

The coalescent for prokaryotes with homologous recombination from external source

Tetsuya Akita^{*,†}, Shohei Takuno^{*}, and Hideki Innan^{*,1}

^{*}SOKENDAI (The Graduate University for Advanced Studies), Hayama, Kanagawa 240-0193, Japan

[†]National Research Institute of Far Seas Fisheries, Fisheries Research Agency, Yokohama, Kanagawa 236-8648, 56 Japan

Manuscript intended for *Genetics*, June 12, 2017

ABSTRACT The coalescent process for prokaryote species is theoretically considered. Prokaryotes undergo homologous recombination not only with other individuals within the same species (intra-specific recombination) but also with other species (inter-specific recombination). This work particularly focuses the latter because the former has been well incorporated in the framework of the coalescent. We here developed a simulation framework for generating patterns of SNPs (single nucleotide polymorphisms) allowing integration of external DNA out of the focal species, and a simulator named **msPro** was developed. We found that the joint work of intra- and inter-specific recombination creates a complex pattern of SNPs. The direct effect of inter-specific recombination is to increase the amount of polymorphism. Because inter-specific recombination is very rare in general, it creates a regions with an exceptionally high level of polymorphisms. Following an inter-specific recombination event, intra-specific recombination chop the integrated foreign DNA into small pieces, making a complicated pattern of SNPs that looks as if foreign DNAs were integrated multiple times. This work with the **msPro** simulator would be useful to understand and evaluate the relative contribution of intra- and inter specific recombination to creating complicated patterns of SNPs in prokaryotes.

1 The coalescent is a population genetic theory, which con- 23
2 sideres the evolutionary process backward in time (Kingman 24
3 1982; Hudson 1983b; Tajima 1983). The coalescent theory 25
4 has been mainly developed by assuming its application to 26
5 higher eukaryotes, perhaps due to a historical reasons: The 27
6 major model species of population genetics have been higher 28
7 eukaryotes such as *Drosophila* and human (*e.g.*, Hartl and 29
8 Clark 2007). The coalescent provides an extremely powerful 30
9 simulation tool for analyzing the pattern of single nucleotide 31
10 polymorphisms (SNPs) in sampled sequences. It is flexible 32
11 enough to incorporate major evolutionary processes includ- 33
12 ing random genetic drift, mutation, recombination and demo- 34
13 graphic history (*e.g.*, Hudson 1990; Nordborg 2001; Wakeley 35
14 2008), whereas it is not very straightforward to incorporate 36
15 complex modes of selection (but see Krone and Neuhauser 37
16 1997; Neuhauser and Krone 1997; Donnelly and Kurtz 1999; 38
17 Fearnhead 2006). **ms** is one of the most popular coalescent 39
18 simulators, which allows to produce patterns of neutral SNPs 40
19 under various settings of demography (Hudson 2002). It in- 41
20 corporates two major outcomes of meiotic recombination, 42
21 that is, meiotic crossing-over and gene conversion. 43
22 Prokaryotes are unique in that they are haploids and do not 44

undergo meiosis, and therefore their recombination mecha-
nisms are quite different from that of meiotic recombination
in eukaryotes. Nevertheless, the coalescent can work with
prokaryotes with a relatively simple modification: Recombi-
nation is treated as an event analogous to meiotic gene con-
version because a prokaryote's circular chromosome needs
double "crossing-over" to exchange a DNA fragment. This
modification can well explain the nature of prokaryotes' ho-
mologous recombination as we will explain below. The ap-
plication of the coalescent theory to bacteria became partic-
ularly popular since McVean *et al.* developed the software
LDhat (McVean *et al.* 2002) for estimating the recombina-
tion rate, which is a modified version of Hudson's compos-
ite likelihood method (Hudson 2001). Because **LDhat** al-
lows recurrent mutations at a single site, it is more suitable
to species with a large population size like bacteria. **LDhat**
has been applied to the multilocus sequence typing (MLST)
data (*e.g.*, Jolley *et al.* 2005; Pérez-Losada *et al.* 2006; Wirth
et al. 2006) and genome-wide SNP data from various species
(*e.g.*, Touchon *et al.* 2009; Donati *et al.* 2010; Haven *et al.*
2011), demonstrating a great variation in the recombination
rate across species. Hudson's **ms** software has also been
successfully used (*e.g.*, Pepperell *et al.* 2010; Thomas *et al.*
2012; Zhang *et al.* 2012; Takuno *et al.* 2012; Cornejo *et al.*
2013; Nell *et al.* 2013; Krause *et al.* 2014; Shapiro 2014;

¹Corresponding author: SOKENDAI (The Graduate University for Advanced Studies), Hayama, Kanagawa 240-0193, Japan. E-mail: hideki.innan@soken.ac.jp

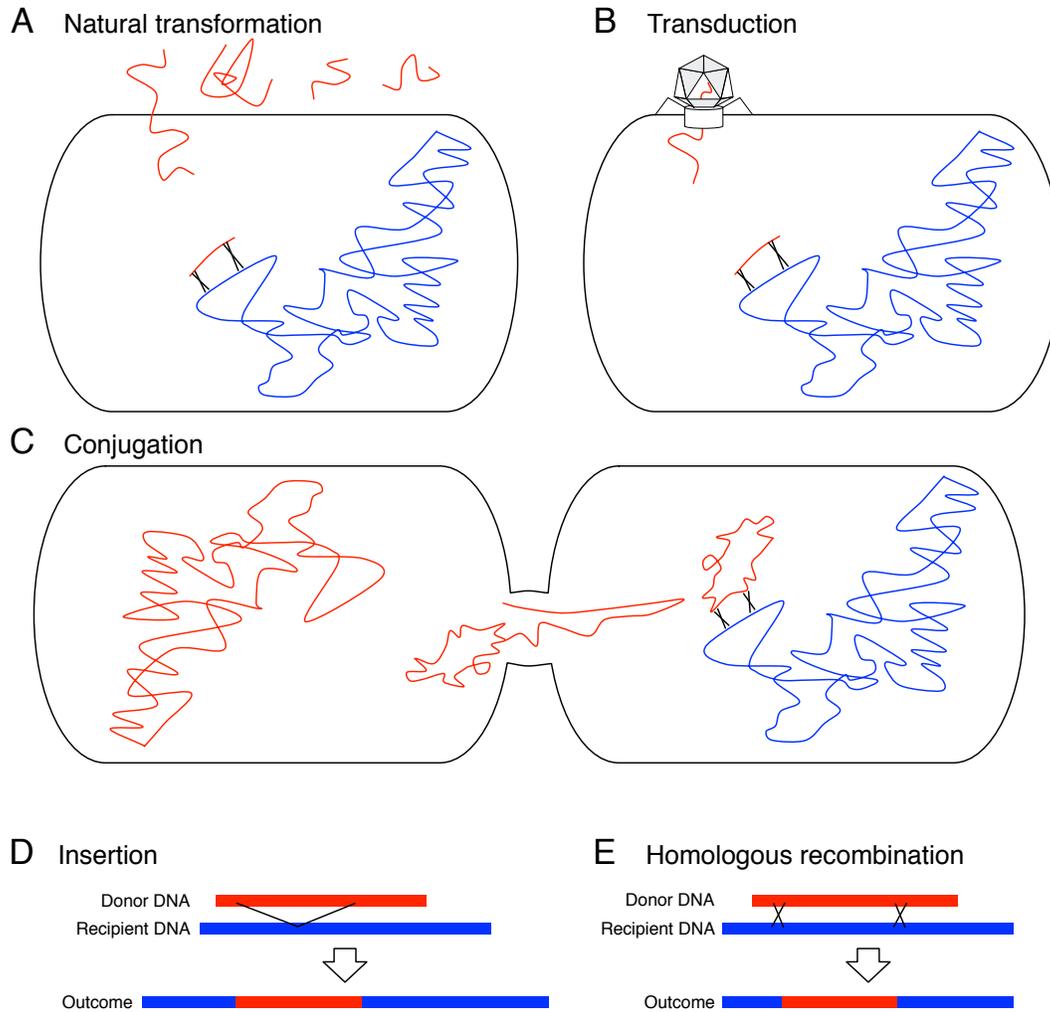


Figure 1 Three major mechanisms for prokaryotes to integrate external DNA: natural transformation (A), transduction (B), and conjugation (C). The host genome and external DNA are presented in blue and red, respectively. (D, E) Two outcomes of recombination via insertion (D) and homologous recombination (E).

48 Rosen *et al.* 2015).

49 Thus, despite the large difference in the recombination
50 mechanism between eukaryotes and prokaryotes, it is
51 technically not very difficult to handle prokaryotes' homo-
52 logous recombination in the coalescent framework. However,
53 this holds only when recombination occurs within a single
54 species. This assumption should hold quite strictly in eukary-
55 otes, but not in prokaryotes for which the concept of species
56 is not as strict as eukaryotes (*e.g.*, Cohan 2002b; Doolittle and
57 Papke 2006; Achtman and Wagner 2008) because of frequent
58 exchanges of DNA between different species due to the nature
59 of their recombination mechanism, as is described in the
60 following.

61 Prokaryotes undergo recombination by incorporating DNA
62 outside of the cell through three major mechanisms: natural
63 transformation, transduction, and conjugation, as illustrated
64 in Figure 1 (*e.g.*, Snyder *et al.* 2013). Natural transforma-

65 tion is a process involving direct uptake of a free extracellu-
66 lar DNA and the integration under natural bacterial growth
67 conditions (Figure 1A). Transduction is a process in which
68 bacterial DNA is introduced into the other bacteria through
69 infection by a phage containing the DNA (Figure 1B). Con-
70 jugation is the transfer of DNA from one bacterial cell to an-
71 other by the transfer functions of a self-transmissible DNA
72 elements, frequently associated with plasmids (Figure 1C).

