Peer Community In Genomics

# A workflow for studying enigmatic non-autonomous transposable elements across bacteria

**Gavin Douglas** based on peer reviews by **Sophie Abby** and 1 anonymous reviewer

Frederic Bertels, Julia von Irmer, Carsten Fortmann-Grote (2023) RAREFAN: A webservice to identify REPINs and RAYTs in bacterial genomes. bioRxiv, ver. 4, peer-reviewed and recommended by Peer Community in Genomics.
https://doi.org/10.1101/2022.05.22.493013

Submitted: 07 June 2022, Recommended: 07 February 2023

**Cite this recommendation as:**
Douglas, G. (2023) A workflow for studying enigmatic non-autonomous transposable elements across bacteria. *Peer Community in Genomics*, 100166. https://doi.org/10.24072/pci.genomics.100166

Published: 07 February 2023

Repetitive extragenic palindromic sequences (REPs) are common repetitive elements in bacterial genomes (Gilson et al., 1984; Stern et al., 1984). In 2011, Bertels and Rainey identified that REPs are overrepresented in pairs of inverted repeats, which likely form hairpin structures, that they referred to as "REP doublets forming hairpins" (REPINs). Based on bioinformatics analyses, they argued that REPINs are likely selfish elements that evolved from REPs flanking particular transposes (Bertels and Rainey, 2011). These transposases, so-called REP-associated tyrosine transposases (RAYTs), were known to be highly associated with the REP content in a genome and to have characteristic upstream and downstream flanking REPs (Nunvar et al., 2010). The flanking REPs likely enable RAYT transposition, and their horizontal replication is physically linked to this process. In contrast, Bertels and Rainey hypothesized that REPINs are selfish elements that are highly replicated due to the similarity in arrangement to these RAYT-flanking REPs, but independent of RAYT transposition and generally with no impact on bacterial fitness (Bertels and Rainey, 2011).

This last point was especially contentious, as REPINs are highly conserved within species (Bertels and Rainey, 2023), which is unusual for non-beneficial bacterial DNA (Mira et al., 2001). Bertels and Rainey have since refined their argument to be that REPINs must provide benefits to host cells, but that there are nonetheless signatures of intragenomic conflict in genomes associated with these elements (Bertels and Rainey, 2023). These signatures reflect the divergent levels of selections driving REPIN distribution: selection at the level of each DNA element and selection on each individual bacterium. I found this observation particularly interesting as I and my colleague recently argued that these divergent levels of selection, and the interaction between them, is key to understanding bacterial pangenome diversity (Douglas and Shapiro, 2021). REPINs could be an excellent system for investigating these levels of selection across bacteria more generally.

The problem is that REPINs have not been widely characterized in bacterial genomes, partially because no bioinformatic workflow has been available for this purpose. To address this problem, Fortmann-Grote et al. (2023) developed RAREFAN, which is a web server for identifying RAYTs and associated REPINs in a set of input genomes. The authors showcase their tool by applying it to 49 Stenotrophomonas maltophilia genomes and providing examples of how to identify and assess RAYT-REPIN hits. The workflow requires several manual steps, but nonetheless represents a straightforward and standardized approach. Overall, this workflow should enable RAYTs and REPINs to be identified across diverse bacterial species, which will facilitate further investigation into the mechanisms driving their maintenance and spread.

***References:***

Bertels F, Rainey PB (2023) Ancient Darwinian replicators nested within eubacterial genomes. BioEssays, 45, 2200085. https://doi.org/10.1002/bies.202200085

Bertels F, Rainey PB (2011) Within-Genome Evolution of REPINs: a New Family of Miniature Mobile DNA in Bacteria. PLOS Genetics, 7, e1002132. https://doi.org/10.1371/journal.pgen.1002132

Douglas GM, Shapiro BJ (2021) Genic Selection Within Prokaryotic Pangenomes. Genome Biology and Evolution, 13, evab234. https://doi.org/10.1093/gbe/evab234

Fortmann-Grote C, Irmer J von, Bertels F (2023) RAREFAN: A webservice to identify REPINs and RAYTs in bacterial genomes. bioRxiv, 2022.05.22.493013, ver. 4 peer-reviewed and recommended by Peer Community in Genomics. https://doi.org/10.1101/2022.05.22.493013

