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 maltophilia*

Abstract

Compared to eukaryotes, repetitive sequences are rare in bacterial genomes and usually do not persist for long. 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 mostly 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 provide solutions on how to find 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 have to 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.

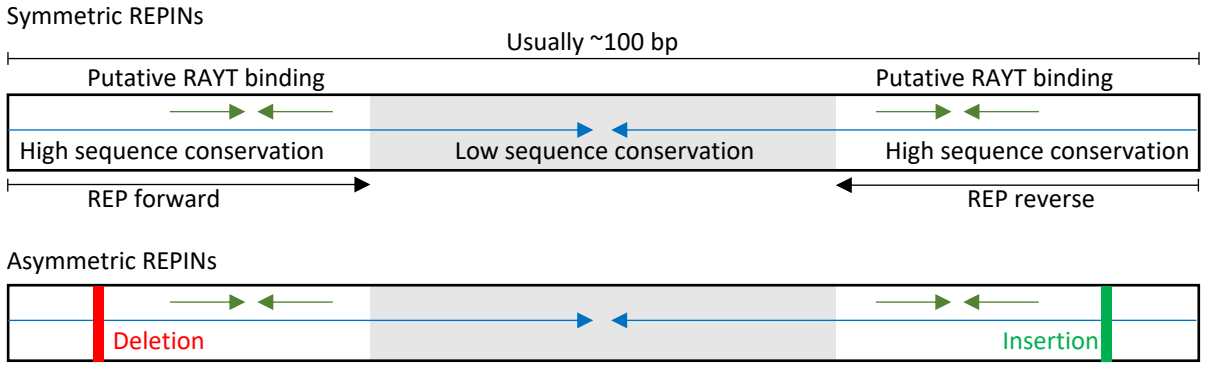


Figure 1. The structure of symmetric and asymmetric REPINs. A typical REPIN consists of two highly conserved regions at the 5' and 3' end of the REPIN (white), separated by a spacer region of lower sequence conservation (grey). The entire REPIN is a palindrome (blue arrows), which means it can form hairpin 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 *Escherichia 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).

REPINs are short (~100 bp) nested palindromic sequences (**Figure 1**) that consist of two inverted REP (repetitive extragenic palindromic (Higgins *et al.* 1982)) sequences that can be present hundreds of times per genome (Bertels, Rainey 2011a). Most REPINs are symmetric where the 5' REP sequences is identical to the 3' REP sequences, with the occasional substitution (Bertels,

Rainey 2011a; b). However, there are also asymmetric REPINs where the 5' REP sequence differs from the 3' REP sequence by a point deletion or insertion (Bertels, Rainey 2011a, 2022), which makes the analysis and detection significantly more difficult (*e.g.*, *Escherichia coli* REPINs). Isolated REP sequences, REP singlets can also be found in the genome. These sequences are decaying remnants of REPINs that are not mobile anymore (Bertels, Rainey 2011a). REPINs are non-autonomous mobile genetic elements, which means they require a RAYT (**R**EP **A**ssociated t**Y**rosine **T**ransposase) transposase gene (also referred to as $tnpA_{REP}$) to replicate inside the genome (Nunvar *et al.* 2010; Bertels, Rainey 2011a; Ton-Hoang *et al.* 2012).

Within a genome, each REPIN population is usually only associated to a single RAYT gene. Hence, RAYT genes occur only in single copies per genome and do not copy themselves, unlike for example insertion sequences where often multiple identical sequences are present inside the genome. Unlike insertion sequences RAYT genes are also only inherited vertically, meaning they are host-beneficial transposases that are coopted by the host (Bertels, Gallie, *et al.* 2017; Bertels, Rainey 2022). The fact that REPINs and their corresponding RAYT genes are confined to a single bacterial lineage makes them very special, in comparison to all other parasitic mobile genetic elements in bacterial genomes (Bertels, Rainey 2022).

Of a total of five different RAYT families, there are only two RAYT families that are associated with REPINs: Group 2 and Group 3 RAYTs (Bertels, Gallie, *et al.* 2017). Group 2 RAYTs are present in most Enterobacteria and usually occur only once per genome associated with a single REPIN population. In contrast, Group 3 RAYTs are found in most *Pseudomonas* species and are usually present in multiple divergent copies per genome, each copy associated with a specific REPIN population (Bertels, Gallie, *et al.* 2017).

REPINs and their corresponding RAYT genes occur exclusively in bacterial genomes and are absent in eukaryotic or archaeal genomes (Bertels, Gallie, *et al.* 2017; Bertels, Rainey 2022). Within bacterial genomes REPINs and RAYTs have been evolving in single bacterial lineages for

77 millions maybe even for a billion years (Bertels, Gallie, *et al.* 2017). The long term persistence of
78 REPINs in single bacterial lineages can also be observed when analyzing REPIN populations
79 (Bertels, Gokhale, *et al.* 2017; Bertels, Rainey 2022).

80

81 Parasitic insertion sequences usually occur in identical copies in bacterial genomes reflecting the
82 fact that insertion sequences persist only briefly in the genome before they are eradicated from
83 the genome or kill their host (Park *et al.* 2021). REPINs in contrast are only conserved at the ends
84 of the sequence (presumably due to selection for function), the rest of the sequence is highly
85 variable and only the hairpin structure is conserved (Bertels, Rainey 2011a). The sequence
86 variability of REPINs within the same genome reflects their long-term persistence in single
87 bacterial lineages (Bertels, Rainey 2022). REPINs cannot simply reinfect another bacterial lineage
88 since they rely for mobility on their corresponding RAYT, which itself is immobile.

89

90 RAYTs and REPINs are distinct from typical parasitic insertion sequences, yet we know very little
91 about their evolution or biology. Currently, it is completely unclear what kind of beneficial
92 function maintains REPINs and RAYTs for millions of years in the genome. The reason for our lack
93 of knowledge is not because REPINs and RAYTs are rare. They are ubiquitously found in many
94 important and well-studied model bacteria such as Enterobacteria, Pseudomonads, Neisseriads,
95 Xanthomonads. Microbial molecular biologists presumably encounter REPINs quite frequently.
96 However, connecting the presence or absence of REPINs/RAYTs with phenotypes is difficult if we
97 do not know when it is a REPIN that is present close to a gene of interest or a different type or
98 repeat sequence. Even if the scientist knows about the presence of a REPIN it is probably also
99 important to know whether a corresponding RAYT is present, since the function of REPINs largely
100 depends on the function of the presence of a corresponding RAYT gene (Bertels, Rainey 2022).

101

102 Yet, the identification of REPIN populations and their corresponding RAYTs can be rather
103 cumbersome if done from scratch. This is particularly true if the microbial molecular biologist is
104 not aware of all the ins and outs of REPIN and RAYT biology. Identifying REPINs starts with an

105 analysis of short repetitive sequences in the genome. If there are excessively abundant short
106 sequences present in the genome, the distribution of these sequences is analyzed next. If they
107 are exclusively identical tandem repeats without sequence variation and present in only one or
108 two loci in the genome then these sequences are probably part of a CRISPR array and not REPINs.

109

110 Here, we present RAREFAN (RAYT/REPIN Finder and Analyzer), a webservice that automates the
111 identification of REPINs and their corresponding RAYTs. RAREFAN is publicly accessible at
112 http://rarefan.evolbio.mpg.de and identifies REPIN populations and RAYTs inside a set of
113 bacterial genomes. RAREFAN also generates graphs to visualize the population dynamics of
114 REPINs, and assigns RAYT genes to their corresponding REPIN groups. Here we will demonstrate
115 RAREFAN's functionality by analyzing REPIN-RAYT systems in the bacterial species
116 *Stenotrophomonas maltophilia*.

