We would like to thank both the editor and the reviewers for the time and effort they spent helping us improve our manuscript and RAREFAN. Every comment was tremendously helpful and the changes we consequently made to the program and manuscript will make RAREFAN more accessible to the community.

Reply to reviewers

Author's Reply:
by Gavin Douglas, 21 Jul 2022 12:18
Manuscript: https://www.biorxiv.org/content/10.1101/2022.05.22.493013v2
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!

Reviews
Reviewed by anonymous reviewer, 15 Jun 2022 15:42
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).

*It would certainly be interesting to extend RAREFAN to automatically annotate genomes. Currently, we do produce *.tab files to view REPIN/RAYT annotations in, for example, the artemis genome browser. However, the reviewer is right with the assessment that the annotation is currently not sufficiently accurate for automatic annotation programs (see merging of REPIN groups at the end of the manuscript etc.). However, RAREFAN can still be useful because it significantly facilitates the analysis of REPIN/RAYT evolution, despite the need for manual curation of the results.*

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.

> *We have significantly extended the introduction to make it more informative for people that are not familiar with REPINs/RAYTs.*

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 tool would certainly work for archaeal genomes. However, there are no RAYT genes present in archaeal genomes and hence there are likely no REPIN populations unless they are amplified by a different type of transposase. We now mention in the introduction that as far as we know REPIN/RAYT systems only exist in eubacterial genomes.*

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.

> *We absolutely agree, this was an unfortunate oversight. We have now made the datasets available. We also changed 50 to 49.*

Fig. 1
It is unclear what kind of threshold is meant in "Determine all 21bp long sequences above a certain threshold".

> *We changed this to make clear that it is occurrence in the genome (default 55).*

"vicinity (<30bp)", however in the legend and in the text it is described that sequences that occur within 15bp are grouped.

> *Thanks. We changed this to be 15bp.*

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".

> *Yes, this formulation is a bit confusing. What we meant is that in order to be grouped together a single occurrence of the two sequences in the genome needs to be in close proximity. Hence, if sequence 1 occurs 55 times and sequence 2 occurs 42 times then only one of these occurrences of sequence 1 needs to be within 15bp of sequence 2 in order to be sorted into the same sequence group. We amended the legend to include this explanation.*

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.

> *Identical 21bp long sequences are not grouped by distance. Different sequences (each of them occurring multiple times in the genome) are grouped into the same sequence group if they occur in close proximity (see above). Highly abundant sequences often occur in close proximity when they are part of a larger sequence that is itself repetitive. In this case lots of different but overlapping (and if the sequence is very long also nonoverlapping) sequences will be classified as being part of the same group because they are actually part of a longer repeat. The most common sequence in each group will be the group representative. Ideally, this would be the most conserved part of the longer sequence repeat. I hope this explanation will make our method clearer. We have also changed the text in the manuscript as well as Figure 1 to reflect this explanation.*

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

> *RAREFAN should also work with draft genomes. But since REPINs or REPIN clusters are repetitive sequences, the results may not be as accurate as for complete genome analyses. We also have not done much testing on draft genomes, hence we decided to delete the "ideally".*

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 reviewer is absolutely right. Scientists still have to curate RAREFAN results to make sure that REPIN groups are not merged; that all REPINs and RAYTs are identified in the reference; that the reference is not too distantly related from the other sequences etc. We also agree it would be great if RAREFAN could automate all these checks and maybe integrate an AI that would pick up inconsistencies in the results and maybe even identify asymmetric REPINs. Unfortunately, as much as we would like to develop RAREFAN further, we do not have the resources or time (due to a number of rejected grant proposals).

While RAREFAN is far from perfect, we still believe it can immensely facilitate REPIN-RAYT research. We have been using RAREFAN in our institute and RAREFAN made it possible for our interns (such as Julia Balk) without any prior bioinformatics experience or knowledge on REPINs or RAYTs to analyse a complex REPIN-RAYT system in a bacterial species. In that sense, RAREFAN is very useful and achieved what we designed it to do.

