertainty associated with riboSeed results is important for potential users to assess its potential applicability.

#### 28. P7, L19R. What is meant by "benchmarking"? This sounds like a very jargon-y term, and I am not sure what is meant here. I assume it means "comparison".

* We do not consider “benchmarking” to be a jargon term. It is found in standard US and British English dictionaries, and has the definition “to measure the quality of something by comparing it with something else of an accepted standard” (Cambridge English Dictionary).

#### 29. P8, L12R. Please check this sentence: "riboSeed was able assist in bringing ".

* This has been corrected in the text.

#### 30. P8, L45. Discussion. "We show that..". This statement appears too strong. I agree with that the authors showed this for E. coli (Figure 2) but is the situation similar in other bacteria? The authors need to either show that this is true for other bacterial species as well (as a minimum for at least one other, unrelated, species, such as Staphylococcus aureus genomes), or refer to other studies showing similar analyses, or clarify here that they showed it "at least for E. coli".

* We have added the entropy figures for all genomes used in the study to the supplementary data (Figure S4) to show the entropy both within the genome as well as across homologous rDNAs in a subset of related strains.