Identification and characterization of domesticated bacterial transposases

Frederic Bertels[1,2,*], Jenna Gallie[2] and Paul B Rainey[1,3,4]

[1]New Zealand Institute for Advanced Study, Massey University at Albany, Auckland, New Zealand

[2]Max Planck Institute for Evolutionary Biology, Department of Evolutionary Theory, Plön, Germany

[3]Max Planck Institute for Evolutionary Biology, Department of Microbial Population Biology, Plön, Germany

[4]Ecole Supérieure de Physique et de Chimie Industrielles de la Ville de Paris (ESPCI ParisTech), CNRS UMR 8231, PSL Research University, 75231 Paris Cedex 05, France

* To whom correspondence should be addressed: bertels@evolbio.mpg.de

ABSTRACT

Selfish genetic elements (SGEs), such as insertion sequences (ISs) and transposons are found in most genomes. Transposons are identifiable by their high copy number and capacity for horizontal transfer. In contrast REP associated tyrosine transposases (RAYTs), a recently described class of bacterial transposases, are typically present at just one copy *per* genome. This suggests that RAYTs no longer copy themselves. Motivated by this possibility we interrogated thousands of fully sequenced bacterial genomes in order to determine patterns of RAYT diversity, their distribution across chromosomes and accessory elements, and rate of duplication. RAYTs encompass exceptional diversity and are divisible into at least five distinct groups. They possess features more similar to housekeeping genes than insertion sequences, are vertically transmitted and have persisted through evolutionary time to the point where they are now found in ~19% of all fully sequenced bacterial genomes. Overall, the evolutionary dynamics of RAYTs suggest that they have been co-opted by host genomes to perform an important but currently unknown function.

Introduction

Mechanisms for the maintenance of selfish genetic elements (SGEs) have been discussed for over 40 years (Nevers and Saedler 1977; Doolittle and Sapienza 1980; Charlesworth et al. 1994). To persist in the long term, the subsequent host generation must contain more SGE copies on average than the previous host generation (Burt and Trivers 2006). This can be achieved even when the element confers a fitness cost, either through mechanisms that disproportionately increase copy number through self-replication and horizontal gene transfer (Doolittle and Sapienza 1980), or by mechanisms that kill offspring devoid of the SGE (*e.g.*, meiotic drive or toxin-antitoxin (TA) systems (Naito et al. 1995; Hurst et al. 1996; Gerdes et al. 2005)).

While SGEs need not benefit the host (Hickey 1982; Bichsel et al. 2013), many SGEs carry genes that do contribute toward enhanced host fitness. In some cases SGEs have even been domesticated. That is, they have evolved to perform functions that provide direct benefit to the host organism. Examples of such fitness enhancing functions include adaptive immunity, where genes derived from transposases mediate rearrangement of gene cassettes involved in antigen recognition (Jones and Gellert 2004), stress response systems derived from

39 bacterial TA systems (Van Melderen and De Bast 2009; Van Melderen 2010), and defence systems against
40 foreign DNA, such as caspases from CRISPRs (Krupovic et al. 2014).
41
42 Recently a new class of bacterial transposase has been described: REP associated tyrosine transposases
43 (RAYTs) are found in a wide range of bacteria (Nunvar et al. 2010; Bertels and Rainey 2011a; Ton-Hoang et al.
44 2012). Unlike typical transposases, which are part of insertion sequences (ISs; (Mahillon and Chandler 1998),
45 RAYTs occur as single copy elements. They are characteristically associated with short repetitive extragenic
46 palindromic sequences (REPs), and are typically arranged as pairs of REP sequences termed REPINs (REP
47 doublets forming hairpINs; (Bertels and Rainey 2011b; Bertels and Rainey 2011a). REPINs are non-
48 autonomous mobile elements that are significantly over-represented in many bacterial genomes, and are likely
49 dependent upon RAYTs for their dissemination (Nunvar et al. 2010; Bertels and Rainey 2011a; Ton-Hoang et
50 al. 2012).
51
52 Recently it has been argued that because REPs perform distinct functions within bacterial cells, ranging from
53 transcriptional termination (Espéli et al. 2001) to regulation of translation (Liang et al. 2015), RAYTs may be
54 domesticated transposable elements (Ton-Hoang et al. 2012; Siguier et al. 2014). While evidence supports the
55 notion that REP sequences have been co-opted to perform diverse functions, to our knowledge there is little
56 evidence to suggest this is also true for RAYTs; to date the only function assigned to RAYTs is movement of
57 REPINs (Nunvar et al. 2010; Bertels and Rainey 2011a; Ton-Hoang et al. 2012). Nonetheless, the idea that
58 RAYTs are domesticated transposons remains an intriguing hypothesis (Bertels and Rainey 2011a; Bertels and
59 Rainey 2011b).
60
61 Here we report an evolutionary analysis of RAYT diversity across thousands of bacterial genomes. Central to
62 our analysis is a comparison of the RAYT family with two families of well-characterized housekeeping genes
63 (*def* and *tpiA*) and two known IS families (IS*200* and IS*110*). We show that RAYTs tend to share characteristics
64 with housekeeping genes rather than with ISs. We also reveal five divergent types of RAYTs. Our data indicate
65 that RAYTs have no capacity for selfish replication and so, in addition to within-genome dissemination of
66 REPINs, RAYTs likely perform some currently unrecognized function that is central to their maintenance
67 within a broad range of bacterial genomes.
68

69 **MATERIAL AND METHODS**

70 **Acquisition of genome sequences**

71 Bacterial genome sequences were downloaded from the NCBI ftp site on the 11th of June 2013
72 (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/). On that day 2,950 bacterial chromosomes and 2,106 plasmids were
73 fully sequenced and available for analysis.

74 **Identification of family members by BLAST**

75 For each gene family (RAYTs, IS*200*, IS*100*, *def*, *tpiA*) one arbitrarily chosen protein was used as a query for a
76 BLAST search. The query sequence for the RAYT family was YafM from *Pseudomonas fluorescens* SBW25
77 (PFLU_RS20900). Query sequences for housekeeping genes were *P. fluorescens* SBW25 peptide deformylase,

Def (PFLU_RS00090), and *P. fluorescens* SBW25 triosephosphate isomerase TpiA (PFLU_RS25840). Query sequences for the ISs were ECIAI1_4438 from *Escherichia coli* IAI1 (IS*200* family) and, ISEc*32* from *E. coli* S88 plasmid pECOS88 (IS*110* family, ECSMS35_RS25240). The sequence of each protein was used to interrogate 2,950 chromosomes and 2,106 plasmids using TBLASTN (protein search against nucleotide database).