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.

The threshold is actually a remnant from a very early analysis of REPINs when we still use 16bp long sequences as seeds. The reason for the 55 threshold was that the most common sequence in a genome without RAYTs/REPINs (P. fluorescens Pf0-1) occurred exactly 55 times. Of course, that number would be lower for 20bp long sequence but we never changed the default settings. One of the reasons we did not change it, is that in most test cases we obtained very good results and as you decrease the threshold the number of false positive groups increases significantly (sequences that should belong to the same group but are split up). In some genomes the number of groups is already very high at that threshold (see for example our Neisseria datasets), which makes it more difficult to identify the number of unique sequence groups.

In summary, selecting a threshold is not trivial because genomes from different species can significantly differ in their repeat content. Some genomes contain lots of repetitive elements that are not REPINs and that will be picked up with a lower threshold. Usually, but not always, REPINs will be the most common short sequence repeat in the genome so most REPINs should make it above the 55 threshold. In our examples, every single REPIN we identified made it across the threshold in at least one of the references. So again, some manual labour is required. But maybe this is not always bad since the repeat landscape of a bacterium can be complex and without closer inspection may be completely misinterpreted.

As to Figure 3B, the number of REPINs is always significantly lower than the number of REP sequences. There were a maximum of 61 REPIN occurrences of Group 1 in S. maltophilia MER1 but the most common 20bp long (REP) sequence occurred a total of

251 times in the same genome. The same REPIN occurred in S. maltophilia Sm53 (the reference) only eight times, but the REP sequence occurred 106 times. The discrepancy between these numbers comes from the fact that there are always a number of REPINs in the genome, which are either not associated with a paired REP sequence or where the paired REP sequence contains mutations (Bertels and Rainey 2011). Furthermore, a REPIN consists of two REP sequences. This means for the REPIN master sequence when it consists of two identical REP sequences the number of REPINs should be at most half of the number of the most abundant REP sequence.

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

The fact that REPIN groups are sometimes merged is an issue with RAREFAN that is caused by REPIN biology. REPINs evolve just like natural populations. Over time REPIN populations diversify and sometimes two diversified REPIN populations end up in the same genome (not sure how, maybe through very rare horizontal transfer of a RAYT gene or very rare duplications of a RAYT gene). When we observe two REPIN populations in the same genome they usually diverged a very long time ago making it very easy to distinguish the populations from each other. However, if the two populations diverged recently then this is not as easy. If we visualize a sequence population as a graph (Figure 6B and C) where each node is a REP sequence and two nodes are connected if they differ in exactly one position then you can see that different REP sequence population can be connected by hybrid REP sequences. Similar to how different natural populations may be distinct in different geographic locations but in intermediate locations we can find hybrids that can breed with both distinct populations.

Since RAREFAN traverses these sequence networks and designates every single sequence it finds in such a connected network to the same REPIN type, it is possible that different REPIN groups are merged into one. One solution would be to just include sequences that differ by at most two nucleotides from the most common seed sequence. But this would lead to a large number of sequences that are never identified (false negative REP/REPIN sequences). Clearly, in the case we presented in Figure 6 it is impossible to perfectly classify each REP sequence to each REP group simply because there seem to be hybrid sequences. And we realize there may be better ways of solving this issue, but we currently do not have the resources or time to pursue them.

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?

Well, we certainly avoid analysing asymmetric REPINs since it makes analyses very hard. But because REPINs are usually the most common repetitive sequences in the bacterial genome, identifying REP sequences should be relatively easy. For example, in E. coli REPINs are in most (all?) cases the most common short repeat in the genome. We hope

that in the future we will find the time to analyse E. coli REPINs in much more detail and maybe we will be able to come up with a way to automatically analyse asymmetric REPINs.

Thanks a lot for all these very helpful and nice comments!

Reviewed by Sophie Abby, 21 Jul 2022 10:33
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.

These are great suggestions. We have significantly extended our introduction to make it more informative.

