Heredity (2018) 120:474–484
https://doi.org/10.1038/s41437-017-0034-1

the geneticssociety

ARTICLE

Coalescent framework for prokaryotes undergoing interspecific homologous recombination

Tetsuya Akita[1,2] · Shohei Takuno[1] · Hideki Innan[1]

Received: 5 July 2017 / Revised: 4 October 2017 / Accepted: 23 October 2017 / Published online: 23 January 2018

Abstract

Coalescent process for prokaryote species is theoretically considered. Prokaryotes undergo homologous recombination with individuals of the same species (intraspecific recombination) and with individuals of other species (interspecific recombination). This work particularly focuses on interspecific recombination because intraspecific recombination has been well incorporated in coalescent framework. We present a simulation framework for generating SNP (single-nucleotide polymorphism) patterns that allows external DNA integration into host genome from other species. Using this simulation tool, msPro, we observed that the joint processes of intra- and interspecific recombination generate complex SNP patterns. The direct effect of interspecific recombination includes increased polymorphism. Because interspecific recombination is very rare in nature, it generates regions with exceptionally high polymorphism. Following interspecific recombination, intraspecific recombination cuts the integrated external DNA into small fragments, generating a complex SNP pattern that appears as if external DNA was integrated multiple times. The insight gained from our work using the msPro simulator will be useful for understanding and evaluating the relative contributions of intra- and interspecific recombination events in generating complex SNP patters in prokaryotes.

Introduction

The coalescent is a population genetic theory that considers the evolutionary processes backward in time (Kingman 1982; Hudson 1983b; Tajima 1983). The coalescent theory has been primarily developed assuming its application to higher eukaryotes, which is perhaps a result of historical factors, including the fact that the major model organisms for population genetics research have historically been higher eukaryotes, such as *Drosophila* and humans (e.g., Hartl and Clark 2007). The coalescent theory provides an extremely powerful framework for analyzing single-nucleotide polymorphism (SNP) patterns in sampled DNA sequences. This framework is sufficiently flexible to incorporate major evolutionary processes, including random genetic drift, mutation, recombination, and demographic history (e.g., Hudson 1990; Nordborg 2001; Wakeley 2008); however, incorporating complex modes of selection is challenging (but see Krone and Neuhauser 1997; Donnelly and Kurtz 1999). The software ms is one of the most popular coalescent simulators, which generates neutral SNP patterns under various demographic settings (Hudson 2002). This incorporates two major outcomes of meiotic recombination: meiotic crossing-over and gene conversion.

Prokaryotes are haploid and do not undergo meiosis; therefore, recombination mechanisms in prokaryotes are considerably different than those in eukaryotes. Nevertheless, the coalescent framework can be applied to prokaryotes with a relatively simple modification, that is, because circular chromosome in prokaryotes requires double "crossing-over" to exchange a DNA fragment, recombination can be considered as an event analogous to meiotic gene conversion. This modification can well explain the nature of homologous recombination in prokaryotes, as explained below. The application of coalescent theory to bacteria has become particularly popular since McVean et al. (2002) developed the software LDhat for estimating

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41437-017-0034-1) contains supplementary material, which is available to authorized users.

✉ Hideki Innan
innan_hideki@soken.ac.jp

[1] Graduate University for Advanced Studies, Hayama, Kanagawa 240-0193, Japan

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

recombination rates; this software is a modified version of Hudson's composite likelihood method (Hudson 2001). Because LDhat allows simulating recurrent mutations at a single site, it is more suitable for organisms with large population sizes, such as bacteria. LDhat has been applied to multilocus sequence typing (MLST) (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 great variation in recombination rates across species. Hudson's ms software has also been successfully used for various population genetics analyzes in prokaryotes (e.g., Fearnhead et al. 2005; Takuno et al. 2012; Cornejo et al. 2013; Rosen et al. 2015).

Thus, despite different recombination mechanisms in eukaryotes and prokaryotes, the theoretical application of the coalescent framework to homologous recombination in prokaryotes is not very difficult. However, this is plausible only in the case of intraspecific recombination. Nevertheless, this assumption may not be true in eukaryotes due to frequent DNA exchanges between different species due to the nature of their recombination mechanism, as described below.

Prokaryotes undergo recombination by integrating free extracellular DNA into host genome through three major mechanisms: natural transformation, transduction, and conjugation (e.g., Snyder et al. 2013), molecular, as illustrated in Supplementary Fig. S1. Natural transformation involves the direct uptake of free extracellular DNA and its integration into host genome under natural bacterial growth conditions (Supplementary Fig. S1a). Transduction involves bacterial DNA integration into another bacterial cell through infection by DNA-containing phage (Supplementary Fig. S1b). Conjugation is the transfer of DNA from one bacterial cell to another through the transfer functions of self-transmissible DNA elements, which are frequently associated with plasmids (Supplementary Fig. S1c). Foreign DNA is usually harmful when integrated into the host genome, and there exist a number of mechanisms to avoid such integration, as summarized in Supplementary Information. These three mechanisms differ in terms of typical rate and tract length (e.g., Snyder et al. 2013); however, most previous studies have summarized single parameters for these in standard coalescent treatment. In Supplementary Information, we have described the conditions under which the standard coalescent treatment should hold, and we have also employed this standard coalescent treatment in our work.

There are two possible outcomes of recombination in prokaryotes (e.g., Lawrence 2013). First, the integrated DNA may be inserted into the genome through a process known as horizontal (or lateral) gene transfer (Supplementary Fig. S1d). The evolutionary role of horizontal gene transfer is emphasized when a novel gene is acquired that contributes to adaptation (Ochman et al. 2000; Dobrindt et al. 2004), but the frequency and importance of such illegitimate recombination are under debate (Shapiro et al. 2012). Alternatively, the incorporated DNA may be exchanged with its homologous part in the genome (if such exists) through a process known as homologous recombination (Supplementary Fig. S1e). Homologous recombination usually involves DNA from the same species because the near-identity requirement (e.g., monitored by RecA-mediated homology search, (Shen and Huang 1986; Majewski and Cohan 1998)) is easily satisfied; however, it is also possible for DNA from different species to be integrated as long as some homology is retained. Homologous recombination between different species sometimes results in unique SNP patterns, and we can search through sequence data to find the footprints of these (e.g., Awadalla 2003; Didelot and Maiden 2010; Azad and Lawrence 2012).