Search results were analysed as follows. First, all database matches with an e-value of less than 1e-2, 1e-5, 1e-15, and 1e-20 were recorded. Second, for each of the recorded database matches all coding sequences (CDS) from Genbank annotations were determined. Third, for overlaps with a single gene, the DNA sequence of the gene and its encoded amino acid sequence were extracted; for multiple overlaps the longest overlapping open reading frame was used for sequence extraction; matches without overlapping genes were ignored. Fourth, in addition to recording the DNA and amino acid sequence of the overlapping gene, the DNA sequence of the 5' and 3' extragenic space immediately flanking the identified gene was stored.

Identification of duplication events (analysis 2)

For all homologs that occur in the same genome, nucleotide sequences were aligned using the Needleman-Wunsch algorithm (Needleman and Wunsch 1970) and pairwise identities (the number of sites identical between two sequences divided by the total number of sites) calculated. All pairs with an identity greater than 95% were deemed duplicates.

Determination of genome-wide frequencies of flanking 16-mers (analysis 4)

Frequencies for all oligonucleotides of length 16 (16mers) from all replicons (chromosomes and plasmids) were determined, according to Bertels and Rainey (2011b). In brief, extragenic 5' and 3' 16mer frequencies were determined for each family member using a sliding window with a step size of one for both leading and lagging strands. With knowledge of this frequency, the most abundant 16mer from each flanking non-coding DNA sequence could be determined.

Calculating pairwise identity of amino acid sequences

Pairwise alignments between protein sequences were computed by applying the Needleman-Wunsch algorithm (Needleman and Wunsch 1970). Pairwise identity is the number of identical sites within the alignment divided by the total number of sites.

Definition of RAYT protein sequence groups

To determine sequence groups a Markov clustering algorithm (MCL) was applied to a matrix of pairwise protein identities greater than 26% (Van Dongen 2000). For proteins longer than 80 amino acids a identity threshold of 24.8% has previously been used for identifying homologs (Sander and Schneider 1991). The resulting sequence clusters were visualized with Cytoscape (www.cytoscape.org) (Cline, et al. 2007).

Phylogenetic analyses

To build a combined RAYT/IS*200* phylogeny, three protein sequences from each RAYT and IS*200* cluster containing more than 30 sequences (six RAYT and six IS*200* groups) were randomly selected. These 36 sequences were then aligned with MUSCLE (Edgar 2004) and a phylogeny was generated using PhyML

(Guindon and Gascuel 2003). The IS*200* Group 6 was excluded because sequences from this group could not be aligned with the rest of the sequences.

RESULTS AND DISCUSSION

Overview of a comparison of RAYTs to IS and housekeeping gene families

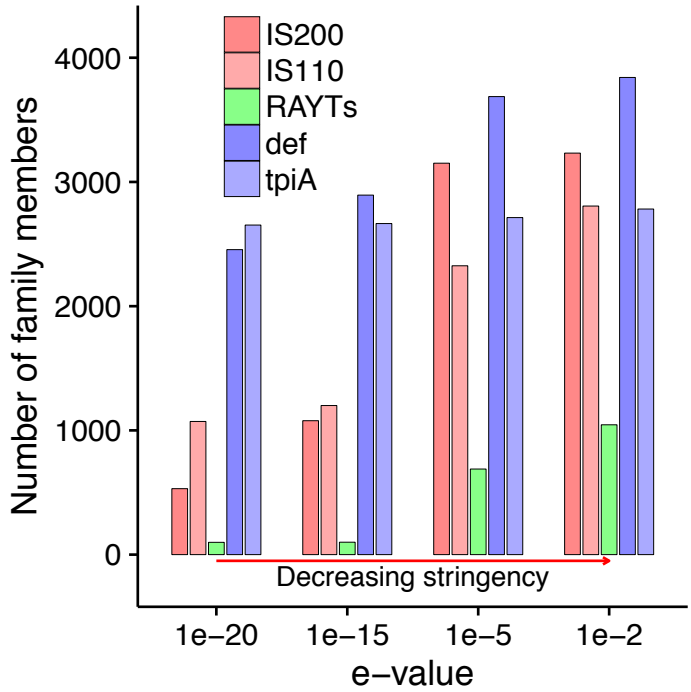
ISs persist by replicating within and between genomes (Doolittle and Sapienza 1980; Bichsel et al. 2013), with dissemination between genomes being mediated by elements capable of horizontal transmission (*e.g.,* plasmids). Hence, ISs are expected to be present in multiple copies within the genomes of individual bacteria, and to be over-represented on transmissible elements such as plasmids. In contrast, genes beneficial for bacteria (*e.g.* housekeeping genes) typically occur in a single copy *per* genome, and are not dependent on horizontal transfer to persist.

In order to assess whether RAYTs show patterns of diversity typical of ISs or housekeeping genes, we first identified members of five gene families and then compared familial characteristics. The five gene families are RAYTs, two IS families (IS*200* and IS*110*), an essential housekeeping gene family (*def*; peptide deformylase), and a non-essential but highly conserved gene family (*tpiA*; triosephosphate isomerase). For each family, we investigated the following four characteristics: (1) the average number of family members *per* replicon (chromosome or plasmid), (2) the proportion of family members for which identical (or nearly identical) copies are present on the same replicon, (3) the proportion of family members present on plasmids, and (4) the average replicon-wide frequency of the most common 16bp sequence from the DNA sequences flanking each family member. The results for the RAYT family were compared to those obtained for the other families.

Identification of members of the RAYT, IS*200*, IS*110*, *def* and *tpiA* gene families

We defined gene families as groups of genes that share a similar sequence with a single representative of that gene family (the reference gene, arbitrarily chosen). To identify all members of a family, the reference protein sequence was used in a tBLASTN (basic local alignment search tool; (Altschul et al. 1990)) search against all publicly available, fully-sequenced bacterial genomes. Each BLAST search was performed at four different e-value thresholds (1e-20, 1e-15, 1e-5 and 1e-2). The e-value is a measure for sequence similarity; low e-values (1e-20) indicate high similarity and high e-values (1e-2) indicate low similarity. Hence gene families defined by less restrictive e-values are expected to be larger as they encompass more divergent genes.

IS and housekeeping gene families are of comparable size (**Figure 1**). At the broadest definition of gene family (*i.e.* 1e-2), the *def* gene family is the largest (3,841 members; members found in 92% of all bacterial genomes searched) followed by IS*200* (3,232; members present in 26% of all bacterial genomes searched), IS*110* (2,806; members found in 26% of all bacterial genomes searched) and *tpiA* (2,782; members present in 93% of all bacterial genomes searched). The RAYT family is considerably smaller, containing 1,045 members with members occurring in 19% of all bacterial genomes analysed.



153

154 **Figure 1. Family size of ISs, RAYTs and housekeeping genes under four decreasingly stringent definitions**
155 **of sequence relatedness**.

156

157 As is expected of functionally constrained housekeeping genes, the diversity of these gene families is low,
158 which is indicated by the small number of additional members added to each family with the addition of
159 increasingly divergent sequences. Conversely, the size of the IS*110*, IS*200* and RAYT families increased
160 considerably upon addition of more divergent members, indicating less sequence conservation. This effect was
161 most notable for the RAYT family, for which the family size increased more than ten-fold from the most to the
162 least stringent family definition.

