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

Carsten Fortmann-Grote, Julia Balk and Frederic Bertels

Max-Planck-Institute for Evolutionary Biology, Department of Microbial Population Biology

Corresponding author: Frederic Bertels, August-Thienemann-Straße 2, 24306 Plön, Germany, bertels@evolbio.mpg.de.

Running title: REPIN/RAYT Finder and ANalyzer

Keywords: sequence analysis – mobile genetic elements – bacterial genomes – *Stenotrophomonas maltophila*

Abstract

Compared to eukaryotes, mobile genetic elements are rare in bacterial genomes and usually do not persist for long in the genome. Yet, there is at least one class of persistent prokaryotic mobile genetic elements: REPINs. REPINs are non-autonomous transposable elements replicated by single-copy transposases called RAYTs. REPIN-RAYT systems are vertically inherited and have persisted in individual bacterial lineages for millions of years. Discovering and analyzing REPIN populations and their corresponding RAYT transposases in bacterial species can be rather laborious, hampering progress in understanding REPIN-RAYT biology and evolution. Here we present RAREFAN, a webservice that identifies REPIN populations and their corresponding RAYT transposase in a given set of bacterial genomes. We demonstrate RAREFAN's capabilities by analyzing a set of 49 *Stenotrophomonas maltophilia* genomes, containing nine different REPIN-RAYT systems. We guide the reader through the process of identifying and analyzing REPIN-RAYT systems across *S. maltophilia*, highlighting erroneous associations between REPIN and RAYTs, and providing solutions on how to find the correct associations. RAREFAN enables rapid, large-scale detection of REPINs and RAYTs, and provides insight into the fascinating world of intragenomic sequence populations in bacterial genomes.

Introduction

Repetitive sequences in bacteria are rare compared to most eukaryotic genomes. In eukaryotic genomes, repetitive sequences are the result of the activities of persistent parasitic transposable elements. In bacteria, in contrast, parasitic transposable elements cannot persist for long periods of time (Park *et al.* 2021; van Dijk *et al.* 2022). To persist in the gene pool, transposable elements must constantly infect novel hosts (Sawyer *et al.* 1987; Lawrence *et al.* 1992; Bichsel *et al.* 2010; Rankin *et al.* 2010; Wu *et al.* 2015; Park *et al.* 2021). Yet, there is at least one exception: a class of transposable elements called REPINs.

REPINs (**REP** doublet forming hair**PINs**) are bacterial repetitive sequences that occur in extragenic spaces (Bertels, Rainey 2011). REPINs are non-autonomous mobile genetic elements that are duplicated by a domesticated, single-copy RAYT transposase (Nunvar *et al.* 2010; Bertels, Rainey 2011; Ton-Hoang *et al.* 2012). In contrast to typical bacterial mobile genetic elements, REPINs have persisted for at least 100 million years in various species (Bertels, Gallie, *et al.* 2017; Park *et al.* 2021; Bertels, Rainey 2022), in the absence of horizontal transfer of RAYT transposases or REPIN populations (Bertels, Gallie, *et al.* 2017; Park *et al.* 2021; Bertels, Rainey 2022). These evolutionary characteristics are consistent with REPIN-RAYT systems providing a benefit to the host (Bertels, Rainey 2022).

The study of REPIN populations and their corresponding RAYTs can be cumbersome. To facilitate REPIN studies, we have developed a webservice called RAREFAN (**RA**YT/**RE**PIN **F**inder and **AN**alyzer). RAREFAN is publicly accessible at http://rarefan.evolbio.mpg.de and identifies REPIN populations and RAYTs inside bacterial genomes. RAREFAN also generates graphs to visualize the population dynamics of REPINs, and assigns RAYT genes to their corresponding REPIN groups. Here we demonstrate RAREFAN's functionality by analyzing REPIN-RAYT systems in the bacterial species *Stenotrophomonas maltophilia*.