In this study, we emphasize on homologous recombination. Due to the mechanism of homologous recombination involving double crossing-over, the outcome of recombination is similar to that of meiotic gene conversion. Therefore, as mentioned earlier, the standard theory was applied for analyzing SNP patterns in bacteria, with the simple modification that the rate of crossing-over is set to zero, and consequently, recombination events (i.e., homologous recombination) were treated analogous to meiotic gene conversion events. This application should be reasonable as long as the donor for any homologous recombination is another individual of the same species. However, it is well known that homologous recombination occasionally involves DNA from other species, and to such a case standard coalescent theory cannot be applied. The purpose of this work was to develop a theoretical coalescent framework for prokaryotes, which allows for homologous recombination both within and between species (intra- and interspecific recombination). We also developed a simulation software, msPro, which will be made available online at http://www.sendou.soken.ac.jp/esb/innan/InnanLab/. A deeper understanding of the role of recombination revealed by this work will be useful for many applications and interpretations of population genetics-based analyzes in prokaryotes, including inferring demography, detecting footprints of selection, and genome-wide association studies (GWAS) (e.g., Shapiro et al. 2012; Rosen et al. 2015; Earle et al. 2016).

It should be noted that there are some existing simulators that allow for interspecific recombination, mainly based on the framework of clonal frame (Didelot and Falush, 2007; Didelot et al., 2010; Ansari and Didelot, 2014; Brown et al. 2016; De Maio and Wilson 2017). In these works, such

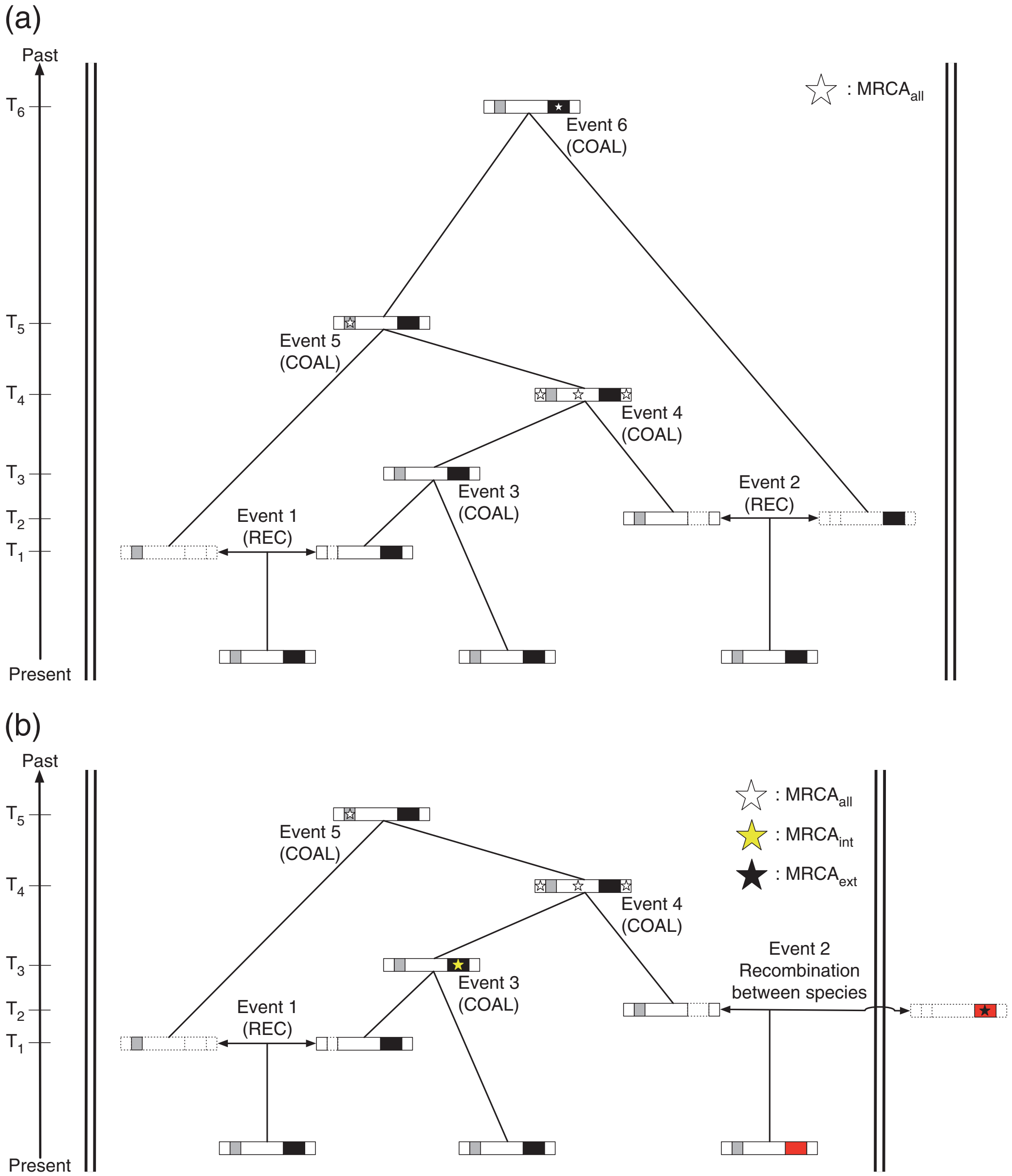

Fig. 1 **a** Ancestral recombination graph with intraspecific recombination. An example with sample size $n = 3$ is illustrated. The sampled three genomes are shown as long boxes, where regions with different histories are presented in different colors. The ancestral lineage splits into two by a recombination event (REC), while a pair of ancestral lineages merges by a coalescent event (COAL). The boxes with dashed lines represent dummy regions whose descendants are not represented in the sample. The two short regions in gray and black are transferred fragments from gene conversion-like recombination events. MRCA_{all} for each region is shown with a star. The white region has MRCA at T_4, the gray region has MRCA at T_5, and the black region has MRCA at T_6. **b** Ancestral recombination graph with intra- and interspecific recombination. The region transferred from an external source is shown in red. Color figure online

recombination was conducted by randomly replacing sequences with a certain level of divergence without a clear biological argument justification. However, interspecific recombination is not a simple process because it involves a number of biological factors, as described below. Thus, one of the aims of this work is to develop a biologically relevant framework to understand interspecific recombination.