163

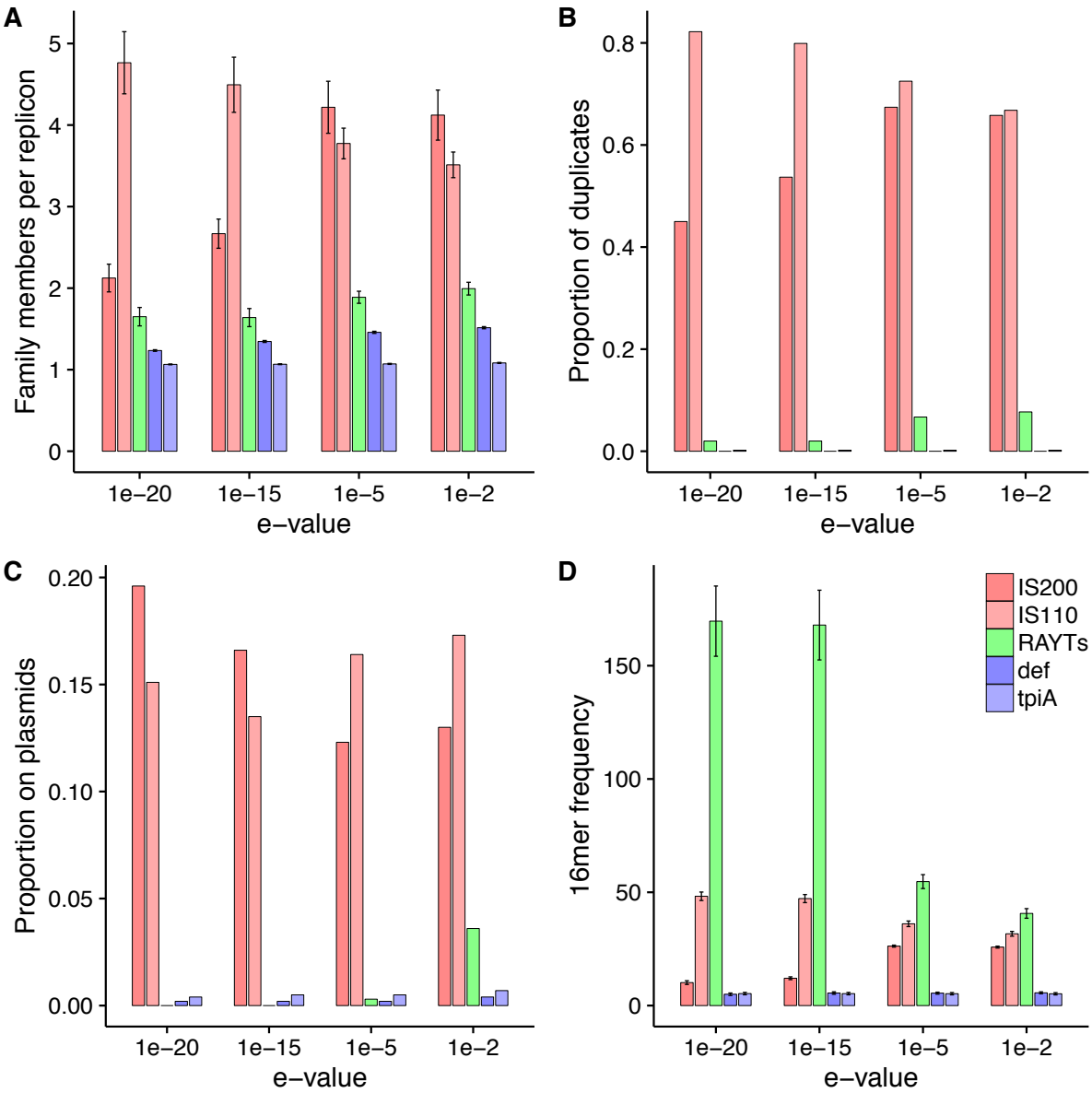


Figure 2. Distribution characteristics of five gene families. (*A*) Average copy number of each gene family *per* replicon (excluding replicons in which no family members occur). (*B*) Proportion of genes for which at least one duplicate (pairwise nucleotide identity of ≥95%) was found on the same replicon. (*C*) Proportion of family members found on plasmids. (*D*) Average frequency of the most abundant 16mer identified in the flanking 5′ and 3′ extragenic regions of each family member. All error bars show standard errors.

Analysis 1: Number of family members *per* replicon

Having determined the number of genes in each family, we next wanted to know how gene family members are distributed across replicons. That is, for replicons that contain at least one member of a particular gene family, how many members are present on average (**Figure 2A**)?

Under the most stringent family definition, replicons that harbour IS*200* or IS*110* contain on average between two and five copies *per* replicon, respectively. Interestingly, the two IS families show opposite trends with respect to e-value classes. The 531 members comprising the most stringently defined IS*200* gene family (**Figure**

178 **1**) occur on 250 different replicons, hence they are represented on average 2.1 times per replicon. This number
179 increases at lower stringency to approximately four occurrences *per* replicon. Contrastingly, the average number
180 of copies *per* replicon for the IS*110* family decreases from 4.8 to 3.5 as familial stringency is lowered. This
181 means that although IS*100* family size increases with decreasing stringency (**Figure 1**), the number of replicons
182 harbouring IS*100* elements increases at a greater rate.
183
184 For the two housekeeping gene families, the number of homologs *per* replicon increases with increasing family
185 size. The extent of the increase, however, is different for the two families. For the *tpiA* family the number of
186 copies *per* replicon increases only slightly (1.06 to 1.08). For the *def* family the increase in the number of family
187 members *per* replicon is higher (1.2 to 1.5). This larger increase may have been caused by an ancient
188 duplication event that led to the addition of a family of *def* paralogs (genes derived from ancient duplication
189 events) for less stringent family definitions (Bergthorsson et al. 2007; Näsvall et al. 2012).
190
191 RAYTs are present a little more than once *per* replicon when family size is defined by the most stringent e-
192 value criterion, and almost two *per* replicon under the least stringent criterion. This hints at diversification of the
193 RAYT family, which we analyse in detail in later sections.
194