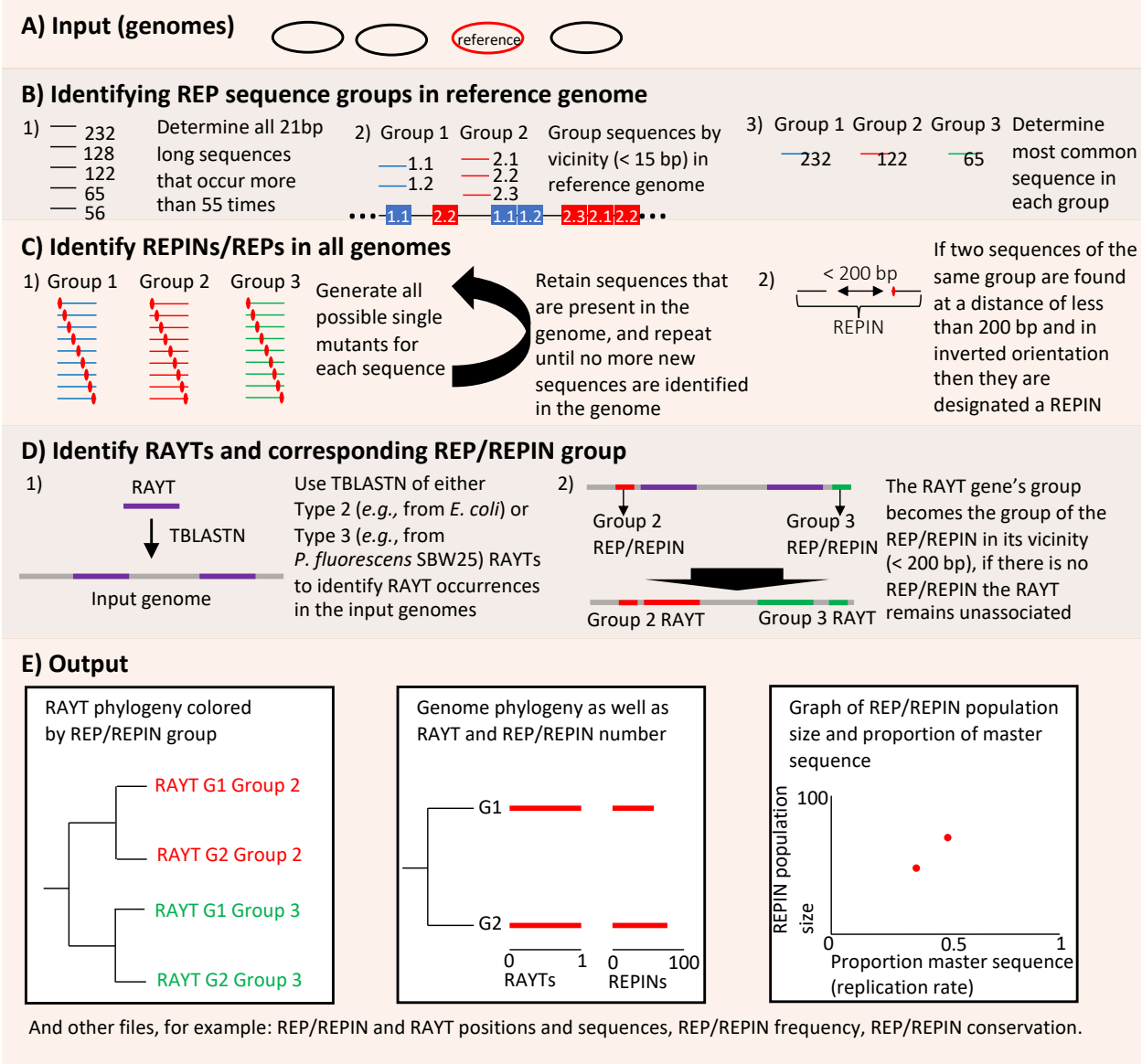


Figure 2. RAREFAN workflow. (**a**) By default RAREFAN requires the user to supply input sequences containing RAYTs and REPINs. These are fully sequenced and complete genomes. (**b**) RAREFAN then identifies seed sequence groups (potential REP sequences) in the reference genome by first isolating all 21 bp (adjustable parameter) long sequences that occur more than 55 times (adjustable parameter) in the reference genome. It is likely that a large number of these sequences belong to the same REPIN sequence type since the conserved part of REPINs is longer than 20bp. Hence, we grouped all sequences together that occur within 15 bp (adjustable parameter) of each other anywhere in the genome. For example, 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 15 bp of 'sequence 2' in order to be sorted into the same sequence group. All further analyses are performed only with the most common sequence in each sequence group. This sequence will be called 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 (e-Value < 1e-30, adjustable parameter) of either *E. coli* (Group 2 RAYT) or from *Pseudomonas fluorescens* SBW25 (Group 3 RAYT) across all submitted genomes. If a RAYT homolog is in the vicinity (default 200 bp, adjustable parameter) of a previously identified REPIN or REP singlet, then this RAYT is designated as associated with this REPIN group. (**e**) The first graph contains a RAYT phylogeny computed from a nucleotide alignment of all identified RAYT genes. The alignment is calculated with MUSCLE (Edgar 2004) and a phylogeny with PHYML (Guindon *et al.* 2010). The RAYT phylogeny indicates what RAYTs are associated with what REPIN populations (largest sequence cluster calculated with MCL) via colour coding. In a second graph the abundance of each REPIN population and RAYT copy number are displayed on a genome phylogeny. If no genome phylogeny is supplied RAREFAN calculates a whole genome phylogeny of the submitted genomes using andi (Haubold *et al.* 2015). In the last graph REPIN population sizes are plotted in relation to the proportion of master sequences. Master sequences are the most abundant REPIN in each population. The REPIN population is the largest sequence cluster that is formed by REPIN sequences (REP sequences are excluded). The largest sequence cluster is identified by applying MCL with an inflation parameter of 1.2 to a sequence matrix where only sequences are connected that differ in exactly one position (Van Dongen 2000). RAREFAN also 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 on 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-scicomp/rarefan.

The public RAREFAN instance at http://rarefan.evolbio.mpg.de runs on a virtual cloud server with 4 single-threaded CPUs and 16GB of shared memory provided and maintained by the Gesellschaft für Wissenschaftliche Datenverarbeitung Göttingen (GWDG) and running the Debian GNU/Linux Operating System (Kleinmann *et al*. 2021).

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 in order to determine REPIN populations.

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 2A**). 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.
- Query RAYT: The RAYT gene that is used to identify homologous RAYTs in the query genomes.
- Tree file: A phylogenetic tree of the reference genomes that can be provided by the user, otherwise the tree will be calculated using andi (Haubold *et al.* 2015).
- Minimum seed sequence frequency: Lower limit on seed sequence frequency in the reference genome to be considered as a REP candidate. Default is 55.
- Seed sequence length: The seed sequence length (in base pairs) is used to identify REPIN candidates from the input genomes. Default is 21 bp.
- Distance group seeds: The maximum distance between a single occurrence of short repetitive sequences to still be sorted into the same sequence group.

- Association distance REPIN-RAYT: The maximum distance at which a REP sequence can be located from a RAYT gene to be linked to that RAYT gene.
- e-value cut-off: Alignment e-value cut-off for identifying RAYT homologs with TBLASTN. Default is 1e-30.
- Analyse REPINs: Ticked REPINs will be analysed (two inverted REP sequences found at a distance of less than 130 bp), if not ticked only short repetitive 21 bp long sequence will be analysed.
- 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. Runs are automatically deleted from the server after six months.

Identification of REPs and REPINs

The algorithm to determine REP sequence groups has been described in previous papers and is slightly improved (Bertels, Rainey 2011a, 2022; Bertels, Gokhale, *et al.* 2017). In our current implementation REPs/REPIN populations are now automatically linked to RAYT genes.