Theoretical framework

Overview

We considered a sample prokaryote DNA sequences with length L bp from n haploid individuals, and traced its ancestral lineages backward in time. Figure 1a shows an

Table 1 List of mathematical symbols

N	Population size
n	Number of sample lineages
L	Length of the simulated region (base pairs)
μ	Mutation rate (per site per generation)
λ	Mean tract length of intraspecific recombination (base pairs)
g	Initiation rate of intraspecific recombination (per site per generation)
ξ	Mean tract length of interspecific recombination (base pairs)
h	Initiation rate of interspecific recombination (per site per generation)
d	Divergence to external DNA sequences
$Q_{\mathrm{int}}(z)$	Distribution of the length of transferred tract of a recombinant within species
$Q_{\mathrm{ext}}(d,z')$	Joint distribution of the divergence and the length of transferred tract of a recombinant between species

ancestral recombination graph under the standard coalescent process, in which all recombination is assumed to be homologous and intraspecific (Hudson 1983a; Griffiths and Marjoram 1996). Under these conditions, homologous recombination is analogous to meiotic gene conversion in the standard coalescent framework for diploid eukaryotes (McVean et al. 2002; Awadalla 2003). A coalescent event merges the ancestral lineages (e.g., events 3, 4, 5, and 6 in Fig. 1a), and homologous recombination separates the lineage into two (e.g., events 1 and 2 in Fig. 1a). Event 1 in Fig. 1a is a homologous recombination, in which a short fragment (denoted by a gray box) is integrated into the recipient genome; as a result, the ancestral lineage is separated into two: one for the recipient genome and another for the integrated fragment.

Then, following the standard treatment, we further traced the ancestral lineages until the lineages of all sampled chromosomes merged to their most recent common ancestor (MRCA). In this study, we defined $\mathrm{MRCA_{all}}$ as the MRCA of all sampled individuals, which is given for each site in focal region. Because of the presence of recombination, different parts of the focal region have different histories; thus, $\mathrm{MRCA_{all}}$ cannot be identical across the focal region. Moreover, different subregions cut by recombination should have their specific $\mathrm{MRCA_{all}}$. For example, $\mathrm{MRCA_{all}}$ for the black region appears at time T_6, whereas $\mathrm{MRCA_{all}}$ for the gray region appears at time T_5 and for the other white regions at time T_4 (Fig. 1a). The ancestral recombination graph contains all historical information for the entire focal region as illustrated in Fig. 1a. On the basis of this ancestral recombination graph, a SNP pattern can be simulated by randomly distributing point mutations on the graph. Thus, the standard coalescent treatment can be applied to prokaryotes undergoing intraspecific homologous recombination (McVean et al. 2002; Awadalla 2003).

A problem arises when DNA from other species is integrated by homologous recombination. Figure 1b illustrates such a case, in which event 2 is assumed to be an interspecific homologous recombination (i.e., the integration of DNA from another species), which is presented as a red box. In this case, the ancestral lineage of the integrated DNA originates from another species; therefore, it is not involved in the coalescent process of the focal species before time T_2. Standard coalescent cannot be applied to such a case. In this work, we propose a simple solution to this problem. Tracing the ancestral lineage of donor species (external source) should be terminated, and the coalescent process should be continued without considering such terminated lineages. Under this treatment, the concept of the MRCA of all sampled sequences ($\mathrm{MRCA_{all}}$) does not apply to regions that have undergone interspecific homologous recombination. The direct donor of the external DNA is called $\mathrm{MRCA_{ext}}$, referring to the most recent common ancestor of donor species. MRCA of the rest is referred to as $\mathrm{MRCA_{int}}$, referring to the most recent common ancestor of internal lineages. In event 2 in Fig. 1b, the gray and white regions that are not involved in the integration of external DNA can be traced back to $\mathrm{MRCA_{all}}$ (at T_5 and T_4, respectively), but we stop tracing the ancestral lineage of the red region at $>T_2$, and the origin of this region is treated as a $\mathrm{MRCA_{ext}}$. Thus, when a region has undergone a homologous recombination with an external source, the sampled sequences have two types of origins: one is $\mathrm{MRCA_{ext}}$ at T_2 as the origin of the red region (shown as a red box with star in Fig. 1b) and another is $\mathrm{MRCA_{int}}$ at T_3 as the origin of the rest (shown as a black box with yellow star in Fig. 1b.

In this study, we consider methods to simulate a SNP pattern in a region that has undergone interspecific homologous recombination. It should be noted that considering the mechanism of homologous recombination, we assumed a reasonable level of sequence identity between the external DNA and the focal species DNA. This indicates that the external lineage should eventually coalesce with the common ancestor of the focal species (on the timescale of species divergence). However, it is very difficult to estimate the probability distribution of time to such an eventual common ancestor, which could be far older than the MRCA of the focal species.

Alternatively, we developed an ad hoc treatment that does not require any unknown ancient demographic history up to species divergence. This treatment is based on a number of empirical studies that demonstrate the rate of successful integration of external DNA largely depends on the nucleotide divergence between the transferred fragment and the recipient sequence; this rate decays almost exponentially with increasing divergence, as reported by many authors (Fraser et al. 2007, and references therein). All symbols used in this paper are summarized in Table 1.

Intraspecific recombination

It is relatively easy to apply coalescent theory to intraspecifc recombination, as mentioned above. Based on previous studies (Wiuf and Hein 2000; McVean et al. 2002), if we assumed that homologous recombination is initiated at any position at a rate g (per site per generation), the elongation process would proceed such that the length of transferred tract z follows a geometric distribution, with mean tract length $= \lambda$ bp, as given by Eq. (1):

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

where $q = 1/\lambda$. This assumption is supported by empirical studies on transformation in many species including Helicobacter pylori (Lin et al. 2009), Streptococcus pneumonia (Croucher et al. 2012), and Haemophilus influenza (Mell et al. 2014). For mathematical convenience, we assumed the unidirectional elongation of conversion tract from 5′ to 3′, which has no quantitative effect on SNP pattern. Based on Eq. (1), the rate at which a region of L bp undergoes homologous recombination per generation is given by:

$$g' = R_{\mathrm{in}} + R_{\mathrm{left}}, \tag{2}$$

where R_{in} is the rate of gene conversion initiating inside the region and R_{left} is the rate initiating outside the region but ending within the observed region. R_{in} and R_{left} are given by gL and $\sum_{i=1}^{L} gQ_{\mathrm{int}}(z \geq i)$, respectively. Assuming all recombination is neutral, this rate (g') is identical to the backward recombination rate, which can be directly incorporated into the coalescent framework. The backward recombination rate per generation is defined as the rate at which a lineage undergoes recombination when a lineage is traced back for a single generation.