195 **Analysis 2: Proportion of family members that are duplicated on a replicon**
196 Analysis 1 encompasses **all** members of a family present on a replicon. This includes recently duplicated genes,
197 genes derived from ancient duplications (paralogs) and laterally acquired genes. Of these, recently duplicated
198 genes are of greatest interest to our study, because they are likely to have arisen from recent self-duplication
199 events (*i.e.* SGE activity). Accordingly, one would expect the relative size of this gene subset to be considerably
200 larger for SGEs than for genes with a beneficial function. Thus, taking advantage of the fact that recent
201 duplication events will have resulted in the presence of gene copies with identical (or nearly so) DNA
202 sequences, we next identified recently duplicated genes for the five gene families at each e-value. Specifically,
203 we determined the proportion of each gene family for which at least one nearly identical copy (≥95%) exists on
204 the same replicon (**Figure 2B**).
205
206 For the least stringent gene family definition (1e-2), about 66% of all IS*200* (2127/3232) and IS*110* (1874/2806)
207 genes appear to have been recently duplicated (**Figure 2B**). For these families, the proportion of duplicated
208 genes correlates with the number of copies *per* replicon (**Figure 2A**; F-statistic for a linear fit of data for both IS
209 families P=6.5 x 10^{-5}). We observe that, with increasing proportion of duplications, there is an increase in the
210 number of gene copies *per* replicon in both IS families. No such correlation exists for either RAYTs or
211 housekeeping genes, as expected from the low proportion of gene duplications (see below). This means that for
212 ISs the change in the average number of family members *per* replicon (**Figure 1**) is not due to ancient
213 duplication events leading to diversification of the IS family; rather, it is due to differences in duplication rates
214 for IS family members more distantly related to the reference gene.
215
216 In contrast to ISs, the two housekeeping gene families show almost no evidence of recent duplication. At 1e-2,
217 there were no duplications among *def* gene family members (0/3841), and only 0.2% (3/2782) for the *tpiA*

218 family for the least stringent family definition. Of all the RAYT genes at 1e-2, 7.7% (80/1045) were found to
219 have recently duplicated with the additional copy being present on the same replicon. The proportion of RAYT
220 duplication events is less than that observed for IS*200* and IS*110*, but nevertheless higher than that of
221 housekeeping gene families. As we show in later sections, this pattern results from the addition of IS*200*
222 sequences to the RAYT family at less stringent e-values.

223 **Analysis 3: Presence of gene family members on plasmids**

224 In the above analyses, both chromosomes and plasmids were interrogated for the presence of homologous
225 genes. We next determined the proportion of family members that are found solely on plasmids (**Figure 2C**).
226 Under the assumption that plasmids facilitate the horizontal transfer of genes, we expect SGEs such as ISs to be
227 present more frequently than housekeeping genes on plasmids (Bichsel et al. 2013).
228
229 ISs are commonly found on plasmids; across all stringency classes, between 13% and 19.6% of IS*200* genes and
230 between 15% and 17.3% of IS*110* genes were found on plasmids. Contrastingly, members of the two
231 housekeeping gene families were almost never found on plasmids - only 0.2% to 0.4% of all *def* genes and 0.4%
232 to 0.7% of all *tpiA* genes occurred on plasmids. For RAYTs, 3.6% of family members occur on plasmids at the
233 least stringent gene family definition (1e-2). This value is between that obtained for ISs and housekeeping
234 genes. However at the slightly more stringent e-value of 1e-5, only 0.3% of all RAYTs occur on plasmids, a
235 value very similar to that found for the housekeeping gene families. Again this observation can be explained by
236 the addition of IS*200* members at higher e-values.

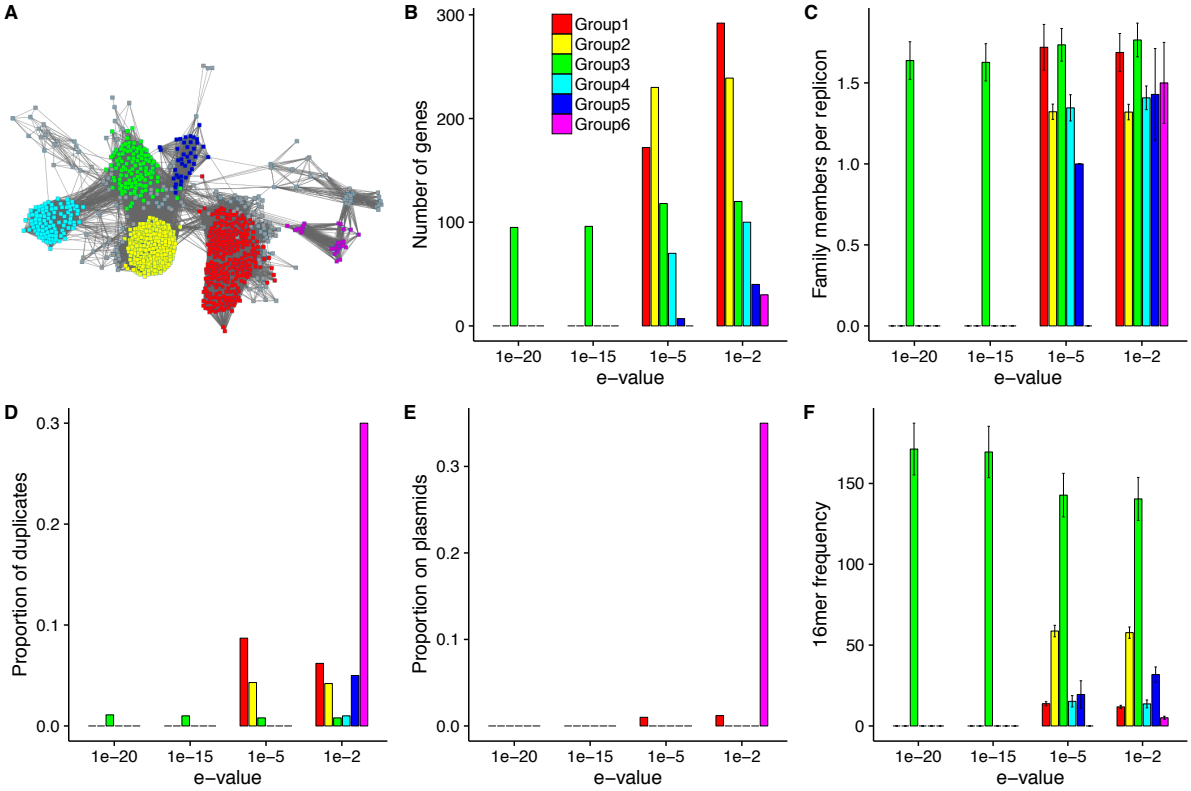
237 **Analysis 4: Maximum 16mer frequencies in the flanking extragenic space**