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 2B**). 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 2C**). 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.

Among all identified REP and REPIN sequences REPIN populations can be isolated. REPIN populations are determined by applying MCL using an inflation parameter of 1.2 (Van Dongen 2000) to a network of REP/REPIN sequences where all sequences that differ by exactly one nucleotide are connected. The clustering results are stored in a file ending in `.mcl`. The sequences of the largest REPIN population (excluding REP singlets) are isolated in a file ending in `largestCluster.nodes`. The largest REPIN populations are shown in the REPIN population plot and the master sequence correlation plot (**Figure 4**).

Identification of RAYTs

RAYTs are identified using TBLASTN (Camacho *et al.* 2009) with either a protein sequence provided by the user, a Group 2 RAYT from *E. 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 (default 200 bp) of a particular REPIN can be used to establish the association between the RAYT gene and a REPIN group (**Figure 2D**). All positions of all REPINs and REP sequences of a REPIN group are checked whether they occur within 200 bp (by default) of a RAYT gene. If so then the RAYT gene is linked to the REPIN group in the file `repin_rayt_association.txt`.

Visualizations

For each REPIN-RAYT group summary plots are generated. These include plots showing the RAYT phylogeny (calculated from a nucleotide alignment using MUSCLE (Edgar 2004) and PHYML (Guindon *et al.* 2010) to generate a phylogeny), REPIN population sizes in relation to the genome phylogeny (provided by the user or if not provided calculated by andi (Haubold *et al.* 2015)) as well as the proportion of master sequences (most common REPIN in a REPIN population) in relation to REPIN population size (**Figure 2E**).

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) and in the file "readme.md" in the output directory.

Sequence analysis and annotation

For verification of RAREFAN results, REPIN-RAYT-systems were analysed in their corresponding genomes using Geneious Prime version 2022.2.2 (Kearse *et al.* 2012). 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. 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

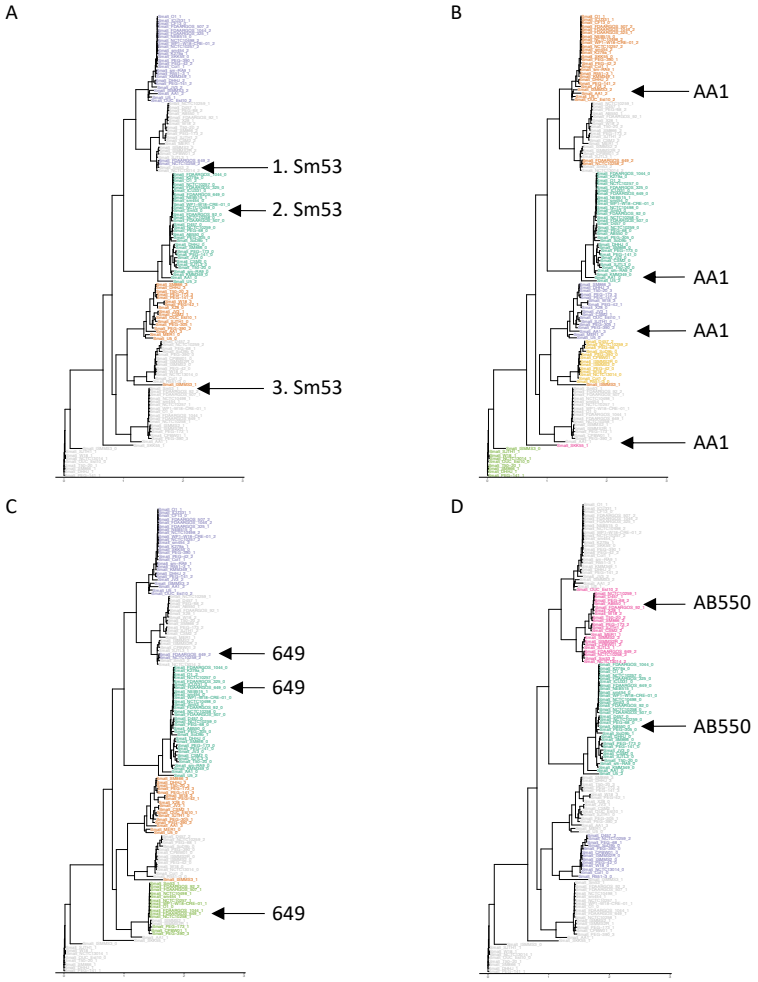
Associated data can accessed by entering the run ID at http://rarefan.evolbio.mpg.de/results.

Results

RAREFAN can identify REPINs and their corresponding RAYTs in a set of fully sequenced bacterial genomes. The RAREFAN algorithm has been used in previous analyses to identify and characterize REPINs and RAYTs in Pseudomonads (Bertels, Rainey 2011a, 2022), Neisseriads (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*.

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

241



242

Figure 3. 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 a REPIN population of a query genome is not present in the reference genome, then the REPIN population cannot be identified in the query genome and the corresponding RAYT gene cannot be linked and 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.

243

244 *Nine different REPIN-RAYT systems in* S. maltophilia

245 REPIN-RAYT systems in *S. maltophilia* are surprisingly diverse compared to other species. For
246 example, *Pseudomonas chlororaphis* contains three separate REPIN populations that are present
247 in all *P. chlororaphis* strains, each associated with its cognate RAYT gene (Bertels, Rainey 2022).
248 *S. maltophilia,* in contrast, contains only one REPIN-RAYT system that is present across almost
249 the entire species (green clade in **Figure 3**), and at least eight REPIN-RAYT systems that are
250 present in subsets of strains (nine clades in **Figure 5**).

251 The patchy presence-absence pattern of REPIN-RAYT systems in *S. maltophilia,* makes the dataset
252 quite challenging to analyse. If a REPIN population is not present in the reference strain then
253 RAREFAN will not be able to detect it in any other strain. Yet, it is possible to detect RAYT genes
254 in all strains of a species independent of the reference strain selection. RAYT genes that are not
255 associated to a REPIN population are displayed in grey (**Figure 3A**). While these RAYT genes are
256 not associated to REPIN populations detected in the reference strain, they might still be
257 associated with a yet unidentified REPIN type present in the genome the unassociated RAYT gene
258 is located in.

259 In order to identify all REPIN populations across a species, we suggest to perform multiple
260 RAREFAN runs with different reference strains. The RAREFAN web interface supports re-
261 launching a given job with modified parameters. To identify as many different REPIN-RAYT
262 systems as possible in each subsequent run the reference should be set to a genome that
263 contains RAYTs that were not associated with a REPIN population previously (*i.e.*, genomes
264 containing grey RAYTs in **Figure 3**). However, this strategy may also fail when the REPIN
265 population size falls below the RAREFAN seed sequence frequency threshold.

266 For example, *S. maltophilia* Sm53 contains three RAYTs only one of which is associated with a
267 REPIN population (RAYT genes indicated in **Figure 3A**). However, the remaining two RAYTs are
268 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 3A**) is associated with the pink REPIN population (**Figure 3D**). If *S. maltophilia* 649 is set as reference RAYT number 3 from Sm53 (**Figure 3A**) is associated with the light green REPIN population (**Figure 3C**). RAYTs from the bottom clade are only associated with REPIN populations when *S. maltophilia* AA1 is chosen as reference (**Figure 3B**). While lower thresholds can guarantee that all REPINs will be identified in the genome, the number of sequence groups that are not REPINs quickly explodes. Especially for genomes that contain large numbers of mobile genetic elements or CRISPRs (Bertels, Rainey 2022).