It is interesting to note that Eq. (2) does not include the probability that a recombination tract covers the entire simulated region. This is because such recombination simply causes a shift of a lineage to another lineage within the same population, which does not essentially affect the coalescent process. However, this does affect the process if the recombination event occurs with different species, as explained below.

As mentioned previously, there are three cellular mechanisms DNA, including transformation, transduction, and conjugation, through which homologous recombination can occur. These occur at different rates, and typical lengths of integrated tracts are different. Nevertheless, based on previous studies (Falush et al. 2001; McVean et al. 2002; Fearnhead et al. 2005; Jolley et al. 2005; Didelot and Falush 2007), we selected a simplified treatment such that all three mechanisms are summarized by a single-backward recombination rate and assumed tract lengths to follow a geometric distribution. We derived $Q_{\mathrm{int}}(z)$ when the three mechanisms are separately considered, as discussed in Supplementary Information.

Intraspecific homologous recombination events in prokaryotes have been easily incorporated in the standard coalescent framework (Wiuf and Hein 2000; McVean et al. 2002; Fearnhead et al. 2005; Jolley et al. 2005). That is, when tracing the ancestral lineage of a certain sequence of L bp, the process assumes either a coalescence or recombination event, with recombination rate (per generation) given by Eq. (2) and coalescence rate given by ${}_{n}\mathrm{C}_2/\mathrm{N}$, where N is the population size and n is the number of lineages.

It should be noted that this simple process can be applied to a single population, in which both coalescence and recombination randomly occur between individuals in the population, and it is easy to incorporate population structure and demographic history into this framework, as incorporated in eukaryote cases. The main differences between the eukaryote and prokaryote cases include factors shaping population structure. In eukaryotes, limited migration between geographic barriers is the major cause, and this also applies to prokaryotes although the situation is more complex. For example, subpopulations of infectious species may form based on host individuals. In addition, there are two major classes of isolation in prokaryotes; ecological and genetic isolation (e.g., Cohan 2002a, b; Lawrence 2013), as summarized in Supplementary Information.

Interspecific recombination

We used the same argument as of backward recombination preciously mentioned. We defined 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 undergoes an interspecific recombination that is initiated at the focal site (the

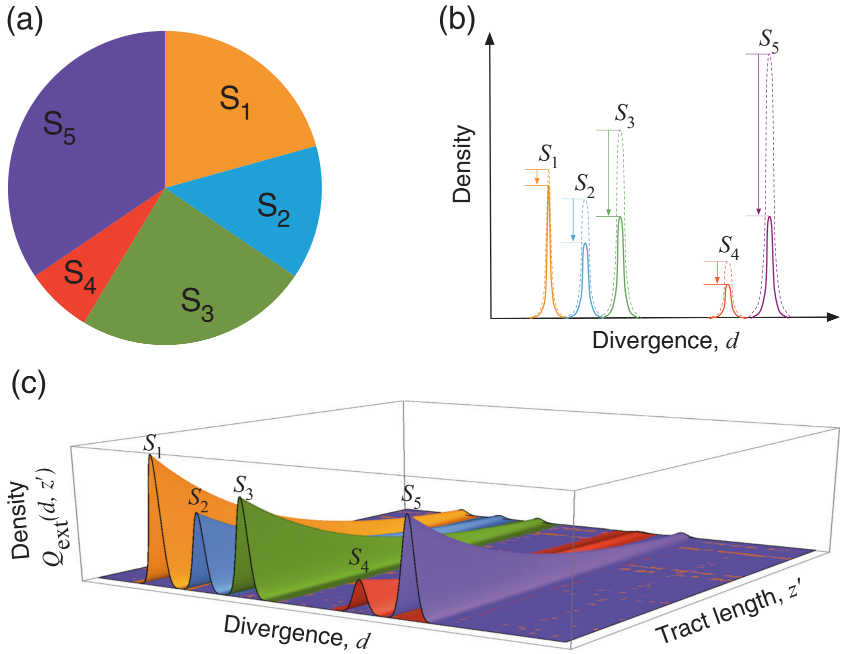

Fig. 2 Illustrating a hypothetical environment with five different species (S_1–S_5) coexisting with the focal species, S_0. **a** The proportion of the five species in the environment. **b** Density distribution of the divergence of external DNA from the focal species. The dashed line represents the actual distribution in the environment, and the solid line represents the weighted distribution taking into account the probability of successful integration (see text for details). **c** The joint density distribution of *d* and successfully integrated tract length (z'). Color figure online

same definition as *g* except for the source of the integrated DNA). Based on *h*, we can compute h', the recombination rate for the simulated region, using an equation similar to Eq. (2) (as given below). With this rate specified, it is very easy to incorporate interspecific homologous recombination into the coalescent framework. That is, when tracing a lineage of a sequence of *L* bp the process considers the next event, including coalescence, intraspecific recombination, or interspecific recombination, with relative backward rates, ${}_nC_2/N$, g', and h', respectively. If interspecific recombination occurs, the length of a transferred region is randomly determined (as given below). Then, the transferred region is replaced by a sequence representing the donor. Thus, the process can be well merged with the backward coalescent treatment; it should be noted that the biological interpretation of *h*, the backward rate interspecific recombination, should be considered carefully, as explained below.