238 Many bacterial genomes contain short, palindromic sequences that are overrepresented in extragenic space.
239 These sequences have been named repetitive extragenic palindromic sequences (REPs) (Higgins et al. 1982;
240 Stern et al. 1984), and have been shown to replicate as part of a doublet termed a REPIN (Bertels and Rainey
241 2011a). REPINs are almost always found in the extragenic space flanking RAYT genes. These flanking REPINs
242 usually contain highly abundant 16bp long sequences (16mers) (Bertels and Rainey 2011a).We generalized this
243 finding and use the abundance of the most common 16mer of the 5′ and 3′ extragenic flanking regions as a
244 feature to distinguish RAYTs from other genes.
245
246 For the IS*200* and IS*110* families, we found higher frequencies of flanking 16mers than for housekeeping genes:
247 the average frequency of flanking 16mers ranged between 10 and 48. The most common flanking 16mer for ISs
248 are typically functionally linked to the IS (Mahillon and Chandler 1998) as they are part of the flanking terminal
249 repeats. Hence, a significant correlation is expected between the proportion of gene family members that are
250 duplicates and 16mer frequencies. The higher the proportion of duplicates the higher the flanking 16mer
251 frequency. Such a correlation is evident from **Figures 2D** and **2B** (F-statistic for a linear fit P=1.7 x 10^{-5})
252 supporting the notion that the most common flanking 16mer is part of the IS.
253
254 We found that 16mers in the adjacent extragenic space of housekeeping genes are repeated infrequently in the
255 remainder of the replicon: average frequencies range from 5.2 to 5.6 (for different e-values). For 100 randomly
256 assembled genomes of ~6.7Mb in length (the length of the *P. fluorescens* SBW25 genome), the most abundant

257 16mer occurs on average 5.08 times in the genome (Bertels and Rainey 2011a). The frequencies for 16mers
258 flanking housekeeping genes are only slightly higher than these values, indicating that duplications do not affect
259 flanking 16mer frequencies for housekeeping genes.
260
261 The RAYT gene family was found to have the highest frequencies of flanking 16mers across replicons, from an
262 average of 170 for the most stringent family definition to 41 for the least stringent. As expected there is no
263 correlation between 16mer frequencies and the proportion of duplicated genes (F-statistic for a linear fit P=0.24,
264 ranging from 2% for the most stringent family definition to 7.7%). Thus gene duplication events are not the
265 cause for the observed high frequencies of 16mers flanking RAYT genes. Instead high flanking 16mer
266 frequencies are caused by RAYTs disseminating neighbouring REPINs throughout the genome.
267

268 **RAYT characteristics show large variations across sequence groups**

269 In the above analyses, some characteristics of RAYT genes change dramatically with the inclusion of more
270 diverse family members. For example, the results of analysis 3 indicate that a subset of RAYTs identified only
271 at the most inclusive e-value (1e-2) is found on plasmids (**Figure 2C**). Further, as family members are added
272 from a stringent to very loose gene family definition, the size of the RAYT family increases ten-fold, more than
273 any other sequence family. Together these results raise the possibility that the increase in RAYT gene family
274 size is attributable to the existence of RAYT subgroups, with more distantly related RAYTs being incorporated
275 with each relaxation of the family definition. In the following section we investigate this hypothesis.



276
277 **Figure 3. Protein similarity network and genomic distribution of RAYT groups.** (A) RAYT protein
278 relationship map constructed by Cytoscape using the organic layout (1e-2; (Cline et al. 2007). RAYT groups

identified by MCL are highlighted according to the legend. Nodes represent RAYT proteins. Nodes are connected if the pairwise amino acid identity is ≥26%. Differences between the Cytoscape protein clustering (proximity of nodes) and the MCL algorithm (colour of nodes) are due to the difference in the information provided to MCL and Cytoscape: Cytoscape only receives information on whether or not nodes are connected, whereas MCL performs coloration based on the exact pairwise similarity between proteins. Graphs (B) to (F) contain the same information as **Figures 1** and **2**, but for the six largest RAYT groups in (A) as opposed to the five gene families.

RAYT protein sequences cluster into distinct subgroups

To determine whether the RAYT family consists of distinct subgroups we applied the MCL clustering algorithm (Van Dongen 2000) to a protein similarity matrix of RAYTs at the least stringent family definition. In the matrix, each node is a protein, and nodes with ≥26% amino acid identity are connected. This similarity matrix contains the amino acid identity for each pair of proteins. The clustering algorithm uses this information to divide the protein network into subgroups of strongly connected nodes. RAYT protein sequences cluster into at least six subgroups, each containing 30 or more members (**Figure 3A**). For each of these subgroups we determined the same genomic distribution characteristics as for the five gene families above, at the same e-values (**Figure 3B-F**).

Characteristics of RAYT protein subgroups

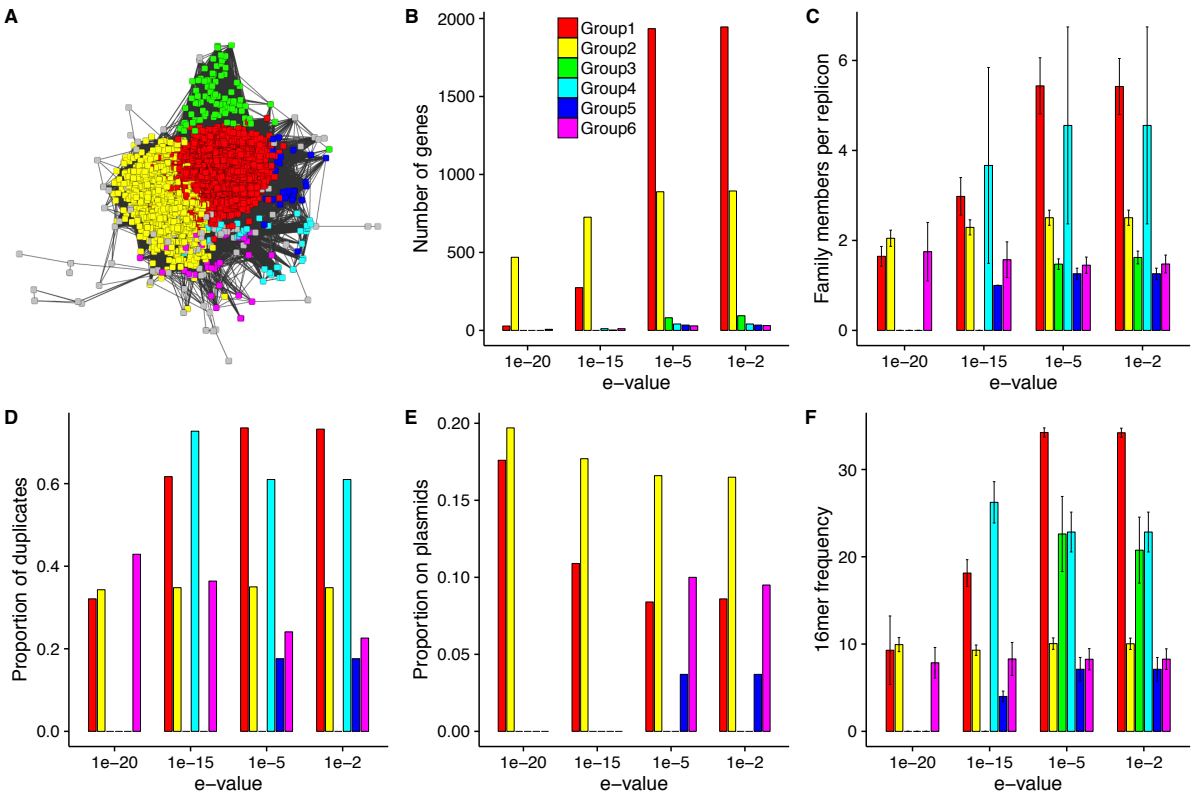
As the protein similarity network in **Figure 3A** shows, the RAYT sequence family is divisible into at least six groups. The different groups show significantly different responses to the inclusion of increasingly diverse RAYT family members.

