Thanks again to both reviewers and the editor for an extremely thorough review of our webserver. We did our best to fix all the issues that were identified.

by Gavin Douglas, 29 Dec 2022 15:38
Manuscript:
[https://www.biorxiv.org/content/10.1101/2022.05.22.493013v3 | https://www.biorxiv.org/content/10.1101/2022.05.22.493013v3] version 3
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.

Done.

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

Done.

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.

RAYTs have been identified originally as transposases that are connected to REPs (Nunvar et al, 2010). Subsequently we have found that there are at least three more RAYT families related to the originally identified RAYT families. But since the newly identified RAYT families are about as closely related to the other two RAYT families as RAYT families 2 and 3 were to each other, they have also been named RAYTs.

*We have changed the sentence to: “*Of a total of five different RAYT families, there are only two RAYT families that are associated with repetitive sequences such as REPIN or REP sequences: Group 2 and Group 3 RAYTs (Bertels, Gallie, *et al.* 2017).”

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.

We double checked and changed inconsistent parameter values.

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).

We have shortened the figure legend.

I was unable to access results under run ID a2ijpkk6.

Results are available now.

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.

That's a very good point. This issue has puzzled me for a long time. A priori you are absolutely right, I think there is no obvious reason why this should be true. Especially considering the wide variety of different RAYT/REPIN associations (single REP vs, REPINs in both flanking regions, REPIN+REP vs …). However, it seems to be true with some rare exceptions that RAYTs are almost always associated with REPIN/REP sequences. I guess this is because REPIN/REP sequences are necessary for accurate RAYT functions, and in the rare cases where RAYTs are not associated with REPs/REPINs the RAYT may already be non-functional. We have made a short statement explaining that we do not know the reason for the strong association between REPINs/REPs and RAYTs.

We added the following sentences to the Identification of RAYTs section: "REPIN or REP sequences are almost always present in the extragenic spaces of RAYT genes and this linkage is consistent across the RAYT phylogeny (as shown in **Figure 5**). However, linkage cause is unclear."

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 suggested change in tense subtly alters the meaning of the sentence; use of the past tense implies that, if the analysis was repeated, a different result would be obtained. This is only true if the input or RAREFAN were changed in some way. However, if the same input and RAREFAN version are applied, the results will be consistent. Therefore, we prefer to keep the present tense.

The authors should use "p-value", "P-value", or "p value", but not "p-Value", which is the current usage in the text.

Done.

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.

We have added a sentence to describe the behavior of RAREFAN.

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.

All done. Thanks a lot!

Reviews
Reviewed by Sophie Abby, 23 Dec 2022 17:08

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?

Yes, example dataset are not deleted.

- Figure 4 legend: “In an equilibrium” => “At equilibrium”?

- Line 281: “The RAREFAN webserver visualizes REPIN population size” => “enables to visualize…”?

We prefer the original phrasing.

- Line 327: “In some RAREFAN runs associations between RAYTs and REPINs are not monophyletic”. Please reformulate, associations are not the ones to be monophyletic.

I am sorry, we do not understand the comment. As you can see in Figure 3, we highlight REPIN-RAYT associations in different colors. Because some of the associations (different colors) are not monophyletic on the RAYT tree, we analyze these in more detail. These analysis lead to the conclusion that RAREFAN can link REPINs and RAYTs erroneously for a number of reasons.

- Figure 6 legend: “connect two sequence cluster.” => clusterS

- Line 180: “have been observed”

All done. Thanks!

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).

> *We fixed this issue. The table is now consistent with the plotted graphs. The issue was that the group numbers were ordered alphabetically and not numerically.*

- 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?

> *We fixed this issue by accepting all sequences containing characters listed on https://en.wikipedia.org/wiki/Nucleic_acid_sequence.*

---------- 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, 07 Dec 2022 08:16

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.

Thanks again for all the time the reviewer invested in reviewing our manuscript!

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.

We agree. We have changed the structure of the methods section.

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?

This is a very good question. Presumably, there are many factors at work. One of the main factors is the loss of RAYT genes. Another important factor is gene duplication followed by differentiation. REPIN-RAYT systems do very rarely duplicate. But when duplication occurs it can allow for differentiation of the RAYT and the associated REPIN group. For some REPIN-RAYT systems this may have happened a very long time ago, leading to very distinct REPIN sequences. However, for some REPIN groups this has happened more recently, which leads to very similar and hard to differentiate REPIN populations (see for example ***Figure 6****). It is possible that some RAYT genes are also horizontally transferred from more distantly related species. This has probably happened when cyanobacteria acquired REPIN-RAYT systems (Bertels, Gallie and Rainey 2017). To understand horizontal transfer of REPIN-RAYT systems in more closely related species much more research is required. We*

are currently trying to understand REPIN-RAYT evolution in P. fluorescens and it looks like gene conversion may play a significant role there.

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?

While CRISPRs have a similar structure to REPINs (inverted repeats +variable spacer), there are some features that allow us to easily distinguish REPINs from CRISPRs. First, CRISPRs are found in tandem repeats usually in a single locus of the genome, while REPINs are distributed across the genome. Second, the long persistence times of REPINs lead to the formation of a REPIN cloud in sequence space (i.e. large REPIN sequence diversity), while all CRISPR repeats in a single genome are almost entirely identical.

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?

To our knowledge there are only two RAYT families that are associated with REPINs (Group 2 RAYTs and Group 3 RAYTs, see Bertels, Gallie, and Rainey 2017). The RAYT from E. coli is part of Group 2 and the RAYT from P. flourescens SBW25 is part of Group 3. When using TBLASTN with a Group 3 RAYT it should identify any other Group 3 RAYT present in the query genome, similarly using the E. coli Group 2 RAYT it should identify any other Group 2 RAYT in the query genome. Hence, REPIN associated RAYTs should be identifiably with using the two RAYTs provided by RAREFAN.