Gilson E, Clément J m., Brutlag D, Hofnung M (1984) A family of dispersed repetitive extragenic palindromic DNA sequences in E. coli. The EMBO Journal, 3, 1417–1421. https://doi.org/10.1002/j.1460-2075.1984.tb01986.x

Mira A, Ochman H, Moran NA (2001) Deletional bias and the evolution of bacterial genomes. Trends in Genetics, 17, 589–596. https://doi.org/10.1016/S0168-9525(01)02447-7

Nunvar J, Huckova T, Licha I (2010) Identification and characterization of repetitive extragenic palindromes (REP)-associated tyrosine transposases: implications for REP evolution and dynamics in bacterial genomes. BMC Genomics, 11, 44. https://doi.org/10.1186/1471-2164-11-44

Stern MJ, Ames GF-L, Smith NH, Clare Robinson E, Higgins CF (1984) Repetitive extragenic palindromic sequences: A major component of the bacterial genome. Cell, 37, 1015–1026. https://doi.org/10.1016/0092-8674(84)90436-7

# Reviews

## Evaluation round #2

DOI or URL of the preprint: https://www.biorxiv.org/content/10.1101/2022.05.22.493013v3
Version of the preprint: 3

### Authors' reply, 26 January 2023

**Download author's reply**
**Download tracked changes file**

### Decision by Gavin Douglas, posted 29 December 2022, validated 03 January 2023

**Additional revisions requested**

Hi Dr. Bertels and colleagues,

Both reviewers assessed your changes and agree that the manuscript is greatly improved. They have raised a few remaining points that warrant some further minor changes and clarifications.

In addition, please see my own minor comments below, which are primarily typo and phrasing fixes.

All the best,
Gavin Douglas

The license information for the source of Figure 1 (Bertels, Rainey, 2022) should be indicated somewhere in the text. There are usually requirements for how to redistribute/modify items from another work. E.g., if this is under a creative commons license, then you should state what license version it corresponds to and give a link to that license.

The title should be changed to “RAREFAN: A webservice”... rather than “RAREFAN: a webservice...”

L67-72 I think many readers will be curious to know what the other RAYT families are associated with, if not REPINs. Since they are defined as “REP-associated” (in their name) I think this deserves at least a quick mention in a sentence or two. Currently at least one parameter value in the Figure 2 mismatches with the Figure legend (distance between inverted sequences being 200bp vs 130bp). The authors should make sure that the values reported represent the current default values and old (or otherwise conflicting) values are not mismatched between the figure and legend, to avoid reader confusion. Also in Figure 2 – I recommend that the figure legend be simplified, as many of the details are already provided in the methods are not really pertinent to interpreting the plot and the take-home messages. I will leave that to the authors’ discretion. However, I do strongly recommend that the references in this legend be removed, as these should be mentioned in the appropriate section of the methods instead (references are generally uncommon in figure legends). I was unable to access results under run ID a2ijpkk6. The authors should clarify the protocol on linking RAYTs to REPINs. Is it generally expected for at least one REP to be within 200bp of the corresponding RAYT? Since the RAYTs act in trans as proteins, there does not seem to be any reason why this necessarily be true, so I think a little additional explanation would be helpful. Should use past tense when discussing specific results. So on L338 for instance, it should be “RAREFAN detected three populations when S. maltophilia Sm54 was selected as the reference strain”. The authors should use “p-value”, “P-value”, or “p value”, but not “p-Value”, which is the current usage in the text. Minor edits L30 – “providing” was actually grammatically correct, and so the revised change to “provide” should be undone. L48 – “REP sequences is” should be “REP sequences are” L53 – I suggest “not mobile anymore” be reworded to “immobile” or “no longer mobile” L58 – “associated to” should be “associated with” L64 – I suggest “very special” be changed to “unique” L77 – I think the year estimate should be clarified. Presumably some RAYT/REPIN groups may have been present in a lineage for less than a million years (or at least this is possible!). So I would re-word to say that “they have been evolving in single bacterial lineages for up to millions of, or perhaps even one billion, years.” L102 – I think “Yet,” should be removed, or perhaps replaced with “Unfortunately,” L104– “ins and outs” should be replaced with less colloquial language, such as “details” or “detailed features” L105 – “the genome” should be “a genome” L106 – “analyzed next” should be “then analyzed” L106-107 – “If they are exclusively” should be something clearer like “If these sequences are exclusively” L107-108 – I would put commas on each side of this sentence fragment: “and present in only one or two loci in the genome” Figure 2 legend “the” should be re-added in front of “seed sequence”. Implementation section of methods – python, java, flask, and shiny should all be capitalized Regarding “Query RAYT” bullet point in implementation methods: above this is described as optional. The authors should clarify the procedure when this protein sequence is not provided, as is currently done for the Tree file option. L180 – “(n-1)” should be “[n-1]” L277 – “Especially” should be “This is especially true” Figure 6 legend – here “group” is capitalized in some but not all cases. In this legend (and in the relevant section of the main text, where this also varies), the authors should consistently write “group” capitalized or lowercase in all instances. L431-432 – “Genbank” should be “GenBank” and I think it would be clearer to say “creating a RAREFAN Galaxy workflow” rather than “integrating RAREFAN into workflows such as Galaxy”, as Galaxy is a means of making workflows available online for easy use, rather than referring to a specific workflow.