The effect of broadening family definition on group size is dramatically different between RAYT groups (**Figure 3B**). Group 3 is essentially invariant to the effects of adding family members of increasing diversity, which is a consequence of the fact that the RAYT query sequence stems from this group. Members of Groups 1, 2, 4, and 5 appear in the family at 1e-5, and are maximal at 1e-2. Members of Group 6, the group most distantly related to the query sequence, are only included at the broadest family definition.

There is little variation in the number of group members *per* replicon at the broadest family definition; the numbers range from 1.3 (Group 2) to 1.8 (Group 3) copies *per* replicon (**Figure 3C**). Despite this, the proportion of group members with a nearly identical copy on the same replicon differs dramatically between groups (**Figure 3D**). Group 6 members (30%) and Group 1 members (6.3%) have the highest proportion of duplication events, whereas Group 3 has the lowest (0.8% at 1e-2). Surprisingly, although Group 3 has the lowest proportion of duplicates, it is the group with the highest number of gene copies *per* replicon, possibly indicating that this group is more functionally diverse than the other RAYT groups.

The proportion of group members found on plasmids shows a similar pattern to the proportion of duplicates observed for each group (**Figure 3E**). Again Group 6 shows by far the highest proportion on plasmids (35% at 1e-2). The only other RAYT group for which any members were found on plasmids was Group 1 (1.2% at 1e-

317 2), which also had the second highest proportion of duplicates (**Figure 3D)**. Hence, of all RAYT subgroups,
318 Group 1 and 6 are the two that most closely resemble ISs in terms of their distribution characteristics.
319
320 The maximum frequency of 16mers found in extragenic space indicates which of the RAYT groups are likely to
321 be associated with REP sequences (**Figure 3F**). Only Group 2 (57 replicon-wide occurrences at 1e-2) and
322 Group 3 (140 occurrences) members are associated with 16mers that are more frequently found across replicons
323 than those associated with ISs. However, Group 1 (12 occurrences) and Group 4 (14 occurrences) are associated
324 with 16mers that are more common than 16mers from housekeeping genes. Group 5 (31 occurrences), despite
325 not showing signs of high duplication rates, is associated with 16mers that are as frequent as those associated
326 with insertion sequences. Group 6 (5), despite relatively high duplication rates is associated with 16mers that
327 show low abundance in the genome.
328
329 Broadly speaking, the results in this section show that RAYT protein relationships (**Figure 3A**) are correlated
330 with the similarities and differences observed in their characteristics. For example, in accordance with the
331 frequency of flanking 16mers, Group 2 and 3 are closely linked in the protein network. Group 1, 4 and 5 are
332 more loosely linked and Group 6 is the most distantly related group.
333
334 The difference in the characteristics of Group 6 to the RAYT family as well as the very loose connection to the
335 RAYT family in the protein relationship network raises the possibility that Group 6 RAYTs also belong to a
336 family of ISs. IS*200* is the most likely close relative (Nunvar et al. 2010; Bertels and Rainey 2011a; Ton-Hoang
337 et al. 2012). To investigate this possibility, the IS*200* family was subjected to identical analyses as the RAYT
338 family.
339



340

341 **Figure 4. Protein similarity network and genomic distribution of IS*200* subgroups.** (A) shows a protein
342 similarity network of the IS*200* family. The map was produced as described in the legend for **Figure 3A**.
343 Graphs (B) to (F) show the same genomic characteristics as in **Figures 2 and 3**.

344 **Characteristics of IS*200* subgroups**

345 The structure of the IS*200* family is much more homogenous than that of the RAYT family (**Figures 4A** and
346 **3A,** respectively). All subgroups are arranged around the largest groups, Group 1 and Group 2. The only
347 subgroup that is more separate is Group 3, a distinction that is also evident in the very different distribution
348 characteristics of Group 3.

349

350 Group sizes in IS*200* are less evenly distributed than RAYT groups (**Figure 4B**). Groups 1 and 2 (the central
351 subgroups in **Figure 4A**) have by far the most members (1946 and 894), whereas the remaining groups all have
352 less than 100 members.

353

354 As expected for ISs, the number of gene copies *per* replicon correlates (**Figure 4C**) well with the observed
355 proportion of duplicates for each group (**Figure 4D**). The proportion of duplicates is high for all IS*200* groups
356 except for Group 3, for which we did not find a single duplicate. Groups with a high proportion of duplicates
357 have a high number of gene copies *per* replicon (F-statistic for a linear fit for all groups $P=10^{-12}$). Again the
358 only outlier is Group 3, which has more gene copies *per* replicon than Group 4 and Group 5, but no duplicates.

359

360 The proportion of genes on plasmids is high for four out of the six IS subgroups (**Figure 4E**). As expected from
361 the previous results, we did not observe any Group 3 members on plasmids. More surprisingly, we also did not

observe Group 4 members on plasmids. This is rather unexpected for an IS with an extremely high replication rate (proportion of duplicates is 61%).

Typical for ISs, the frequency of flanking 16mers again correlates with the proportion of duplicates (**Figure 4F;** F-statistic for a linear fit for all groups $P=10^{-6}$), except for Group 3, which is associated with highly abundant 16mers (21 occurrences across replicons at 1e-2) but shows no signs of duplication.

Our analyses show that RAYTs and IS*200* sequences have very different genomic distribution characteristics. RAYTs show features more typical of housekeeping genes and IS*200* shows features typical of ISs. However, in each family we identified one subgroup that does not fit into the overall pattern. It is possible that the IS-like RAYT group belongs to the IS*200* family, considering its distant relationship to the RAYT family. Similarly, although not as distantly related to the IS*200* query sequence, it is possible that IS*200* Group 3 belongs to the RAYT family. To investigate these possibilities we performed a combined phylogenetic and network analysis for the RAYT and IS*200* Groups.