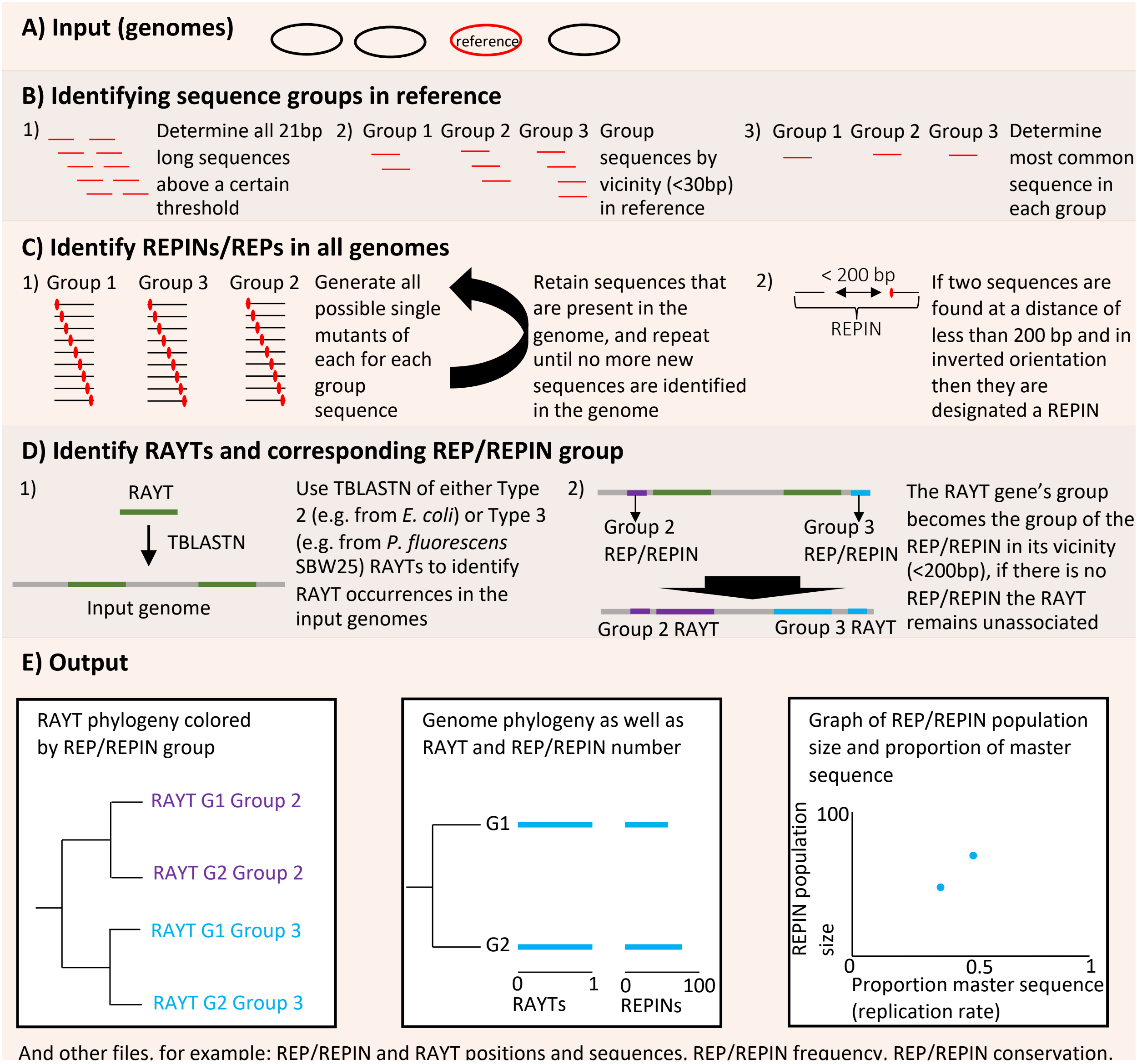


Figure 1. RAREFAN workflow. (**a**) By default, RAREFAN requires the user to supply input sequences containing RAYTs and REPINs. These are, ideally, fully sequenced and complete genomes. (**b**) RAREFAN then identifies seed sequence groups in the reference genome by first isolating all 21 bp long sequences that occur more than 55 times in the reference genome. It is likely that a large number of these sequences belong to the same REPIN type. Hence, we grouped all sequences that occur within 15 bp of each other, anywhere in the genome. All further analyses are performed only with the most common sequence in each sequence group. This sequence will be called the seed sequence. (**c**) The occurrences of the seed and mutated seed sequences are identified in all submitted genomes. If a mutated seed sequence is identified in a genome, then all single mutants of that seed sequence are searched recursively in the same genome. All identified sequences that occur within 130 bp in inverted orientation of each other are designated REPINs. All other identified seed sequences and mutated seed sequences are REP singlets. (**d**) TBLASTN is used to identify RAYT homologs across all submitted genomes. If a RAYT homolog is in the vicinity of a previously identified REPIN or REP singlet, then this RAYT is

designated as associated with this REPIN group. (**e**) Finally, RAREFAN plots three different summary graphs and generates various files containing, for example, REP, REPIN, or RAYT sequences and their positions in the query genomes.

Methods

Implementation

RAREFAN is a modular webservice. It consists of a web frontend written in the python programming language (Van Rossum, Drake Jr 1995) using the flask framework (Grinberg 2018), a java (Arnold *et al.* 2005) backend for genomic sequence analysis and an R (R Core Team 2016) shiny app (RStudio, Inc 2013) for data visualization. The software is developed and tested under the Debian GNU/Linux operating system (Kleinmann *et al.* 2021). All components are released under the MIT opensource license (Initiative 2021) and can be obtained from our public GitHub repository at https://github.com/mpievolbio-scicom/rarefanhttps://github.com/mpievolbio-scicomp/rarefan.

The java backend drives the sequence analysis. It makes system calls to TBLASTN (Altschul *et al.* 1990) to identify RAYT homologs and to MCL (Van Dongen 2000) for clustering REPIN sequences.

Jobs submitted through the web server are queued and executed as soon as the required resources become available. Users are informed about the status of their jobs. After job completion, the user can trigger the R shiny app to visualize the results.

The java backend can also be run locally *via* the command line interface (available for download at https://github.com/mpievolbio-scicomp/rarefan/releases).

Usage of the webservice

The front page of our webservice allows users to upload their bacterial genomes in FASTA (.fas) format (**Figure 1A**). Optionally, users may also provide RAYT protein FASTA sequences (.faa) or phylogenies in NEWICK (.nwk) format. After successful completion of the upload process, the user fills out a web form to specify the parameters of the algorithm:

- Reference sequence: Which of the uploaded genome sequences will be designated as reference genome (see below for explanations). Defaults to the first uploaded filename in alphabetical order.
- Seed sequence length: The seed sequence length (in base pairs) is used to identify REPIN candidates from the input genomes. Default is 21 bp.
- Minimum seed sequence frequency: Lower limit on seed sequence frequency in the reference genome to be considered as a REP candidate. Default is 55.
- e-value cut-off: Alignment e-value cut-off for identifying RAYT homologs with TBLASTN. Default is 1e-30.
- User email (optional): If provided, then the user will be notified by email upon run completion.

The job is then ready for submission to the job queue. Upon job completion, links to browse and to download the results, as well as a link to a visualization dashboard are provided. If a job runs for a long time then users may also come back to RAREFAN at a later time, query their job status and eventually retrieve their results by entering the run ID into the search field at http://rarefan.evolbio.mpg.de/results. Relevant links and the run ID are communicated either on the status site or by email if the user provided their email address during run configuration.

Identification of REPs and REPINs

The algorithm to determine REP sequence groups has been described in previous papers and is slightly improved in our implementation (Bertels, Rainey 2011, 2022; Bertels, Gokhale, *et al.* 2017). First, all N bp (21 bp by default) long seed sequences that occur more than M times (55 by default) are extracted from the reference genome. N and M are the seed sequence length and minimum seed sequence frequency, respectively (**Figure 1B**). All sequences occurring within the reference genome at least once within 15 bp of each other are then grouped together into n REP sequence groups (numbered 0-(n-1)). The most common sequence in each group, named REP seed sequence, is used for further analyses in each input genome.

In the next step all possible point mutants of the seed sequences are generated and searched for in the genome (**Figure 1C**). If a sequence is found in the genome, then all possible point mutations are generated for this sequence as well and so on until no more sequences can be identified. If two sequences are found within 130 bp of each other in inverted orientation, then these are designated REPINs.

Identification of RAYTs

RAYTs are identified using TBLASTN (Camacho *et al.* 2009) with either a protein sequence provided by the user or a Group 2 RAYT from *Eschericha coli* (yafM, Uniprot accession Q47152) or a Group 3 RAYT from *P. fluorescens* SBW25 (yafM, Uniprot accession C3JZZ6). 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 (**Figure 1D**).

Visualizations

For each REPIN-RAYT group summary plots are generated. These include plots showing the RAYT phylogeny, REPIN population sizes in relation to the genome phylogeny as well as the proportion of master sequences (most common REPIN in a REPIN population) in relation to REPIN population size (**Figure 1E**).

Other outputs

Identified REPINs, REP singlets as well as RAYTs are written to FASTA formatted sequence files and to tab formatted annotation files that can be read with the Artemis genome browser (Rutherford *et al.* 2000). The REPIN-RAYT associations as well as the number of RAYT copies per genome are written to tabular data files. A detailed description of all output files is provided in the manual (http://rarefan.evolbio.mpg.de/manual).

Sequence analysis and annotation

For verification of RAREFAN results, REPIN-RAYT-systems were analyzed in their corresponding genomes using Geneious Prime (Java Version 11.0.12+7 (64 bit); Biomatters). Nucleotide sequences and positions of REP singlets, REPINs, and RAYTs were extracted from output files generated by RAREFAN and mapped in the relevant *S. maltophilia* genome. The association of a RAYT gene to a REPIN population has been validated when the corresponding seed sequence is flanking both ends of the RAYT gene within 130 bp. Complete RAREFAN data used for analysis can be accessed by using the run IDs listed in **Table 1.**

Table 1. RAREFAN IDs linking to the raw data of the presented analyses.

Run ID	Reference genome
_1a8l7wu	*S. maltophilia* Sm53
mknhxpv8	*S. maltophilia* AA1
pgfmaxx5	*S. maltophilia* FDAARGO_649
yy72i755	*S. maltophilia* AB550
78eu9zl0	*S. maltophilia* ISMMS3

Associated data can accessed by entering the run ID at http://rarefan.evolbio.mpg.de/results.

Results

RAREFAN can identify REPINs and their corresponding RAYTs in a set of – ideally fully sequenced – bacterial genomes. The RAREFAN algorithm has been used in previous analyses to identify and characterize REPINs and RAYTs in *Pseudomonas* (Bertels, Rainey 2011, 2022), *Neisseria* (Bertels, Rainey 2022), and Enterobacteria (Bertels, Gallie, *et al.* 2017; Park *et al.* 2021). To demonstrate RAREFAN's capabilities, we are presenting an analysis of 49 strains belonging to the opportunistic pathogen *S. maltophilia*.

