

Discovering complete quasispecies in bacterial genomes

Frederic Bertels^{*1}, Chaitanya S. Gokhale^{*} and Arne Traulsen^{*}

^{*}Department of Evolutionary Theory, Max Planck Institute for Evolutionary Biology, August-Thienemann-Str. 2, 24306, Plön, Germany

ABSTRACT Mobile genetic elements can be found in almost all genomes. Possibly the most common non-autonomous mobile genetic elements in bacteria are REPINs that can occur hundreds of times within a genome. The sum of all REPINs within a genome are an evolving populations because they replicate and mutate. We know the exact composition of this population and the sequence of each member of a REPIN population, in contrast to most other biological populations. Here, we model the evolution of REPINs as quasispecies. We fit our quasispecies model to ten different REPIN populations from ten different bacterial strains and estimate duplication rates. We find that our estimated duplication rates range from about 5×10^{-9} to 37×10^{-9} duplications per generation per genome. The small range and the low level of the REPIN duplication rates suggest a universal trade-off between the survival of the REPIN population and the reduction of the mutational load for the host genome. The REPIN populations we investigated also possess features typical of other natural populations. One population shows hallmarks of a population that is going extinct, another population seems to be growing in size and we also see an example of competition between two REPIN populations.

KEYWORDS REP sequences; mobile genetic elements; evolution; bacteria; quasispecies

Introduction

Repetitive sequences are common in most bacterial genomes, but rare compared to most eukaryotic genomes (Jurka *et al.* 2007; Versalovic *et al.* 1991). A large proportion of repetitive sequences in bacterial genomes are the result of self-replicating DNA sequences. These sequences usually encode an enzyme called a transposase that specifically copies its own sequence (Mahillon and Chandler 1998). There are also repetitive sequences that do not encode a transposase themselves, but are copied by a transposase that is encoded elsewhere in the genome. These elements are referred to as MITEs (Miniature Inverted repeat Transposable Elements) (Wessler *et al.* 1995). MITEs were first described in plant genomes (Bureau and Wessler 1994) and later also in bacteria (Oggioni and Claverys 1999). Recently, it has been shown that REP (Repetitive Extragenic Palindromic) sequences (Higgins *et al.* 1982) or more specifically REPINs (REP doublets forming hairpINs) (Bertels and Rainey 2011b), one of

the most abundant repeat families in bacteria, are also MITEs (Nunvar *et al.* 2010; Bertels and Rainey 2011b,a; Ton-Hoang *et al.* 2012).

REP sequences are about 25 bp long sequences that are highly abundant in bacterial genomes (Higgins *et al.* 1982; Aranda-Olmedo *et al.* 2002; Silby *et al.* 2009). They contain a short imperfect palindromic sequence that can form short hairpins in single stranded DNA or RNA. REP sequences mostly occur in non-coding DNA between genes and are part of REPINs. REPINs in most *Pseudomonas* strains consist of two REP sequences in inverted orientation separated by a highly diverse nucleotide sequence (Bertels and Rainey 2011b). REPINs are a replicative unit and are mobilized by RAYTs (REP Associated tYrosine Transposases) (Nunvar *et al.* 2010; Bertels and Rainey 2011b; Ton-Hoang *et al.* 2012). Although the structure of REPINs in *Pseudomonas* is well defined, for REPINs in *E. coli* there has not been an extensive study on what exactly comprises the replicative unit.

The occurrence of REP sequences and associated functions have been described in many different bacterial genomes (Higgins *et al.* 1982; Aranda-Olmedo *et al.* 2002; Silby *et al.* 2009). However, their evolution has rarely been studied in detail (Bertels and Rainey 2011a,b) and nothing is known about the dupli-

54 cation rates of REPINs. Although, we know that closely related
55 *E. coli* strains contain varying numbers of REP sequences, this
56 may not be a direct result of replication. Instead it may be more
57 likely that it is a consequence of the extremely dynamic genome
58 composition of *E. coli* (Touchon *et al.* 2009), where REP sequences
59 get deleted or inserted together with other parts of the genome.
60 However, the lack of evidence for novel REPIN insertions proba-
61 bly means that duplication rates are low, despite the presence
62 of hundreds of REPINs in some genomes (Bertels and Rainey
63 2011b).

64 As it is difficult to study the evolution of the complete REPIN
65 sequence due to the highly diverse loop region (which is proba-
66 bly strongly affected by recombination), we model the evolution
67 of the most conserved 25bp at each end of the REPIN. Here we
68 infer REPIN duplication rates by modeling the most abundant
69 REPINs in a bacterial genome as a quasispecies in equilibrium.
70 The beauty of studying REPINs in bacterial genomes is that we
71 know the exact composition of the population at the time of
72 genome sequencing, something that is impossible to achieve for
73 almost any other population study.

74 We first fit the equilibrium of our quasispecies model for a
75 REPIN population from *Pseudomonas fluorescens* SBW25 and later
76 for nine other bacterial genomes. Our results show that despite
77 the large divergence between the bacterial strains, our inferred
78 duplication rates are very similar and very low. All rates fall into
79 a narrow margin between one replication in about 31×10^6 and
80 200×10^6 host divisions. Hence, if a bacterium were to divide
81 every 40 minutes, it would take about 2359 years for a specific
82 REPIN duplication to fix in the population. The astonishing
83 rarity of these events may explain the lack of evidence for novel
84 REPIN insertions in bacterial genomes.

85 Materials and Methods

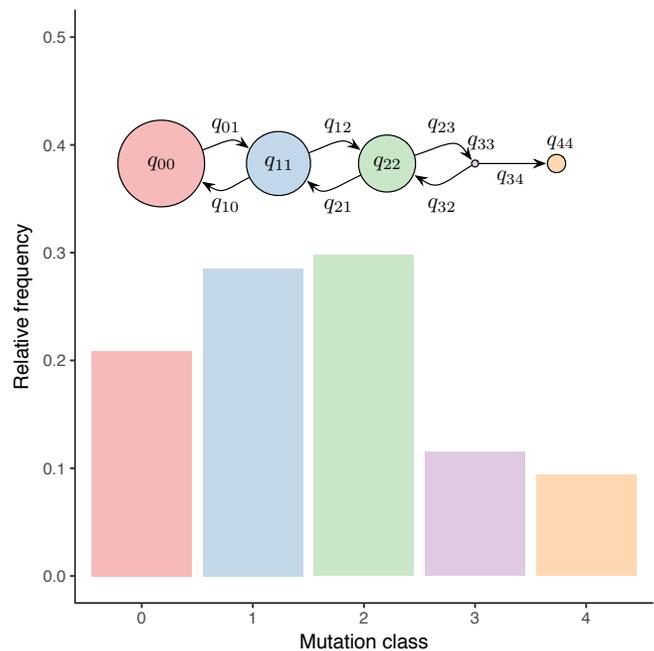
86 Quasispecies model

87 The quasispecies model describes the mutation-selection balance
88 of a set of similar sequences that evolve on a fitness landscape.
89 On this landscape, each sequence has a certain fitness. Sequences
90 with high fitness leave many offspring, sequences with low
91 fitness leave few offspring. The fitness landscape is traversed
92 by acquiring mutations (Eigen 1971; Eigen and Schuster 1977;
93 Nowak 1992).