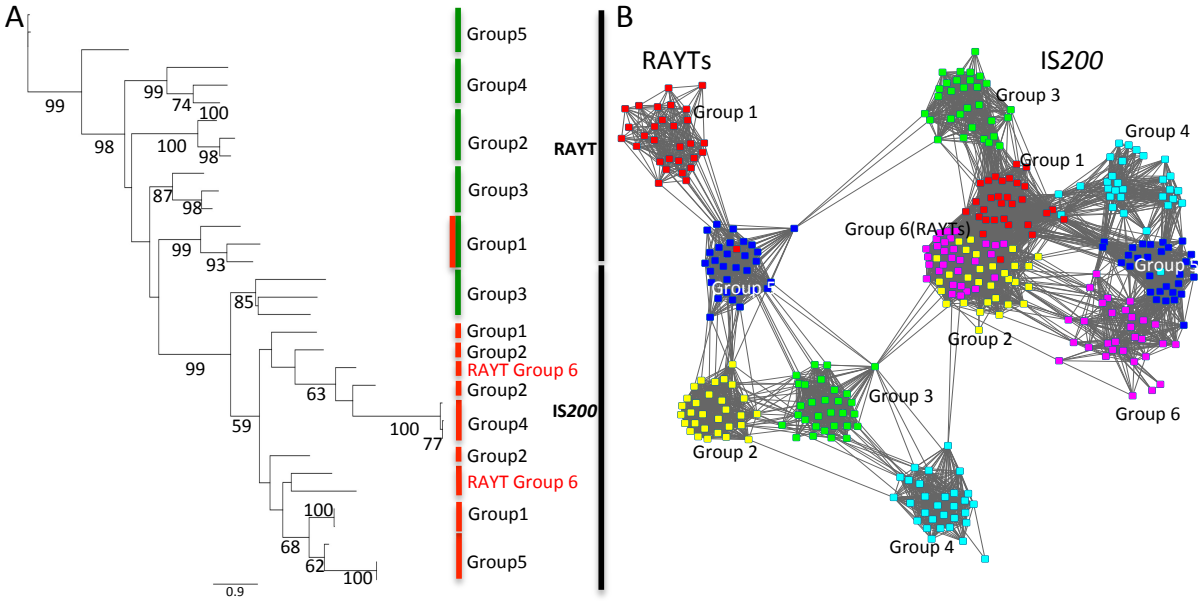


Figure 5. Phylogeny of RAYT and IS*200* family. (A) Phylogenetic tree. From each IS*200* and RAYT group, three random proteins were selected and aligned with MUSCLE (Edgar 2004). We excluded IS*200* Group 6, because the alignments could not be calculated correctly when members of this subgroup were included. The phylogeny was built using PHYML (Guindon and Gascuel 2003). All RAYTs are found in the same clade except for RAYT Group 6, which clusters within the IS*200* family. This is plausible considering the IS-like genome distribution characteristics. Green bars indicate low duplication rates and a low presence on plasmids; red bars indicate high duplication rate and high presence on plasmids. RAYT Group 1 is marked with a red/green bar as within the RAYT family it has the highest duplication rate and the highest presence on plasmids. (B) Protein similarity network of the RAYT (left) and IS*200* (right) gene family. We chose 30 random proteins from the six largest RAYT and IS200 groups and displayed them with Cytoscape identically to **Figures**

388 **3F** and **4F**. RAYT Group 6 clusters within IS*200* Group 1 in line with the phylogeny in (A) and the genomic
389 distribution characteristics in **Figure 3**.

390 **Combined RAYT and IS*200* analysis**

391 The combined RAYT and IS*200* phylogeny shows that the IS*200* and RAYT families are distinct with the
392 exception of RAYT Group 6, which clusters within the IS*200* phylogeny (**Figure 5A**), In contrast, the RAYT-
393 like IS*200* Group 3 clusters as a distinct group with the IS*200* family. However, according to the phylogenetic
394 tree, IS*200* Group 3 is the most closely related IS*200* group to the RAYT family. This raises the possibility that
395 RAYTs have been domesticated only once and that the IS*200* Group 3 has maintained sequence features similar
396 to IS*200* sequences. This hypothesis, however, needs to be treated with caution as the high divergence between
397 the different genes, potential convergent evolution and horizontal gene transfer hinders ability to infer the
398 phylogenetic history with confidence.
399
400 The separation between RAYTs and IS*200* in the phylogenetic tree is not as clear in the protein network (**Figure**
401 **5B**). Although RAYTs seem to form relatively coherent clusters, these clusters do not include the RAYT-like
402 Group 3 from the IS*200* family.
403

404 **Conclusion**

405 RAYTs are a recently described class of bacterial transposases that share a distant evolutionary relationship with
406 the IS*200* gene family (Lam and Roth 1983; Lam and Roth 1986; Barabas et al. 2008). The suggestion of a
407 distant evolutionary relationship comes from the fact that both RAYTs and IS*200* share an identical catalytic
408 centre (Nunvar et al. 2010; Bertels and Rainey 2011a; Ton-Hoang et al. 2012). Despite this similarity, the data
409 presented here show that the RAYT and IS*200* families display different characteristics. Members of the IS*200*
410 gene family show characteristics typical of ISs, whereas members of the RAYT family are, in many respects,
411 almost indistinguishable from housekeeping genes. The housekeeping-like nature of RAYTs provides a firm
412 basis for the earlier suggestion that RAYTs are domesticated transposons (Bertels and Rainey 2011a; Bertels
413 and Rainey 2011b; Ton-Hoang et al. 2012; Siguier et al. 2014).
414
415 The housekeeping role of RAYTs is unclear, but the existence of five distinct subfamilies (**Figures 3A & 5B**)
416 raises the possibility of five different functions. Biochemical studies on RAYTs from *E. coli* provide evidence
417 of *in vitro* endonuclease and DNA binding activity (Ton-Hoang et al. 2012). Such activity is thought to be
418 integral to the capacity of RAYTs to catalyse transposition of non-autonomous REP / REPIN elements *via* an
419 IS*200*-like mechanism (Barabas et al. 2008; Ton-Hoang et al. 2010). But this role alone is insufficient to explain
420 the maintenance of RAYTs within genomes. In the absence of RAYTs catalysing their own transposition (for
421 which there is no evidence), there is no mechanism for RAYTs to escape mutational decay, leading to the
422 prediction that RAYTs must perform some function that benefits the host cell (Bertels and Rainey 2011a;
423 Bertels and Rainey 2011b).
424
425 Although there are various plausible scenarios, it is challenging to conceive of a single RAYT function that
426 defines both its interaction with REPINs and its beneficial effect on the host bacterium. One possibility is that

427 the various previously described (and likely co-opted) functions of REPs and REPINs, including effects on
428 translation, transcription and DNA organization (Newbury et al. 1987; Yang and Ames 1988; Espéli et al. 2001;
429 Voineagu et al. 2008; Liang et al. 2015), depend on some functional interaction with RAYTs. For example,
430 binding of the RAYT protein to a particular REP or REPIN sequence may modulate its effects, with the
431 different classes of RAYT possibly reflecting different specialisations in different bacterial groups. An alternate
432 idea that we find intriguing — and which draws upon the endonuclease function of RAYTs — is that RAYTs
433 function in concert with REPINs in a CRISPR-like way (Horvath and Barrangou 2010), with the palindromic
434 REPIN structure acting as bait for incoming foreign DNA (Tobes and Pareja 2006; Treangen et al. 2009) that is
435 then inactivated or degraded *via* the endonuclease activity of the RAYT. It is not inconceivable that such a
436 scenario might also generate antagonistic co-evolution between bait and RAYT promoting the evolution of
437 divergent RAYT subfamilies.
438

439 **REFERENCES**

440 Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol
441 215:403–410.

