Review of:
"RAREFAN: a webservice to identify REPINs and RAYTs in bacterial genomes"
For PCI Genomics

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 (TnpA_{REP} 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”.