94 The quasispecies model has been applied previously mostly
95 to model viral populations (Seifert *et al.* 2015; Domingo and
96 Schuster 2016). Here, we model REPIN sequences that mutate
97 and duplicate: the fitness in the quasispecies model corresponds
98 to the REPIN duplication rate and the model's mutation rate to
99 the genome mutation rate. We assume that the REPIN popula-
100 tion in our genome is a quasispecies in equilibrium. The most
101 abundant sequence in our population is our master sequence.
102 With increasing genetic distance to the master sequence, fitness
103 changes. For our model we assume five discrete fitness classes.
104 The 0th class contains the master sequence. Sequences differing
105 in 1, 2 or 3 positions are in the next three classes. The remaining
106 sequences are in the 4th fitness class. The frequencies of the se-
107 quences belonging to each of these classes i are given by x_i . The
108 population evolves to a mutation-selection balance as described
109 by the standard quasispecies equation (Page and Nowak 2002;
110 Bull *et al.* 2005)

$$\dot{x}_i = \sum_{j=0}^n x_j f_j q_{ji} - x_i \phi. \quad (1)$$

111 In our case n equals 4. The fitness of sequences belonging to
112 each class j is given by f_j and the average fitness of the popu-
113 lation by $\phi = \sum_{i=0}^n x_i f_i$. The probability that a sequence from
114 class j mutates into i is given by q_{ji} . In our model, sequences
115 can only acquire a single mutation per time step. Hence, \mathbf{Q} is a
116 tri-diagonal matrix with non-zero entries in the main diagonal
117 (no mutation) the first diagonal above (sequence acquires an ad-
118 ditional mutation) and the first diagonal below (back mutation).
119 For a mutation rate μ and a sequence length L , the probability
120 of transitioning to the next mutation class $i + 1$ is $\mu(L - i\frac{1}{3})$ and
121 to the previous mutation class $i - 1$ is $i\mu\frac{1}{3}$. The fourth mutation
122 class is the only class where we assume a back mutation rate of
123 zero — the exact value would depend on the frequency distribu-
124 tion of the sequences that differ by more than three mutations
125 to the master sequence. We also assume that the mutation rate
126 of REPINs only depends on the host mutation rate. Mutations
127 that occur during the duplication process are assumed to be
128 negligible.



129 **Figure 1 Exemplar results for a quasispecies model.** For a
130 mutation rate of $\mu = 8.9 \times 10^{-11}$, and the fitnesses as given
131 in Table 1 (1+scaled duplication rate), we illustrate the equi-
132 librium distribution of the relative frequencies of *P. fluorescens*
133 SBW25 REPINs. The radii of the circles indicate the duplica-
134 tion rate, which is the quasispecies fitness subtracted by one.
135 Note that the actual fitness differences are extremely minute
136 at the level of 10^{-9} . The cartoon merely illustrates the archi-
137 tecture of the fitness landscape. The mutation probabilities are
138 given by (q_{ij}) while self-replication occurs with probability q_{ii} .

139 Parameterizing the quasispecies model

140 We set the fitness of the highest mutation class to one, $f_4 = 1$. For
141 a given set of equilibrium sequence frequencies, we can then cal-
142 culate the relative fitness of the remaining four mutation classes
143 for a given mutation rate (see File S7). For all our bacteria we
144 assume a host mutation rate of 8.9×10^{-11} , which was inferred
145 for *E. coli* (Wielgoss *et al.* 2011). The duplication rate is then the
146 calculated fitness for each mutation class subtracted by one.

137 **Stochastic simulations**

138 For each REPIN population, we performed a stochastic simu- 198
139 lation to determine the extent of stochastic fluctuation on the 199
140 equilibrium frequencies. These fluctuations mainly depend on 200
141 the REPIN population size. As we cannot simulate evolution 201
142 for the genome mutation rate, we scaled our fitness values up 202
143 to fit a mutation rate of 10^{-4} . With the new mutation rate, each 203
144 discrete time step corresponds to $g = \frac{10^{-4}}{8.9 \times 10^{-11}} \approx 10^6$ bacte- 204
145 rial generations. Because we assume multiplicative fitness, the 205
146 fitness values at a mutation rate of 10^{-4} are comparable to $(f_i)^g$. 206

147 We modeled evolution with a Wright-Fisher process (Ewens 206
148 1979). We start the simulation with a clonal population of the 207
149 master sequence at carrying capacity, which is set to the num- 208
150 ber of REPINs observed in the genome. The number of off- 209
151 spring each sequence leaves in each generation is equal to the 210
152 sequence's fitness. If the number of offspring exceeds the carry- 211
153 ing capacity, a random selection of the same size as the carrying 212
154 capacity survives to the next time point. We modeled a total of 213
155 10^9 generations.

156 We repeated each simulation 100 times and measured the 214
157 proportion of simulations where the 0th mutation class persisted 215
158 at a frequency of more than 10%. 216

159 **Determining REPIN populations**

160 We extracted REPIN populations from 10 bacterial genomes 217
161 the following way: For each of these genomes we determined 218
162 the most common 25 bp long sequence. We then recursively 219
163 searched the genome for all sequences that have a Hamming 220
164 distance of 2 to all identified sequences until no more sequences 221
165 were found. We call these sequences REP sequences. For all 222
166 REP sequences we determined whether they were part of a se- 223
167 quence cluster by checking whether there were any additional 224
168 occurrences in a vicinity of 130bp. From these sequence clus- 225
169 ters we extracted REPINs. REPINs consist of two adjacent REP 226
170 sequences that are found in opposite directions (one on the pos- 227
171 itive strand the other on the negative DNA strand, also called 228
172 inverted repeats) in the DNA sequence. The REPINs we found 229
173 were extracted and joined together facing the same direction in 230
174 alphabetical order. REP sequences found as direct repeats or as 231
175 singlets in the genome were also extracted (as single sequences). 232
176 We added another 25bp of adenine nucleotides at the end of 233
177 each REP singlet to make them easily comparable with REPINs. 234