S. maltophilia strains contain Group 3 RAYTs, which are also commonly found in plant-associated *Pseudomonas* species such as *P. fluorescens* or *P. syringae* (Bertels, Rainey 2011, 2022). Similar to Group 3 RAYTs in other species, *S. maltophilia* contains multiple REPIN-RAYT systems per genome. Group 2 RAYTs, in contrast, contain only ever one REPIN-RAYT system per genome (Bertels, Rainey 2022).

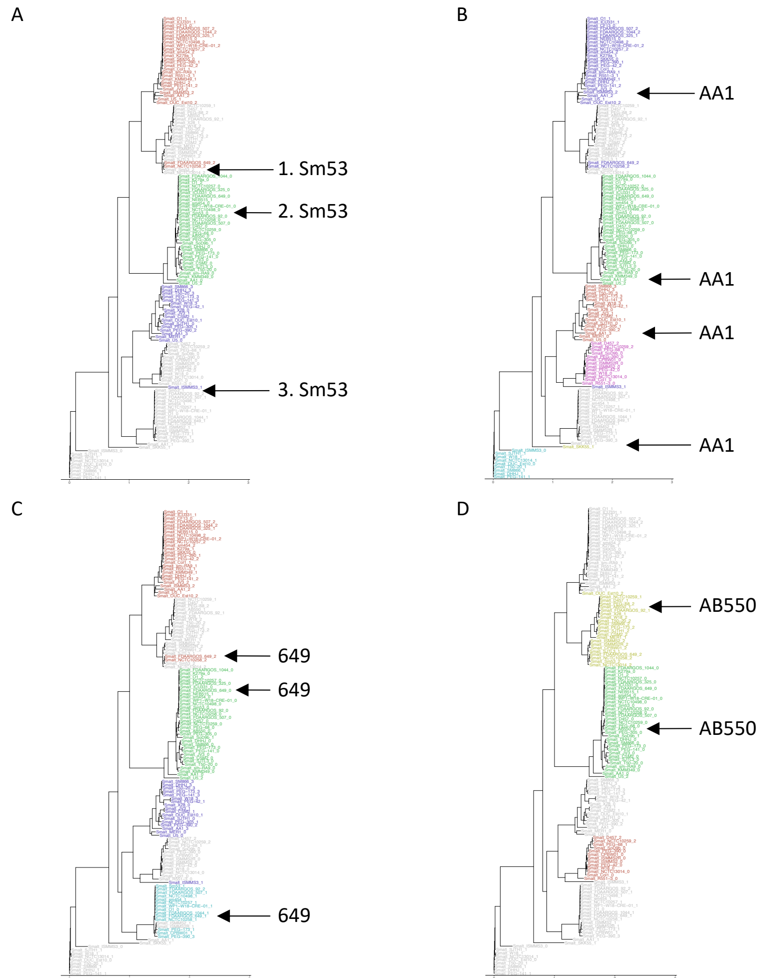


Figure 2. Phylogenetic trees built from RAYT genes extracted from *S. maltophilia* genomes. RAYT genes are coloured according to their association with REPIN populations in the reference genome. If no REPIN population present in the reference genome is associated with a RAYT gene the RAYT gene is coloured in grey. The four panels **A-D** show phylogenies for four different reference strains. *S. maltophilia* strains Sm53, AA1, 649 and AB550 were used in panels **A** to **D**, respectively. Locations of a reference strain's RAYT genes in the tree are indicated by arrows. An association between almost all RAYTs and REPIN populations could be made by using four different reference genomes. Most of the RAYT genes are coloured (associated to a REPIN group) in at least one of the trees. The three numbered RAYT genes from the Sm53 RAREFAN run are referenced in the text.

Nine different REPIN-RAYT systems in S. maltophilia

REPIN-RAYT systems in *S. maltophilia* are surprisingly diverse compared to other species. For example, *P. chlororaphis* contains three separate REPIN populations that are present in all *P.*

chlororaphis strains, each associated with its cognate RAYT gene (Bertels, Rainey 2022). *S. maltophilia,* in contrast*,* contains only one REPIN-RAYT system that is present across almost the entire species (green clade in **Figure 2**), and at least eight REPIN-RAYT systems that are present in subsets of strains (nine clades in **Figure 4**).

The patchy presence-absence pattern of REPIN-RAYT systems in *S. maltophilia,* makes the dataset quite challenging to analyse. If a REPIN population is not present in the reference strain then RAREFAN will not be able to detect it in any other strain. Yet, it is possible to detect RAYT genes in all strains of a species independent of the reference strain selection. RAYT genes that are not associated to a REPIN population are displayed in grey (**Figure 2A**). While these RAYT genes are not associated to REPIN populations detected in the reference strain, they might still be associated with a yet unidentified REPIN type. In order to identify all REPIN populations across a species, we suggest to perform multiple RAREFAN runs with different reference strains. The RAREFAN web interface supports re-launching a given job with modified parameters.

To identify as many different REPIN-RAYT systems as possible in each subsequent run the reference should be set to a genome that contains RAYTs that were not associated with a REPIN population previously (i.e., genomes containing grey RAYTs in **Figure 2**). However, this strategy may also fail when the REPIN population size falls below the RAREFAN seed sequence frequency threshold.

For example, *S. maltophilia* Sm53 contains three RAYTs only one of which is associated with a REPIN population (RAYT genes indicated in **Figure 2A**). However, the remaining two RAYTs are indeed associated with a REPIN population, but these REPIN populations are too small to be detected in *S. maltophilia* Sm53 (the seed sequence frequency threshold is set to 55 by default). In other *S. maltophilia* strains the REPIN populations are large enough to exceed the threshold. For example, if *S. maltophilia* AB550 is set as reference, RAYT number 1 from Sm53 (**Figure 2A**) is associated with the yellow REPIN population (**Figure 2D**). If *S. maltophilia* 649 is set as reference RAYT number 3 from Sm53 (**Figure 2A**) is associated with the turquoise REPIN population (**Figure 2C**). RAYTs from the bottom clade are only associated with REPIN populations when *S. maltophilia* AA1 is chosen as reference (**Figure 2B**).