In order to define *h*, we considered the coalescent process of a particular species (population) that coexists with a number of other species. The focal species potentially undergoes recombination with these species, and the rate of such recombination would be determined by a number of genetic and ecological factors, as mentioned above. Figure 2 illustrates the hypothetical situation of a certain species, S_0, that coexists with five other species (S_1–S_5); their proportion is presented as a pie-chart (Fig. 2a). The five species are ordered on the basis of their divergence (*d*) from S_0, which might follow some distributions as illustrated in Fig. 2b. The dashed lines in Fig. 2b represent the density distributions of the divergence of DNA sequences that could recombine with the focal species. As mentioned above, the rates of successful integration of these DNA sequences into the focal species would vary depending on the species because of genetic and ecological barriers against recombination. Furthermore, even in the case of successful recombination, the integrated DNA may be deleterious to the host species and can be immediately selected out of the population. The distributions represented by solid lines in Fig. 2b take these effects into account, and the degree of reduction for each species is indicated by an arrow. Noting that the tract length of homologous recombination follows an approximate geometric distribution, we obtained $Q_{ext}(d, z')$, which is the joint distribution of *d* and the successfully integrated tract length (z'), as illustrated in Fig. 2c. We defined *h* as the rate of successful interspecific recombination per site. This value (*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 assumed that successfully incorporated foreign DNA is neutral in the population of the focal species. Under this condition, interspecific recombination can be simply incorporated into the coalescent framework as described above. That is, interspecific recombination at a rate h' is included together with coalescence and intraspecific recombination at rates ${}_nC_2/N$ and g', respectively. When interspecific recombination occurs, the tract length (z') and nucleotide divergence within the tract (*d*) can be determined as a random variable from Q_{ext} (d, z').

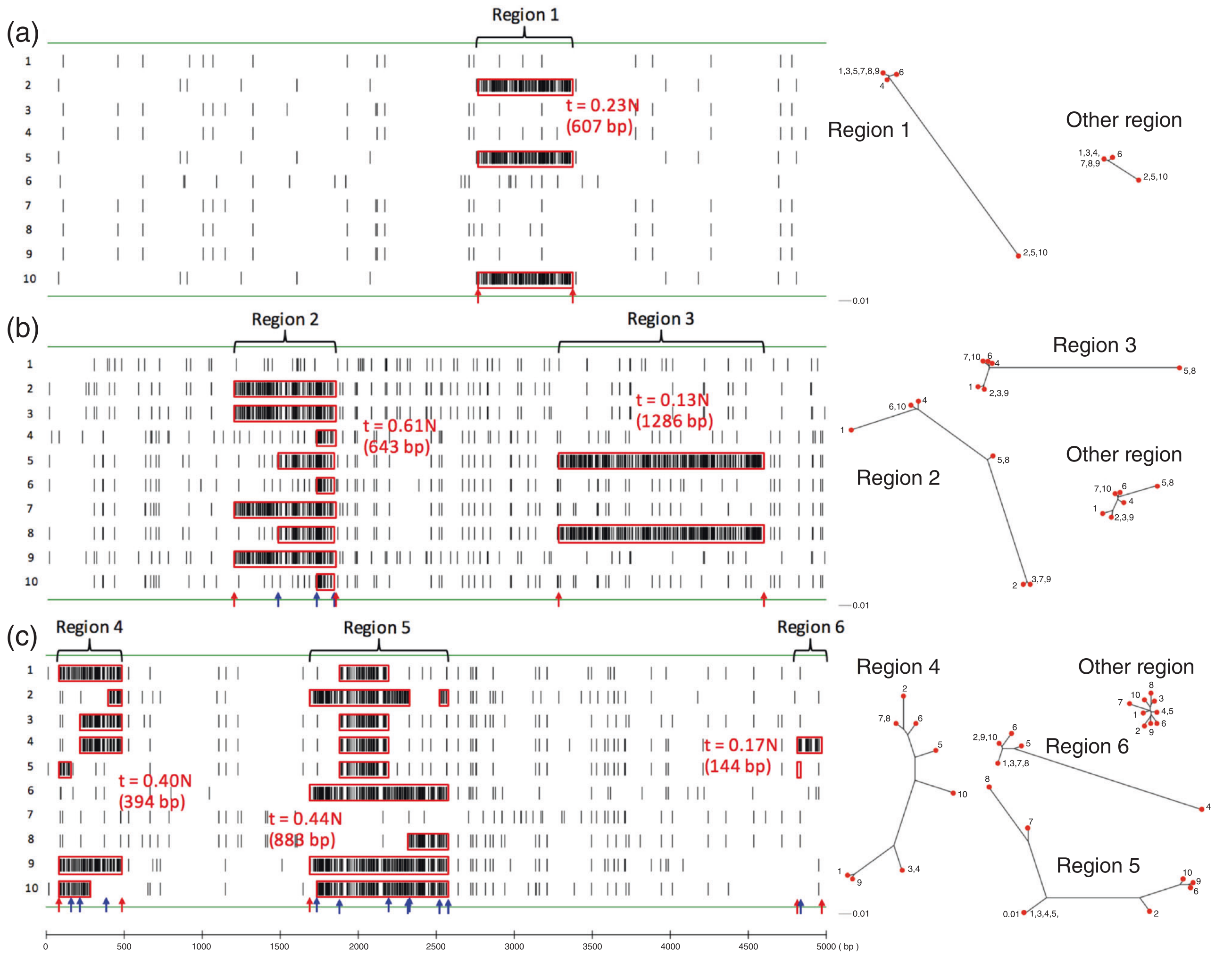

Fig. 3 Typical SNP patterns with interspecific recombination with no intraspecific recombination (**a**; $2Ng = 0$), with a moderate level of intraspecific recombination (**b**; $2Ng = 0.001$), and with a high recombination rate (**c**; $2Ng = 0.005$). Vertical bars indicate the locations of point mutations in the simulated region with $L = 5000$ bp. The regions that underwent interspecific recombination are specified (Regions 1–6), and neighbor-jointing trees for these regions are shown in comparison with other regions with no interspecific recombination. The breakpoints of interspecific recombination events are shown as red arrows; intraspecific recombination events are shown as blue arrows, and events integrating fragmented foreign DNA are shown as red boxes. Color figure online

The computation of h' from h is slightly different from that employed for the treatment of intraspecific recombination (Eq. (2)) because we cannot ignore the recombination event that encompassed the entire simulated region. Thus, h' is given by the following Eq. (3):

$$h' = R_{\text{in}}^{h} + R_{\text{left}}^{h} + R_{\text{all}}^{h}, \tag{3}$$

where R_{in}^{h} is the rate of recombination initiating inside the region, R_{left}^{h} is the rate initiating outside the region but ending within the focal sequence, and R_{all}^{h} is the rate initiating in the 5′ upstream of the region but ending in the 3′ downstream of the region. R_{in}^{h}, R_{left}^{h}, and R_{all}^{h} are given by hL, $\sum_{i=1}^{L} hQ_{\text{ext}}(z' \geq i)$, and $\sum_{i=L+1}^{\infty} hQ_{\text{ext}}(z' \geq i)$, respectively, where $Q_{\text{ext}}(z')$ is calculated by integrating the joint probability distribution ($Q_{\text{ext}}(d, z')$) over d. Then, h' is given as:

$$h' = h\left(L + \sum_{i=1}^{\infty} Q_{\text{ext}}(z' \geq i)\right). \tag{4}$$