178 **Clustering REPIN sequences**

179 REPIN populations can be represented as sequence networks. 235
180 In these networks, each node represents a sequence. Vertices 236
181 between nodes exist if the Hamming difference between the se- 237
182 quence pair is one. Because REPIN populations in *Pseudomonas* 238
183 do not always evolve on a single peak due to the presence of mul- 239
184 tiple RAYTs (transposases) in the genome, we extracted subpop- 240
185 ulations clustered around the master sequence. We determined 241
186 these subpopulations for all *Pseudomonas* strains by applying a 242
187 Markov clustering algorithm implemented in the MCL package 243
188 (van Dongen 2000) with the inflation parameter set to 1.2 to the 244
189 sequence network. The MCL algorithm simulates random walks 245
190 on a stochastic graph by alternating between expansion and 246
191 inflation operations, where larger inflation parameters will lead 247
192 to more fragmented networks

193 We used the largest REPIN cluster for our analyses. Since 248
194 these clusters exclude decayed sequences far from the master 249
195 sequences, we also included all sequences with a Hamming dis- 250
196 tance of two to any sequence in the cluster. Of the sequences 251

197 identified in the last step we only included instances that oc- 252
198 curred less than three times in the genome. Sequences that occur 253
199 more than three times in the genome are likely to have been 254
200 duplicated by other RAYTs.

201 **Inferring an error threshold**

202 The error threshold defines a critical point in a quasispecies 255
203 where with the given fitness values and mutation rate it is im- 256
204 possible to maintain the master sequence. Here we deviate 257
205 slightly from this definition as we define the error threshold as 258
206 the point where the master sequence cannot be maintained at a 259
207 relative frequency of more than 1%. To determine the duplica- 260
208 tion rate at which we reach our error threshold, we decrease all 261
209 fitness values in increments of 1×10^{-12} . As soon as one of the 262
210 five fitness parameter reaches one, this parameter will remain 263
211 constant for the remainder of the procedure. We performed this 264
212 procedure for the fitness landscape of each species separately.

213 **Data Availability**

214 All genomes are publicly available on Genbank 265
215 (<https://www.ncbi.nlm.nih.gov/genbank/>) under the 266
216 following accession numbers:

Species Name	NCBI Accession number
<i>P. syringae</i> pv. <i>tomato</i> DC3000	NC_004632.1
<i>P. synxantha</i> BG33R	CM001514
<i>P. fluorescens</i> A506	NC_017911
<i>P. fluorescens</i> SBW25	NC_012660.1
<i>P. putida</i> GB1	NC_010322.1
<i>E. coli</i> 536	NC_008253.1
<i>E. coli</i> K-12 MG1655	CP014225.1
<i>E. coli</i> UTI89	NC_007946.1
<i>E. coli</i> B REL606	NC_012967.1
<i>E. coli</i> UMN026	NC_011751.1

218 We included eight supplemental files. File S1 contains de- 267
219 tailed descriptions of all supplemental files. File S2 contains 268
220 the sequence and frequency of the most common 25 bp long 269
221 sequence, the gene name of the flanking RAYT and the number 270
222 of RAYTs, in all of the bacteria analyzed in this study. File S3 271
223 contains the modeling and simulation results for all ten REPIN 272
224 populations we analyzed in our study. File S4 contains the Pro- 273
225 portion of symmetric REPINs in all identified sequences from 274
226 all studied strains. File S5 contains the duplication rates and 275
227 equilibrium frequencies for each of the 10 REPIN populations 276
228 at the error threshold. File S6 contains the Mathematica code 277
229 we used to calculate equilibrium frequencies, fitness values and 278
230 error thresholds for all 10 REPIN populations. File S7 contains 279
231 the same Mathematica code as pdf. File S8 contains the sequence 280
232 frequencies of the different mutation classes for all 10 REPIN 281
233 populations. 282

234 **Results and Discussion**

235 **REPINs in *Pseudomonas fluorescens* SBW25.**

236 In *Pseudomonas fluorescens* SBW25 REPINs consist of two in- 283
237 verted highly conserved sequences that are 25 bp in length, 284

238 separated by a sequence of varying length that shows low levels
239 of conservation (Bertels and Rainey 2011b,a). The processes that
240 lead to the varying levels of conservation in REPINs are not
241 very well understood. Hence, we will focus our analysis only on
242 the most conserved 25 bp flanking the REPIN. These sequences
243 have been discovered a long time ago in *E. coli* and have been
244 called REP sequences (Stern *et al.* 1984). To find the most con-
245 served parts of the REPIN, we determined the most common
246 25bp long sequence in the SBW25 genome. This sequence occurs
247 265 times and is usually part of a REPIN (Bertels and Rainey
248 2011b). We then add all sequences that differ in no more than
249 two positions to this sequence. For the identified sequences
250 we do the same and so on, until we can find no more new se-
251 quences in the genome. The resulting REP population contains
252 932 REP sequences. For these sequences, we determine whether
253 they are part of a REP cluster, by looking for all occurrences in
254 the vicinity of 130bp. From these clusters, we extract adjacent
255 pairs of inverted REP sequences or REPINs. REP singlets were
256 also extracted but marked with a 25bp long adenine sequence.
257 The relationship between REPINs is visualized as a sequence
258 network (Figure 2).

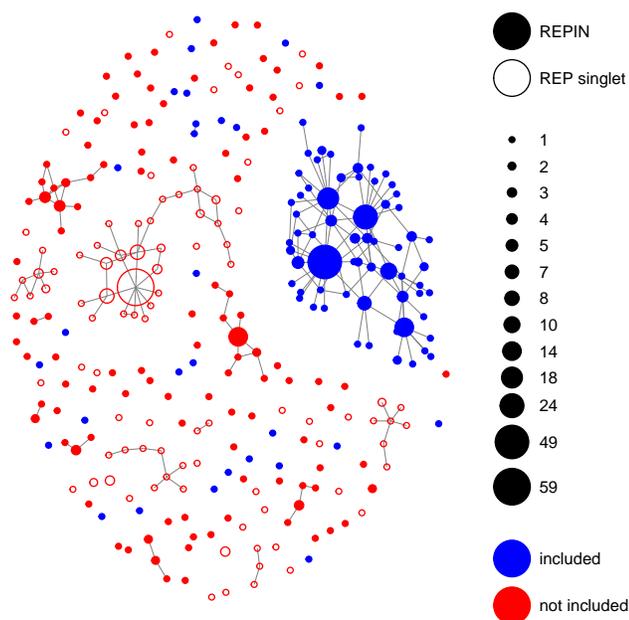


Figure 2 Structure of the REPIN population in SBW25. REPINs that differ in exactly one position are connected. REP sequences that do not form REPINs (e.g. singlets) are shown as empty circles. Blue “included” nodes belong to the REPIN population for which we infer duplication rates. Red (“not included”) nodes were excluded from the analysis because they likely evolve on a more complex fitness landscape that is more difficult to model. The size of the nodes indicates the frequency of the corresponding sequence in the SBW25 genome.