- 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.

> *We generally like to avoid mentioning other names for RAYTs in order not to confuse readers. tnpA$_{REP}$ was introduced 2 years after Nunvar et al. termed the gene RAYT. We think that creating new names for the same gene for no good reason does not help the community, and really just leads to confusion. Avoiding confusion is also the reason why we used the term RAYT in our publication in 2011 instead of using our working name (RAP) for the gene at the time. Nevertheless, we have included the name tnpA$_{REP}$ in brackets in the introduction. But personally I would recommend not using this term anymore.*

- 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?)

> *In our opinion, the fact that the user can provide a predetermined phylogeny will not make the RAREFAN algorithm easier to understand.*

--- B step => “Identifying REP sequence groups” this title would be more explanatory (if I’m correct?). Otherwise please clarify what are “sequence groups”.

> *All sequences that are more frequent in the genome than the set threshold can be divided into sequence groups. At this point it is not clear whether these groups are REP sequences or other repetitive sequences. Highly abundant sequences are categorized into groups based on their location in the genome. If any of the sequences occur in close proximity to each other anywhere in the genome (< 15bp by default), then the two sequences are placed in the same group.*
>
> *Hence, this comment is slightly tricky, but we decided that calling it “Identifying REP sequence groups” as Sophie suggests does make it easier to understand the figure.*

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

Yes. We changed the text.

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.

Yes, it is vicinity in the reference sequence. Sequence similarity is not taken into account. We have explained this in a question by reviewer one and have changed the text. Briefly, short 20bp long sequences that are part of a larger repeat sequence will share a certain part of their sequence and hence will overlap or occur very closely to each other. So when we check whether we find two different highly abundant sequences close to each other we assume that these sequences are part of the same larger repeat. This is of course not always true, which is why changing the vicinity parameter may lead to different results. A larger vicinity parameter will lead to fewer sequence groups, which may lead to two different sequence groups being combined. When two different sequence groups are combined, one of the two sequence groups will not be analysed since only the most common sequence in each group will be used for further analyses. Conversely, if a parameter is used that is too small, then real sequence groups will be chopped into pieces and the same sequence group will be analysed multiple times.

We have changed the text of the figure legend to make this clearer.

--- B step => performed on a reference “genome” add “genome”?

Done.

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

We have added colours and additional information. We hope it is a bit clearer now.

--- C step => “of each for each” typo?

Gone.

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

Yes, only REPs from the same group. We made this clear now.

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

We have now indicated which parameters are adjustable.

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

We added that a whole genome phylogeny is calculated with andi (Haubold et al. 2015) when no genome phylogeny is supplied.

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.

Currently, MCL results are used as a basis for the REPIN population plot and the "master sequence" plot. We have now added an explanation to Figure 1 (now Figure 2) and the methods section.

--- 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?

Yes, it is. We have corrected the mistake and hopefully made this clearer in the figure as well.

--- 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)?

Yes, this is the user's choice. If there is no a priori knowledge then the user performs a run first with one of the genes and if this run fails to identify RAYTs then performs a rerun with the other RAYT group gene as query. Reruns are an option under RAREFAN and can be performed easily.

We have added the e-Value threshold again.

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

Done.

--- 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).

Thanks! We added explanations and citations.

- 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?).

> *There are a few differences. And the program that ended up being RAREFAN really evolved over the years. Here are the key points I remember us changing.*
>
> *First, we changed the default seed sequence length from 16bp (Bertels and Rainey 2011) to 20bp. The idea here was that it was less likely that different REPIN groups are merged. Although, as we show in the paper REPIN group mergers still happen but it is much less likely than when we use 16bp long sequences. Of course, there is a trade-off here. If the conserved region of the REPIN were to be shorter than 20bp it would be difficult to identify REPINs and REP sequences.*
>
> *Second, in previous RAREFAN derivatives we did not automatically link a REPIN group with a RAYT gene. Linking was always done manually.*
>
> *Third, we made the vicinity parameter an adjustable parameter to prevent merging of REPIN groups when the parameter is too large.*
>
> *Fourth, the ability to automatically identify REPINs was added after 2011 (Bertels and Rainey 2011) for the Quasispecies paper in 2017 (Bertels et al. 2017).*
>
> *Last, the main advantage of RAREFAN is that it is now a webserver and that it should be usable by anyone, which is a work in progress and hopefully usability will improve as more and more users try RAREFAN.*
>
> *We have highlighted the, in our opinion, main difference (RAYTs and REPINs are automatically linked) to previous versions of the method in the text.*

- 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?