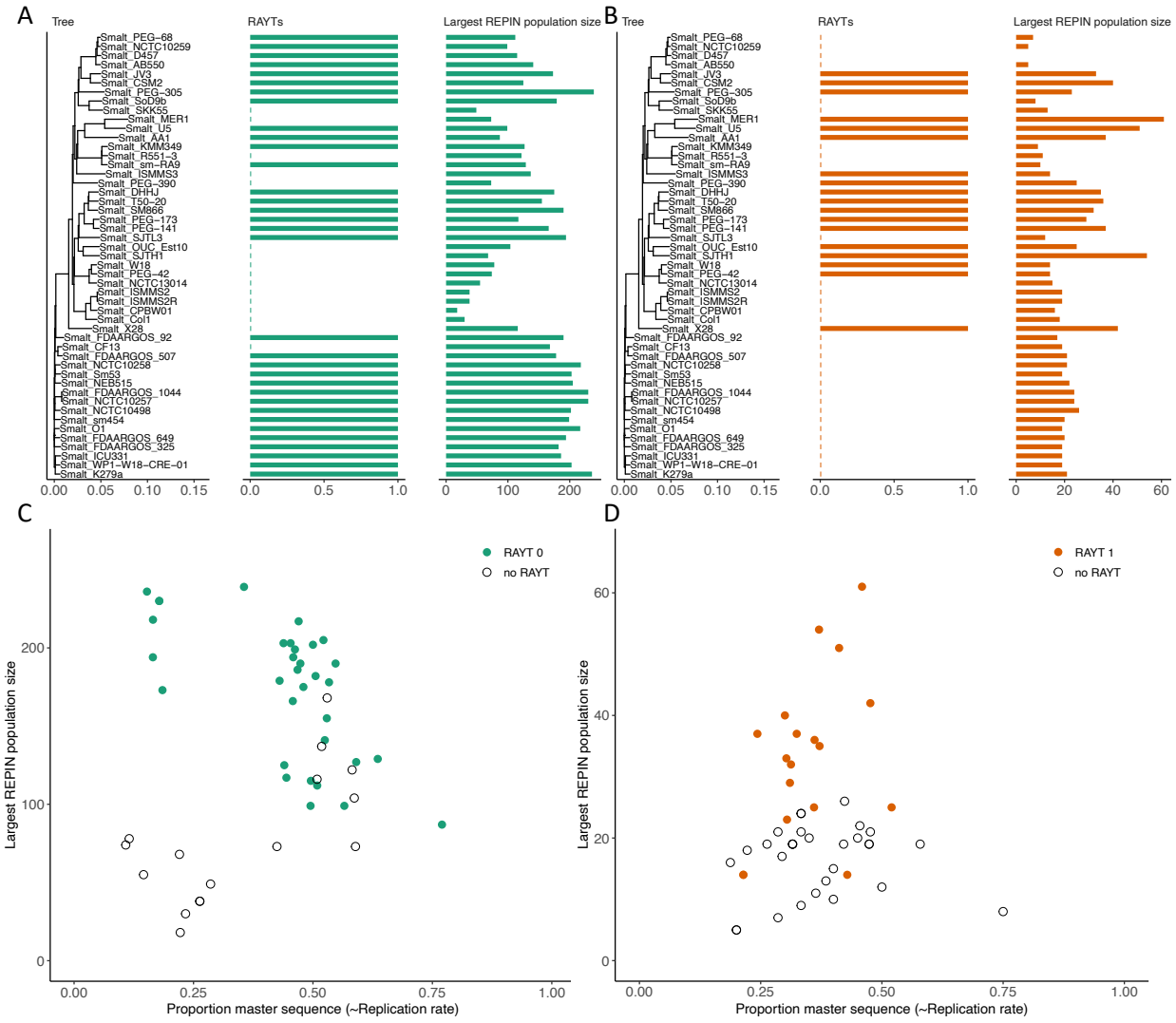


Figure 4. REPIN population sizes and conservation. The plots show two REPIN populations and their associated RAYTs that were identified in *S. maltophilia* using *S. maltophilia* Sm53 as reference. (**A**) The phylogenetic tree on the left side is a whole genome phylogeny generated by andi (Haubold *et al.* 2015). Shown on the right are REPIN population sizes (which is the largest REPIN cluster calculated by MCL) 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 orange 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 orange. (**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. In an equilibrium 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 potential replication rate

The RAREFAN webserver visualizes REPIN population size and RAYT numbers in barplots. Barplots are 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 3A**). The largest REPIN population (calculated by MCL from all REPINs of that type) has a corresponding RAYT gene in almost all strains (first barplot in **Figure 4A**) and most REPIN populations contain more than 100 REPINs (second barplot in **Figure 4A**). The second largest REPIN population in Sm53 (orange population in **Figure 4B**) 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 4C** 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 if the population is in mutation selection balance (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

297 identical master sequences because the time between replication events is not sufficient to
298 accumulate mutations and because quick extinction of the element usually prevents the
299 accumulation of mutations after replication (Park *et al.* 2021; Bertels, Rainey 2022). REPIN
300 populations in contrast replicate slowly and persist for long periods of time, which means that a
301 high proportion of master sequences suggests a high REPIN replication rate.

302 In *S. maltophilia* the proportion of master sequences in the population does not seem to correlate
303 well with REPIN population size, both in the green and the orange population (**Figures 4C and D**).
304 Similar observations have been made in *P. chlororaphis* (Bertels, Rainey 2022), and may suggest
305 that an increase in population size is not caused by an increase in replication rate. Population size
306 is likely to be more strongly affected by other factors such as the loss of the corresponding RAYT
307 gene, which leads to the decay of the REPIN population. One could even speculate that high
308 REPIN replication rates are more likely to lead to the eventual demise of the population due to
309 the negative fitness effect on the host (Bertels, Rainey 2022).

310 The presence of RAYTs and the size of the corresponding REPIN population do correlate
311 surprisingly well (**Figure 4A** and **B**, p-Value = 0.008 of the linear model of independent contrasts
312 (Felsenstein 1985) of green RAYT and REPIN number, p-Value = 0.003 for orange REPIN
313 populations). Green RAYTs are absent from an entire *S. maltophilia* clade (middle of **Figure 4A**).
314 This clade has also lost a significant amount of green REPINs, and the proportion of the master
315 sequences in these populations is low (**Figure 4C**). Similarly, genomes without orange RAYTs have
316 smaller REPIN populations in the orange population than genomes with the corresponding RAYT
317 (**Figure 4D**). A similar observation has been made previously in *E. coli, P. chlororaphis, N.*
318 *meningitidis* and *N. gonorrhoeae* where the loss of the RAYT gene is followed by a decay of the
319 associated REPIN population (Park *et al.* 2021; Bertels, Rainey 2022).