259 The population network in Figure 2 has many sequence hubs
260 distantly related and not connected to the master sequence. In-
261 stead of a very rugged activity landscape of a single RAYT (the
262 transposase responsible for duplicating REPINs), we think it
263 is more likely that these hubs were created by the concurrent
264 activities of multiple RAYT transposases (the SBW25 genome
265 contains three RAYT genes). As it is impossible to accurately

266 model this complexity for small REPIN populations, we decided
267 to reduce the REPIN population to all sequences that are part of
268 the largest cluster as well as all sequences that are at most 2bp
269 different from any sequence that is part of the cluster.

270 The “included” subpopulation selected in Figure 2 has 235
271 members. We will model this subpopulation as quasispecies,
272 with five sequence classes, that are 0, 1, 2, 3 and more than 3
273 mutations away from the master sequence. In our model we
274 will also assume that the population is in equilibrium and the
275 frequencies of the sequences we observe are steady state fre-
276 quencies. The mutation rate in our model was chosen to be high
277 to facilitate stochastic simulations of the evolutionary process.
278 The fitness values for each mutation class were calculated from
279 the quasispecies equation for the sequence frequencies observed
280 in SBW25 (Table 1).

281 The quasispecies equation provides us with a set of fitness
282 values that perfectly recapitulate the observed frequencies for
283 infinitely large populations (Figure 3A). However, REPIN popu-
284 lations are relatively small, which means that population size
285 will have a strong effect during REPIN evolution. To estimate
286 stochastic effects, we used the calculated fitness parameters for
287 each mutation class to perform a stochastic Wright-Fisher simu-
288 lation with a maximum of 235 individuals (Figure 3B). Our
289 simulation shows that the distributions of the mutation classes
290 are wide, particularly for the master sequence, which is probably
291 an effect of the small population size (Figure 3C).

292 The rate at which duplications occur can be calculated from
293 the inferred fitness values. We calculate the duplication rate from
294 these fitness values by subtracting one, as “one” is the part of the
295 fitness in our model that corresponds to REPIN maintenance.
296 The duplication rate we inferred for the master sequence in
297 SBW25 is 9.8×10^{-9} per generation and per sequence.

298 However, this means that for the 3rd mutation class, we infer
299 negative duplication rates (Table 1). Unless there is an active
300 deletion process for these mutation classes, these duplication
301 rates are unlikely to be accurate. Alternatively, it is possible that
302 members of the 4th mutation class are more likely to replicate
303 than members of the 3rd mutation class. This could be true
304 as it is possible that these sequences are also recognized by a
305 second RAYT transposase in the SBW25 genome. To alleviate
306 this problem, we can simply scale up all mutation classes so
307 the lowest fitness is 1. This leads to a higher duplication rate of
308 the master sequence’s mutation class of 11.3×10^{-9} instead of
309 9.8×10^{-9} (Table 1).

310 If we assume one cell division to take 40 minutes and novel
311 REPIN insertions to be selectively neutral then it would take
312 about 6734 years until a novel REPIN master sequence fixes in
313 the SBW25 population. This seems to be a surprisingly long
314 time, but it would explain, why, to our knowledge, there is no
315 report of novel REPIN insertions within genomes. It may also
316 explain why REPINs can be maintained for long times within
317 a genome without being selected against because due to the
318 rarity of duplication events the negative fitness effects resulting
319 from transposition (e.g. transposition is likely to disrupt genes
320 because about 88% of the SBW25 genome are coding regions
321 (Silby *et al.* 2009)) are probably negligible.

322 **REPIN duplication rates in other bacteria.**

323 We also calculated duplication rates for four more *Pseudomonas*
324 strains and five more *E. coli* strains. The *E. coli* strains we chose
325 were quite distantly related to each other and belong to phy-
326 logroups A, B2 and D. The *Pseudomonas* strains we chose are very

Table 1 Inferred REPIN duplication rates in *P. fluorescens* SBW25.

Mutation class	Inferred Duplication Rate $\lambda_i (\times 10^{-9})^a$	Scaled Duplication Rate $\tilde{\lambda}_i (\times 10^{-9})^b$
0	9.8	11.3
1	6.5	8.1
2	5.5	7.1
3	-1.6	0
4	0	1.6

^a We identified a master sequence in the data and inferred the frequency of the different mutation classes. We use the equilibrium of our quasispecies model to calculate the associated fitness values f_i and setting f_4 to 1, where λ_i is $f_i - 1$.

^b The scaled duplication rate is: $\tilde{\lambda}_i = \frac{f_i}{\min(f_j)} - 1$.