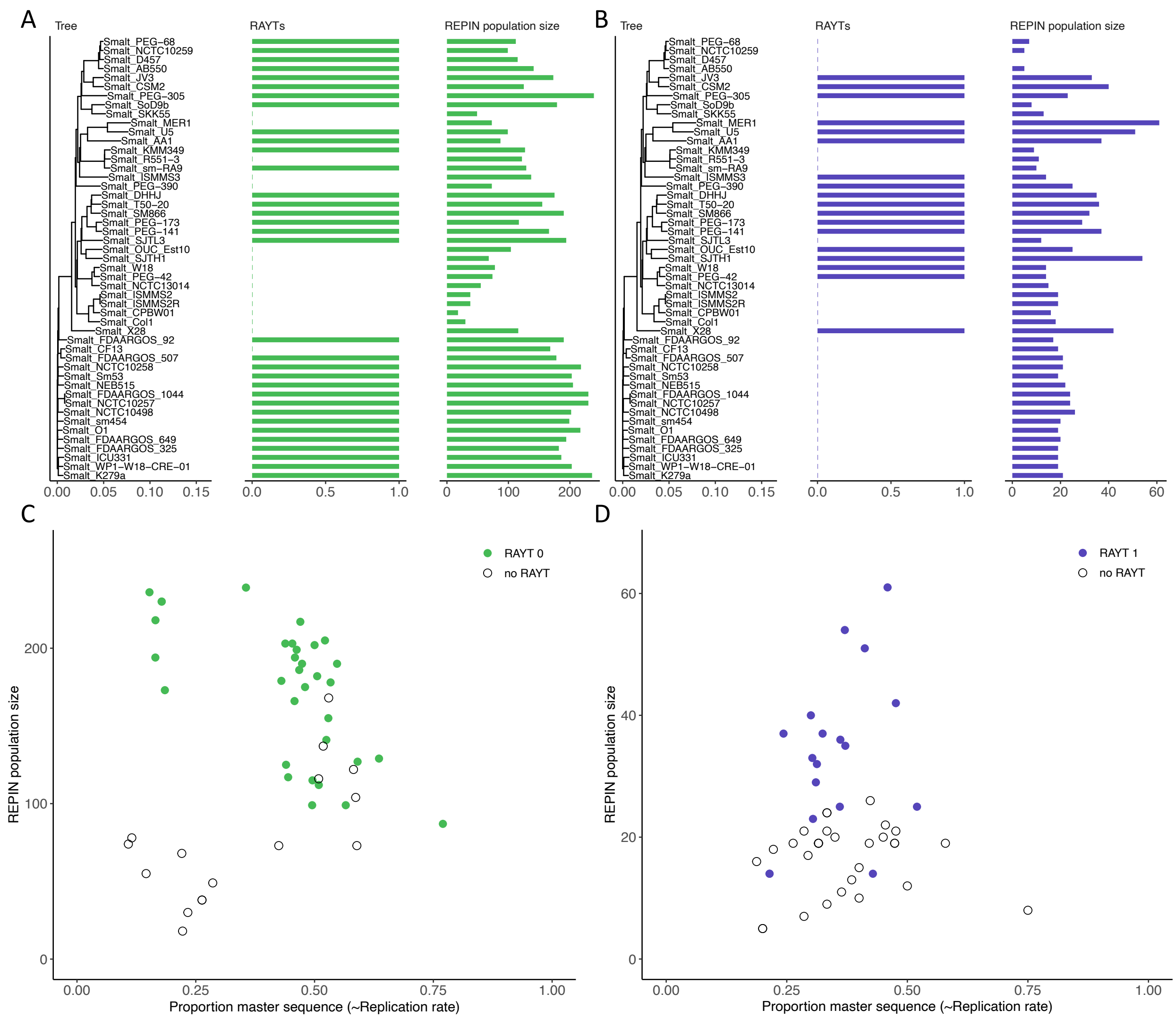


Figure 3. REPIN population sizes and conservation. The plots show two REPIN populations and their associated RAYTs that were identified in *S. maltophilia* using *S. maltophilia* Sm53 as reference. (**A**) The phylogenetic tree on the left side is a whole genome phylogeny generated by andi (Haubold *et al.* 2015). Shown on the right are REPIN population sizes and the number of associated RAYTs sorted by the genome phylogeny. The green REPIN populations and associated RAYTs are present in most strains in high abundance (maximum of 239 occurrences in *S. maltophilia* K279a, left panel). (**B**) The blue population in contrast is present in much lower numbers (maximum of 61 occurrences in *S. maltophilia* MER1, right panel). Note, REPIN populations are assigned consistent colours based on their abundance in the reference genome. For example, the most abundant REPIN population in the reference is always coloured in green, and the second most abundant population is always coloured in blue. (**C and D**) Proportion of master sequence in *S. maltophilia* REPIN populations. The master sequence in a REPIN population is the most common REPIN sequence. The higher the proportion of the master sequence in the population the higher the replication rate (Bertels, Gokhale, *et al.* 2017). The presence and

absence of an associated RAYT is also indicated by the colours of the dots. Empty circles indicate that the REPIN population is not associated with a RAYT gene in that genome.

RAREFAN visualizes REPIN population size and replication rate

The RAREFAN webserver visualizes REPIN population size and RAYT numbers in barplots. The barplot is ordered by the phylogenetic relationship of the submitted bacterial strains (Yu *et al.* 2018). RAREFAN detects three populations when *S. maltophilia* Sm53 is selected as reference strain (**Figure 2A**). The largest REPIN population has a corresponding RAYT gene in almost all strains (first barplot in **Figure 3A**) and most REPIN populations contain more than 100 REPINs (second barplot in **Figure 3A**). The second largest REPIN population in Sm53 (purple population in **Figure 3B**) is significantly smaller and contains no more than 61 REPINs in any strain and most strains do not contain a corresponding RAYT for this population.

RAREFAN also provides information on REPIN replication rate (**Figure 3C** and **D**). REPIN replication rate can be estimated by dividing the number of the most common REPIN sequence (master sequence) by the REPIN population size (Bertels, Gokhale, *et al.* 2017). If a REPIN replicates very fast most of the population will consist of identical sequences because mutations do not have enough time to accumulate between replication events. Hence, the proportion of master sequences will be high in populations that have a high replication rate. Transposable elements such as insertion sequences consist almost exclusively of identical master sequences because the time between replication events is not sufficient to accumulate mutations and because quick extinction of the element usually prevents the accumulation of mutations after replication (Park *et al.* 2021; Bertels, Rainey 2022). REPIN populations in contrast replicate slowly and persist for long periods of time, which means that a high proportion of master sequences suggests a high REPIN replication rate.

In *S. maltophilia* the proportion of master sequences in the population does not seem to correlate well with REPIN population size, both in the green and the purple population (**Figures 3C and D**). Similar observations have been made in *P. chlororaphis* (Bertels, Rainey 2022), and may suggest that an increase in population size is not caused by an increase in replication rate. Population size is likely to be more strongly affected by other factors such as the loss of the corresponding RAYT gene, which leads to the decay of the REPIN population. One could even speculate that high

REPIN replication rates are more likely to lead to the eventual demise of the population due to the negative fitness effect on the host (Bertels, Rainey 2022).

The presence of RAYTs and the size of the corresponding REPIN population do correlate surprisingly well. RAYTs are absent from an entire *S. maltophilia* clade (middle of **Figure 3A**). This clade has also lost a significant amount of green REPINs, and the proportion of the master sequences in these populations is low (**Figure 3C**). Similarly, genomes without RAYTs have smaller REPIN populations in the purple population than genomes with the corresponding RAYT (**Figure 3D**). A similar observation has been made previously in *E. coli, P. chlororaphis* and *Neisseria* where the loss of the RAYT gene is followed by a decay of the REPIN population (Bertels, Rainey 2022).