> *This is certainly an excellent idea and we will consider it in a forthcoming RAREFAN release, time and resources permitting.*

- 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...)?

> *This is a very good question. There is probably no fool proof way, but we did the following. As you can see in Figure 2 for every RAREFAN run only a selection of identified RAYTs is associated with REPINs. We identified the RAYTs that were not linked to REPIN*

populations and reran RAREFAN with the genomes as reference that contained a RAYT that was unassociated. While this did not always work we quickly identified five genomes that resulted in a link with a REPIN population and all of the identified RAYT genes.

In Figure 4 we show the four different reference sequence runs that link REPIN populations to all main RAYT clades. We also decided to do an additional run with the genome S. maltophilia ISMMS3 because one RAYT from S. maltophilia ISMMS3 RAYTs was still not linked after those four RAREFAN runs. In the ISMMS3 genome we identified sequence palindromes that resemble those of REP sequences. However, these sequences are not found at a high frequency in the genome and hence were not identified by RAREFAN.

In the current version Table 1 shows the reference genomes we used in our analysis and the run_id, which can be used to access these analyses on RAREFAN.

- 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?

Lower sequence thresholds quickly lead to an explosion of the number of sequence groups. Especially in genomes with a lot of MGE activity. But you are right, most of the sequence repeats that are identified do not form REPINs. So this issue should not significantly affect an only REPIN analysis.

Lowering the seed sequence frequency should not significantly affect computation times.

- 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?

*We have included a short paragraph about measurements of elapsed time for complete RAREFAN runs. We timed runs for various species, varying the number of submitted genomes and randomly selecting the reference strain. Results for two different query RAYTs were averaged. The data is presented in two supplementary figures (Supplementary Figure 3 and Supplementary Figure 4). The elapsed time for a given dataset of N genomes of average sequence length L (in megabases) can be estimated as T = (8-10) seconds * N * L.*

- 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.

We now have performed a linear model of independent contrasts of RAYT and REPIN number. The p-Values of those regressions are between 0.003 and 0.008. The p-Value is

still surprisingly high, since only transitions between strains that have no RAYT and strains that have a RAYT contribute to the linear regressions. Differences (independent contrasts) between nodes that either both contain a RAYT or where the RAYT is absent in both strains are always 0.

- 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?

By default, RAREFAN links every REPIN group that is found in the vicinity of a RAYT gene to a RAYT. So even if the wrong REPIN population is linked to a RAYT gene the correct one should also be linked. However, when the correct REPIN population is not found in a reference genome then only the incorrect one will be linked to the RAYT. This is the issue we observed and described in the manuscript. We have rewritten the paragraphs to make this clearer.

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

We have changed RAREFAN to make this an adjustable parameter.

- 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?

The problem of merged sequence groups identified on line 272-273/280 can and has been solved by decreasing the vicinity parameter. However, the more difficult issue to solve is shown in Figure 6B. When two sequences are too closely related to be separated. In the past we have solved this issue by increasing the seed sequence length from 16bp to 20bp. But as you can see this still is a problem in some instances. I am not sure what you mean with “clustering and dereplicating”. However, we have applied MCL in the past, which sorts connected networks into clusters. Unfortunately, MCL still requires the correct parameter to identify the “correct” clusters and maybe there is an automatic way to learn those parameters, but we do not have the expertise here.