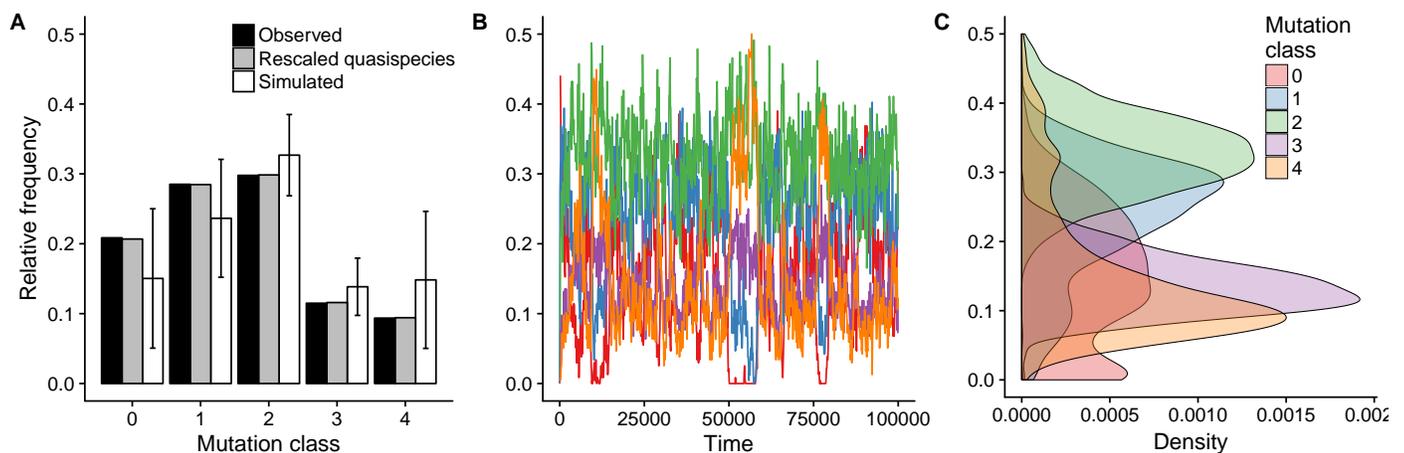


Figure 3 Inferred and observed steady state REPIN frequencies in *P. fluorescens* SBW25. (A) Shows the observed frequencies at a mutation rate of 8.9×10^{-11} . We rescaled time to allow us to do simulations at a mutation rate of 10^{-4} . The resulting quasispecies equilibria agree almost perfectly with the observed frequencies. A simulation of a single Wright-Fisher process (10^5 generations) with the same fitness values allows us to infer the variation of these frequencies. (B) Relative frequencies obtained from the Wright-Fisher process using the scaled fitness values for 10^5 generations. (C) Density plot of the relative frequencies of the mutation classes from the Wright-Fisher process.

327 distantly related to each other as well as to *E. coli* (Figure 4A). To
328 get an idea about how distantly related the individual strains
329 are, we gauge the time that has passed since the strains diverged
330 by measuring the 16S rDNA divergence (Ochman and Wilson
331 1987; Ochman *et al.* 1999). Ochman *et al.* estimated that it takes
332 about 50 million years for the 16S rDNA to diverge by 1%. Ac-
333 cording to these estimates, the most recent common ancestor
334 (MRCA) of the *E. coli* strains lived approximately 15 million
335 years ago (mya). The MRCA of the *Pseudomonas* strains lived
336 approximately 100 mya and *E. coli* and *Pseudomonas* diverged
337 about 600 mya. Hence, the REPIN populations in our selected
338 bacteria have been evolving independently of each other for a
339 very long time. RAYTs, the genes that mobilize REPINs in *E.*
340 *coli* and *Pseudomonas*, are also very different in *E. coli* and *Pseu-*
341 *domonas* and belong to two different gene classes (Bertels and
342 Rainey 2011b). There is no detectable sequence conservation in
343 the nucleotide sequence and very little sequence conservation
344 in the aminoacid sequence apart from the catalytic center of the
345 protein.

346 **Divergent bacteria have divergent REPIN populations**

347 The divergence between the different bacterial strains is also
348 reflected in the similarity between the most abundant 25bp long
349 sequences (REP sequences). The most common sequences in
350 *E. coli* are almost all identical, except for that of UTI89, where
351 the most common sequence is shifted by one nucleotide with
352 respect to the other *E. coli* sequences (File S2). But all *E. coli*
353 REP sequences are very different to all of the *Pseudomonas* REP
354 sequences. Among the *Pseudomonas* strains, the REP sequences
355 from *P. fluorescens* A506 and *P. fluorescens* BG33R are almost iden-
356 tical (again shifted by one nucleotide), which are also the most
357 closely related strains. Despite this similarity, the population
358 sizes and structures are completely different between the two
359 strains (see population networks in File S3). This observation
360 highlights the opportunity to study the evolution of entire pop-
361 ulations instead of single strains, which is basically impossible
362 for any other natural population.

363 REPIN populations in *E. coli* form relatively simple networks,
364 consistent with a single fitness peak. In contrast, REPIN popula-
365 tions from *Pseudomonas* form more complex networks, which is
366 more consistent with a rugged fitness landscape (see sequence
367 networks in File S3). The differences in the complexity of the se-
368 quence network may stem from the fact that there is only a single
369 RAYT gene in *E. coli*, but there are usually multiple RAYT genes
370 in *Pseudomonas*. If we assume that the activities of multiple RAYT
371 genes can interfere with each other, then generalist sequences
372 that can be moved by multiple RAYT genes will evolve, and give
373 rise to a complex sequence network.

374 Although the the divergence between *E. coli* and *Pseudomonas*
375 are very large and the differences between the structure of the
376 REPIN (File S4, whether the REPIN is symmetric as in *Pseu-*
377 *domonas* or not as in *E. coli*) and the corresponding transposase
378 are tremendous (Bertels and Rainey 2011b) the inferred REPIN
379 population sizes are surprisingly similar (Figure 4B). REPIN
380 populations in *E. coli* range between 165 (UMN026) and 242
381 (MG1655) members. REPIN populations in *Pseudomonas* are
382 spread more widely and range between 23 (DC3000) and 309
383 (A506) members. The population size has a strong effect on
384 whether the master sequence can persist within the population
385 or whether it will die out. Our simulations show that among all
386 *Pseudomonas* REPIN populations only that of *P. fluorescens* A506
387 and *P. fluorescens* SBW25 are large enough to persist over long

388 periods of time. In *E. coli*, in contrast, most populations persist
389 over 10^5 time steps (Figure 4C).

390 **Small REPIN populations in *Pseudomonas***

391 *P. syringae* DC3000 is different from the other *Pseudomonas* strains
392 not only the REPIN population is particularly small (only 23
393 members), which leads to a particularly unstable REPIN popula-
394 tion (Figure 4C). Another notable feature of the DC3000 REPIN
395 population is that a large part of the repetitive sequences does
396 not form REPINs (File S4). This suggests to us that the DC3000
397 REPIN population may be a dead or dying population, which
398 is slowly disintegrating due to genetic drift. This hypothesis
399 is further supported by the observation that the only RAYT in
400 DC3000 is not flanked by the most common 25bp long sequence
401 in the genome, which is the case for all other population we have
402 analyzed (File S2) and has been a defining feature of the REPIN-
403 RAYT system (Bertels and Rainey 2011b). Together, our data
404 suggests that the reason for the small and unstable REPIN pop-
405 ulation in DC3000 is that it is slowly disintegrating over time.
406 Hence the population is probably not in equilibrium, which
407 means that the inferred duplication rates may not be accurate.