73 It is known that such incorporated DNA from outside of the
74 cell is usually harmful when integrated into the host genome,
75 so that there are a number of mechanisms to avoid integra-
76 tion (*e.g.*, Lorenz and Wackernagel 1994; Majewski 2001;
77 Cohan 2002a; Chen and Dubnau 2004; Thomas and Nielsen
78 2005; Marraffini and Sontheimer 2010; Vasu and Nagaraja
79 2013). Because DNAs from different species should be much
80 more harmful than those from the same species, most mech-
81 anisms involve some kind of self-recognition systems, in

82 which markers are distributed through the genome to dis- 136
83 tinguish from those originating external source (*i.e.*, differ- 137
84 ent species). In some bacterial species, such as *Neisseria* 138
85 *gonorrhoeae* and *Haemophilus influenzae*, efficient natural 139
86 transformation requires the presence of short sequence motifs 140
87 (~ 10 bp), called as DNA uptake sequences (DUS) or up- 141
88 take signal sequences (USS), that is interspersed among the 142
89 genome, which may prevent the incoming DNA from differ- 143
90 ent species (or strains) integrating into their genomes. Phage 144
91 defense mechanisms may also work against incoming DNA 145
92 from external source. The restriction-modification system is 146
93 a common mechanism or degrading DNA that is not prop- 147
94 erly modified (*e.g.*, through DNA methylation), which have 148
95 been identified in $\sim 90\%$ of prokaryote species (Roberts *et al.* 149
96 2010). Clustered, regularly interspaced short palindromic re- 150
97 peat (CRISPR) loci and their associated proteins (Cas) are 151
98 found in the genomes of $\sim 90\%$ of archaea and $\sim 50\%$ of eu- 152
99 bacteria (Grissa *et al.* 2007; Rousseau *et al.* 2009). These 153
100 sequences are separated by short sequences of DNA (23–50 154
101 bp) known as spacers, most of which exhibit homology to 155
102 previously encountered phage or plasmid genomes, suggest- 156
103 ing that these loci provide memory for the bacteria to prevent 157
104 repeated incoming encounters. 158

105 In addition, when these mechanisms do not work perfectly, 159
106 there is another round of screening process to prevent inte- 160
107 gration of external DNA through homologous recombination 161
108 (Majewski 2001). For example, recombination requires near 162
109 identical regions (*e.g.*, monitored by RecA mediated homol- 163
110 ogy search), (Shen and Huang 1986; Majewski and Cohan 164
111 1998) so that external DNA has less chance to be integrated. 165
112 In addition, it is also pointed out that the mismatch repair sys- 166
113 tem is effective in preventing recombination between highly 167
114 mismatched sequences (Claverys and Lacks 1986; Majewski 168
115 2001; Overballe-Petersen *et al.* 2013). Thus, there are a num- 169
116 ber of molecular mechanisms to prevent integrating external 170
117 DNA to the host genome. Nevertheless, it has been repeat- 171
118 edly demonstrated that prokaryote genomes undergo recom- 172
119 bination, not only within the same species but also with dif- 173
120 ferent species (reviewed in Majewski 2001). 174

121 There are two possible outcomes of recombination as 175
122 shown in Figures 1D and E (see Lawrence 2013, for a re- 176
123 view). One is that the incorporated DNA is inserted into the 177
124 genome (Fig. 1D), and the other is that the incorporate DNA 178
125 is exchanged with its homologous part of the genome if any 179
126 (Fig. 1E). The former is known as horizontal gene transfer or 180
127 lateral gene transfer, and its evolutionary role is emphasized 181
128 when a novel gene is acquired and contributes to adaptation 182
129 (Ochman *et al.* 2000; Dobrindt *et al.* 2004; Fraser *et al.* 2009; 183
130 Polz *et al.* 2013), although the frequency and importance 184
131 of such illegitimate recombination is under debate (de Vries 185
132 *et al.* 2001; Shapiro *et al.* 2012). The latter is known as ho- 186
133 mologous recombination, and it usually involves DNAs from 187
134 the same species because the near-identity requirement of the 188
135 RecA mediated homology search criteria is easily satisfied, 189

136 whereas it is also possible that DNA from different species 137
138 is integrated as long as it retains some homology. Homolo- 139
140 gous recombination between different species sometimes re- 141
142 mains unique patterns of SNPs, from which we can search for 143
144 their footprints in the sequence data (reviewed in Awadalla 145
2003; Didelot and Maiden 2010; Azad and Lawrence 2012; 146
Nakhleh 2013). 147

148 The focus of this article is the latter, homologous recom- 149
150 bination. Considering the mechanism of homologous recom- 151
152 bination involving double crossing-over, the outcome 153
154 is similar to meiotic gene conversion. Therefore, as men- 155
156 tioned earlier, the standard coalescent has been commonly 157
158 applied for analyzing patterns of SNPs in bacteria with a sim- 159
160 ple modification in the setting; the rate of crossing-over is 160
161 set to zero, so that all recombination events (*i.e.*, homolo- 161
162 gous recombination) are treated as if they are meiotic gene 162
163 conversion. This application should be reasonable as long 163
164 as the donor of homologous recombination is always an- 164
165 other individual in the same species. However, it is well 165
166 known that homologous recombination occasionally involves 166
167 DNA from other species, and this is the case that the stan- 167
168 dard coalescent cannot handle. The purpose of this work is 168
169 to develop the theoretical framework of the coalescent for 169
170 prokaryotes, which allows homologous recombination both 170
171 within and between species. We also developed a simu- 171
172 lation software named **msPro**, which will be available at 172
173 <http://www.sendou.soken.ac.jp/esb/innan/InnanLab/>. 173

163 Theoretical Framework

164 **Overview:** Consider a sample of prokaryote DNA sequences 164
165 with length L bp from n haploids, and trace their ancestral 165
166 lineages backward in time. Figure 2A shows an example of 166
167 an ancestral recombination graph under the standard coales- 167
168 cent, in which all recombination is assumed to be homolo- 168
169 gous recombination within the same species (Hudson 1983a; 169
170 Griffiths and Marjoram 1996). Under this setting, the process 170
171 is analogous to meiotic gene conversion in the standard co- 171
172 alescent for diploid eukaryotes (McVean *et al.* 2002; Awadalla 172
2003). A coalescent event merges the ancestral lineages (*e.g.*, 173
174 events 3, 4, 5, and 6 in Fig. 2A), and homologous recom- 174
175 bination separates the lineage into two (*e.g.*, event 1, and 2 175
176 in Fig. 2A). For example, event 1 in Fig. 2A is a homolo- 176
177 gous recombination, in which a short fragment (presented by 177
178 a gray box) is integrated into the recipient genome, so that 178
179 the ancestral lineage is separated into two; one for the recip- 179
180 ient genome and the other is for the integrated fragment. Then, 180
181 following the standard treatment, we further trace their an- 181
182 cestral lineages until the lineages of all sampled chromosome 182
183 merge to their MRCA (Most Recent Common Ancestor), 183
184 which is referred to as $MRCA_{all}$ in this article. It should be 184
185 noted that with the presence of recombination, different parts 185
186 of the region have different histories, so that $MRCA_{all}$ cannot 186
187 be identical across the region; different subregions chopped 187

188 by recombination should have their specific $MRCA_{all}$. For 242
189 example, $MRCA_{all}$ for the black region appears at time T_6 , 243
190 while that for the gray region is at time T_5 and for the other 244
191 white regions are at T_4 (Fig. 2A). The ancestral recombina- 245
192 tion graph has all historical information for the entire regions 246
193 as illustrated in Fig. 2A. With this ancestral recombination 247
194 graph, a pattern of SNPs can be simulated by randomly dis- 248
195 tributing point mutations on the graph. Thus, the standard 249
196 coalescent treatment works for prokaryotes with homologous 250
197 recombination within species (McVean *et al.* 2002; Awadalla 251
198 2003). 252

199 The problem is when DNA from other species is integrated 253
200 by homologous recombination. Figure 2B illustrates such a 254
201 situation, in which event 2 is assumed to be a homologous 255
202 recombination event from external source (*i.e.*, integration of 256
203 DNA from other species), which is presented in a red box. In 257
204 this case, the ancestral lineage of the transferred DNA orig- 258
205 inates from outside of the focal species, so that it is not in- 259
206 volved in the coalescent process of the focal species before 260
207 time T_2 . This is the situation that the standard coalescent can- 261
208 not handle. We here propose a simple solution to this prob- 262
209 lem: Tracing the ancestral lineage of external source should 263
210 be terminated, and the coalescent process should be contin- 264
211 ued without considering such terminated lineages. Under this 265
212 treatment, the concept of the MRCA of all sampled sequences 266
213 ($MRCA_{all}$) does not apply to such a region that experienced 267
214 homologous recombination from external source. The direct 268
215 donor of the external DNA is called $MRCA_{ext}$, most recent 269
216 common ancestor from external source, and MRCA of the 270
217 rest is referred to as $MRCA_{int}$, most recent common ancestor 271
218 of internal lineages. In event 2 in Figure 2B, while the gray 272
219 and white regions that are not involved in the integration of 273
220 external DNA can be traced back to $MRCA_{all}$ (at T_5 and T_4 , 274
221 respectively), we may stop tracing the ancestral lineage of the 275
222 red region at $> T_2$, and the origin of this region is treated as 276
223 a $MRCA_{ext}$. Thus, when a region experienced a homologous 277
224 recombination from external source, the sampled sequences 278
225 have two kinds of origins, one is $MRCA_{ext}$ at T_2 as the origin 279
226 of the red part (shown by a red box with a star in Figure 2B) 280
227 and the other is $MRCA_{int}$ at T_3 as the origin of the rest, shown 281
228 by a black box with a yellow star in Figure 2B. 282

229 We here consider how to simulate a pattern of SNPs in 283
230 such a region that experienced homologous recombination with 284
231 external source. Note that provided the mechanism of homolo- 285
232 gous recombination, we assume a reasonable level of 286
233 sequence identity between the external DNA and the focal 287
234 species. This means that the external lineage should eventu- 288
235 ally coalesce with the common ancestor of the focal species 289
236 (on the time scale of species-divergence). However, it is very 290
237 difficult to know the probability distribution of the time to 291
238 such eventual common ancestor, which could be far older 292
239 than the MRCA of the focal species. 293

240 Alternatively, we develop an ad-hoc treatment that does 284
241 not require any unknown ancient demographic history up to 285

species divergence. The idea of our treatment is based on 242
a number of empirical demonstrations that the rate of suc- 243
cessful integration of external DNA heavily depends on the 244
nucleotide divergence between the transferred fragment and 245
the recipient sequence; the rate decays almost exponentially 246
with increasing divergence as demonstrated by many authors 247
(Albritton *et al.* 1984; Roberts and Cohan 1993; Vulić *et al.* 248
1997; Zahrt and Maloy 1997; Lorenz and Sikorski 2000; Ma- 249
jowski *et al.* 2000). See below for details. 250