442 Barabas O, Ronning DR, Guynet C, Hickman AB, Ton-Hoang B, Chandler M, Dyda F. 2008. Mechanism of
443 IS*200*/IS*605* family DNA transposases: activation and transposon-directed target site selection. Cell
444 132:208–220.

445 Bergthorsson U, Andersson DI, Roth JR. 2007. Ohno's dilemma: evolution of new genes under continuous
446 selection. Proc Natl Acad Sci U S A 104:17004–17009.

447 Bertels F, Rainey PB. 2011a. Within-Genome Evolution of REPINs: a New Family of Miniature Mobile DNA
448 in Bacteria. PLoS Genet. 7:e1002132.

449 Bertels F, Rainey PB. 2011b. Curiosities of REPINs and RAYTs. Mob Genet Elements 1:262–268.

450 Bichsel M, Barbour AD, Wagner A. 2013. Estimating the fitness effect of an insertion sequence. J Math Biol
451 66:95–114.

452 Burt A, Trivers R. 2006. Genes in Conflict: The Biology of Selfish Genetic Elements. Cambridge,
453 Massachusetts: Belknap Press of Harvard University Press

454 Charlesworth B, Sniegowski P, Stephan W. 1994. The evolutionary dynamics of repetitive DNA in eukaryotes.
455 Nature 371:215–220.

456 Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech
457 M, Gross B, et al. 2007. Integration of biological networks and gene expression data using Cytoscape. Nat
458 Protoc 2:2366–2382.

459 Doolittle WF, Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature
460 284:601–603.

461 Espéli O, Moulin L, Boccard F. 2001. Transcription attenuation associated with bacterial repetitive extragenic
462 BIME elements. J Mol Biol 314:375–386.

463 Gerdes K, Christensen SK, Løbner-Olesen A. 2005. Prokaryotic toxin-antitoxin stress response loci. Nat Rev
464 Micro 3:371–382.

465 Hickey DA. 1982. Selfish DNA: a sexually-transmitted nuclear parasite. Genetics 101:519–531.

466 Higgins CF, Ames GF, Barnes WM, Clement JM, Hofnung M. 1982. A novel intercistronic regulatory element
467 of prokaryotic operons. Nature 298:760–762.

468 Horvath P, Barrangou R. 2010. CRISPR/Cas, the immune system of bacteria and archaea. Science 327:167–
469 170.

470 Hurst LD, Atlan A, Bengtsson BO. 1996. Genetic conflicts. Q Rev Biol 71:317–364.

471 Jones JM, Gellert M. 2004. The taming of a transposon: V(D)J recombination and the immune system. Immunol
472 Rev 200:233–248.

473 Krupovic M, Makarova KS, Forterre P, Prangishvili D, Koonin EV. 2014. Casposons: a new superfamily of
474 self-synthesizing DNA transposons at the origin of prokaryotic CRISPR-Cas immunity. BMC Biol 12:36.

475 Lam S, Roth JR. 1983. IS*200*: a *Salmonella*-specific insertion sequence. Cell 34:951–960.

476 Lam S, Roth JR. 1986. Structural and functional studies of insertion element IS*200*. J Mol Biol 187:157–167.

477 Liang W, Rudd KE, Deutscher MP. 2015. A Role for REP Sequences in Regulating Translation. Mol Cell
478 58:431–439.

479 Mahillon J, Chandler M. 1998. Insertion sequences. Microbiol Mol Biol Rev 62:725–774.

480 Naito T, Kusano K, Kobayashi I. 1995. Selfish behavior of restriction-modification systems. Science 267:897–
481 899.

482 Näsvall J, Sun L, Roth JR, Andersson DI. 2012. Real-time evolution of new genes by innovation, amplification,
483 and divergence. Science 338:384–387.

484 Nevers P, Saedler H. 1977. Transposable genetic elements as agents of gene instability and chromosomal
485 rearrangements. Nature 268:109–115.

486 Newbury SF, Smith NH, Robinson EC, Hiles ID, Higgins CF. 1987. Stabilization of translationally active
487 mRNA by prokaryotic REP sequences. Cell 48:297–310.

488 Nunvar J, Huckova T, Licha I. 2010. Identification and characterization of repetitive extragenic palindromes
489 (REP)-associated tyrosine transposases: implications for REP evolution and dynamics in bacterial
490 genomes. BMC Genomics 11:44.

491 Sander C, Schneider R. 1991. Database of homology-derived protein structures and the structural meaning of
492 sequence alignment. Proteins 9:56–68.

493 Siguier P, Gourbeyre E, Chandler M. 2014. Bacterial insertion sequences: their genomic impact and diversity.
494 FEMS Microbiol Rev 38:865–891.

495 Stern MJ, Ames GF, Smith NH, Robinson EC, Higgins CF. 1984. Repetitive extragenic palindromic sequences:
496 a major component of the bacterial genome. Cell 37:1015–1026.

497 Tobes R, Pareja E. 2006. Bacterial repetitive extragenic palindromic sequences are DNA targets for Insertion
498 Sequence elements. BMC Genomics 7:62.

499 Ton-Hoang B, Pasternak C, Siguier P, Guynet C, Hickman AB, Dyda F, Sommer S, Chandler M. 2010. Single-
500 stranded DNA transposition is coupled to host replication. Cell 142:398–408.

501 Ton-Hoang B, Siguier P, Quentin Y, Onillon S, Marty B, Fichant G, Chandler M. 2012. Structuring the bacterial
502 genome: Y1-transposases associated with REP-BIME sequences. Nucleic Acids Res 40:3596–3609.

503 Treangen TJ, Abraham A-L, Touchon M, Rocha EPC. 2009. Genesis, effects and fates of repeats in prokaryotic
504 genomes. FEMS Microbiol Rev 33:539–571.

505 Van Dongen S. 2000. A cluster algorithm for graphs. Report-Information systems:1–40.

506 Van Melderen L, De Bast MS. 2009. Bacterial toxin-antitoxin systems: more than selfish entities? PLoS Genet.
507 5:e1000437.

508 Van Melderen L. 2010. Toxin-antitoxin systems: why so many, what for? Curr Opin Microbiol 13:781–785.

509 Voineagu I, Narayanan V, Lobachev KS, Mirkin SM. 2008. Replication stalling at unstable inverted repeats:
510 interplay between DNA hairpins and fork stabilizing proteins. Proc Natl Acad Sci U S A 105:9936–9941.

511 Yang Y, Ames GF. 1988. DNA gyrase binds to the family of prokaryotic repetitive extragenic palindromic
512 sequences. Proc Natl Acad Sci U S A 85:8850–8854.

513