408 The populations found in BG33R and GB1 are also too small
409 to persist for extended periods of time. However, in contrast
410 to DC3000, they are also the two populations with the highest
411 inferred duplication rate, and in both cases the most common
412 25bp long sequence does flank a RAYT gene and both popula-
413 tions consist mostly of REPINs (File S4). Hence there is no sign
414 of population disintegration. The inferred high duplication rates
415 are likely to evolve for small populations, because the mutation
416 load for small populations is comparatively small. This suggests
417 that these two populations may be growing.

418 **REPIN populations in competition**

419 The population network in BG33R is particularly interesting as
420 it contains two similar sized population (126 and 147 members)
421 and the REPIN master sequence consists in both cases of two
422 identical 25mers that occur both exactly 160 times in the genome
423 and differ in 5 nucleotide positions (i.e. the REPIN master se-
424 quence differs in 10 positions). When inferring the fitness of
425 the master sequence for both populations, then we also get very
426 similar and extremely high duplication rates of 32×10^{-9} and
427 37×10^{-9} . One would expect the evolution of high duplication
428 rates not only for growing populations but also for populations
429 that are competing for space in the genome. With space we are
430 referring to regions in the genome that are fitness neutral, i.e.
431 regions of the genome that incur no fitness cost when inserted
432 into.

433 **REPIN populations in *E. coli***

434 In *E. coli* the most abundant 25bp long sequences do not form
435 symmetric REPINs as observed in *Pseudomonas* (File S4). This
436 could lead us to the conclusion, as for DC3000, that *E. coli* does
437 not contain any REPIN populations that are alive. However
438 there are a few differences to DC3000. First of all, RAYTs in *E.*
439 *coli* are very distantly related to RAYTs in most *Pseudomonas*,
440 which leaves the possibility that REPINs in *E. coli* are structured
441 differently to REPINs in *Pseudomonas*. Second, there is not a
442 single instance of a REPIN in any of the five *E. coli* populations.
443 If *E. coli* REPIN populations were dying populations, then all
444 populations in *E. coli* were already dead. This either happened
445 about 15mya, when the last common ancestor of the five *E.*
446 *coli* strains lived or it happened recently simultaneously. If it

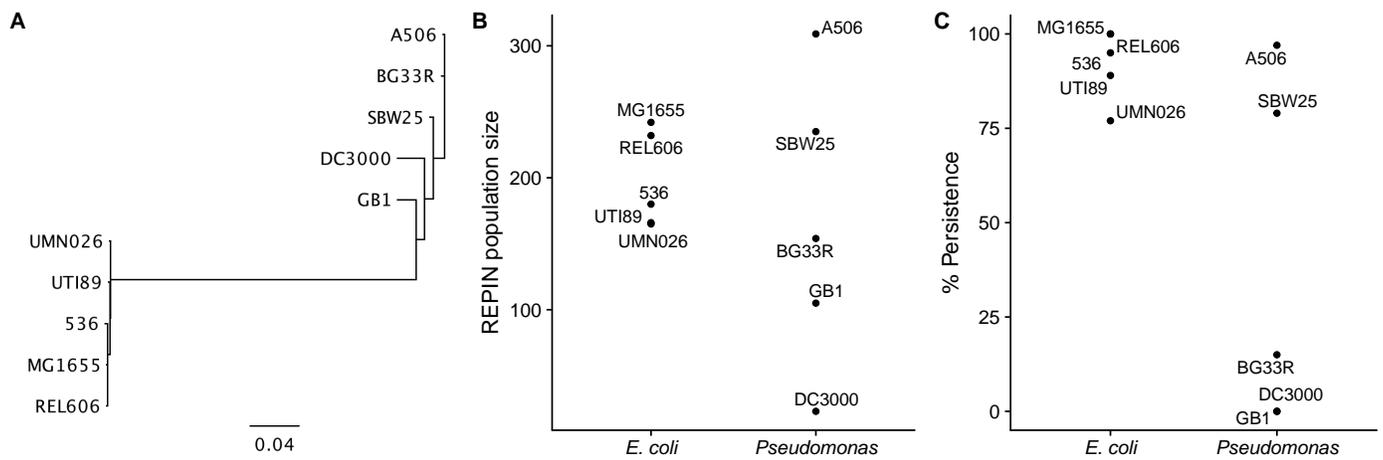
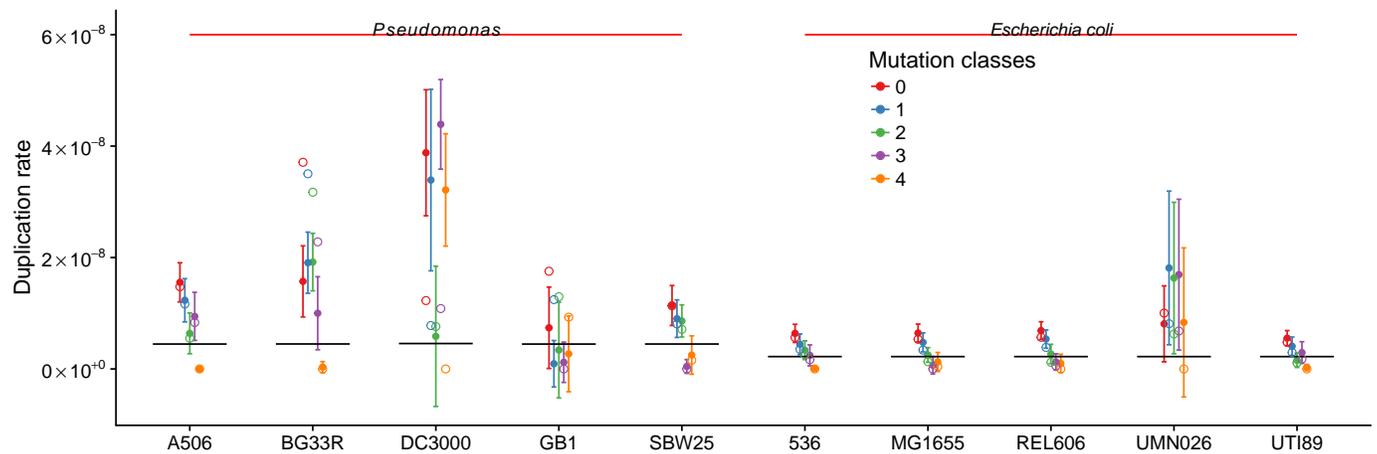


Figure 4 REPIN populations in other bacteria. (A) 16S tree showing the phylogenetic relationship between REPIN containing bacteria selected in our study. The scale bar shows the number of substitutions per nucleotide site. (B) REPIN population sizes in *E. coli* and *Pseudomonas*. (C) Proportion of 100 simulations where at least 10% of all sequences are maintained as master sequence at the end of the simulation.