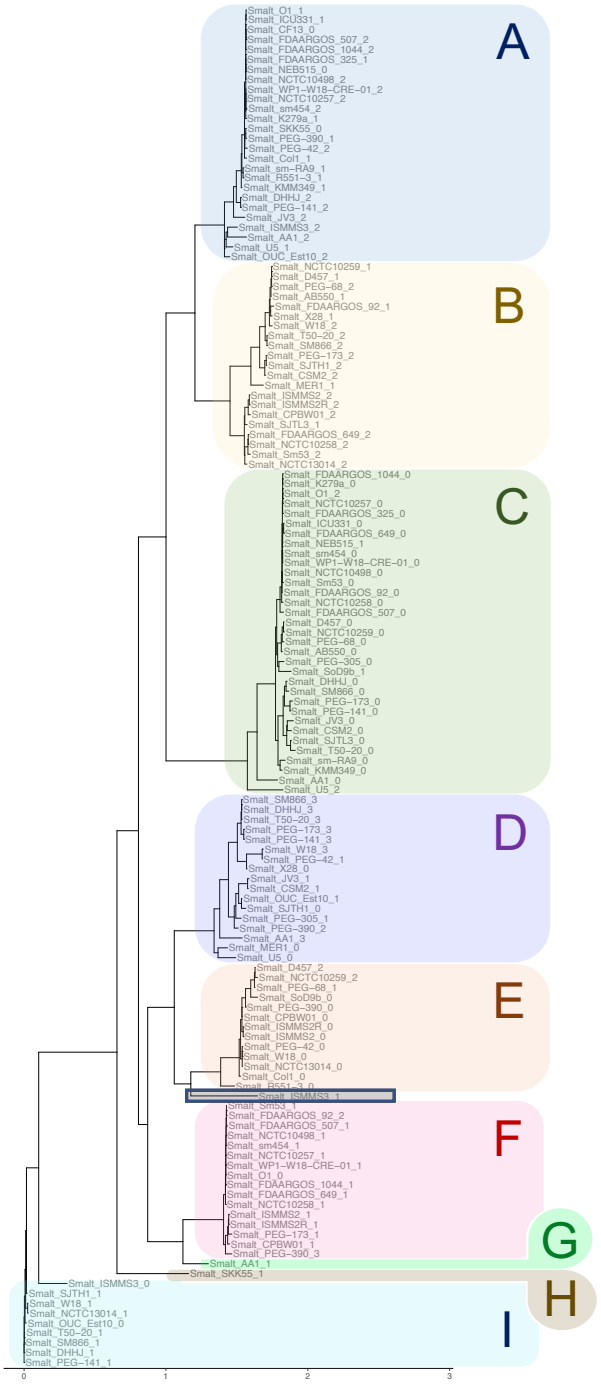


Figure 5. Phylogeny of RAYT genes and their associated REPINs. The tree shows RAYT genes from 49 *S. maltophilia* strains. Colours of clades A-I are assigned according to their association with a REPIN found within 130 bp of the RAYT gene (see **Table 2**). Except for a single RAYT gene ISMMS3_1 (grey 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 5** is associated with a unique imperfect palindrome that is present at the 5' and/or 3' end of the RAYT gene.

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. In some RAREFAN runs associations between RAYTs and REPINs are not monophyletic, as for example the red RAYTs in **Figure 3A**. However, the same clade of RAYTs is uniformly coloured in yellow in **Figure 3D**, suggesting that the entire RAYT clade is associated with the same REPIN group.

An analysis of all REPIN groups that were identified by RAREFAN across four different RAREFAN runs (**Table 1**, one additional analysis was performed with ISMMS3) 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 monophyletic RAYT group (**Figure 5**). Only a single RAYT is not associated with a REPIN population (ISMMS3_1).

RAREFAN could not link a REPIN to the RAYT gene ISMMS3_1 (**Figure 5**, 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, and the REPIN population has not spread to a considerable size yet.

There are two more cases where RAREFAN failed to link RAYT genes with any REPINs (ISSMS2_ and ISSMS2R_1**, Supplementary Figure 1 D** and **E**). Detailed sequence analyses showed that the respective REPINs are located at a distance of more than 130 bp (an adjustable parameter in RAREFAN). These REPINs are ignored by RAREFAN by default. However, this parameter can be adjusted manually and when set to a distance of 200 bp, RAREFAN correctly links these REPINs to the RAYT gene.

In three cases the wrong REPIN population was linked to a RAYT gene. In our dataset this can happen when RAYTs 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. This problem can be alleviated by performing analyses with multiple reference genomes and comparing the results.

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 2B**). 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 in each group 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 distance parameter is set to 15 bp and Sm53 is used as a reference. When the parameter is set to 30 bp 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.

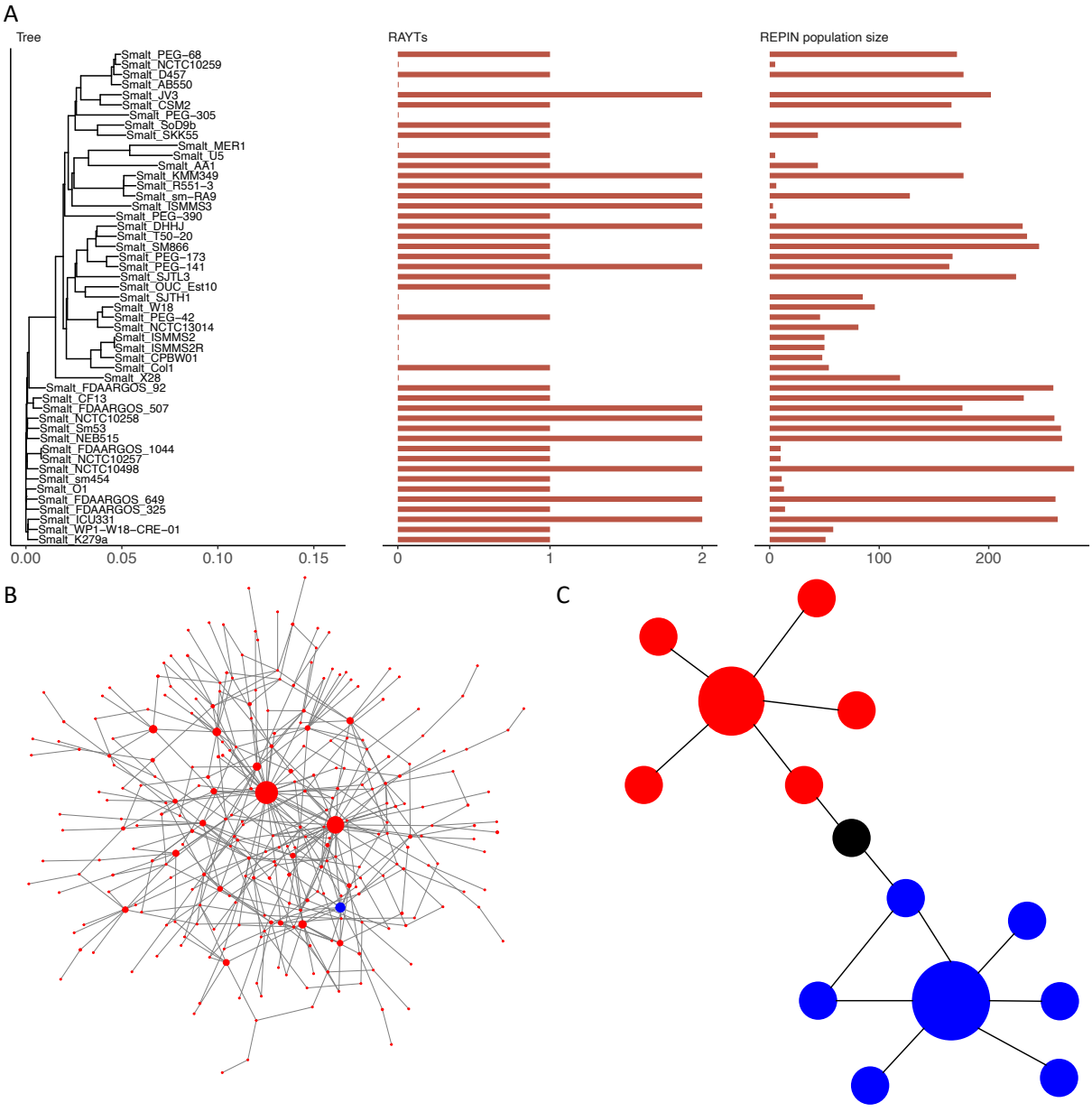


Figure 6. Closely related REPIN populations may be merged by RAREFAN. (A) REPIN group 2 identified in a *S. maltophilia* Sm53 RAREFAN run. The RAREFAN result suggests that REPIN group 2 is sometimes associated with two RAYTs. (**B**) A closer inspection of the data shows that Group 2 is a combination 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.