**Homologous recombination within species (intra-specific 253
recombination):** It is relatively straightforward to incorpo- 254
rate homologous recombination within species as mentioned 255
above. Following previous studies (Wiuf and Hein 2000; 256
McVean *et al.* 2002), we assume that a homologous recom- 257
bination event is initiated at any position at rate g per site per 258
generation. Then, it is assumed that the elongation process 259
proceeds such that the length of transferred tract, z , follows a 260
geometric distribution, with mean tract length $= \lambda$ bp:

$$Q_{int}(z) = q(1 - q)^{z-1}, \quad (1)$$

260 where $q = 1/\lambda$. This assumption is supported by empiri- 261
cal studies on transformation of many species including *He-* 262
licobacter pylori (Lin *et al.* 2009), *Streptococcus pneumoniae* 263
(Croucher *et al.* 2012), and *Haemophilus influenzae* (Mell 264
et al. 2014). For mathematical convenience, we assume uni- 265
directional elongation of conversion tract from 5' to 3', which 266
has no quantitative effect on the pattern of SNPs. Given 267
Equation 1, the rate that a region of L bp undergoes homolo- 268
gous recombination per generation is given by

$$g' = R_{in} + R_{left}, \quad (2)$$

269 where R_{in} is the rate of gene conversion initiating inside the 270
region and R_{left} is the rate outside the region but ending 271
within the observed sequence. R_{in} and R_{left} are given by

$$\begin{aligned} R_{in} &= gL, \\ R_{left} &= \sum_{i=1}^L gQ_{int}(z \geq i) = \sum_{i=1}^L g(1 - q)^i, \end{aligned} \quad (3)$$

272 Assuming all recombination is neutral, this rate (g') is iden- 273
tical to the backward recombination rate, which can be di- 274
rectly incorporated into the coalescent framework. The back- 275
ward recombination rate per generation is defined as the rate 276
at which a lineage undergo recombination when a lineage is 277
traced back for a single generation. 278

279 It is interesting to note that Equation 2 does not include the 280
probability that a recombination tract cover the entire simu- 281
lated region. This is because such recombination simply 282
causes a shift of a lineage to another lineage within the same 283
population, which does not essentially affect the coalescent 284
process. However, this does affect the process if the recombi- 285
nation event occurs with different species as we will explain 286
in the next section (see below). 287

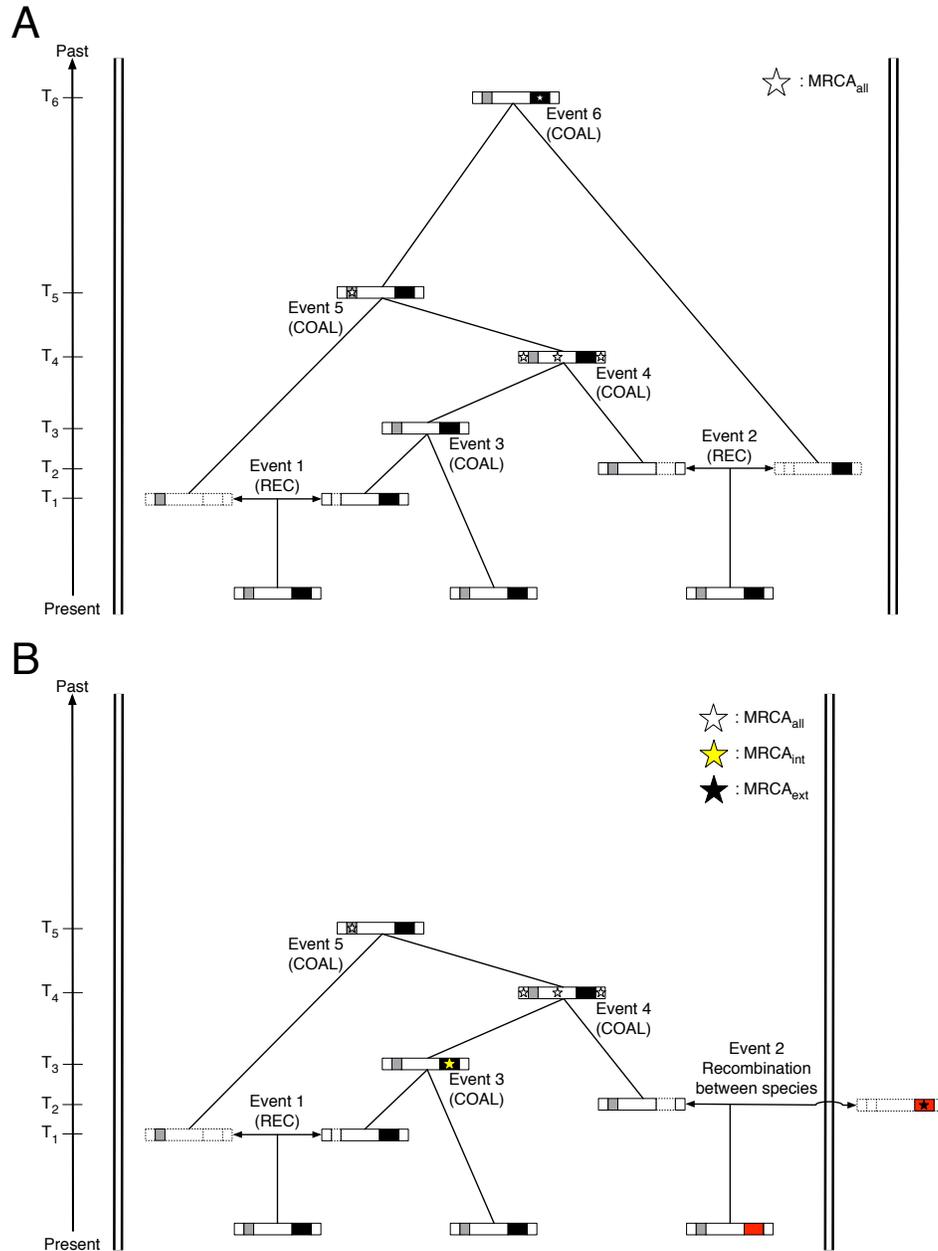


Figure 2 (A) Ancestral recombination graph with recombination within species alone. An example with sample size $n = 3$ is illustrated. The sampled three genomes are shown by long boxes, where regions with different histories are presented in different colors. The ancestral lineage is split into two by a recombination event (REC), while a pair of ancestral lineages merges by a coalescent event (COAL). The boxes with dashed lines represent dummy regions whose descendants do not show up in the sample. The two short regions in gray and black are transferred fragments by gene conversion-like recombination events. MRCA_{all} for each region is shown by a star. The white part has MRCA at T_4 , the gray part has MRCA at T_5 and the black part has MRCA at T_6 . (B) Ancestral recombination graph with recombination within species and from external source. The region transferred from external source is shown in red.

286 It should be noted that there are three mechanisms for a cell
 287 to incorporate DNA, transformation, transduction and conju-
 288 gation, through which homologous recombination can occur.
 289 They occur at different rates and typical lengths of integrated
 290 tracts should be different. Quite short fragments are usually

291 integrated through transformation, while relatively large frag-
 292 ments may be involved in recombination through transduc-
 293 tion and conjugation (Cohan 2002a). Therefore, it is biologi-
 294 cally reasonable to model these processes separately, and this
 295 is what was done in earlier studies (see Maynard Smith 1994;

296 Hudson 1994).

297 However, in some studies (particularly in coalescent-based
298 studies), all three recombination processes are not specified
299 (e.g., Falush *et al.* 2001; McVean *et al.* 2002; Awadalla 2003;
300 Fearnhead *et al.* 2005). This should be partly because the
301 three mechanisms are commonly summarized by a single
302 backward recombination rate and the tract length is simply
303 assumed to follow a geometric distribution (although not
304 specifically described in these literatures to the best of our
305 knowledge). This may work perhaps because the possible
306 outcome of the three recombination mechanisms are similar
307 in that they can be described as a double-recombination event
308 (Wiuf 2001) even when the typical tract lengths and rates
309 are different (Maynard Smith 1994; Hudson 1994), but we
310 should remember that this is a conventional approximation.

311 More strictly, if we consider the three mechanisms sep-
312 arately, denote by g_{tf} , g_{td} , and g_{cj} , respectively, the initia-
313 tion rates of transformation-, transduction- and conjugation-
314 oriented recombination per site per generation, and for each,
315 let us assume that the tract length follows a geometric dis-
316 tribution (the mean lengths are λ_{tf} , λ_{td} , and λ_{cj} for the
317 three mechanisms). Then, the total initiation rate per site
318 is $g_{\text{total}} = g_{\text{tf}} + g_{\text{td}} + g_{\text{cj}}$, but the density distribution of
319 tract length is not a simple geometric distribution with a single
320 parameter, rather given by an average of three geometric
321 distributions:

$$Q_{\text{int}}(z) = \frac{1}{g_{\text{total}}} \times \left[g_{\text{tf}} q_{\text{tr}} (1 - q_{\text{tr}})^{z-1} + g_{\text{td}} q_{\text{td}} (1 - q_{\text{td}})^{z-1} + g_{\text{cj}} q_{\text{cj}} (1 - q_{\text{cj}})^{z-1} \right], \quad (4)$$

322 where $q_{\text{tr}} = 1/\lambda_{\text{tr}}$, $q_{\text{td}} = 1/\lambda_{\text{td}}$, and $q_{\text{cj}} = 1/\lambda_{\text{cj}}$. Thus,
323 strictly speaking, there should be situations where the ad-
324 hoc treatment using a single geometric distribution may not
325 hold. Nevertheless, the simplified treatment may work fairly
326 well if we assume that one of the three mechanisms dom-
327 inates the other two. For example, it is well known that
328 many of *Bacillus* species show a very high transformation
329 rate (especially in laboratory strains of *B. subtilis*, Earl *et al.*
330 2008), whereas some species are not naturally transformable
331 (e.g., *Escherichia coli*, *Salmonella typhimurium*; Lorenz and
332 Wackernagel 1994) and conjugation and/or transduction may
333 be should be the major cause of recombination.

334 Thus, although it is mathematically correct to model the
335 three mechanisms separately, there should be many cases
336 where it is reasonable to use the simplified treatment. This
337 is convenient to apply the coalescent theory to real polymor-
338 phism data for estimating the rate of homologous recombi-
339 nation, especially when the relative contributions of the three
340 mechanisms are unknown. In this work, therefore, we em-
341 ploy the simplified treatment with a single rate of homo-

342 gous recombination (Equation 2) with a single geometric dis-
343 tribution with parameter q (Equation 1), following previous
344 theoretically studies (Falush *et al.* 2001; McVean *et al.* 2002;
345 Fearnhead *et al.* 2005; Jolley *et al.* 2005; Didelot and Falush
346 2007).

347 As mentioned above, a homologous recombination event
348 within prokaryote species is easily incorporated in the stan-
349 dard framework of the coalescent (Wiuf and Hein 2000;
350 McVean *et al.* 2002; Fearnhead *et al.* 2005; Jolley *et al.*
351 2005). That is, when tracing the ancestral lineage of a certain
352 sequence with L bp, the process waits for either coalescent
353 or recombination event, and the per-generation rate for the
354 latter is given by Equation 2, while the rate of coalescence is
355 given by $\binom{n}{2}/N$, where N is the population size and n is the
356 number of lineages.