447 happened 15mya, then we would expect the population to have
448 vanished by now and not consist of up to 242 members. It
449 also seems unlikely that it happened recently in all strains at
450 the same time and within the same time frame all the REPINs
451 vanished but the singlets remained. Finally, the most common
452 25bp long sequences in the five strains does still flank the RAYT
453 gene something that is not the case for DC3000 but for all other
454 REPIN populations in our study (File S2).

455 **REPIN duplication rate is close to the error threshold**

456 The duplication rates of the master sequences are in the range
457 of 5×10^{-9} and 37×10^{-9} and 5×10^{-9} and 15×10^{-9} when ex-
458 cluding unstable populations. Considering that the rates were in-
459 ferred for very different species and the species contain very dif-
460 ferent transposases that disperse the REPIN populations, these
461 values are very similar. This may be due to at least two reasons.
462 First, the duplication rate is very close to its lower possible limit,
463 because the number of mutations that occur on average between
464 two duplication events is between 0.12 and 0.39 for *Pseudomonas*
465 (0.29 and 0.39 without unstable populations) and between 0.22
466 and 0.46 for *E. coli* (0.39 and 0.46 without UMN026). If on aver-
467 age one mutation occurs between two duplication events, then it
468 is impossible to maintain a master sequence. For our model
469 a master sequence cannot be maintained above a frequency of
470 1% when the duplication rate of the master sequence and all
471 other sequences is equal or lower than 2.2×10^{-9} for *E. coli* and
472 4.4×10^{-9} for *Pseudomonas* (File S5 and Figure 5). Second, each
473 duplication event can be seen as a mutation that is introduced at
474 a random position in the genome. This means that an increase
475 in the duplication rate would also increase the mutational load
476 for the host organism. Hence, similar to selection for replica-
477 tion fidelity (Lynch *et al.* 2016), selection will favor organisms
478 with decreased REPIN duplication rates, but is limited by the
479 power of random genetic drift. The REPIN duplication rates we
480 inferred are probably the result of these two opposing forces.

481 **Maintenance of the REPIN-RAYT system**

482 The low duplication rate we inferred for all REPIN populations
483 also suggests that REPIN sequences have been part of bacterial
484 genomes for a very long time. This again raises the question of
485 how and why they are maintained. There are two explanations:
486 (1) the REPIN-RAYT system is frequently transmitted horizon-
487 tally or (2) they provide a benefit to the host organism (Bichsel
488 *et al.* 2013).

489 It is possible that the REPIN-RAYT system does get horizon-
490 tally transferred from time to time. However, horizontal trans-
491 fers are likely to be rare, because in order to establish a novel
492 REPIN population in a new host both the transposase (RAYT)
493 and the REPIN have to be transferred. This process is probably
494 facilitated by the fact that RAYTs are usually flanked by REPINs
495 (Bertels and Rainey 2011b). However, the rarity of these events
496 is consistent with the observation that the establishment of a
497 population that is as diverse as the REPIN population in SBW25
498 will take thousands of years. Hence it seems unlikely that hori-
499 zontal transfers are frequent enough to explain the ubiquitous
500 presence of the REPIN-RAYT system in bacteria.

501 Alternatively, the REPIN-RAYT system may be maintained
502 because it provides a selective advantage to the host bacterium.
503 For individual REP sequences there have been many studies on
504 potential benefits (Liang *et al.* 2015; Higgins *et al.* 1988; Espéli
505 *et al.* 2001). However, local benefits are unlikely to outweigh
506 the detrimental effects of transposition into genes or regulatory

507 regions let alone explain the maintenance of the REPIN-RAYT
508 system. It seems more likely that the REPIN-RAYT system pos-
509 sesses a function other than the dispersion of REPINs that is
510 beneficial for the host bacterium.

511 **Acknowledgements.** All authors acknowledge the generous
512 funding from the Max Planck Society.

513 **Literature Cited**

- 514 Aranda-Olmedo, I., R. Tobes, M. Manzanera, J. L. Ramos, and
515 S. Marqués, 2002 Species-specific repetitive extragenic palin-
516 dromic (REP) sequences in *Pseudomonas putida*. *Nucleic*
517 *acids research* **30**: 1826–1833.
- 518 Bertels, F. and P. B. Rainey, 2011a Curiosities of REPINs and
519 RAYTs. *Mobile genetic elements* **1**: 262–268.
- 520 Bertels, F. and P. B. Rainey, 2011b Within-genome evolution of
521 REPINs: a new family of miniature mobile DNA in bacteria.
522 *PLoS Genetics* **7**: e1002132.
- 523 Bichsel, M., A. D. Barbour, and A. Wagner, 2013 Estimating the
524 fitness effect of an insertion sequence. *Journal of mathematical*
525 *biology* **66**: 95–114.
- 526 Bull, J. J., L. A. Meyers, and M. Lachmann, 2005 Quasispecies
527 made simple. *PLoS Computational Biology* **1**: 450–460.
- 528 Bureau, T. E. and S. R. Wessler, 1994 Stowaway: a new family
529 of inverted repeat elements associated with the genes of both
530 monocotyledonous and dicotyledonous plants. *The Plant cell*
531 **6**: 907–916.
- 532 Domingo, E. and P. Schuster, 2016 *Quasispecies: From Theory to*
533 *Experimental Systems*, volume 392. Springer.
- 534 Eigen, M., 1971 Selforganization of matter and the evolution
535 of biological macromolecules. *Die Naturwissenschaften* **58**:
536 465–523.
- 537 Eigen, M. and P. Schuster, 1977 The hypercycle. a principle of
538 natural self-organization. part a: Emergence of the hypercycle.
539 *Die Naturwissenschaften* **64**: 541–565.
- 540 Espéli, O., L. Moulin, and F. Boccard, 2001 Transcription atten-
541 uation associated with bacterial repetitive extragenic BIME
542 elements. *Journal of Molecular Biology* **314**: 375–386.
- 543 Ewens, W. J., 1979 *Mathematical Population Genetics*. Springer,
544 Berlin.
- 545 Higgins, C. F., G. F.-L. Ames, W. M. Barnes, J. M. Clement, and
546 M. Hofnung, 1982 A novel intercistronic regulatory element
547 of prokaryotic operons. *Nature* **298**: 760–762.
- 548 Higgins, C. F., R. S. McLaren, and S. F. Newbury, 1988 Repetitive
549 extragenic palindromic sequences, mRNA stability and gene
550 expression: evolution by gene conversion? — a review. *Gene*
551 **72**: 3–14.
- 552 Jurka, J., V. V. Kapitonov, O. Kohany, and M. V. Jurka, 2007 Repet-
553 itive sequences in complex genomes: structure and evolution.
554 *Annual review of genomics and human genetics* **8**: 241–259.
- 555 Liang, W., K. E. Rudd, and M. P. Deutscher, 2015 A Role for
556 REP Sequences in Regulating Translation. *Molecular cell* **58**:
557 431–439.
- 558 Lynch, M., M. S. Ackerman, J.-F. Gout, H. Long, W. Sung, W. K.
559 Thomas, and P. L. Foster, 2016 Genetic drift, selection and the
560 evolution of the mutation rate. *Nature Reviews Genetics* **17**:
561 704–714.
- 562 Mahillon, J. and M. Chandler, 1998 Insertion sequences. *Microbi-*
563 *ology and Molecular Biology Reviews* **62**: 725–774.
- 564 Nowak, M. A., 1992 What is a quasispecies? *Trends in Ecology*
565 *and Evolution* **7**: 118–121.
- 566 Nunvar, J., T. Huckova, and I. Licha, 2010 Identification and
567 characterization of repetitive extragenic palindromes (REP)-