375

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

377 Incorrect merging of REPIN groups can occur when two REPIN groups are closely related. We
378 identified merged REPIN groups in *S. maltophilia* because RAREFAN linked some REPIN groups
379 with two RAYT genes in the same genome (**Figure 6A**). While REPIN groups linked to two RAYTs
380 has been observed before in *Neisseria meningitidis* (Bertels, Rainey 2022), it is particularly
381 unusual in *S. maltophilia* due to some key differences between REPIN-RAYT in the two bacterial
382 species. First, RAYTs in *N. meningitidis* belong to Group 2 and RAYTs in *S. maltophilia* belong to
383 Group 3 (Bertels, Gallie, *et al.* 2017), two very divergent RAYT groups. Second, RAYTs that are
384 associated to the same REPIN group in *N. meningitidis* are almost identical, since they are copied
385 by an insertion sequence *in trans* (Bertels, Rainey 2022), something that is not the case for *S.*
386 *maltophilia*, where the two RAYTs are very distinct from each other (green and red clade in **Figure**
387 **3A**, or clade A and C in **Figure 5**).

388 A closer inspection of all sequences identified in REPIN group 2 shows that it also contains
389 sequences belonging to REPIN group 0 (palindromes linked to clade A and C in **Table 2**). The
390 relationship between the sequences shows that there is a chain of sequences that all differ by at
391 most a single nucleotide between the most abundant sequence in group 2 to the most abundant
392 sequence in group 0 (**Figure 6B and C**). Hence, the reason group 0 and group 2 are merged is that

they are too closely related to each other and hybrids of the two REPIN groups exist. 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`); (3) sometimes it may be possible to rerun RAREFAN with a different reference strain where the issue does not occur; or (4) increase the length of the seed sequence.

Performance

We measured the elapsed time for a complete RAREFAN run for three different species and for 5, 10, 20, and 40 genomes with randomly selected reference strains and the two query RAYTs (yafM_Ecoli and yafM_SBW25). For a given number N of submitted genomes of average sequence length L (in megabases), a RAREFAN run completes in approximately T = (8-10 seconds) * N * L on our moderate server hardware (4 CPU cores, 16 GB shared RAM) (**Supplementary Figure 3** and **4)**.

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 1**). 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

421 can still be used to analyze REP singlet populations. To date, the only known asymmetric REPIN
422 populations are found in *E. coli.* However, it is likely that asymmetric REPINs also exist in other
423 microbial species.

424 RAREFAN sometimes cannot correctly divide REPINs into REPIN groups. Either because REPINs
425 from different groups occur in close proximity in the genome, an issue that can easily be solved
426 by adjusting a RAREFAN parameter, or because two REPIN groups are very closely related (**Figure**
427 **6**). Unfortunately, RAREFAN is not able to automatically detect and resolve the assignment of
428 closely related REPINs into groups yet. Hence it is advisable to manually check associations
429 between REPIN groups and RAYT genes by analyzing the composition of REPIN groups.

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

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

436 **Acknowledgements**

437 We would like to thank Prajwal Bharadwaj for assisting us with the sequence analysis and Jenna
438 Gallie for valuable feedback on the manuscript.

439 **References**

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

445 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool.
446 *Journal of Molecular Biology*, **215**, 403–410. https://doi.org/10.1006/jmbi.1990.9999
447 Arnold K, Gosling J, Holmes D (2005) *The Java programming language*. Addison Wesley
448 Professional.
449 Bertels F, Gallie J, Rainey PB (2017) Identification and Characterization of Domesticated Bacterial
450 Transposases. *Genome Biology and Evolution*, **9**, 2110–2121.
451 https://doi.org/10.1093/gbe/evx146
452 Bertels F, Gokhale CS, Traulsen A (2017) Discovering Complete Quasispecies in Bacterial
453 Genomes. *Genetics*, **206**, 2149–2157. https://doi.org/10.1534/genetics.117.201160
454 Bertels F, Rainey PB (2011a) Within-Genome Evolution of REPINs: a New Family of Miniature
455 Mobile DNA in Bacteria. *PLoS genetics*, **7**, e1002132.
456 https://doi.org/10.1371/journal.pgen.1002132
457 Bertels F, Rainey PB (2011b) Curiosities of REPINs and RAYTs. *Mobile Genetic Elements*, **1**, 262–
458 268. https://doi.org/10.4161/mge.18610
459 Bertels F, Rainey PB (2022) Ancient Darwinian replicators nested within eubacterial genomes. ,
460 2021.07.10.451892. https://doi.org/10.1101/2021.07.10.451892
461 Bichsel M, Barbour AD, Wagner A (2010) The early phase of a bacterial insertion sequence
462 infection. *Theoretical Population Biology*. https://doi.org/10.1016/j.tpb.2010.08.003
463 Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+:
464 architecture and applications. *BMC Bioinformatics*, **10**, 421–9.
465 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

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, **32**, 1792–1797. https://doi.org/10.1093/nar/gkh340

Felsenstein J (1985) Phylogenies and the comparative method. *American Naturalist*, 1–15.

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

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology*, **59**, 307–321. https://doi.org/10.1093/sysbio/syq010

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*.

488 Lawrence JG, Ochman H, Hartl DL (1992) The evolution of insertion sequences within enteric
489 bacteria. *Genetics*, **131**, 9–20. https://doi.org/10.1093/genetics/131.1.9
490 Nunvar J, Huckova T, Licha I (2010) Identification and characterization of repetitive extragenic
491 palindromes (REP)-associated tyrosine transposases: implications for REP evolution and
492 dynamics in bacterial genomes. *BMC Genomics*, **11**, 44. https://doi.org/10.1186/1471-
493 2164-11-44
494 Park HJ, Gokhale CS, Bertels F (2021) How sequence populations persist inside bacterial genomes.
495 *Genetics*, **217**. https://doi.org/10.1093/genetics/iyab027
496 R Core Team (2016) R: A Language and Environment for Statistical Computing.
497 Rankin DJ, Bichsel M, Wagner A (2010) Mobile DNA can drive lineage extinction in prokaryotic
498 populations. *Journal of Evolutionary Biology*. https://doi.org/10.1111/j.1420-
499 9101.2010.02106.x
500 RStudio, Inc (2013) Easy web applications in R.
501 Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, Rajandream MA, Barrell B (2000) Artemis:
502 sequence visualization and annotation. *Bioinformatics*, **16**, 944–945.
503 https://doi.org/10.1093/bioinformatics/16.10.944
504 Sawyer SA, Dykhuizen DE, DuBose RF, Green L, Mutangadura-Mhlanga T, Wolczyk DF, Hartl DL
505 (1987) Distribution and Abundance of Insertion Sequences Among Natural Isolates of
506 Escherichia coli. *Genetics*, **115**, 51–63. https://doi.org/10.1093/genetics/115.1.51
507 Ton-Hoang B, Siguier P, Quentin Y, Onillon S, Marty B, Fichant G, Chandler M (2012) Structuring
508 the bacterial genome: Y1-transposases associated with REP-BIME sequences. *Nucleic*
509 *Acids Research*, **40**, 3596–3609. https://doi.org/10.1093/nar/gkr1198