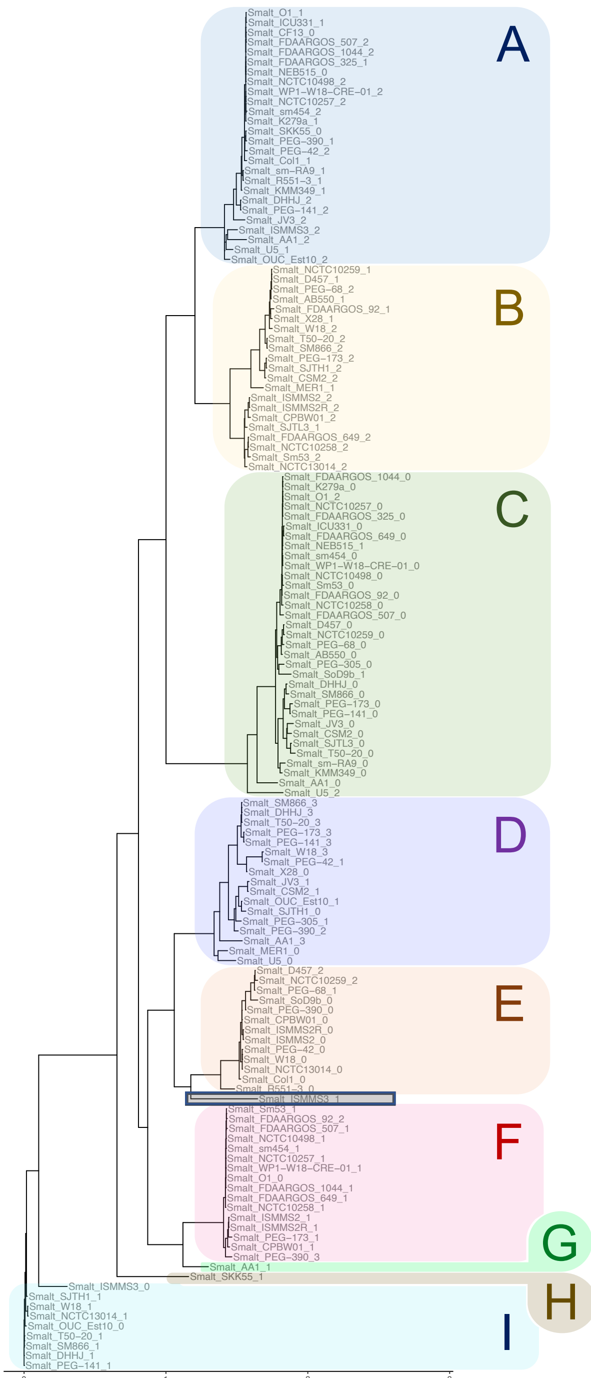


Figure 4. Phylogeny of RAYT genes and their associated REPINs. The tree shows RAYT genes from 49 *S. maltophilia* strains. Colours of clades A-I are assigned according to their association with a REPIN found within 130 bp to the RAYT gene (see **Table 2**). Except for a single RAYT gene ISMMS3_1 (grey box), which could not be linked to a REPIN population.

Table 2. REPIN palindromes associated with each RAYT clade.

RAYT population	REPIN palindromes

A	CCGACC**AAC**GGTCGG
B	CCAACC**AA**GGTTGGC
C	CCGGCC**AGC**GGCCGG
D	TCCACGC**ATG**GCGTGGA
E	CCGAGC**CCAT**GCTCGG
F	TCGACT**AAC**AGTCGA
G	TCGACC**AAC**GGTCGA
H	GCCGGGC**ATG**GCCCGGC
I	AGTCGAGC**TT**GCTCGACT

Each RAYT clade from **Figure 4** is associated with a unique imperfect palindrome that is present at the 5' and 3' end of the REPIN (**Figure 5**).

Linking REPIN populations with RAYT genes can be challenging

Unfortunately RAREFAN is not always able to link the correct REPIN population with the correct RAYT gene. As shown in **Figure 2** some RAREFAN runs indicate that sometimes associations between RAYTs and REPINs seem to not be monophyletic, as for example RAYTs highlighted in red in **Figure 2A**. However, the same clade of RAYTs is uniformly coloured in yellow in Figure 2D, suggesting that the entire RAYT clade is associated with the same REPIN group. To investigate the true relationships between REPINs and RAYTs we first mapped REPIN groups to RAYT genes.

An analysis of all REPIN groups that were identified by RAREFAN across five different RAREFAN runs showed that there are a total of nine different REPIN groups, each defined by an individual central palindrome (**Table 2**). Each REPIN group is associated with a mono-phyletic RAYT group (**Figure 4**). There is only a single RAYT that we could not identify a REPIN for (ISMMS3_1). Although there seems to be a one to one mapping between RAYT clades and REPIN groups, the question remains why RAREFAN sometimes links RAYT genes to the wrong REPIN group.

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 **(Supplementary Figure 1 A-C)**. A single REP sequence from the "wrong" (non-monophyletic RAYT) clade occurs together with multiple REP or REPIN sequences from the "right" (monophyletic in a different RAREFAN run) clade. REPINs are linked to the "wrong" RAYT when the correct REPIN population is absent in the chosen reference genome. Moreover, the "wrongly" associated REP singlets always show up as belonging to the REPIN population of a RAYT sister group and show high sequence similarity.

Additionally to linking the wrong REPINs and RAYTs, RAREFAN sometimes failed to link RAYT genes with any REPINs. Detailed sequence analyses showed that in two out of a total of three such cases the REPIN was located at a distance of more than 130bp (a parameter set in RAREFAN that could be modified) **(Supplementary Figure 1 D-E)**. REPINs that are located at a distance of more than 130bp are ignored by RAREFAN. In a third case there was no REPIN that could be linked to the RAYT gene ISMMS3_1 (**Figure 4**, grey box). While there is a sequence that resembles the A palindrome as well as variants of the C-palindrome flanking both sides of the RAYT gene, (**Supplementary Figure 2**), none of the sequences formed REPIN populations large enough to be identified by RAREFAN. Presumably the RAYT ISMMS3_1, which is only present in a single *S. maltophilia* strain, is at the early stages of establishing a REPIN population. Based on our findings, RAREFAN users should always critically analyse RAREFAN results, particularly when the results require unusual evolutionary explanations.

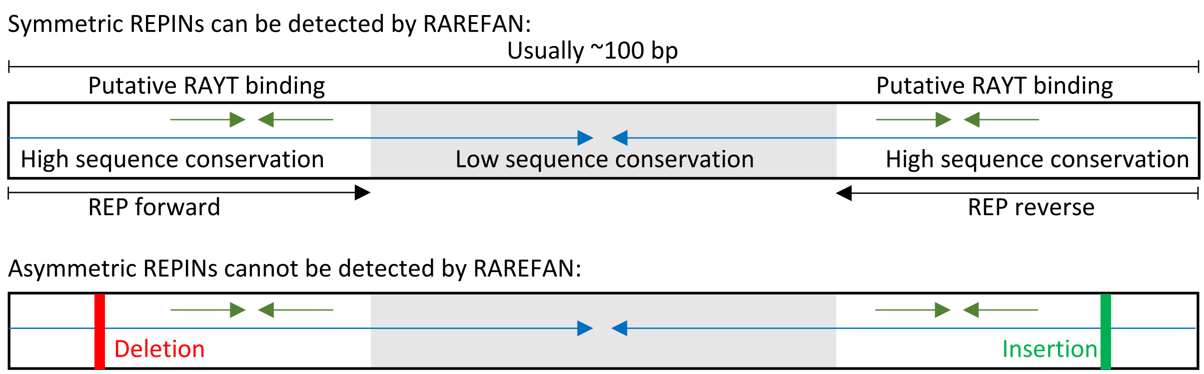


Figure 5. The structure of symmetric and asymmetric REPINs. A typical REPIN consists of two highly conserved regions at the 5' and 3' end of the REPIN (white), separated by a spacer region of lower sequence conservation (grey). The entire REPIN is a palindrome (blue arrows), which means it can form hairpin structures in single stranded DNA or RNA. Each 5' and 3' region contains a nested imperfect palindrome, which is referred to as REP (repetitive extragenic palindromic)

sequence and has first been described in *E. coli* (Higgins *et al.* 1982). REPINs can be either symmetric or asymmetric. Asymmetric REPINs have a deletion and a corresponding insertion in the highly conserved 5' or 3' end, which leads to "bubbles" in the hairpin structure. REPINs in *E. coli* are asymmetric, which makes analyses with RAREFAN more challenging. Figure adapted from (Bertels, Rainey 2022).

REPIN groups may be lost when the seed distance is too large

The seed distance parameter determines whether two highly abundant sequences are sorted into the same or different REPIN groups (**Figure 1B**). If two REPINs from two different groups occur next to each other, at a distance of less than the seed distance parameter, then the two seeds are erroneously sorted into the same group. If two different REPIN groups are sorted into the same group then one of the groups will be ignored by RAREFAN, because only the most abundant seed will be used to identify REPINs.