Mutation

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

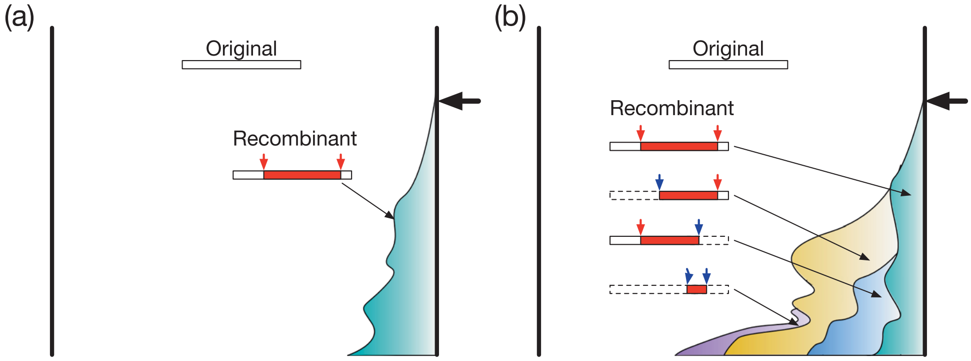

Fig. 4 Cartoon of the typical behavior of population frequency of foreign DNA, **a** without intraspecific recombination and **b** with intraspecific recombination. The time of foreign DNA integration into the population, denoted by a thick black arrow, producing a recombinant haplotype in which the integrated DNA is specified with a red box and arrows. **a** When there is no intraspecific recombination, the population consists of two haplotypes, the original and recombinant haplotype. **b** When intraspecific recombination is involved, the integrated DNA could be fragmented by recombination, thereby creating various types of recombinant haplotypes each of which should have only a part of the integrated DNA. Additional breakpoints by intraspecific recombination are shown as blue arrows. In such a case, the number of individuals containing at least part of the integrated DNA is much larger than the case with no recombination, and the length of integrated DNA is shorter. Color figure online

Results and discussion

We performed simulations to generate a number of SNP patterns to examine the effect of interspecific homologous recombination. The mutation rate $\theta = 0.01$ was constant throughout the simulations. For intraspecific recombination, the mean tract length was constant at $\lambda = 1000$ bp but initiation rate of recombination (g) varied. We first considered a relatively low interspecific recombination rate ($2Nh = 0.00005$). We used a simplified assumption to demonstrate the point, that is, the average divergence to external DNA was constant at 20% and the tract length followed a geometric distribution with a constant mean ξ (i.e., $Q_{\mathrm{ext}}(d = 0.2, z') = \xi^{-1}(1 - \xi^{-1})^{z'-1}$). Typical SNP patterns generated from the simulations with $n = 10$ and $L =$ 5000 bp are shown in Fig. 3. The positions of SNPs are shown as solid vertical lines along the simulated region. In Fig. 3a, no intraspecific recombination ($2Ng = 0$) is assumed. One interspecific recombination (607 bp) occurred $t = 0.23N$ generations ago on the ancestral lineages of individuals 2, 5, and 10, and the positions of two breakpoints of the recombination event are shown with red arrows. The region that originates from foreign DNA can be clearly recognized as a cluster of SNPs due to large divergence ($d = 0.2$, 20 times larger than the population mutation rate θ). This region is referred to as Region 1 and shown as red box. The neighbor-jointing tree for this region is completely different from that for the other region; individuals 2, 5, and 10 are highly diverged from the other seven individuals in Region 1.

In our simulation results, polymorphism increased as interspecific recombination events increased. As shown in Supplementary Fig. S2a, polymorphism increases with increased initiation rate (h), mean tract length (ξ), and divergence (d), where polymorphism is measured by π, which indicates the average number of nucleotide differences per site. This result agrees with the theoretical prediction that the expectation of π is given by a simple function of h, ξ, and d:

$$\frac{\theta + \phi}{1 + 2(\theta + \phi)}, \tag{5}$$

where $\phi = 2Nh\xi d$ (Supplementary Information for the derivation). It is evident that the most important parameter is the product of three recombination-associated parameters, $h\xi d$, which represents the probability that the allelic state at a single site is inverted by recombination, as Supplementary Fig. S2b demonstrates that π is approximately given by a simple liner function of $h\xi d$.

In Fig. 3b, a moderate intraspecific recombination is introduced ($2Ng = 0.001$). Two external DNA fragments are integrated (Regions 2 and 3). In Region 2, a 643 bp fragment of foreign DNA was integrated $t = 0.61N$ generations ago. It should be noted that while individuals 2, 3, 7, and 9 contain the entire fragment, only a part of the integrated fragment is observed in individuals 4, 5, 6, 8, and 10. This is due to intraspecific recombination that occurred after the integration; that is, the integrated fragment was cut into smaller fragments and distributed into the population. Such intraspecific recombination events were detected on the basis of simulated ancestral recombination graph (Fig. 3b, blue arrows indicate breakpoints). By contrast, due to the recent origin of Region 3 ($t = 0.13N$ generations), no intraspecific recombination occurred and the entire integrated region (1286 bp) remained intact in individuals 5 and 8, similar to Region 1 in Fig. 3a.