## Reviewed by Sophie Abby, 23 December 2022

The new version of this manuscript is much improved, with a more detailed biological background and the provided clarifications on methods in main text and figures, as well as a new section on performances. Even though it is clear that manual curation might still be needed to assess the relevance of the provided results, the limits of the approaches are outlined with examples discussed and possible contingency plans. Therefore, and given the fact that there is so far no resource available to investigate REPIN-RAYT, I believe that the RAREFAN tool is valuable to the microbiologists community, and could in principle be supportive of its recommendation by *PCI Genomics*, provided that the pointed issues on the webserver are sorted out.

**On main text:**

Here are a few minor points/typos:

- It might be good to add the link to the webserver at the end of the abstract.
- Line 165. “Association distance REPIN-RAYT”. Please provide the default values.
- Line 187: “Among all identified REP and REPIN sequences REPIN populations can be isolated.” Something is wrong with this sentence. Is a word missing?
- Line 223: “Complete RAREFAN data used for analysis can be accessed by using the run IDs listed in Table 1.” You mentioned that the results from users are stored for 180 days. Would the IDs listed in Table 1 be stably kept over time?
- Figure 4 legend: “In an equilibrium” => “At equilibrium”?
- Line 281: “The RAREFAN webserver visualizes REPIN population size” => “enables to visualize…”?
- Line 327: “In some RAREFAN runs associations between RAYTs and REPINs are not monophyletic”. Please reformulate, associations are not the ones to be monophyletic.
- Figure 6 legend: “connect two sequence cluster.” => clusterS
- Line 180: “have been observed”

**Testing the webserver: http://rarefan.evolbio.mpg.de**

I could test the webserver, after the issue described in the email exchange reported below was solved.

- I could submit 8 complete genomes of Klebsiella for a run (z9hgj3ld, results accessible here: http://rarefan.evolbio.mpg.de/results?run_id=z9hgj3ld ) which found no results. I thus changed the threshold of occurrences “min_nmer_occurrence” to 5 (re-run job y8586vnk – results accessible here: http://rarefan.evolbio.mpg.de/results?run_id=y8586vnk ), and could obtain some “RAYT” occurrences.

I found it difficult to interpret the master table of results on the summary page. For instance for the run y8586vnk, it seemed from the table that there were only the 0 group for which there were REP/REPINs occurrences. However, when clicking on “Plot data” (results accessible here: http://rarefan.evolbio.mpg.de//shiny/analysis/?run_id=y8586vnk ) there were REPINs identified for group 4 (and not only group 0 as reported in the table).

- Also, I came across a minor issue, in another run (sd0oyhv1 `http://rarefan.evolbio.mpg.de/results?run_id=sd0oyhv1`): I submitted six genomes, but one was judged unfit under this error message: “GCA_000009985.1_ASM998v1_genomic.fna contains non-DNA sequences and will be removed”, and dropped out of the analysis. However, this is a genomic FASTA file obtained from the NCBI/Refseq database.
I looked into the file and found a few “N” characters, a standard letter to represent “any nucleotide”. Maybe could the authors take into account that N characters could be present in some genomes, and more thoroughly test the way the nature of the FASTA files are provided?

**———————- ADDENDUM ———————-**

Email exchange with Dr. Frederic Bertels

===== Dr Abby to Dr Bertels, 20th of Dec 2022 =====

“I am writing to you to follow-up on the revised version of your article on the RAREFAN webserver.