A manual analysis (e.g., multiple sequence alignment) of sequences in the `groupSeedSequences` folder of the RAREFAN output can identify erroneously merged REPIN groups. In *S. maltophilia*, groups are separated well when the parameter is set to 15 and Sm53 is used as a reference. When the parameter is set to 30 instead, one of the REPIN groups will be missed by RAREFAN.

A small seed distance parameter will separate seed sequences belonging to the same REPIN group into different groups. Hence, RAREFAN will analyse the same REPIN group multiple times. While this will lead to increased RAREFAN runtimes, these errors, are easy to spot, because (1) the same RAYT gene will be associated to multiple REPIN groups, (2) the central palindrome between the group is identical and (3) the master sequence between the groups will be very similar.

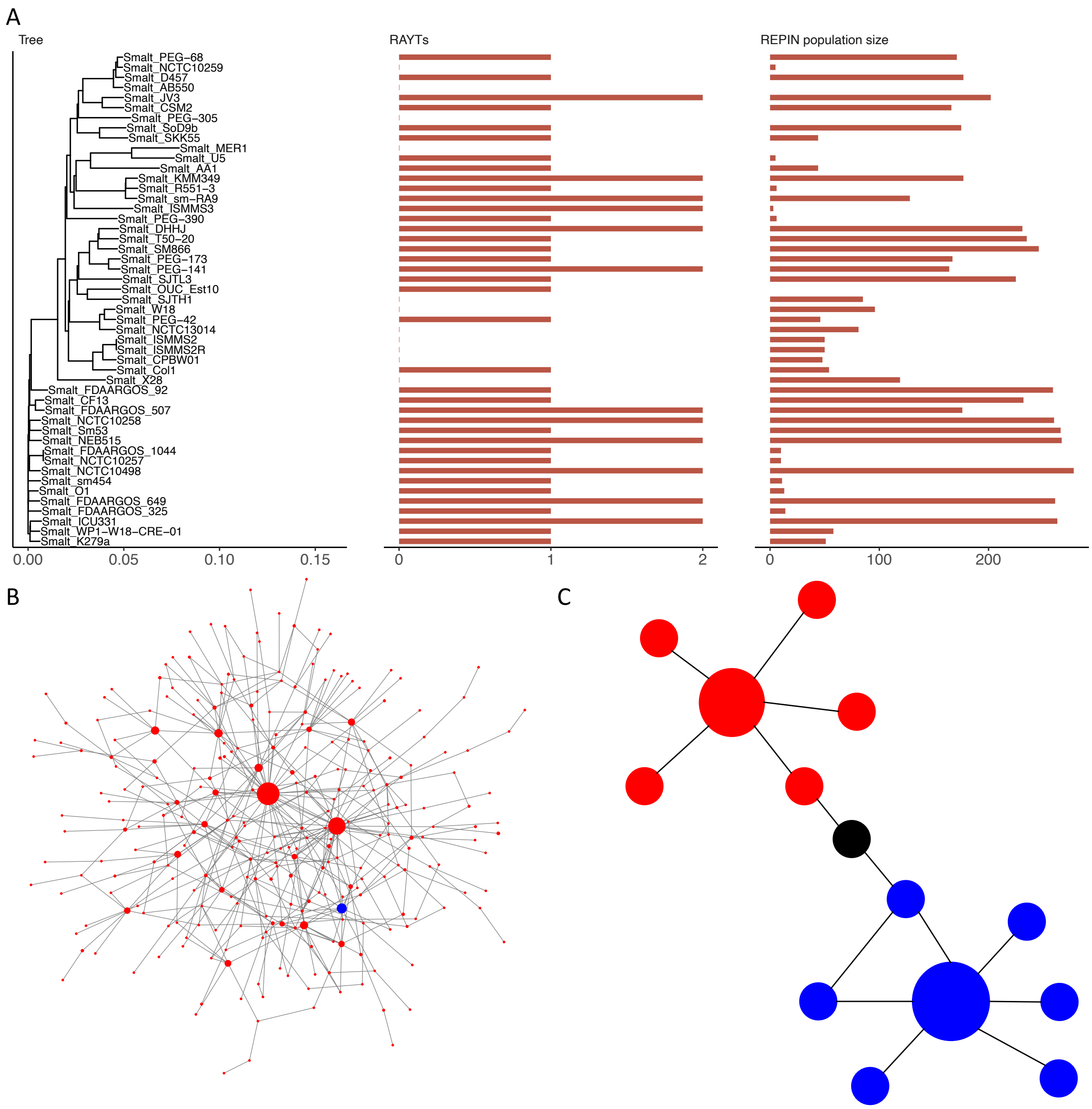


Figure 6. Closely related REPIN populations may be merged by RAREFAN. (A) REPIN group 2 identified in a *S. maltophilia* Sm53 RAREFAN run. The RAREFAN result suggests that REPIN group 2 is sometimes associated with two RAYTs. (**B**) A closer inspection of the data shows that Group 2 is a combination of two different REPIN groups, the real Group 2 and Group 0. The network shown, visualizes all REP sequences identified as Group 2. Nodes in the network represent 21 bp long REP sequences. Two nodes are connected if the sequences they represent differ by exactly one nucleotide. The node size indicates the abundance of the sequence in the genome. The blue node represents the most common Group 2 sequence, occurring 65 times in the genome. The largest red node occurs 407 times in the genome and resembles a Group 0 REP sequence. (**C**) Illustration of how small changes to a single sequence can connect two sequence cluster. The most common 21 bp long sequence in Group 0 differs in only four positions from the most

common 21 bp long sequence in Group 2. There is a set of sequences that connects these two groups that only differ in exactly one position each (nodes connected by an edge), which passes through the black node. If there is such an unbroken path between REP sequences, then REPIN groups will be merged.

Closely related REPIN groups may be merged into a single group by RAREFAN

Incorrect merging of REPIN groups can occur when two REPIN groups are closely related. We identified merged REPIN groups in *S. maltophilia* because RAREFAN linked some REPIN groups with two RAYT genes in the same genome (**Figure 6A**). While REPIN groups linked to two RAYTs has been observed before in *Neisseria meningitidis* (Bertels, Rainey 2022), it is particularly unusual in *S. maltophilia* due to some key differences between REPIN-RAYT in the two bacterial species. First, RAYTs in *N. meningitidis* belong to Group 2 and RAYTs in *S. maltophilia* belong to Group 3 (Bertels, Gallie, *et al.* 2017), two very divergent RAYT groups. Second, RAYTs that are associated to the same REPIN group in *N. meningitidis* are almost identical, since they are copied by an insertion sequence *in trans* (Bertels, Rainey 2022), something that is not the case for *S. maltophilia*, where the two RAYTs are very distinct from each other (green and red clade in **Figure 2A**, or clade A and C in **Figure 4**).

A closer inspection of all sequences identified in REPIN group 2 shows that it also contains sequences belonging to REPIN group 0 (palindromes linked to clade A and C in **Table 2**). The relationship between the sequences shows that there is a chain of sequences that all differ by at most a single nucleotide between the most abundant sequence in group 2 to the most abundant sequence in group 0 (**Figure 6B and C**). Because sequence groups are built by identifying all related sequences in the genome recursively, closely related groups (the REPIN group 0 seed only differs in four nucleotides from the REPIN group 2 seed sequence) can be merged into a single REPIN group. REPIN population size and RAYT number are the sum of REPIN group 0 and 2. There are various possibilities to resolve this issue: (1) subtract sequences from group 0 (which does not contain group 2) from REPIN group 2; (2) use a different sequence seed from the group 2 seed collection in the seed sequence file (`groupSeedSequences/Group_Smalt_Sm53_2.out`); or (3) sometimes it may be possible to rerun RAREFAN with a different reference strain where the issue does not occur.