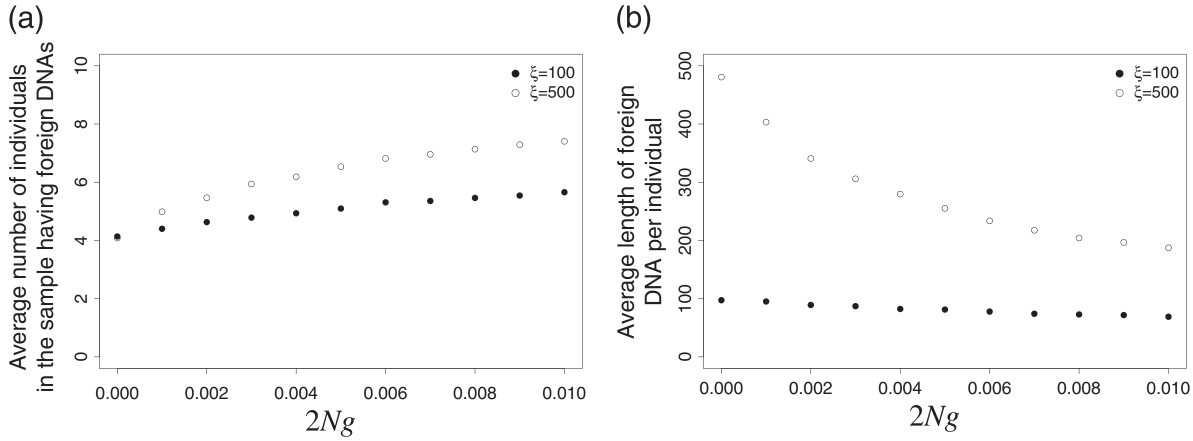

Fig. 5 The effect of intraspecific recombination on **a** average number of individuals in the sample containing foreign DNA and **b** average length of integrated foreign DNA, from 10,000 simulations runs with $n = 15$ and $L = 10{,}000$ bp. Results up to $2Ng = 0.01$ are shown here. For $2Ng$ 0.01, these increasing/decreasing trends last monotonically (data not shown)

With an even higher intraspecific recombination rate ($2Ng = 0.005$) in Fig. 3c, fragmentation is more apparent. Three regions underwent interspecific recombination (Regions 4, 5, and 6), and all of them underwent intraspecific recombination. An intriguing pattern exists in Region 5, where only a part of the integrated fragment was observed in the sample. The recombination occurred $t = 0.44N$ generations ago. The actual length of the integrated foreign fragment was ≥883 bp; however, none of the sampled ten individuals contained the 5′ breakpoint. This process is described by the cartoon in Fig. 4, which illustrates the typical behavior, in terms of population frequency, of a foreign DNA fragment integrated at time 0 with and without intraspecific recombination. Without intraspecific recombination, the entire integrated DNA can be vertically transmitted to the following generations (Fig. 4a). By contrast, with intraspecific recombination, the integrated DNA is fragmented into fragments of various lengths (Fig. 4b). Consequently, more individuals may contain a part of the integrated DNA, but the average length of the integrated DNA in each individual would be shorter: some might lose the 5′ breakpoint and some might contain only a short region in the middle. One potential caveat when interpreting data is that, even when there was only one interspecific recombination event involved, the data might look as if multiple interspecific events have incorporated foreign DNA independently. Indeed, when applied to one of our simulated datasets, GENECONV (Sawyer 1989), a commonly used software to detect gene conversion tracts, identified a number of gene conversion tracts surrounding the region that had underwent a single interspecific recombination event (Supplementary Fig. S3), indicating this region to be a hotspot of integration.

Given this effect of interspecific recombination, we predicted that with increased intraspecific recombination rate, the number of individuals with foreign DNA would increase and the average length of foreign DNA fragments would decrease. Our simulations (Fig. 5) quantitatively demonstrate our prediction. The number of individuals with some proportion of foreign DNA increased as intraspecific recombination rate ($2Ng$) increased (Fig. 5a), whereas the average length of each foreign DNA in the sample decreased (Fig. 5b). This effect of intraspecific recombination ($2Ng$) is larger when ξ is larger. These findings would be useful for improving algorithms to identify genomic regions that have undergone homologous recombination (Didelot and Falush 2007; Didelot et al. 2009; Ansari and Didelot 2014; Yahara et al. 2014).

We thus demonstrated that the joint processes of intra- and interspecific recombination can generate complex SNP patterns, and a better theoretical understanding of these effects is required to improve the interpretation of SNP data in prokaryotes. Considering that interspecific recombination is rather common in prokaryotes and that strong impact this event on SNP patterns, as we have shown in this study, it is important to avoid misleading interpretations of data when examining recombination events, which could potentially result in the misevaluation of the relative contribution of demography and selection.

In this study, we developed a fast simulator for producing a number of SNP iterations by incorporating both intra- and interspecific recombination. We developed a software, msPro, on the basis of Hudson's commonly used software ms (msPro indicates ms for prokaryotes). The input commands for msPro are very similar to those for ms, and the form of output is exactly the same as ms, assuming a finite-site model (not for an infinite-site model). msPro can incorporate various forms of demographic history as ms can. It should be noted that our simulator runs after specifying the density distribution of external DNA, Q_{ext}. When there is no prior knowledge regarding external DNA, it is difficult to set Q_{ext}. Considering such a case, the default settings for msPro are provided in Supplementary Information.

Acknowledgements The authors very much thank the three reviewers for their useful and kind comments that improved the manuscript. This work was supported in part by the Japan Society for the Promotion of Science (JSPS).

Compliance with ethical standards

Conflict of interest The authors declare that they have no competing interests.

References

Ansari MA, Didelot X (2014) Inference of the properties of the recombination process from whole bacterial genomes. Genetics 196:253–265

Awadalla P (2003) The evolutionary genomics of pathogen recombination. Nat Rev Genet 4:50–60

Azad RK, Lawrence JG (2012) Detecting laterally transferred genes. In: Anisimova M (ed.) Evolutionary Genomics, Humana Press, Methods in Molecular Biology, Clifton, NJ, p 281–308

Brown T, Didelot X, Wilson DJ, De Maio N (2016) SimBac: simulation of whole bacterial genomes with homologous recombination. Microb Genom 2:1–6

Cohan FM (2002a) Sexual isolation and speciation in bacteria. Genetica 116:359–370

Cohan FM (2002b) What are bacterial species? Annu Rev Microbiol 56:457–487

Cornejo OE, Lefébure T, Bitar PDP, Lang P, Richards VP, Eilertson K et al. (2013) Evolutionary and population genomics of the cavity causing bacteria *Streptococcus mutan*s. Mol Biol Evol 30:881–893

Croucher NJ, Harris SR, Barquist L, Parkhill J, Bentley SD (2012) A high-resolution view of genome-wide pneumococcal transformation. PLoS Pathog 8:e1002745

De Maio N, Wilson DJ (2017) The bacterial sequential Markov coalescent. Genetics 206:333–343

Didelot X, Falush D (2007) Inference of bacterial microevolution using multilocus sequence data. Genetics 175:1251–1266

Didelot X, Lawson D, Falush D (2009) SimMLST: simulation of multi-locus sequence typing data under a neutral model. Bioinformatics 25:1442–1444