I’ve been trying to test the webserver. But I could not manage to obtain results. Therefore I am unable to complete my review.

I’ve submitted three runs. Unfortunately, I made a mistake with the 1st run and submitted the entire proteome of an organism instead of the entire genome (run ID gvnnqk79). This is a silly mistake on my side, however you might expect such common mistakes to be made on a public server.

I don’t know if it is related to this, but then I’ve submitted two more runs with appropriate genome files, but they have been stuck in the queue since yesterday, while the first job (gvnnqk79) seems to be stuck at the ”Rarefan - started” stage.

The ID for the jobs are the following: gvnnqk79; _xwiv2us; clj3vckt

Could you please have a look, and let us know what is going on and when we will be able to test the server?”

===== Dr Bertels to Dr abby, 21st of Dec 2022 =====

“we fixed the issue you encountered. It was actually caused by a full error stream buffer that was filled by tons of error messages from the BLAST formatdb command. We have not encountered this error previously since the formatdb command does generally not produce large error messages. The buffer is large enough to store small to medium sized error messages that can be read once formatdb is finished. However, the error messages produced by generating a DNA BLAST database from protein sequences completely filled up the buffer, the program was then paused and waited for the buffer to be emptied so it could continue writing the error messages. The emptying never happened and the program did not finish. The waiting program in turn clogged up the server queue and prevented other jobs from being run.

We changed RAREFAN so the buffer is now continuously read, which should prevent a deadlock (at least at that position in the code). We are also testing whether the submitted sequence is a DNA sequence, if it is not a DNA sequence then an error message is thrown. We also are implementing a function that kills any job that has been running for more than 3h.

We hope that these changes will prevent the problems that you have experienced in the future.”

## Reviewed by anonymous reviewer 1, 07 December 2022

I want to thank the authors for their thorough response to the reviewer’s comments. In my opinion, the manuscript improved substantially, and I have only few comments left.

In the methods, it might work better to first describe the Identification of the REPs, REPINs, and RAYTs, and afterwards the implementation and usage of the webserver.

The S. maltophilia example is very interesting due to the patchy presence-absence patterns of REPIN-RAYT systems. Do the authors have any idea how this patchiness evolved, given that the systems evolve vertically?

The risk of confusing CRISPRs with REPINs is mentioned in the introduction and methods. Is it recommended to run a CRISPR detection tool and remove the identified regions from the REPIN candidates? Could this be integrated into the pipeline?

As I understand, the RAYT sequence needs to be known by the user or one of the 2 known sequences needs to be chosen. However, there might not be previous knowledge on the RAYT sequence in the organism. Would it be feasible to include blasting against all known RAYT variants in the pipeline?

## Evaluation round #1

DOI or URL of the preprint: https://www.biorxiv.org/content/10.1101/2022.05.22.493013v2

### Authors' reply, 15 November 2022

**Download author's reply**
**Download tracked changes file**

### Decision by Gavin Douglas, posted 21 July 2022

**Revisions needed**

Two reviewers have now finished their reports and they have highlighted numerous points that should be addressed. The main critique appears to be that further clarification is needed, both in terms of the motivation for annotating these elements in particular and regarding various technical details of your approach.The second reviewer also highlighted several practical issues (as well as discrepancies in the results themselves) that they ran into when trying to run the tool, which I found especially concerning.

I think all of the points that were raised are constructive and should help to improve the manuscript substantially. I look forward to seeing the next version!

### Reviewed by anonymous reviewer 1, 15 June 2022

This preprint presents a tool to identify a particular class of mobile elements in bacterial genomes. Such a tool will make these elements more easy to detect and will allow a wider audience to annotate them. However, some manual steps in the annotation are still necessary which might limit the application of the tool.

The broad scope of annotating REPINs is not completely clear to me. The manuscript gives the impression that manual steps are still needed to annotate REPINs and to link them to RAYTs. Thus, it is currently not possible to include this annotation into pipelines for prokaryote genome annotation (such as PROKKA).

The introduction could be more explicit on the motivation of the study? Why do researchers want to identify REPINs? What kind of studies could this identification support?