Discussion

RAREFAN allows users to quickly detect REPIN populations and RAYT transposases inside bacterial genomes. It also links the RAYT transposase genes to the REPIN population it duplicates. These data help the user to study REPIN-RAYT dynamics in their strains of interest without a dedicated bioinformatician, and hence will render REPIN-RAYT systems widely accessible.

One limitation of RAREFAN is that REPINs can only be identified in genomes when they are symmetric (**Figure 5**). Symmetric REPINs have seed sequences that can morph into each other by a series of single substitutions (intermediate sequences need to be present in the genome). A REPIN consists of a 5' and a 3' REP sequence. If one of these REP sequences contains an insertion or deletion, which the other REP sequence does not contain then RAREFAN will not recognize the second repeat of the seed sequence. In this case, RAREFAN will not be able to identify REPINs but can still be used to analyze REP singlet populations. To date, the only known asymmetric REPIN population are *E. coli* REPINs. However, it is likely that asymmetric REPINs also exist in other microbial species.

RAREFAN sometimes cannot correctly divide REPINs into REPIN groups. Either because REPINs from different groups occur in close proximity in the genome, an issue that can easily be solved by adjusting a RAREFAN parameter, or because two REPIN groups are very closely related (**Figure 6**). Unfortunately, RAREFAN is not able to automatically detect and resolve the assignment of closely related REPINs into groups yet. Hence it is advisable to manually check associations between REPIN groups and RAYT genes by analyzing the composition of REPIN groups.

In the future we aim to make RAREFAN even more versatile and easier to use by, for example, automatically integrating data from public databases such as Genbank, and integrating RAREFAN into workflows such as Galaxy (Afgan *et al.* 2018).

RAREFAN makes the study of REPIN-RAYT systems more accessible to any biologist or bioinformatician interested in studying intragenomic sequence populations. Our tool will help understand the purpose and evolution of REPIN-RAYT systems in bacterial genomes.

Acknowledgements

We would like to thank Prajwal Bharadwaj for assisting us with the sequence analysis.

References

Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, Chilton J, Clements D, Coraor N, Grüning BA, Guerler A, Hillman-Jackson J, Hiltemann S, Jalili V, Rasche H, Soranzo N, Goecks J, Taylor J, Nekrutenko A, Blankenberg D (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. *Nucleic Acids Res.*, **46**, W537-W544-W537–W544. https://doi.org/10.1093/nar/gky379

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410. https://doi.org/10.1006/jmbi.1990.9999

Arnold K, Gosling J, Holmes D (2005) *The Java programming language*. Addison Wesley Professional.

Bertels F, Gallie J, Rainey PB (2017) Identification and Characterization of Domesticated Bacterial Transposases. *Genome Biology and Evolution*, **9**, 2110–2121. https://doi.org/10.1093/gbe/evx146

Bertels F, Gokhale CS, Traulsen A (2017) Discovering Complete Quasispecies in Bacterial Genomes. *Genetics*, **206**, 2149–2157. https://doi.org/10.1534/genetics.117.201160

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

Bertels F, Rainey PB (2022) Ancient Darwinian replicators nested within eubacterial genomes. , 2021.07.10.451892. https://doi.org/10.1101/2021.07.10.451892

Bichsel M, Barbour AD, Wagner A (2010) The early phase of a bacterial insertion sequence infection. *Theoretical Population Biology*. https://doi.org/10.1016/j.tpb.2010.08.003

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421–9. https://doi.org/10.1186/1471-2105-10-421

van Dijk B, Bertels F, Stolk L, Takeuchi N, Rainey PB (2022) Transposable elements promote the evolution of genome streamlining. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **377**, 20200477. https://doi.org/10.1098/rstb.2020.0477

Grinberg M (2018) *Flask web development: developing web applications with python*. O'Reilly Media, Inc.

Haubold B, Klötzl F, Pfaffelhuber P (2015) andi: fast and accurate estimation of evolutionary distances between closely related genomes. *Bioinformatics*, **31**, 1169–1175. https://doi.org/10.1093/bioinformatics/btu815

Higgins CF, Ames GF, Barnes WM, Clement JM, Hofnung M (1982) A novel intercistronic regulatory element of prokaryotic operons. *Nature*, **298**, 760–762. https://doi.org/10.1038/298760a0

Initiative TOS (2021) The MIT License.

Kearse M, Moir R, Wilson A, Stones-Havas S (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. ….

Kleinmann SG, Rudolph S, Vila S, Rodin J, Peña JF-S (2021) *The Debian GNU/Linux Operating System Manual*.

Lawrence JG, Ochman H, Hartl DL (1992) The evolution of insertion sequences within enteric bacteria. *Genetics*, **131**, 9–20. https://doi.org/10.1093/genetics/131.1.9

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

Park HJ, Gokhale CS, Bertels F (2021) How sequence populations persist inside bacterial genomes. *Genetics*, **217**. https://doi.org/10.1093/genetics/iyab027

R Core Team (2016) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rankin DJ, Bichsel M, Wagner A (2010) Mobile DNA can drive lineage extinction in prokaryotic populations. *Journal of Evolutionary Biology*. https://doi.org/10.1111/j.1420-9101.2010.02106.x

RStudio, Inc (2013) *Easy web applications in R.*

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B (2000) Artemis: sequence visualization and annotation. *Bioinformatics*, **16**, 944–945. https://doi.org/10.1093/bioinformatics/16.10.944

Sawyer SA, Dykhuizen DE, DuBose RF, Green L, Mutangadura-Mhlanga T, Wolczyk DF, Hartl DL (1987) Distribution and Abundance of Insertion Sequences Among Natural Isolates of Escherichia coli. *Genetics*, **115**, 51–63. https://doi.org/10.1093/genetics/115.1.51

Ton-Hoang B, Siguier P, Quentin Y, Onillon S, Marty B, Fichant G, Chandler M (2012) Structuring the bacterial genome: Y1-transposases associated with REP-BIME sequences. *Nucleic Acids Research*, **40**, 3596–3609. https://doi.org/10.1093/nar/gkr1198