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

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

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

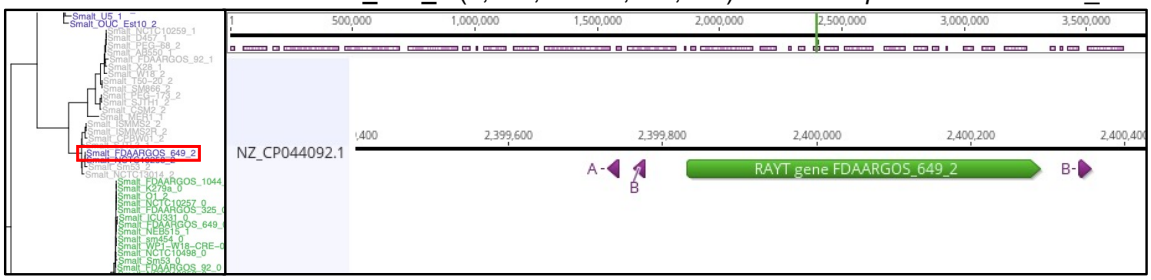
516 Yu G, Lam TT-Y, Zhu H, Guan Y (2018) Two Methods for Mapping and Visualizing Associated Data
517 on Phylogeny Using Ggtree. (FU Battistuzzi, Ed,). *Molecular biology and evolution*, **35**,
518 3041–3043. https://doi.org/10.1093/molbev/msy194

519

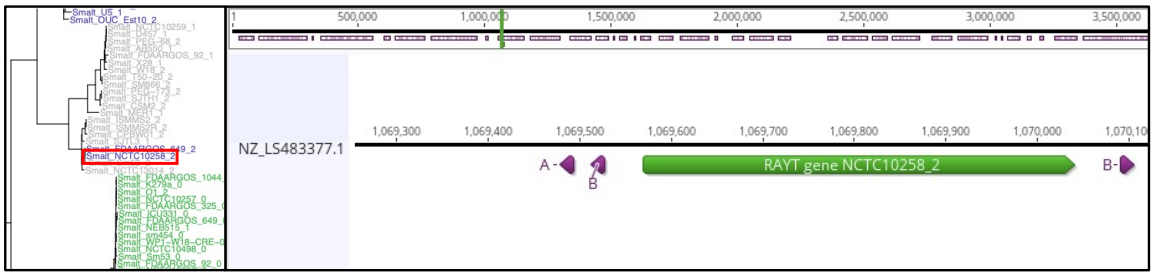
520

Supplementary Figures

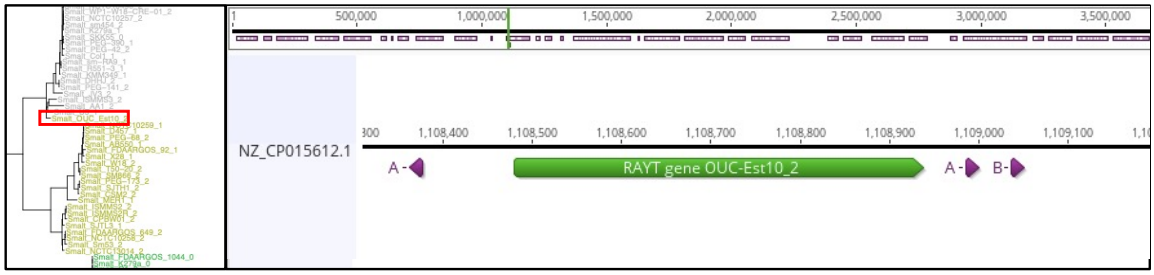
A **FDAARGOS_649_2** (2,399,832 - 2,400,290) in *S. maltophilia* FDAARGOS_649



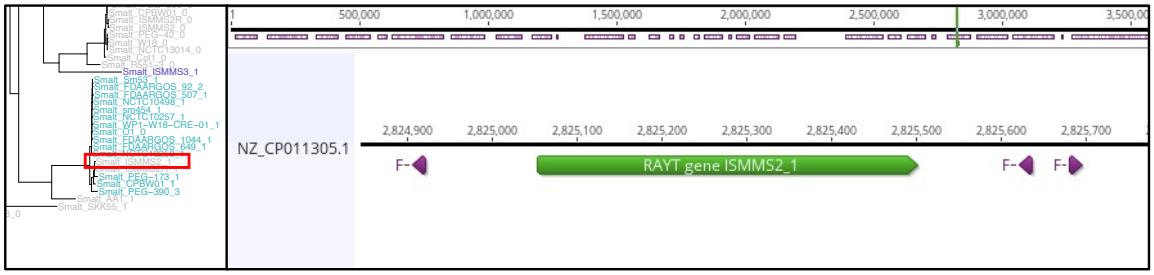
B **NCTC10258_2** (1,069,570 – 1,070,040) in *S. maltophilia* NCTC10258



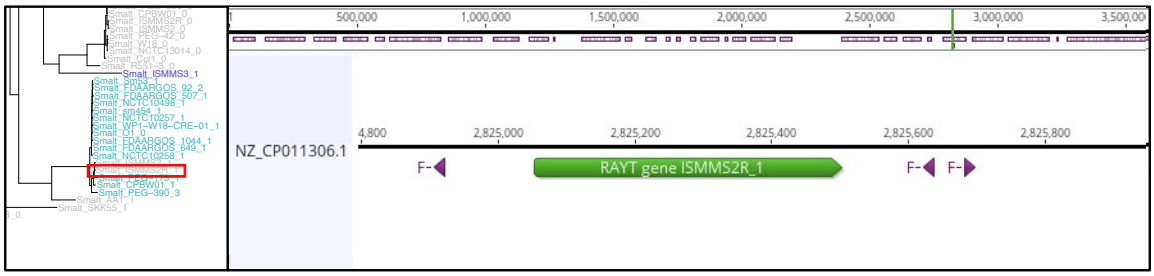
C **OUC_Est10_2** (1,108,480 – 1,108,938) in *S. maltophilia* OUC_Est10



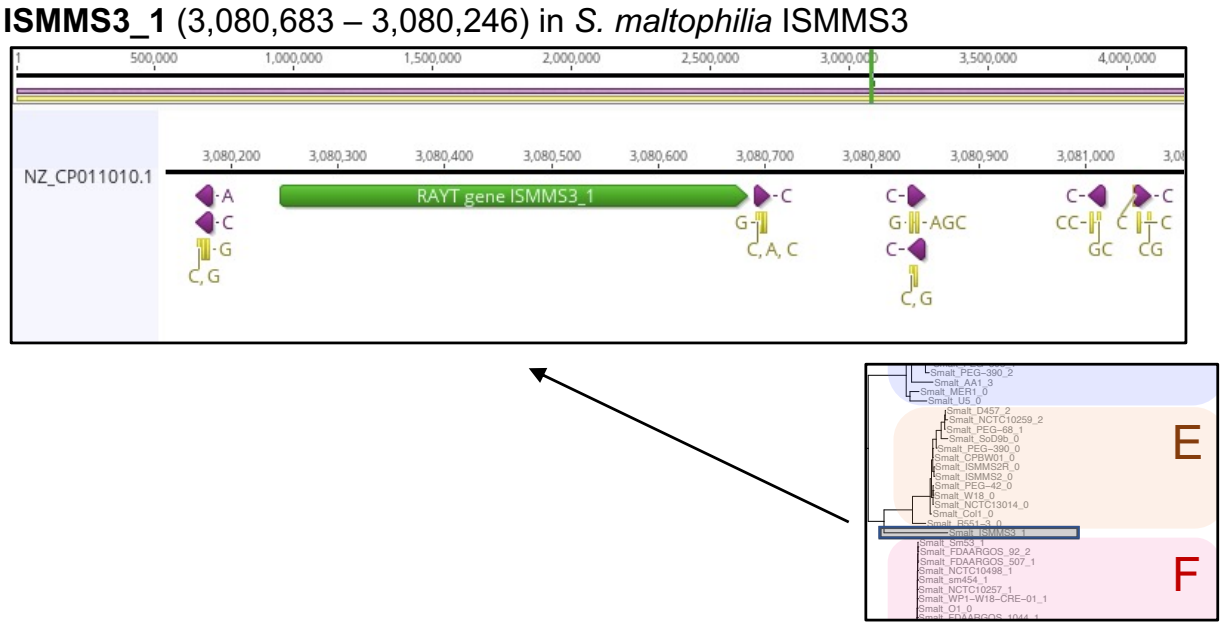
D **ISMMS2_1** (2,825,502 – 2,825,053) in *S. maltophilia* ISMMS2