357 Note that this simple process holds in a single population,
358 in which coalescence occurs randomly between any individ-
359 uals in the population and so does recombination, but it is
360 straightforward to incorporate population structure and de-
361 mographic history into this framework as is done for eukary-
362 ote cases. The difference between eukaryote and prokary-
363 ote is the causes of population structure. In eukaryotes, lim-
364 ited migration between geographic barriers should be the
365 major cause, and this also applies to prokaryotes although
366 more complicated. For example, subpopulations of infectious
367 species may form based on host individuals.

368 In addition, there are two major classes of isolation in
369 prokaryote, ecological and genetic isolation. Ecological iso-
370 lation is defined as a difference of niche that can reduce the
371 rate of recombination between bacterial populations (Cohan
372 2002a,b). Physiological difference between donor and recip-
373 ient would decrease the chance of recombination. For ex-
374 ample, *Vibrio splendidus* lives in coastal bacterioplankton,
375 exhibiting resource partitioning (specific season and/or free-
376 living size fraction) among strains and phylogenetic diver-
377 gence corresponding to each niche. It is suggested that eco-
378 logical isolation is working as a barrier of DNA exchanges
379 between niches (Hunt *et al.* 2008). Genetic isolation is de-
380 fined as the establishment of mutation accumulation that pre-
381 vents one strain from integrating foreign DNA of other strains
382 (e.g., Lawrence 2013). As described in the Introduction, the
383 rate of successful integration of DNA of other strains de-
384 pends on a number of self-recognition mechanisms, includ-
385 ing short-specific sequences (i.e., DUS or USS), restriction-
386 modification systems, RecA-mediated homology search, and
387 mismatch correction system. While both ecological and ge-
388 netic isolation are often used in the context of homologous
389 recombination with different species (or strains) involved in
390 sexual isolation (Cohan 2002a; Fraser *et al.* 2009), these con-
391 cept should work for homologous recombination within the
392 same species but between different populations (or strains).
393 Thus, there are many factors to cause isolation within the
394 same species, which should heavily affect the pattern of the
395 coalescent process. Therefore, when analyzing data with coa-

lescent simulations, past demographic history including such isolations should be well taken into account accordingly (e.g., Kreitman 2000; Nordborg 2001; Rosenberg and Nordborg 2002; Nordborg and Innan 2002; Sousa and Hey 2013).

Homologous recombination with different species (inter-specific recombination): We again use the backward argument. We define h as the backward recombination initiation rate per site. That is, when tracing the ancestral lineage of a single generation backward in time, h is the rate at which the lineage experiences a recombination event from external source that is initiated at the focal site (the same definition as g except for the source of integrated DNA). Given h , we can compute h' , the rate for the simulated region, using a similar equation to (2) (see below for details). With this rate specified, it is very straightforward to incorporate homologous recombination with different species into the coalescent framework: When tracing a lineage of a sequence with L bp, the process considers which is the next event, coalescence, recombination within species or recombination from external source, with relative backward rates, $\binom{n}{2}/N$, g' and h' , respectively. If a recombination event from external source occurs, the length of a transferred region is randomly determined (see below). Then, the transferred region is replaced by a sequence representing external source. Thus, the process can be well merged with the backward treatment of the coalescent, except that the biological interpretation of h , the backward rate recombination from external source, should be considered carefully, as we explain in the following.

In order to define h , let us consider the coalescent process of a particular species (population), around which there are a number of different species. The focal species potentially undergoes recombination with these species, and the rate of such recombination should be determined by a number of genetic and ecological factors as mentioned above. Figure 3 illustrates a hypothetical situation of a certain species, S_0 , around which there are five other species ($S_1 - S_5$), and their proportion is shown in the pie-chart (Figure 3A). The five species are in the order based on the divergence (d) from S_0 . d in each species might follow some distribution as illustrated in Figure 3B. Then, the dashed line (lines in five colors combined) in Figure 3B can be considered to represent the density distribution of divergence of DNA sequences that could recombine with the focal species. As mentioned above, the rate of successful integration of these DNA to the focal species varies depending on the species due to the genetic and ecological barriers against recombination. Furthermore, even when recombination successfully occurred, integrated DNA may be deleterious to the host individual and could be immediately selected out of the population. The distribution in the solid line in Figure 3B takes these effects into account, and the degree of reduction for each species is shown by an arrow. Then, noting that the tract length of homologous recombination roughly follows a geometric distribution, we obtain $Q_{\text{ext}}(d, z')$, the joint distribution of d and

successfully integrated tract length (z') as illustrated in Figure 3C. The definition of h is the per-site rate of such successful recombination from external source. h is much smaller than the forward recombination rate because we assume that deleterious recombinations are immediately purged from the population. In other words, we here assume that successfully incorporate foreign DNAs are neutral in the population of the focal species. Under this setting, recombination from external source can be simply incorporated in the coalescent framework as described above: The event of recombination from external source is included at rate h' together with coalescence and recombination within species that occur at rates $\binom{n}{2}/N$ and g' , respectively. When recombination from external source occurs, the tract length (z') and nucleotide divergence within the tract (d) can be determined as a random variable from $Q_{\text{ext}}(d, z')$.

The computation of h' from h is slightly different from the treatment for recombination within the same species (see Equation 4), because we cannot ignore the recombination event that encompassed the entire simulated region. That is, h' is given by

$$h' = R_{\text{in}}^h + R_{\text{left}}^h + R_{\text{all}}^h, \quad (5)$$

where R_{in}^h is the rate of recombination initiating inside the region, R_{left}^h is the rate initiating outside the region and ending within the focal sequence, and R_{all}^h is the rate initiating the 5' upstream of the region and ending in the 3' downstream of the region. R_{in}^h and R_{left}^h and R_{all}^h are given by

$$\begin{aligned} R_{\text{in}}^h &= hL, \\ R_{\text{left}}^h &= \sum_{i=1}^L hQ_{\text{ext}}(z' \geq i), \\ R_{\text{all}}^h &= \sum_{i=L+1}^{\infty} hQ_{\text{ext}}(z' \geq i), \end{aligned} \quad (6)$$

where $Q_{\text{ext}}(z')$ is calculated by integrating the joint probability distribution ($Q_{\text{ext}}(d, z')$) over d . From Equations 5-6, h' is written as

$$h' = h \left(L + \sum_{i=1}^{\infty} Q_{\text{ext}}(z' \geq i) \right). \quad (7)$$

Mutation: Once an ancestral recombination graph is constructed, neutral point mutations are distributed on it. Our model assumes a finite length of sequence (L bp), and mutation occurs symmetrically between two allelic states at rate μ per site per generation, and the population mutation rate is defined as $\theta = 2N\mu$.

Results and Discussion

We carried out simulations to generate a number of patterns of SNPs to demonstrate the effect of homologous re-

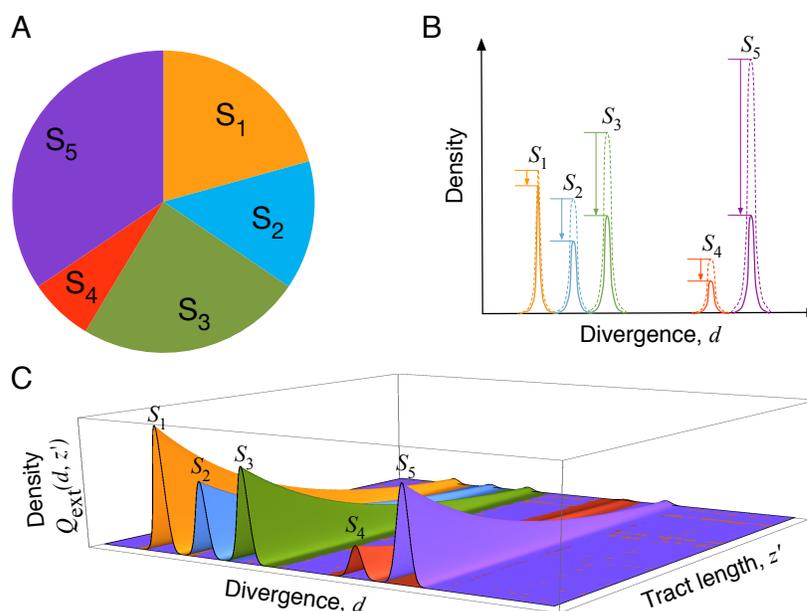


Figure 3 Illustrating a hypothetical environment where five different species ($S_1 - S_5$) are there around the focal species, S_0 . (A) The proportion of the five species in the environment. (B) Density distribution of the divergence of environmental DNA from the focal species (dashed line) and the waited distribution according to the probability of successful integration in the genome of the focal species. (C) The joint density distribution of d and successfully integrated tract length (z').

488 combination from external source (inter-specific recombina- 516
 489 tion). The mutation rate $\theta = 0.01$ was fixed throughout 517
 490 this work. For recombination within the focal species (intra- 518
 491 specific recombination), the mean tract length was fixed to 519
 492 be $\lambda = 1000$ bp, and the rate (g) was changed. We first 520
 493 considered a relatively low recombination rate from external 521
 494 source ($2Nh = 0.00005$). We here used a simplified assump- 522
 495 tion to demonstrate the point, that is, the average divergence 523
 496 to external DNA was fixed to be 20% and the tract length 524
 497 followed a geometric distribution with a fixed mean ξ (i.e., 525
 498 $Q_{\text{ext}}(d = 0.2, z') = \xi^{-1}(1 - \xi^{-1})^{z'-1}$). Figure 4 shows typi- 526
 499 cal patterns of SNPs from the simulation results with $n = 10$, 527
 500 and $L = 5,000$. The positions of SNPs are presented by 528
 501 solid vertical lines along the simulated region. In Figure 4A, 529
 502 no recombination within species is assumed ($2Ng = 0$). One 530
 503 recombination event (607 bp) from external source occurred 531
 504 $t = 0.23N$ generations ago on the ancestral lineage of indi- 532
 505 viduals 2, 5 and 10, and the positions of two breakpoints of 533
 506 the recombination event are shown by red arrows. The region 534
 507 that originates from foreign DNA can be clearly recognized 535
 508 as a cluster of SNPs due to large divergence ($d = 0.2$, 20 536
 509 times larger than θ). This region is referred to as Region 1 537
 510 and boxed in red. Neighbor-joining tree for this region is 538
 511 completely different form that for the other region: Individ- 539
 512 uals 2, 5 and 10 are highly diverged from the other seven 540
 513 individuals in Region 1.