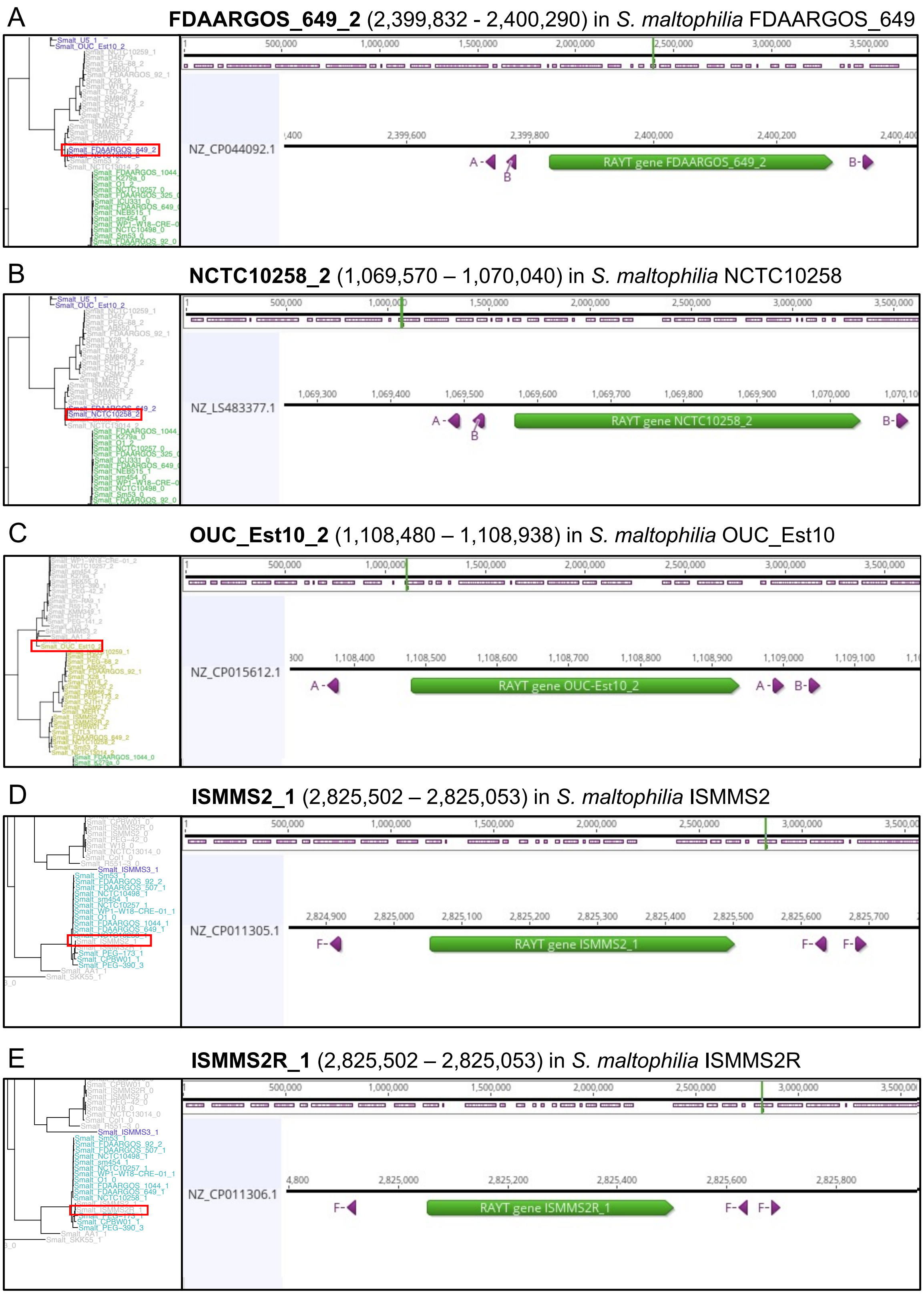
Van Dongen S (2000) A cluster algorithm for graphs. *Report-Information systems*, 1–40.

Van Rossum G, Drake Jr FL (1995) *Python reference manual*. Centrum voor Wiskunde en Informatica Amsterdam.

Wu Y, Aandahl RZ, Tanaka MM (2015) Dynamics of bacterial insertion sequences: can transposition bursts help the elements persist? *BMC Evolutionary Biology*, **15**, 288–12. https://doi.org/10.1186/s12862-015-0560-5

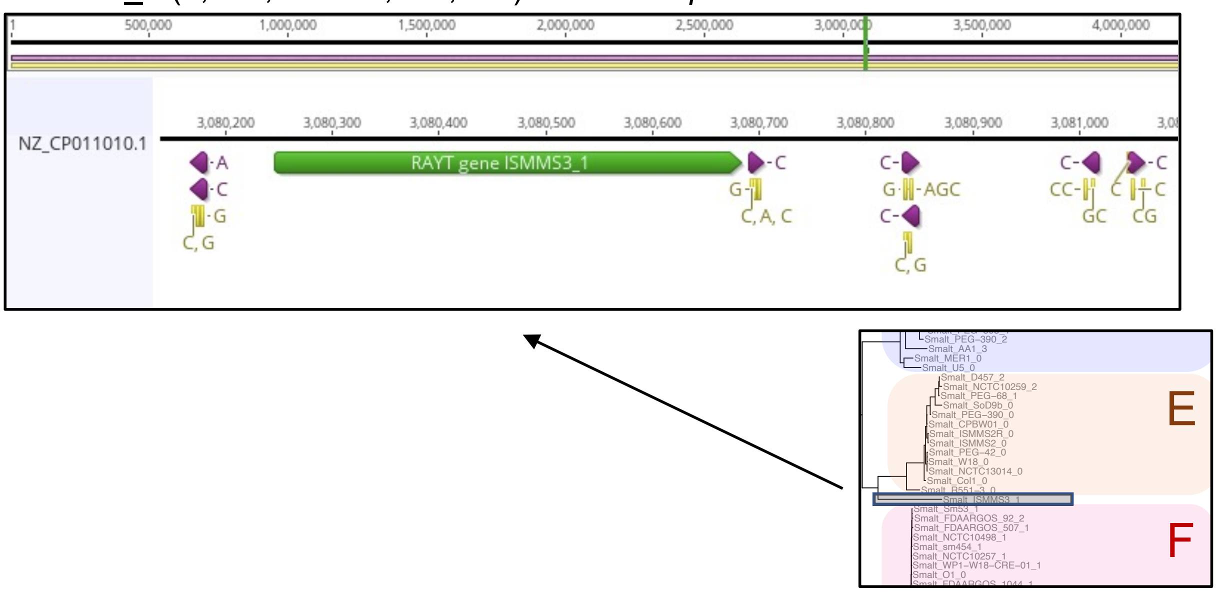
Yu G, Lam TT-Y, Zhu H, Guan Y (2018) Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree. (FU Battistuzzi, Ed,). *Molecular biology and evolution*, **35**, 3041–3043. https://doi.org/10.1093/molbev/msy194

Supplementary Figures



Supplementary Figure 1: Sequence analysis shows REPIN groups are indeed associated with monophyletic RAYTs. Non-monophyletic or missing associations to REPIN populations identified by RAREFAN were investigated in the corresponding genomes using Geneious (Kearse *et al.* 2012). Red boxes mark the position of the atypical RAYT that is being analyzed in detail. Mapping of REPIN palindromes A-I (with zero mismatches) shows FDAARGOS_649_2 (A), NCTC10258_2 (B), and OUC_Est_2 (C) are linked to the wrong REPIN group because REP singlets that are ordinarily linked to a RAYT sister clade are found in close proximity to the RAYT. These wrong associations between REPIN and RAYT usually occur when the correct REPIN population is absent from the reference genome. ISMMS2R_1 (D) and ISMMS2_1 (E) were not linked to REPIN population by RAREFAN because the corresponding seed sequences were located at a distance of more than 130 bp from the RAYT gene. Nucleotide sequences and positions were extracted from output files generated by RAREFAN. Complete genome sequences are available in NCBI Nucleotide Database using Accessions: (A) NZ_CP044092.1, (B) NZ_LS483377.1, (C) NZ_CP015612.1, (D) NZ_CP011306.1, (E) NZ_CP011305.1.

ISMMS3_1 (3,080,683 – 3,080,246) in *S. maltophilia* ISMMS3



Supplementary Figure 2: RAYT gene ISMMS3_1 cannot be linked to a REPIN population. The sequence of the RAYT gene ISMMS3_1 and its flanking sequences was analyzed in Geneious (Kearse *et al.* 2012). The inset shows the location of ISMMS3_1 in the RAYT phylogeny (grey box). When mapping all of the identified palindromes to the RAYT region and allowing up to four mismatches (yellow annotations), various mutants of palindrome C were found in close proximity of the RAYT gene. However, we could not identify a corresponding REPIN population, which may indicate that the population has not yet expanded in the genome.