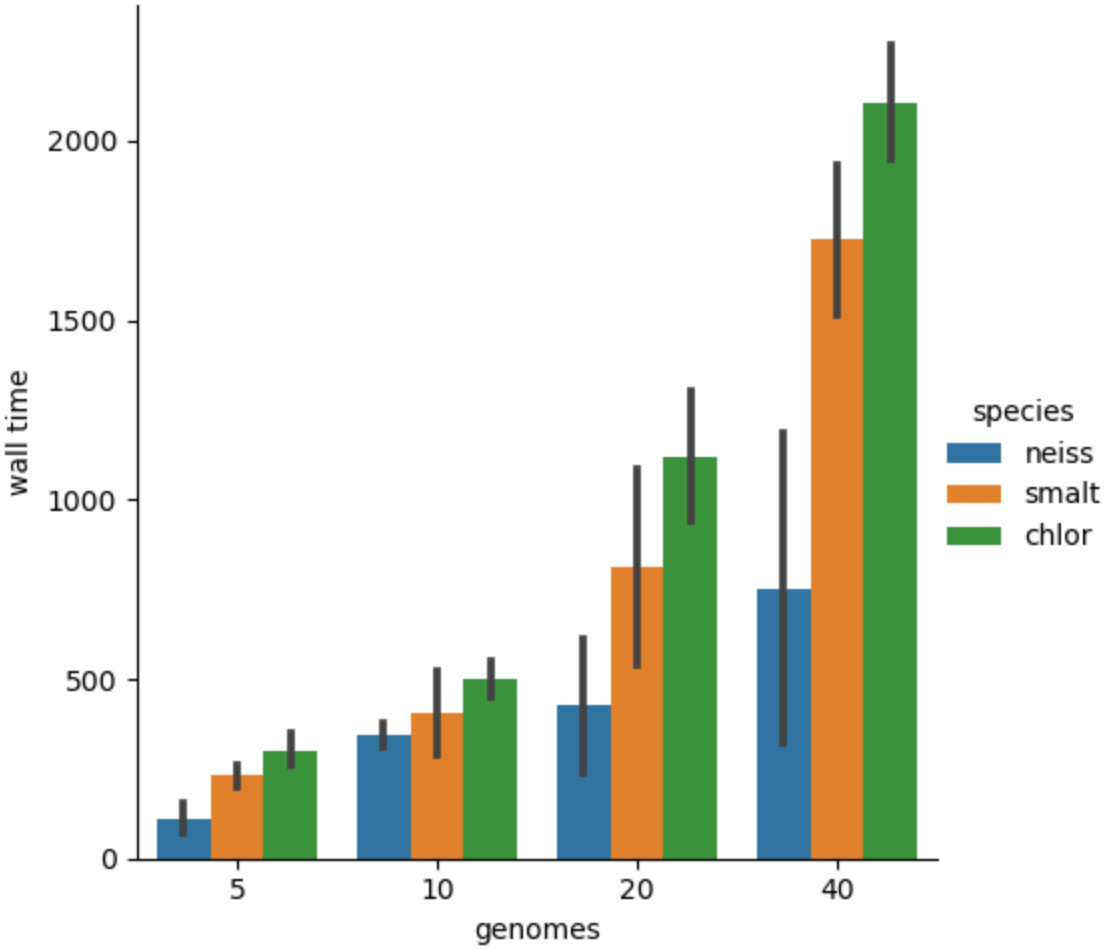
E **ISMMS2R_1** (2,825,502 – 2,825,053) in *S. maltophilia* ISMMS2R



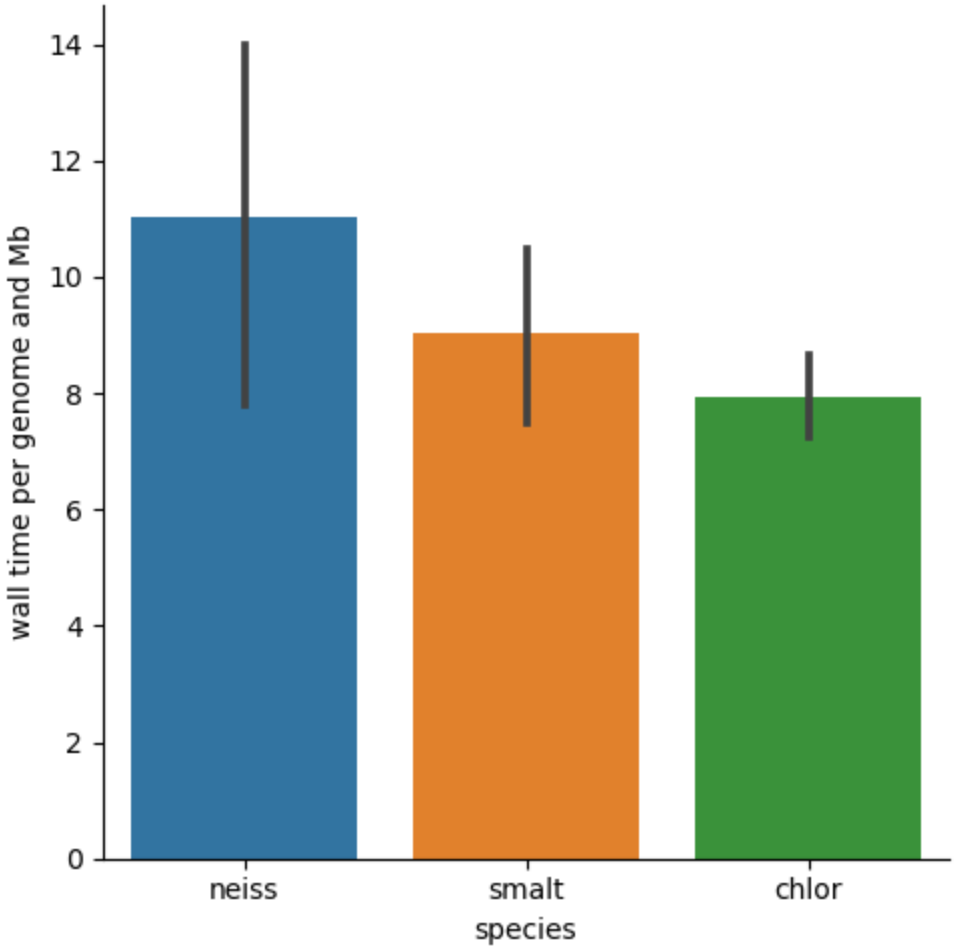
Supplementary Figure 1. Sequence analysis shows REPIN groups are indeed associated with monophyletic RAYTs. Non-monophyletic or missing associations to REPIN populations identified by RAREFAN were investigated in the corresponding genomes using Geneious (Kearse *et al.* 2012). Red boxes mark the position 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.



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 were analysed 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.



Supplementary Figure 3. Average time (in seconds) it takes RAREFAN to complete for different genome numbers from three bacterial species (*N. meningitidis, N. gonorrhoeae*, *S. maltophilia*, *Pseudomonas* chlororaphis). Black bars indicate the 95% CI across four runs, where two runs share the same query RAYT. For each run reference and query strains were randomly selected. All measurements were performed on 4CPU cores with 16 GB of shared memory.



552

553 **Supplementary Figure 4.** Approximate elapsed run time per megabase sequence length
554 calculated from the same runtime data generated in **Supplementary Figure 3**.

555

556