The manuscript lacks an introduction into REPINs. How are they defined and how do they look like? E.g., it is mentioned that they are repetitive sequences. How long can the repeats be, how many repeats are there, are they 100% identical, are they consecutive? Although this information is present in previous papers, it is crucial for this manuscript and I suggest to include it in the introduction. Also, it only becomes clear in the discussion that there are symmetric and asymmetric REPINs and the tool only identifies the former ones. Such limitations should be stated in the introduction or methods.

The introduction states ”The study of REPIN populations and their corresponding RAYTs can be cumbersome.” The authors might want to mention the particular challenges in the introduction.

The paper focusses on bacterial REPINs. Do these elements also occur in archaea? Would the tool work for archaeal genomes? That would be interesting to mention in the introduction.

The data set is linked on the RAREFAN website (but 50 strains are mentioned, whereas there are 49 in the manuscript). However, the access is restricted. The access should be unrestricted for review.

Fig. 1
It is unclear what kind of threshold is meant in ”Determine all 21bp long sequences above a certain threshold”. ”vicinity (<30bp)”, however in the legend and in the text it is described that sequences that occur within 15bp

are grouped.
legend: ”Hence, we grouped all sequences that occur within 15 bp of each other, anywhere in the genome.” It is unclear what ”anywhere in the genome” means in that context. As I understand, they are within 15bp, which is not consistent with ”anywhere”.

As described, identical 21bp long sequences are grouped by distance. Then the seed sequence is extracted as the most common sequence in each group (line 105). How can there be multiple different sequences within each group? As I understood, they should all be identical.

It is mentioned that the genomes should ”ideally” be fully sequenced and complete. Does the tool also work with contig-state draft genomes?

The results demonstrate very well how the results depend on the reference genome. The authors then suggest to run the tool with multiple different reference genomes. However, this needs to be done manually, and the potentially different links between REPINs and RAYT are currently resolved manually by the authors (Fig. 2). I got the impression that expert knowledge on REPINs is still required to resolve these multiple runs. Thus, I wonder, whether this process could be automated. I.e. could the analysis be run iteratively with each genome as a reference genome and the results are then merged? This would allow for a fully automated analysis given a set of strains and would largely improve the usability of the tool.

The results also nicely demonstrate how the results depend on the chosen parameters, e.g., the frequency threshold 55. This number looks indeed quite high given the results presented in Fig. 3B. Why is such a high threshold chosen? Do false positive findings increase with lower thresholds? It would be very interesting to discuss this.

An example is described where REPIN groups can be merged (line 300). It is unclear if that is done automatically by the tool.

The authors mention that ”the only known asymmetric REPIN population are E. coli REPINs.” I wonder if that is due to the difficulty in the identification of asymmetric REPINs? Might they have been overlooked?

## Reviewed by Sophie Abby, 21 July 2022

**Review of:**

**“RAREFAN: a webservice to identify REPINs and RAYTs in bacterial genomes”**

In this article, Fortmann-Grote and colleagues present a webservice to identify in bacterial genomes a class of repetitive elements and the associated transposase, namely the REPIN and RAYT. These mobile elements are quite intriguing as they seem to be largely vertically transmitted (i.e. the transposase seems to be rather immobile). Their function is still to be determined. Beyond these elements detection, the webservice also provides some graphs to analyse the search results. As a test case, the authors applied the search engine to a set of 49 genomes of the bacterium Stenotrophomonas maltophilia. The results and limitations of the search are discussed, and some guidelines provided for the users to obtain the most relevant pictures of these elements distribution in the genomes of interest.

The webservice provided could prove useful to microbiologists in need to analyse characteristics of their genomes, and could speed up research on these particular mobile elements. However overall, I found that the description of the method proposed could be largely improved. And I report several inconsistencies observed when running the webservice on authors-provided or original genome datasets, making the webservice results difficult to interpret. I give more details on these aspects and more, in the following review.

**Manuscript review: major points**

- The introduction lacks the necessary biological background to understand the choices made for the search engine implementation. For instance, how many copies of a given REP are usually found in genomes? Of a given REPIN? Why a default number of 55 copies to consider a REP for further search? Are REPs found in REPINs structures always that abundant in genomes? Or are there some REPINs that do correspond to lowly abundant

REP? How long are REPs in REPIN? How long are REPINs? Why use REPs of 21 bp when previous papers by the authors use for instance 16mer searches (Bertels & Bainey 2011)? How many RAYTs are usually found in a genome, are they genetically linked to REPINs? etc... Adding such a paragraph could help the readers to understand the method proposed for REPIN+RAYT detection.