514 It is thus obvious that the level of polymorphism increases 540
 515 as recombination events from external source increases. As 541

516 shown in Figure 5A, the level of polymorphism increases 517
 518 with increasing the initiation rate (h), mean tract length (ξ) 519
 520 and divergence (d), where the amount of polymorphism is 521
 522 measured by π , the average number of nucleotide differences 523
 524 per site. This simulation result agrees with theoretical predic- 525
 526 tion that the expectation of π is given by a simple function of 527
 528 h, ξ, d :

$$\frac{\theta + \phi}{1 + 2(\theta + \phi)}, \quad (8)$$

529 where $\phi = 2Nh\xi d$. See Appendix for the derivation. It is 530
 531 obvious that the most important parameter is the product of 532
 533 three recombination-associated parameters, $h\xi d$, which rep- 534
 535 represents the probability that the allelic state at a single site 536
 537 is flipped by recombination (see Appendix), as Figure 5B 538
 539 clearly demonstrates that π is given by a simple liner func- 540
 541 tion of $h\xi d$.

540 In Figure 4B, a moderate level of recombination within 541
 542 species (intra-specific recombination) is introduced ($2Ng = 543$
 544 0.001). Two external DNA fragments are integrated (Regions 545
 546 2 and 3). In Region 2, a 643 bp of foreign DNA was inte- 547
 548 grated $t = 0.61N$ generations ago. It is important to notice 549
 550 that whereas individuals 2, 3, 7 and 9 have the entire frag- 551
 552 ment, only a part of the integrated fragment is observed in 552
 553 individuals 4, 5, 6, 8 and 10. This is due to intra-specific re- 554
 555 combination that occurred after the integration; the integrated 556
 557 fragment was chopped into pieces and distributed into the 558
 559 population. By looking at the simulated ancestral recombina- 560
 561 tion graph, we found three such intra-specific recombination

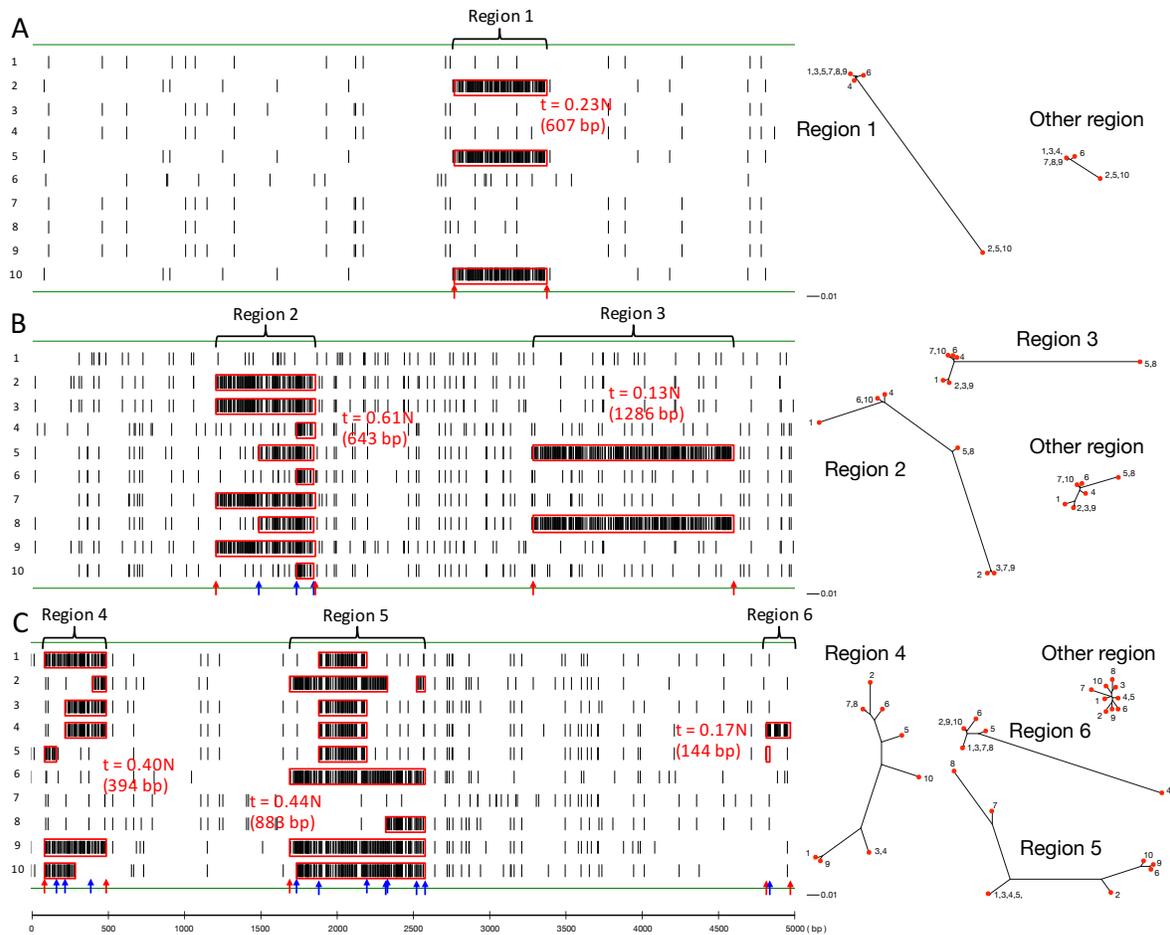


Figure 4 Typical patterns of SNPs with inter-specific recombination from external source with no intra-specific recombination (A; $2Ng = 0$), with a moderate level of intra-specific recombination (B; $2Ng = 0.001$), and with a high recombination rate (C; $2Ng = 0.005$). Vertical bars indicate the locations of point mutations in the simulated region with $L = 5000$ bp. The regions that experienced inter-specific recombination are specified (Regions 1-6), and neighbor-joining trees for these regions are shown in comparison with other regions with no inter-specific recombination. The breakpoints of inter-specific recombination events are presented by red allows, while blue ones exhibits intra-specific recombination events that fragmented the integrated foreign DNAs shown in red boxes.

542 events occurred (blue arrows the breakpoints). By contrast, 557
 543 in Region 3, due to its recent origin ($t = 0.13N$ generations 558
 544 ago), no intra-specific recombination was involved so that the 559
 545 entire integrated region (1286 bp) remains intact in individu- 560
 546 als 5 and 8, similar to Region 1 in Figure 4A.

547 With even a higher intra-specific recombination rate 557
 548 ($2Ng = 0.005$) in Figure 4C, fragmentation is more 558
 549 enhanced. There are three regions that experienced recombi- 559
 550 nation from external source (Regions 4, 5 and 6), and all 560
 551 of them involved intra-specific recombination. An intriguing 561
 552 pattern is seen in Region 5, where only a part of the inte- 562
 553 grated fragment is observed in the sample. The recombination 563
 554 occurred $t = 0.44N$ generations ago. The actual length of 564
 555 the integrated foreign fragment was more than 883 bp, but 565
 556 none of the sampled ten individuals have the 5' breakpoint. 566

This process can be well understood with the cartoon in Fig-
 ure 6, which illustrates the typical behaviors of population
 frequency of a foreign DNA integrated at time 0 with and
 without intra-specific recombination. With no intra-specific
 recombination, the entire integrated DNA can be vertically
 transmitted in the following generations (Figure 6A). By
 contrast, with intra-specific recombination, the integrated
 DNA is fragmented into various lengths (Figure 6B). As a
 consequence, more individuals have chances to have a part
 of the integrated DNA, but the length of the integrated DNA
 in each individual is on average short; some might lose the
 5' breakpoint and some might have only a short region in
 the middle. One potential caveat when interpreting data is
 that, when there was only one inter-specific recombination
 event, one might think multiple inter-specific events have
 incorporated

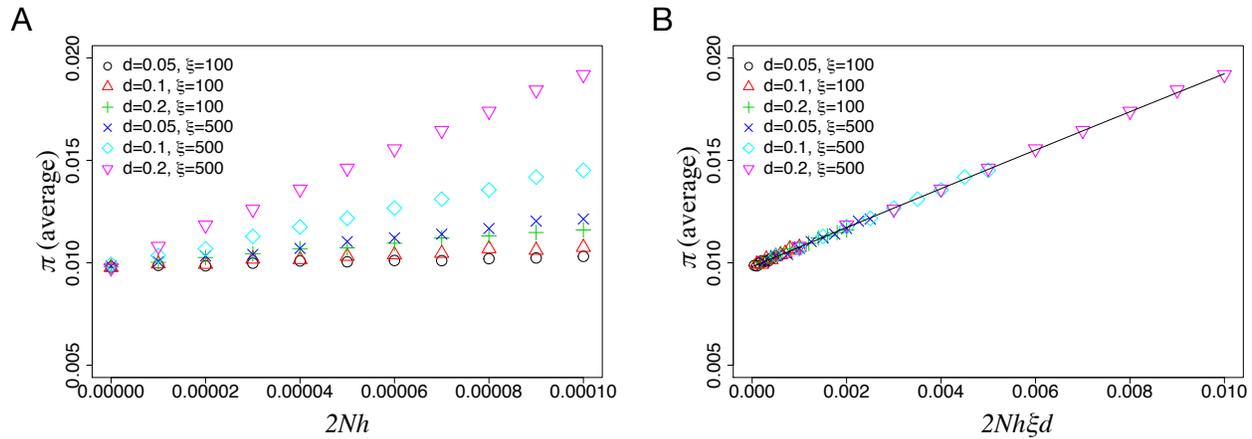


Figure 5 The effect of recombination from external source on the amount of polymorphism measure by π (A) π as a function of $2Nh$. (B) π is in a clear linear correlation with $2Nh\xi d$ (Equation 8). The averages π over 10,000 runs of simulations with $n = 15$ and $L = 10,000$ are shown.