- 568 associated tyrosine transposases: implications for REP evolu-
569 tion and dynamics in bacterial genomes. *BMC Genomics* **11**:
570 44.
- 571 Ochman, H., S. Elwyn, and N. A. Moran, 1999 Calibrating bac-
572 terial evolution. *Proceedings of the National Academy of Sci-*
573 *ences* **96**: 12638–12643.
- 574 Ochman, H. and A. C. Wilson, 1987 Evolution in bacteria: ev-
575 idence for a universal substitution rate in cellular genomes.
576 *Journal of Molecular Evolution* **26**: 74–86.
- 577 Oggioni, M. R. and J.-P. Claverys, 1999 Repeated extragenic se-
578 quences in prokaryotic genomes: a proposal for the origin and
579 dynamics of the RUP element in *Streptococcus pneumoniae*.
580 *Microbiology* **145**: 2647–2653.
- 581 Page, K. M. and M. A. Nowak, 2002 Unifying evolutionary dy-
582 namics. *Journal of Theoretical Biology* **219**: 93–98.
- 583 Seifert, D., F. Di Giallonardo, K. J. Metzner, H. F. Günthard, and
584 N. Beerenwinkel, 2015 A framework for inferring fitness land-
585 scapes of patient-derived viruses using quasispecies theory.
586 *Genetics* **199**: 191–203.
- 587 Silby, M. W., A. M. Cerdeño-Tárraga, G. S. Vernikos, S. R. Gid-
588 dens, R. W. Jackson, G. M. Preston, X.-X. Zhang, C. D. Moon,
589 S. M. Gehrig, S. A. Godfrey, C. G. Knight, J. G. Malone,
590 Z. Robinson, A. J. Spiers, S. Harris, G. L. Challis, A. M. Yaxley,
591 D. Harris, K. Seeger, L. Murphy, S. Rutter, R. Squares, M. A.
592 Quail, E. Saunders, K. Mavromatis, T. S. Brettin, S. D. Bentley,
593 J. Hothersall, E. Stephens, C. M. Thomas, J. Parkhill, S. B. Levy,
594 P. B. Rainey, and N. R. Thomson, 2009 Genomic and genetic
595 analyses of diversity and plant interactions of *Pseudomonas*
596 *fluorescens*. *Genome biology* **10**: R51.
- 597 Stern, M. J., G. F.-L. Ames, N. H. Smith, E. C. Robinson, and C. F.
598 Higgins, 1984 Repetitive extragenic palindromic sequences: A
599 major component of the bacterial genome. *Cell* **37**: 1015–1026.
- 600 Ton-Hoang, B., P. Siguier, Y. Quentin, S. Onillon, B. Marty,
601 G. Fichant, and M. Chandler, 2012 Structuring the bacte-
602 rial genome: Y1-transposases associated with REP-BIME se-
603 quences. *Nucleic acids research* **40**: 3596–3609.
- 604 Touchon, M., C. Hoede, O. Tenaillon, V. Barbe, S. Baeriswyl,
605 P. Bidet, E. Bingen, S. Bonacorsi, C. Bouchier, O. Bouvet, A. Cal-
606 teau, H. Chiapello, O. Clermont, S. Cruveiller, A. Danchin,
607 M. Diard, C. Dossat, M. El Karoui, E. Frapy, L. Garry, J. M.
608 Ghigo, A. M. Gilles, J. Johnson, C. Le Bouguenec, M. Lescat,
609 S. Mangenot, V. Martinez-Jehanne, I. Matic, X. Nassif, S. Oztas,
610 M. A. Petit, C. Pichon, Z. Rouy, C. Saint Ruf, D. Schneider,
611 J. Tourret, B. Vacherie, D. Vallenet, C. Medigue, Rocha, Ed-
612 uardo P. C., and E. Denamur, 2009 Organised Genome Dynam-
613 ics in the *Escherichia coli* Species Results in Highly Diverse
614 Adaptive Paths. *PLoS Genetics* **5**.
- 615 van Dongen, S., 2000 A Cluster algorithm for graphs. *Report -*
616 *Information systems* pp. 1–40.
- 617 Versalovic, J., T. Koeuth, and J. R. Lupski, 1991 Distribution of
618 repetitive DNA sequences in eubacteria and application to
619 fingerprinting of bacterial genomes. *Nucleic acids research* **19**:
620 6823–6831.
- 621 Wessler, S. R., T. E. Bureau, and S. E. White, 1995 LTR-
622 retrotransposons and MITEs: important players in the evolu-
623 tion of plant genomes. *Current opinion in genetics & develop-*
624 *ment* **5**: 814–821.
- 625 Wielgoss, S., J. E. Barrick, O. Tenaillon, S. Cruveiller, B. Chane-
626 Woon-Ming, C. Medigue, R. E. Lenski, and D. Schneider, 2011
627 Mutation Rate Inferred From Synonymous Substitutions in a
628 Long-Term Evolution Experiment With *Escherichia coli*. *G3:*
629 *Genes, Genomes, Genetics* **1**: 183–186.