- I know it is "only" a matter of nomenclature but could the authors also mention other names attributed to RAYT? From the Ton-Hoang 2012 paper for instance (TnpAREP if I'm correct)? That could help researchers that are unfamiliar with the literature and the field of repetitive elements to understand exactly what RAREFAN is about.

- As described in Figure 1 and in the main text, I could not properly understand how the REPIN search functions. Please clarify considerably both the figure and the text.

In particular:

1) On Fig. 1:

— A step => add perhaps optional input files (for instance a genome phylogeny if I got it right?)

— B step => "Identifying REP sequence groups" this title would be more explanatory (if I'm correct?). Otherwise please clarify what are "sequence groups".

Step 1) "Determine 21bp long sequences above a certain threshold" of what (number of occurrences, right?)? etc...

Step 2) It is unclear the difference between the groups. Sequences are grouped by vicinity on the reference genome sequence? based on sequence similarity? Please clarify the text.

— B step => performed on a reference "genome" add "genome"?

— B step overall schema could probably be improved to increase clarity.

— C step => "of each for each" typo?

— C step => step 2) REPins are identified from pairs of REPs from within a same group? Or not necessarily? Please clarify.

— The parameters that can be changed by the user could be mentioned on Fig. 1.

— Add at which step is the genome phylogeny computed (and with what). Is this an optional or mandatory step? etc...

2) In main text:

— Line 70, it is mentioned that MCL is used to cluster REPIN sequences. When is this used in RAREFAN? It does not seem to appear on Figure 1.

— Line 104 "All sequences occurring... at least once within 15bp of each other" => I don't understand, could you please clarify? Where does this appear on Fig. 1? Is it rather the 30bp vicinity of step B2?

— Lines 113-114: it is unclear to me whether Group 2 or Group 3 RAYT reference sequences would be used, or both. Please clarify. Is that the user choice? Can both be used if no a priori knowledge is held on which type to find in the genomes to analyse? Also, could you remind here which tblastn parameter is used (cf. line 88)?

— Line 117: please add more explanations on how REPIN populations and RAYT are linked.

— Line 120: please add that it is a user-provided genome phylogeny or a computed one (it was unclear to me, I only got it when going through the webservice pages).

- The authors state that the described method to detect REP sequences has already been described elsewhere (in articles by the authors themselves), but that the present implementation is "slightly improved". Could the authors clarify what is different from the previous methodology, and how this is an improvement? How do the results compare to previous genome analyses performed in some of the cited papers (for instance 1st paragraph of results?).

- Line 171: the authors “suggest to perform multiple RAREFAN runs with different reference strains.” Could there be a relevant way to automatically merge the results from different runs?

- In relation to above comment: Please state in the methods which genomes were used as a reference for the five different runs mentioned in Line 239. How did the authors choose these 5 genomes (sometimes, four are mentioned?), and could there be some hints on how to choose them (ANI-based? based on the genome phylogeny...)?

- Line 180-181: what happens if the seed sequence frequency threshold is lowered for REP search? Would that result in many false positives for REPINs? Or would the obtained candidate REPs naturally be expunged as not part of REPINs? And in terms of computation, would that be considerably slower?

- On the same note, could the authors give a hint about the computational time required and how it scales with the size of the genome dataset to analyse?

- Line 218-225: Interesting observations about the presence of RAYT and REPIN population sizes, but please provide numbers and statistics for the statements in this paragraph.

- Line 244-245: “A detailed analysis of the extragenic space of “wrongly” associated RAYT genes showed that these genes are flanked by seed sequences from two different REPIN populations”.

So how is this handled by RAREFAN? How is this decided which REPIN population is assigned to a RAYT exactly? On Line 117 it is simply written that ”The presence of RAYTs in the vicinity of a particular REPIN can be used to establish the association between the RAYT gene and a REPIN group”. Could this be possible to assign to a RAYT the REPIN population that is most often found next to it? Could this be signified in the log or output files that there are some ambiguities to help guide the user?

- Line 254-256: can the user change the 130bp parameter between a RAYT and REPIN to consider them associated? Please clarify in the text.