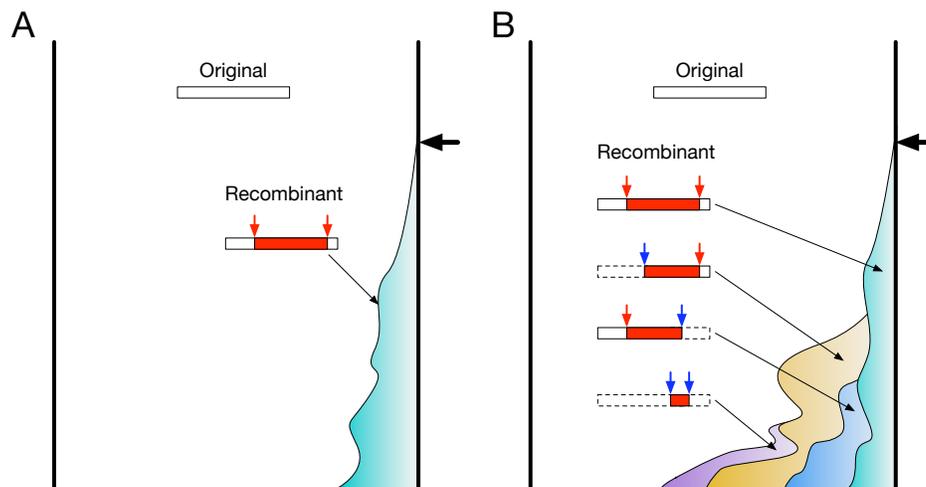


Figure 6 Cartoons of the typical behavior of population frequency of a foreign DNA, (A) without intra-specific recombination and (B) with intra-specific recombination. The time of the foreign DNA introduced into the population is denoted by a thick black arrow, producing a recombinant haplotype in which the integrated DNA is specified by a red box and arrows. (A) When there is no intra-specific recombination, the population consists of two haplotypes, the original and recombinant haplotype. (B) When intra-specific recombination is involved, the integrated DNA could be fragmented by recombination, thereby creating various kinds of recombinant haplotypes, each of which should have only a part of the integrated DNA. Additional breakpoints by intra-specific recombination are shown by blue arrows. In such a situation, the number of individuals having at least a part of part of the integrated DNA is much larger than the case with no recombination (A), while the length of integrated DNA is shorter.

572 foreign DNA independently. Indeed, when applied to one of our simulated data, **GENECONV** (Sawyer 1989), a commonly used software to detect gene conversion tracts, identifies a number of gene conversion tracts around the region that experienced a single time of inter-specific recombination, that incorporated a 2000 bp of foreign DNA at $t = 0.94N$ (Figure 7). The tracts inferred by **GENECONV** are presented by purple lines, showing as if there is a hotspot of integration.

580 Given this effect of inter-specific recombination, it is pre-

581 dicted that with increasing the rate of intra-specific recombination, (i) the number of individuals having foreign DNA increases and (ii) the length of foreign DNA decreases. This is quantitatively demonstrated by simulations (Figure 8). Figure 8 shows that the number of individuals that have at least a part of foreign DNA increase as the rate of intra-specific recombination ($2Ng$) increases (Figure 8A), whereas the average length of each foreign DNA in the sample decreases (Figure 8B). This effect of intra-specific recombination ($2Ng$) is

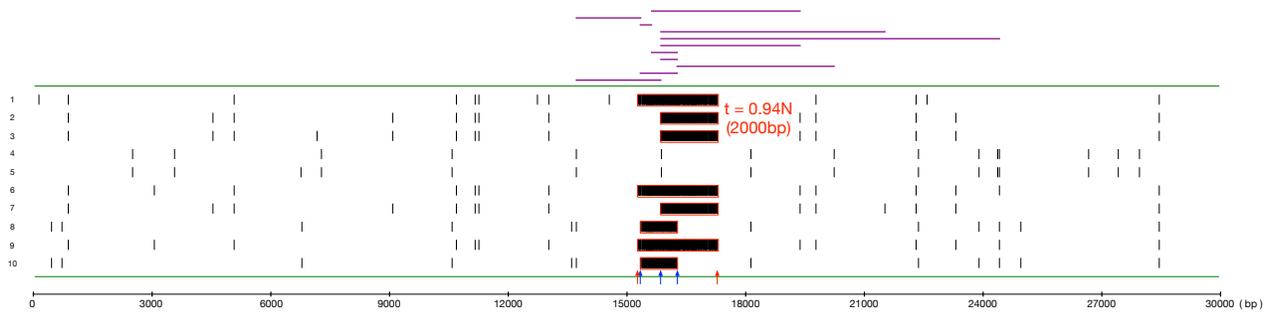


Figure 7 Application of **GENECONV** to a simulated region ($n = 10$, $L = 30,000$ bp), in which a 2000 bp of foreign DNA was integrated $0.94N$ generation ago (red boxed), followed by three additional intra-specific recombinations that fragmented the foreign DNA. The breakpoints of inter-specific recombination events are presented by red allows, while blue ones exhibits intra-specific recombination events that fragmented the integrated foreign DNAs. Vertical bars indicate the locations of point mutations. **GENECONV** with the default setting identified 11 integrated tracts (purple horizontal bars), making it look as if there is a hotspot of integration.

larger when ξ is larger. These findings should be useful to improve the algorithms to identify genomic regions that underwent homologous recombination (Didelot and Falush 2007; Didelot *et al.* 2009; Ansari and Didelot 2014; Yahara *et al.* 2014).

We thus demonstrated that the joint work of intra- and inter-specific recombination could create a complicated pattern of SNPs and it is needed to obtain full theoretical understanding of this for interpreting SNP data from prokaryotes. Given quite common homologous recombination from external source in prokaryotes and strong impact on the pattern of SNPs as we have shown here, we have to avoid a misleading interpretation of observed data due to recombination, potentially resulting in misevaluation of the relative contribution of demography and selection. We here developed a fast simulator for producing a number of realizations of SNPs with both intra- and inter-specific recombination. The software named **msPro** was developed based on Hudson's commonly used software **ms** (**msPro** means **ms** for prokaryotes), and the input command and the form of output are very similar to **ms**. **msPro** can incorporate various forms of demographic history as **ms** does. **msPro** will be available upon request.

It should be noted that our simulator runs after specifying the density distribution of external DNA, Q_{ext} . When there is no prior knowledge on the environmental DNA, it is difficult to set Q_{ext} . Considering such a case, the default setting of **msPro** is given as follows. A first approximation is that the density distribution of tract length (z') follow a geometric distribution $\xi^{-1}(1 - \xi^{-1})^{z'-1}$, regardless of divergence. According to empirical studies (Zawadzki and Cohan 1995; Linz *et al.* 2000), typical lengths of integrated DNA may be a few kb, so we assume $\xi = 1000$ bp. If we assume a uniform distribution of divergence in the external DNA in the environment, the density distribution of d (*i.e.*, divergence of successfully integrated DNA) simply follows the rate of successful integration, which may be approximated by an exponential distribution (see Fraser *et al.*

2007, and references therein), namely, $\alpha \exp[-\alpha d]$, where α is a parameter to specify the decay. According to Figure 1A in Fraser *et al.* (2007), $\alpha \sim 20$ might fit the observed data from some bacterial species. Therefore, we set $Q_{\text{ext}}(d, z') = \alpha \exp[-\alpha d] \times \xi^{-1}(1 - \xi^{-1})^{z'-1}$.

Acknowledgements

This work was supported in part by the Japan Society for the Promotion of Science (JSPS).

Literature Cited

- Achtman, M., and M. Wagner, 2008 Microbial diversity and the genetic nature of microbial species. *Nat. Rev. Microbiol.* 6: 431–440.
- Albritton, W., J. Setlow, M. Thomas, F. Sottnek and A. Steigerwalt, 1984 Heterospecific transformation in the genus *Haemophilus*. *Mol. Gen. Genet.* 193: 358–363.
- Ansari, M. A., and X. Didelot, 2014 Inference of the properties of the recombination process from whole bacterial genomes. *Genetics* 196: 253–65.
- Awadalla, P., 2003 The evolutionary genomics of pathogen recombination. *Nat. Rev. Genet.* 4: 50–60.
- Azad, R. K., and J. G. Lawrence, 2012 Detecting laterally transferred genes, pp. 281–308 in *Evolutionary Genomics*, edited by M. Anisimova. Volume 1. Humana Press, New York.
- Beaumont, M. A., W. Zhang and D. J. Balding, 2002 Approximate bayesian computation in population genetics. *Genetics* 162: 2025–2035.

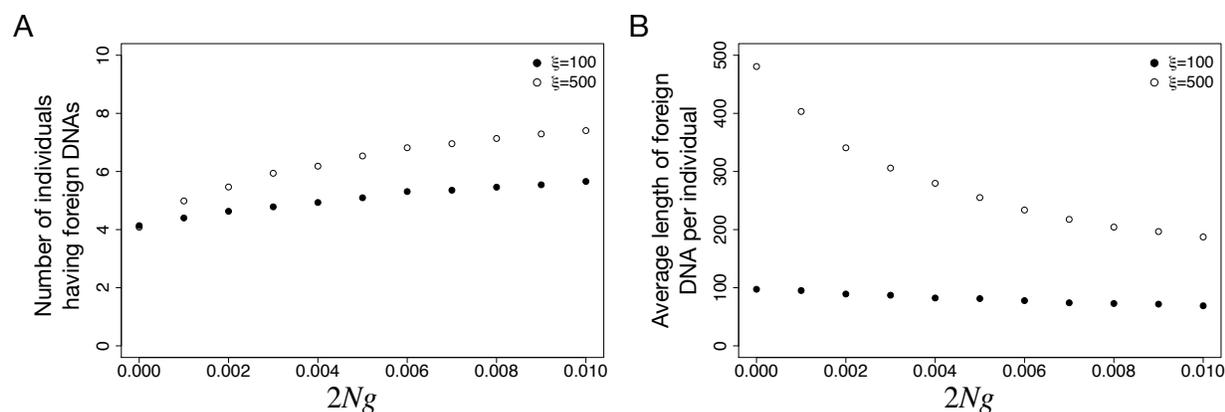


Figure 8 The effect of intra-specific recombination on (A) the number of individuals having foreign DNA and (B) average length of integrated foreign DNA, from 10,000 runs of simulations with $n = 15$ and $L = 10,000$.