Didelot X, Lawson D, Darling A, Falush D (2010) Inference of homologous recombination in bacteria using whole-genome sequences. Genetics 186:1435–1449

Didelot X, Maiden MCJ (2010) Impact of recombination on bacterial evolution. Trends Microbiol 18:315–322

Dobrindt U, Hochhut B, Hentschel U, Hacker J (2004) Genomic islands in pathogenic and environmental microorganisms. Nat Rev Microbiol 2:414–424

Donati C, Hiller NL, Tettelin H, Muzzi A, Croucher NJ, Angiuoli SV et al (2010) Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. Genome Biol 11:R107

Donnelly P, Kurtz TG (1999) Genealogical processes for fleming-viot models with selection and recombination. Ann Appl Probab 9:1091–1148

Doolittle WF, Papke RT (2006) Genomics and the bacterial species problem. Genome Biol 7:e116

Earle SG, Wu CH, Charlesworth J, Stoesser N, Gordon NC, Walker TM et al. (2016) Identifying lineage effects when controlling for population structure improves power in bacterial association studies. Nat Microbiol 1:16041

Falush D, Kraft C, Taylor NS, Correa P, Fox JG, Achtman M et al. (2001) Recombination and mutation during long-term gastric colonization by *Helicobacter pylor*i: estimates of clock rates, recombination size, and minimal age. Proc Natl Acad Sci USA 98:15056–15061

Fearnhead P, Smith NGC, Barrigas M, Fox A, French N (2005) Analysis of recombination in *Campylobacter jejuni* from MLST population data. J Mol Evol 61:333–340

Fraser C, Hanage WP, Spratt BG (2007) Recombination and the nature of bacterial speciation. Science 315:476–480

Griffiths RC, Marjoram P (1996) Ancestral inference from samples of DNA sequences with recombination. J Comput Biol 3:479–502

Hartl DL, Clark AG (2007) Principles of population genetics. Sinauer Associates, Sunderland

Haven J, Vargas LC, Mongodin EF, Xue V, Hernandez Y, Pagan P et al (2011) Pervasive recombination and sympatric genome diversification driven by frequency-dependent selection in *Borrelia burgdorferi*, the lyme disease bacterium. Genetics 189:951–966

Hudson RR (1983a) Properties of a neutral allele model with intragenic recombination. Theor Popul Biol 23:183–201

Hudson RR (1983b) Testing the constant-rate neutral allele model with protein sequence data. Evolution 37:203–217

Hudson RR (1990) Gene genealogies and the coalescent process. Oxf Surv Evol Biol 7:1–43

Hudson RR (2001) Two-locus sampling distributions and their application. Genetics 159:1805–1817

Hudson RR (2002) Generating samples under a wright-fisher neutral model of genetic variation. Bioinformatics 18:337–338

Jolley KA, Wilson DJ, Kriz P, McVean G, Maiden MCJ (2005) The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in *Neisseria meningitidis*. Mol Biol Evol 22:562–569

Kingman JF (1982) On the genealogy of large populations. J Appl Probab 19:27–43

Krone SM, Neuhauser C (1997) Ancestral processes with selection. Theor Popul Biol 51:210–237

Lawrence JG (2013) Gradual speciation: Further entangling the tree of life. In: Gophna U (ed.) Lateral Gene Transfer in Evolution. Springer, New York, NY, p 243–262

Lin EA, Zhang XS, Levine SM, Gill SR, Falush D, Blaser MJ (2009) Natural transformation of *Helicobacter pylori* involves the integration of short dna fragments interrupted by gaps of variable size. PLoS Pathog 5:e1000337

Majewski J, Cohan FM (1998) The effect of mismatch repair and heteroduplex formation on sexual isolation in Bacillus. Genetics 148:13–18

McVean G, Awadalla P, Fearnhead P (2002) A coalescent-based method for detecting and estimating recombination from gene sequences. Genetics 160:1231–1241

Mell JC, Lee JY, Firme M, Sinha S, Redfield RJ (2014) Extensive cotransformation of natural variation into chromosomes of naturally competent *Haemophilus influenzae*. G3 4:717–731

Nordborg M (2001) Coalescent theory. In: Balding DJ, Bishop M, Cannings C (eds.) Handbook of statistical genetics. Wiley-Blackwell, Chichester, UK

Ochman H, Lawrence JG, Groisman EA (2000) Lateral gene transfer and the nature of bacterial innovation. Nature 405:299–304

Pérez-Losada M, Browne EB, Madsen A, Wirth T, Viscidi RP, Crandall KA (2006) Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. Infect Genet Evol 6:97–112

Rosen MJ, Davison M, Bhaya D, Fisher DS (2015) Fine-scale diversity and extensive recombination in a quasisexual bacterial population occupying a broad niche. Science 348:1019–1023

Sawyer S (1989) Statistical tests for detecting gene conversion. Mol Biol Evol 6:526–538

Shapiro BJ, Friedman J, Cordero OX, Preheim SP, Timberlake SC, Szabó G et al. (2012) Population genomics of early events in the ecological differentiation of bacteria. Science 336:48–51

Shen P, Huang HV (1986) Homologous recombination in *Escherichia coli*: dependence on substrate length and homology. Genetics 112:441–457

Snyder L, Peters JE, Henkin TM, Champness W (2013) Molecular genetics of bacteria. ASM Press, Washington, DC

Tajima F (1983) Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460

Takuno S, Kado T, Sugino RP, Nakhleh L, Innan H (2012) Population genomics in bacteria: a case study of *Staphylococcus aureus*. Mol Biol Evol 29:797–809

Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P et al (2009) Organised genome dynamics in the *Escherichia col*i species results in highly diverse adaptive paths. PLoS Genet 5: e1000344

Wakeley J (2008) Coalescent theory: An introduction. Roberts and Company, Greenwood Village, Colorado

Wirth T, Falush D, Lan R, Colles F, Mensa P, Wieler LH et al. (2006) Sex and virulence in *Escherichia coli*: an evolutionary perspective. Mol Microbiol 60:1136–1151

Wiuf C, Hein J (2000) The coalescent with gene conversion. Genetics 155:451–462

Yahara K, Didelot X, Ansari MA, Sheppard SK, Falush D (2014) Efficient inference of recombination hot regions in bacterial genomes. Mol Biol Evol 31:1593–1605