- Lines 272-273 and 280: Couldn’t the problem of merged seed groups or split seed groups be sorted automatically by using a sequence clustering and ”dereplication” approach to identify seed sequence to be used for the search (or is this already the case and I didn’t get it)? More generally, what improvements could the authors envision for their tools? Could this be discussed in the Discussion section?

**Manuscript review: minor points**

- “Stenotrophomonas maltophilia” is misspelled line 15 in the keyword list on page 1.

- Line 18 in the abstract: saying that “mobile genetic elements are rare in bacterial genomes” may be a bit strong. Maybe could this more specifically only refer to repetitive elements? If the authors agree with this?

- Line 21: instead of “are vertically inherited”, could the authors consider changing to “seem mostly vertically inherited”? To nuance a bit, as these elements have not been thoroughly studied in many genomes so far?

- Line 92: could this be specified on which servers is RAREFAN run? Is it stably maintained?

- Line 121, you define what is a “master sequence”. Could this concept also appear on Fig. 1 for homogeneity sake?

- Line 212, “P. chlororaphis” please spell out the entire genus name upon first appearance.

**Test of the webservice http://rarefan.evolbio.mpg.de/**

Overall I found difficult to understand the results. Also, I found confusing/inconsistent some of the output sentences on the main Results page and error/warning messages, when faced to the output files results. I also had server connexion issues when accessing the Plot data section. Whether this was a temporary issue with the server or something recurrent, I could not say. Here are the details:

- On the main Results page, regarding REPINs appears the number of REPINs detected in the reference genome. Could it be possible to display the number of REPIN groups and how they distribute among genomes? On the form of a simple table for instance?

- I ran RAREFAN using the “Dodkonia” test dataset provided on the website (from Zenodo) with default parameters (including reference genome chosen by default, dsw-1) and sequence data contained in the “in” folder, there were warnings or errors raised:

“Status: complete with warnings

There have been warning or errors during the postprocessing of your results. Please inspect the output data and logfile (out/rarefan.log) carefully.”

Is this related to the first line of the rarefan.log file reading: “Wrong letter in DNA sequence: |”? I obtained this error with multiple input datasets, is this a bug?

- Using the same “Dodkonia” test dataset, there were no RAYT identified. But several REPIN groups. However, I don’t understand in the Plot data, why the histogram of the REPIN population size (“REPINs” tab in the analysis toolbox) shows only for REPIN population 0, but does not show along trees starting from REPIN group 1? How many REPIN populations were proposed? Where is this information is provided (see also my comment above)?

- When using a dataset I chose (5 Kingella kingae genomes, ran with different reference genomes: runs IDs 92cx136, b2ecb95l and _v6qq4vm), I had the following message on the Results page:

“REPINs

. There was a problem with the REP(IN) analysis output data. Please check your results carefully.”

When is this message provided, and could it be more explicit? Is it linked to the following sentence?

“We detected 0 REPINs in the reference genome.”

- I got the following message on the Kingella dataset:

“Seed sequences

There are 0 21bp long sequences in the reference genome that occur more frequently than 55 times.”

I don’t understand this, as there were several REPIN proposed subsequently? Including in the reference genome? Arent’ the REPIN searches based on REP found in the reference genome, as suggested by Fig. 1? Moreover, there were >70 sequences listed as overrepresented in the file “.overrep”. (example of runs “_v6qq4vm”, or run “b2ecb95l”).

- I could not find the output file called “prox.stats” in both runs (Dodkonia and Kingella) in the downloaded folders. However, they were available in the Dodkonia “out” folder provided on Zenodo.

- I don’t understand why certain maxREPIN_[0-5] files are empty? Could the reason be added to the output file description? Goes the same for presAbs_[0-5].txt files

- When clicking the “Plot data” link, I repeatedly had issues with accessing these. It said: “Disconnected from the server. Reload ”

- Just an observation, in “results.txt”, it seems that the names of the genome files on the form of “GCF_11612705” have been parsed, resulting in 5 columns whenever there are 4 columns in the same output file for the Dodkonia dataset.

- Could the run number be reported in the rarefan.log file? It would be convenient to the user to access previous runs’ results stored on the server. For how long are these runs’ results stored?

- When downloading the Results data as an archive, would it be possible to add to the archive a README file describing the output files? It could for example be directly taken from the text of the `http://rarefan.evolbio.mpg.de/manual` page, section “File output”.

**Download the review**