- 654 Brown, T., X. Didelot, D. J. Wilson and N. De Maio, 2016 683 Didelot, X., D. Lawson and D. Falush, 2009 SimMLST:
655 SimBac: simulation of whole bacterial genomes with ho- 684 simulation of multi-locus sequence typing data under a
656 mologous recombination. *Microbial genomics* 2: 685 neutral model. *Bioinformatics* 25: 1442–4.
- 657 Chen, I., and D. Dubnau, 2004 DNA uptake during bacterial 686 Didelot, X., and M. C. J. Maiden, 2010 Impact of recombini-
658 transformation. *Nat. Rev. Microbiol.* 2: 241–9. 687 nation on bacterial evolution. *Trends Microbiol.* 18: 315–
688 22.
- 659 Claverys, J. P., and S. A. Lacks, 1986 Heteroduplex de- 689 Didelot, X., and D. J. Wilson, 2015 ClonalFrameML:
660 oxyribonucleic acid base mismatch repair in bacteria. *Mi- 690 efficient inference of recombination in whole bacterial
661 crobiol. Rev.* 50: 133–165. 691 genomes. *PLoS Comput Biol* 11: e1004041.
- 662 Cohan, F. M., 2002a Sexual isolation and speciation in bac- 692 Dobrindt, U., B. Hochhut, U. Hentschel and J. Hacker, 2004
663 teria. *Genetica* 116: 359–370. 693 Genomic islands in pathogenic and environmental mi-
664 Cohan, F. M., 2002b What are bacterial species? *Annu. 694 croorganisms. Nat. Rev. Microbiol.* 2: 414–424.
- 665 *Rev. Microbiol.* 56: 457–87. 695 Donati, C., N. L. Hiller, H. Tettelin, A. Muzzi, N. J. Croucher,
666 Cornejo, O. E., T. Lefébure, P. D. P. Bitar, P. Lang, V. P. 696 *et al.*, 2010 Structure and dynamics of the pan-genome
667 Richards, *et al.*, 2013 Evolutionary and population ge- 697 of *Streptococcus pneumoniae* and closely related species.
668 nomics of the cavity causing bacteria *Streptococcus mu- 698 Genome Biol.* 11:R107.
- 669 *tans. Mol. Biol. Evol.* 30: 881–93. 699 Donnelly, P., and T. G. Kurtz, 1999 Genealogical processes
670 Croucher, N. J., S. R. Harris, L. Barquist, J. Parkhill and S. D. 700 for fleming-viot models with selection and recombination.
671 Bentley, 2012 A high-resolution view of genome-wide 701 *Ann Appl. Probab.* 9:1091–1148.
- 672 pneumococcal transformation. *PLoS Pathog.* 8: e1002745. 702 Doolittle, W. F., and R. T. Papke, 2006 Genomics and the
673 de Vries, J., P. Meier and W. Wackernagel, 2001 The natural 703 bacterial species problem. *Genome Biol.* 7: e116.
- 674 transformation of the soil bacteria *Pseudomonas stutzeri* 704 Doroghazi, J. R., and D. H. Buckley, 2011 A model for the
675 and *Acinetobacter* sp. by transgenic plant DNA strictly 705 effect of homologous recombination on microbial diversi-
676 depends on homologous sequences in the recipient cells. 706 fication. *Genome Biol. Evol.* 3: 1349–56.
- 677 *FEMS Microbiol. Lett.* 195: 211–215. 707 Earl, A. M., R. Losick and R. Kolter, 2008 Ecology and
678 De Maio, N., and D. J. Wilson, 2017 The bacterial sequen- 708 genomics of *Bacillus subtilis*. *Trends Microbiol.* 16: 269–
679 tial Markov coalescent. *Genetics* 206: 333–343. 709 75.
- 680 Didelot, X., and D. Falush, 2007 Inference of bacterial 710 Falush, D., C. Kraft, N. S. Taylor, P. Correa, J. G. Fox, *et al.*,
681 microevolution using multilocus sequence data. *Genetics* 711 2001 Recombination and mutation during long-term gas-
682 175: 1251–66. 712 tric colonization by *Helicobacter pylori*: estimates of clock

- 713 rates, recombination size, and minimal age. *Proc. Natl.* 757
714 *Acad. Sci. USA* 98: 15056–15061. 758
- 715 Falush, D., M. Torpdahl, X. Didelot, D. F. Conrad, D. J. 759
716 Wilson, *et al.*, 2006 Mismatch induced speciation in 760
717 *salmonella*: model and data. *Philos. Trans. R. Soc. B* 361:
718 2045–53. 761
- 719 Fearnhead, P., 2006 Perfect simulation from nonneutral pop- 762
720 ulation genetic models: variable population size and popu- 763
721 lation subdivision. *Genetics* 174: 1397–406. 764
- 722 Fearnhead, P., N. G. C. Smith, M. Barrigas, A. Fox and 765
723 N. French, 2005 Analysis of recombination in *Campy-* 766
724 *lobacter jejuni* from MLST population data. *J. Mol. Evol.* 767
725 61: 333–40. 768
- 726 François, O., M. G. B. Blum, M. Jakobsson and N. A. Rosen- 769
727 berg, 2008 Demographic history of european populations 770
728 of *Arabidopsis thaliana*. *PLoS Genet.* 4: 1–15. 771
- 729 Fraser, C., W. P. Hanage and B. G. Spratt, 2007 Recombi- 772
730 nation and the nature of bacterial speciation. *Science* 315:
731 476–80. 773
- 732 Fraser, C., E. J. Alm, M. F. Polz, B. G. Spratt and W. P. Han- 774
733 age, 2009 The bacterial species challenge: making sense 775
734 of genetic and ecological diversity. *Science* 323: 741–746. 776
- 735 Fraser, C., W. P. Hanage and B. G. Spratt, 2005 Neutral mi- 777
736 croepidemic evolution of bacterial pathogens. *Proc. Natl.* 778
737 *Acad. Sci. USA* 102: 1968–1973. 779
- 738 Griffiths, R. C., and P. Marjoram, 1996 Ancestral inference 780
739 from samples of DNA sequences with recombination. *J.* 781
740 *Comput. Biol.* 3: 479–502. 782
- 741 Grissa, I., G. Vergnaud and C. Pouchel, 2007 The CRISPRdb 783
742 database and tools to display CRISPRs and to generate dic- 784
743 tionaries of spacers and repeats. *BMC Bioinformatics* 8:
744 172. 785
- 745 Hartl, D. L., and A. G. Clark, 2007 *Principles of population* 786
746 *genetics*. Sinauer Associates, Sunderland. 787
- 747 Haven, J., L. C. Vargas, E. F. Mongodin, V. Xue, Y. Hernan- 788
748 dez, *et al.*, 2011 Pervasive recombination and sympatric 789
749 genome diversification driven by frequency-dependent se- 790
750 lection in *Borrelia burgdorferi*, the lyme disease bac- 791
751 terium. *Genetics* 189: 951–66. 792
- 752 Hudson, R. R., 1983a Properties of a neutral allele model 793
753 with intragenic recombination. *Theor. Popul. Biol.* 23:
754 183–201. 794
- 755 Hudson, R. R., 1983b Testing the constant-rate neutral allele 795
756 model with protein sequence data. *Evolution* 37: 203–217. 796
- 757 Hudson, R. R., 1990 Gene genealogies and the coalescent 797
758 process, pp. 1–43 in *Oxford surveys in evolutionary biol-* 798
759 *ogy*, edited by D. Futuyma and J. Antonovics. Volume 7. 799
760 Oxford Univ Press, Oxford. 800
- 761 Hudson, R. R., 1994 Analytical results concerning linkage 801
762 disequilibrium in models with genetic transformation and 802
763 conjugation. *J. evol. biol.* 7: 535–548. 803
- 764 Hudson, R. R., 2001 Two-locus sampling distributions and 804
765 their application. *Genetics* 159: 1805–1817. 805
- 766 Hudson, R. R., 2002 Generating samples under a wright- 806
767 fisher neutral model of genetic variation. *Bioinformatics* 807
768 18: 337–338. 808
- 769 Hunt, D. E., L. A. David, D. Gevers, S. P. Preheim, E. J. Alm, 809
770 *et al.*, 2008 Resource partitioning and sympatric differ- 810
771 entiation among closely related bacterioplankton. *Science* 811
772 320: 1081–5. 812
- 773 Jolley, K. A., D. J. Wilson, P. Kriz, G. McVean and M. C. J. 813
774 Maiden, 2005 The influence of mutation, recombina- 814
775 tion, population history, and selection on patterns of ge- 815
776 netic diversity in *Neisseria meningitidis*. *Mol. Biol. Evol.* 816
777 22: 562–9. 817
- 778 Kingman, J. F., 1982 On the genealogy of large populations. 818
779 *J. Appl. Probab.* 19: 27–43. 819
- 780 Krause, D. J., X. Didelot, H. Cadillo-Quiroz and R. J. 820
781 Whitaker, 2014 Recombination shapes genome architec- 821
782 ture in an organism from the archaeal domain. *Genome* 822
783 *Biol. Evol.* 6: 170–8. 823
- 784 Kreitman, M., 2000 Methods to detect selection in popula- 824
785 tions with applications to the human. *Annu. Rev. Genom.* 825
786 *Hum. Genet.* 1: 539–559. 826
- 787 Krone, S. M., and C. Neuhauser, 1997 Ancestral processes 827
788 with selection. *Theor. popul. biol.* 51: 210–237. 828
- 789 Lawrence, J. G., 2013 Gradual speciation: Further entan- 829
790 gling the tree of life, pp. 243–262 in *Lateral Gene Transfer* 830
791 *in Evolution*, edited by U. Gophna. Springer, New York. 831
- 792 Lin, E. A., X.-S. Zhang, S. M. Levine, S. R. Gill, D. Falush, 832
793 *et al.*, 2009 Natural transformation of *Helicobacter pylori* 833
794 involves the integration of short dna fragments interrupted 834
795 by gaps of variable size. *PLoS Pathog.* 5: e1000337. 835
- 796 Linz, B., M. Schenker, P. Zhu and M. Achtman, 2000 Fre- 836
797 quent interspecific genetic exchange between commensal 837
798 neisseriae and neisseria meningitidis. *Mol. Microbiol.* 36:
799 1049–1058. 838
- 800 Lorenz, M. G., and J. Sikorski, 2000 The potential for 839
801 intraspecific horizontal gene exchange by natural genetic 840
802 transformation: sexual isolation among genomovars of 841
803 *Pseudomonas stutzeri*. *Microbiology* 146: 3081–3090. 842