As we have done before, it may be possible to increase the seed sequence length again in those situations. But as discussed above this may also lead to an increasing number of REPINs that will not be identified by RAREFAN.

Manuscript review: minor points

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

Changed.

- 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?

Changed.

- 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?

We have changed this to mostly. However, we have studied REPIN-RAYT systems in many hundreds of genomes across Enterobacteria, Pseudomonas, Stenotrophomonas and Neisseria. Within species the pattern so far is always the same: there is basically no evidence of significant horizontal transfer or at least no more than we would expect for housekeeping genes. But of course, it would be great if these studies could be verified/repeated by scientists other than our group. Hopefully, RAREFAN will facilitate this.

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

We added a sentence to the Methods section specifying also the hardware and operating system. The server is currently stably maintained.

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

We have amended the figure legend for Figure 1.

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

Thanks.

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?

> *Changed: The results summary page now shows a table that lists the number of REPINs and number of REP singlets (if "Analyse REPINs" was ticked) or number of REPs for each strain and each group. For runs submitted previous to this change, only the sum of REP/REPINs is displayed.*

- 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?

> *We removed these confusing error messages from the results summary page. Instead the job is now marked as "postprocessing" as soon as the REP/REPIN finder returns without error. The job status then changes to "complete" when RAYT and genome phylogenies have been produced and the data is prepared for download. Instead of one log file for the entire pipeline, we now provide one log file per pipeline stage, making it easier for the user to trace back the cause of unexpected results. If the REP/REPIN finder returns an error, the job is marked as "failed". The results summary page also shows the status of each stage (setup, queued, running, failed, or complete) and the error code if the stage failed. In case of failure, the error log of each failed stage is printed out.*
>
> *The REP/REPIN finder is applied to a single character sequence. Individual sequences are separated by the "|" character. The warning "Wrong letter in DNA sequence" signals that for the subsequence containing "|" REP/REPINs will not be identified, which is the intended behaviour. It is not related to other warning that are printed on the results summary page. We have now removed this message from the RAREFAN log output.*

We have also removed any other confusing and unspecific outputs from the log files to make it easier for the user to interpret the log file. The log from MCL is also removed.

- 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)?

We changed both the results page and the plot page. The results page now contains a table with the number of REP and REPIN sequences that were identified. The plot page now provides a message for empty groups that states, that no REPINs/REPs/RAYTs could be identified.

- 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."

*As outlined above, we reworked the entire log and error handling. The general message "*There was a problem with the REP(IN) analysis output data. Please check your results carefully*" does not appear anymore, instead a detailed report of each stage in the RAREFAN pipeline is generated.*

- 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").

We changed this message and now show a summary table of all the different REPs/REPINs identified in each of the submitted strains.

- 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.

> *The data was generated with an earlier version of RAREFAN. The file “prox.stats” does no longer exist.*

- 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

> *If there are no REPINs identified for a sequence group then these are not shown in the presAbs file. In the Kingella dataset the sequences from the first few sequence groups do not form REPINs so there is no data to be shown in the presAbs files. I hope now that we show a summary table this is clearer. But please let us know if you can think of a better way to communicate these results.*

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

> *There have been several problems with the R shiny-app server software that we have now fixed.*

- 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.

> *Thanks a lot! We fixed this bug now.*

- 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?

> *Results are currently stored for 180 days. We have changed the log file mechanisms such that the run id is reported in the first line of rarefan.log*

- 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”.

> *A file “readme.md” is now added to each output directory explaining the content of each file in a table. The table is identical to the one found in the RAREFAN online manual.*

Thanks a lot for all the work you have put into testing RAREFAN! We really appreciate this. I had a look at the Kingella dataset. One of the genomes does indeed carry a RAYT gene and it seems to be associated with a repetitive sequence that forms an unusually structured "REPIN". However, this repetitive sequence seems to be spread throughout the genome in very regular intervals (about 1000bp) and it occurs over a 1000 times. I have never seen anything like this before and these sequences do not really look like REPINs. So I do not really know what is going on in Kingella.