- 804 Lorenz, M. G., and W. Wackernagel, 1994 Bacterial gene 850
805 transfer by natural genetic transformation in the environ- 851
806 ment. *Microbiol. Rev.* 58: 563–602. 852
- 807 Majewski, J., 2001 Sexual isolation in bacteria. *FEMS* 853
808 *Microbiol. Lett.* 199: 161–169. 854
- 809 Majewski, J., and F. M. Cohan, 1998 The effect of mismatch 855
810 repair and heteroduplex formation on sexual isolation in 856
811 *Bacillus*. *Genetics* 148: 13–18. 857
- 812 Majewski, J., P. Zawadzki, P. Pickerill, F. M. Cohan and C. G. 858
813 Dowson, 2000 Barriers to genetic exchange between bac- 859
814 terial species: *Streptococcus pneumoniae* transformation. 860
815 *J. Bacteriol.* 182: 1016–1023. 861
- 816 Marjoram, P., J. Molitor, V. Plagnol and S. Tavaré, 2003 862
817 Markov chain monte carlo without likelihoods. *Proceed-* 863
818 *ings of the National Academy of Sciences* 100: 15324– 864
819 15328. 865
- 820 Marraffini, L. A., and E. J. Sontheimer, 2010 CRISPR inter- 866
821 ference: RNA-directed adaptive immunity in bacteria and 867
822 archaea. *Nat. Rev. Genet.* 11: 181–90. 868
- 823 Maynard Smith, J., 1994 Estimating the minimum rate of 869
824 genetic transformation in bacteria. *J. Evol. Biol.* 7: 525– 870
825 534. 871
- 826 McVean, G., P. Awadalla and P. Fearnhead, 2002 A 872
827 coalescent-based method for detecting and estimating re- 873
828 combination from gene sequences. *Genetics* 160: 1231– 874
829 1241. 875
- 830 Mell, J. C., J. Y. Lee, M. Firme, S. Sinha and R. J. Redfield, 876
831 2014 Extensive cotransformation of natural variation 877
832 into chromosomes of naturally competent *Haemophilus in-* 878
833 *fluenzae*. *G3 (Bethesda)* 4: 717–31. 879
- 834 Nakhleh, L., 2013 Computational approaches to species 880
835 phylogeny inference and gene tree reconciliation. *Trends* 881
836 *Ecol. Evol.* 28: 719–728. 882
- 837 Nell, S., D. Eibach, V. Montano, A. Maady, A. Nkwescheu, 883
838 *et al.*, 2013 Recent acquisition of *Helicobacter pylori* by 884
839 baka pygmies. *PLoS Genet.* 9: e1003775.
- 840 Neuhauser, C., and S. M. Krone, 1997 The genealogy of 885
841 samples in models with selection. *Genetics* 145: 519–534. 886
- 842 Nordborg, M., 2001 Coalescent theory in *Handbook of sta-* 887
843 *tistical genetics*, edited by D. J. BALDING, M. Bishop, and 888
844 C. Cannings. Wiley-Blackwell, Chichester, UK. 889
- 845 Nordborg, M., and H. Innan, 2002 Molecular population 890
846 genetics. *Curr. Opin. Plant Biol.* 5: 69–73. 891
- 847 Ochman, H., J. G. Lawrence and E. A. Groisman, 2000 Lat- 892
848 eral gene transfer and the nature of bacterial innovation. 893
849 *Nature* 405: 299–304. 894
- Overballe-Petersen, S., K. Harms, L. A. A. Orlando, J. V. M. 895
Mayar, S. Rasmussen, *et al.*, 2013 Bacterial natural trans- 896
formation by highly fragmented and damaged DNA. *Proc.* 897
Natl. Acad. Sci. USA 110: 19860–19865. 898
- Pepperell, C., V. H. Hoepfner, M. Lipatov, W. Wobeser, 899
G. K. Schoolnik, *et al.*, 2010 Bacterial genetic signa- 900
tures of human social phenomena among *M. tuberculosis* 901
from an Aboriginal Canadian population. *Mol. Biol. Evol.* 902
27: 427–40. 903
- Pérez-Losada, M., E. B. Browne, A. Madsen, T. Wirth, 904
R. P. Viscidi, *et al.*, 2006 Population genetics of micro- 905
bial pathogens estimated from multilocus sequence typing 906
(MLST) data. *Infect. Genet. Evol.* 6: 97–112. 907
- Polz, M. F., E. J. Alm and W. P. Hanage, 2013 Horizontal 908
gene transfer and the evolution of bacterial and archaeal 909
population structure. *Trends Genet.* 29: 170–175. 910
- Roberts, M., and F. Cohan, 1993 The effect of DNA se- 911
quence divergence on sexual isolation in bacillus. *Genetics* 912
134: 401–408. 913
- Roberts, R. J., T. Vincze, J. Posfai and D. Macelis, 2010 914
REBASE—a database for DNA restriction and modifica- 915
tion: enzymes, genes and genomes. *Nucleic Acids Res.* 916
38: D234–236. 917
- Rosen, M. J., M. Davison, D. Bhaya and D. S. Fisher, 2015 918
Fine-scale diversity and extensive recombination in a qua- 919
sisexual bacterial population occupying a broad niche. *Sci-* 920
ence 348: 1019–1023. 921
- Rosenberg, N. A., and M. Nordborg, 2002 Genealogical 922
trees, coalescent theory and the analysis of genetic poly- 923
morphisms. *Nat. Rev. Genet.* 3: 380–90. 924
- Rousseau, C., M. Gonnet, M. Le Romancer and J. Nicolas, 925
2009 CRISPI: a CRISPR interactive database. *Bioinform-* 926
atics 25: 3317–8. 927
- Sawyer, S., 1989 Statistical tests for detecting gene conver- 928
sion. *Mol. Biol. Evol.* 6: 526–538. 929
- Shapiro, B. J., 2014 Signatures of natural selection and eco- 930
logical differentiation in microbial genomes, pp. 339–359 931
in *Ecological Genomics*. Springer, New York. 932
- Shapiro, B. J., J. Friedman, O. X. Cordero, S. P. Preheim, 933
S. C. Timberlake, *et al.*, 2012 Population genomics of 934
early events in the ecological differentiation of bacteria. 935
Science 336: 48–51. 936
- Shen, P., and H. V. Huang, 1986 Homologous recombination 937
in *Escherichia coli*: dependence on substrate length and 938
homology. *Genetics* 112: 441–457. 939

- 895 Snyder, L., J. E. Peters, T. M. Henkin and W. Champness, 1994
 896 2013 *Molecular genetics of bacteria*. ASM Press, Wash- 942
 897 ington, DC. 943
- 898 Sousa, V., and J. Hey, 2013 Understanding the origin of 944
 899 species with genome-scale data: modelling gene flow. *Nat.* 945
 900 *Rev. Genet.* 14: 404–14. 946
- 901 Tajima, F., 1983 Evolutionary relationship of DNA se- 947
 902 quences in finite populations. *Genetics* 105: 437–460. 948
- 903 Takuno, S., T. Kado, R. P. Sugino, L. Nakhleh and H. Innan, 949
 904 2012 Population genomics in bacteria: A case study of 950
 905 *Staphylococcus aureus*. *Mol. Biol. Evol.* 29: 797–809. 951
- 906 Thomas, C. M., and K. M. Nielsen, 2005 Mechanisms of, 952
 907 and barriers to, horizontal gene transfer between bacteria. 953
 908 *Nat. Rev. Microbiol.* 3: 711–21. 954
- 909 Thomas, J. C., P. A. Godfrey, M. Feldgarden and D. A. 955
 910 Robinson, 2012 Candidate targets of balancing selection 956
 911 in the genome of *Staphylococcus aureus*. *Mol. Biol. Evol.* 957
 912 29: 1175–86. 958
- 913 Touchon, M., C. Hoede, O. Tenaillon, V. Barbe, S. Baeriswyl, 959
 914 *et al.*, 2009 Organised genome dynamics in the *Es-* 960
 915 *cherichia coli* species results in highly diverse adaptive 961
 916 paths. *PLoS Genet.* 5: e1000344. 962
- 917 Vasu, K., and V. Nagaraja, 2013 Diverse functions of 963
 918 restriction-modification systems in addition to cellular de- 964
 919 fense. *Microbiol. Mol. Biol. Rev.* 77: 53–72. 965
- 920 Vulić, M., F. Dionisio, F. Taddei and M. Radman, 1997 966
 921 Molecular keys to speciation: DNA polymorphism and the 967
 922 control of genetic exchange in enterobacteria. *Proc. Natl.* 968
 923 *Acad. Sci. USA* 94: 9763–9767. 969
- 924 Wakeley, J., 2008 *Coalescent theory: An introduction*. 970
 925 Roberts and Company, Greenwood Village, Colorado. 971
- 926 Wirth, T., D. Falush, R. Lan, F. Colles, P. Mensa, *et al.*, 2006 972
 927 Sex and virulence in *Escherichia coli*: an evolutionary per- 973
 928 spective. *Mol. Microbiol.* 60: 1136–51. 974
- 929 Wiuf, C., 2001 Recombination in human mitochondrial 975
 930 DNA? *Genetics* 159: 749–756. 976
- 931 Wiuf, C., and J. Hein, 2000 The coalescent with gene con- 977
 932 version. *Genetics* 155: 451–462. 978
- 933 Zahrt, T. C., and S. Maloy, 1997 Barriers to recombination 979
 934 between closely related bacteria: MutS and RecBCD in- 980
 935 hibit recombination between *Salmonella typhimurium* and 981
 936 *Salmonella typhi*. *Proc. Natl. Acad. Sci. USA* 94: 9786– 982
 937 9791. 983
- 938 Zawadzki, P., and F. M. Cohan, 1995 The size and continu- 984
 939 ity of dna segments integrated in bacillus transformation. 985
 940 *Genetics* 141: 1231–1243. 986
- Zhang, L., J. C. Thomas, X. Didelot and D. A. Robinson, 2012
 Molecular signatures identify a candidate target of
 balancing selection in an *arcD*-like gene of *Staphylococcus*
epidermidis. *J. Mol. Evol.* 75: 43–54.
- Yahara, K., X. Didelot, M. A. Ansari, S. K. Sheppard and
 D. Falush, 2014 Efficient inference of recombination hot
 regions in bacterial genomes. *Mol. Biol. Evol.* 31: 1593–
 1605.

Appendix

Consider a certain site of two samples and changes of their state in one generation backward in time. Let P_t be their current diversity and P_{t-1} be that of before generation. Assuming that *de novo* mutation and recombination between species does not occur simultaneously and occur once at most, recursion of the state of their diversity under a finite two-states model can be written,

$$P_{t-1} = (1 - 2\lambda h) \times \left\{ (P_t(1 - 2\mu) + 2\mu(1 - P_t)) \left(1 - \frac{1}{N}\right) + \frac{2\mu}{N} \right\} + 2\lambda h \{ (1 - P_t)d + P_t(1 - d) \}. \quad (9)$$

The expression in the first curly bracket means the case that the recombination does not occur, while the expression in the second curly bracket means the opposite case. At equilibrium (*i.e.*, $P_{t-1} = P_t$), the recursion can be solved and then the states (denoted by P^*) is,

$$P^* = \frac{2N(\mu + h\lambda(d - 2\mu))}{1 - 2h\lambda + 4Nh\lambda d - 4\mu + 4N\mu + 8h\lambda\mu - 8N\mu\lambda h} \approx \frac{\theta + \phi}{1 + 2(\theta + \phi)}, \quad (10)$$

where $\theta = 2N\mu$ and $\phi = 2Nh\lambda d$, and the terms with h or μ as factors are ignored. P^* is corresponding to the expectation of π in our framework.