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.

The coalescent is a population genetic theory, which considers the evolutionary process backward in time (Kingman 1982; Hudson 1983b; Tajima 1983). The coalescent theory has been mainly developed by assuming its application to higher eukaryotes, perhaps due to a historical reasons: The major model species of population genetics have been higher eukaryotes such as Drosophila and human (*e.g.,* Hartl and Clark 2007). The coalescent provides an extremely powerful simulation tool for analyzing the pattern of single nucleotide polymorphisms (SNPs) in sampled sequences. It is flexible enough to incorporate major evolutionary processes including random genetic drift, mutation, recombination and demographic history (*e.g.,* Hudson 1990; Nordborg 2001; Wakeley 2008), whereas it is not very straightforward to incorporate complex modes of selection (but see Krone and Neuhauser 1997; Neuhauser and Krone 1997; Donnelly and Kurtz 1999; Fearnhead 2006). **ms** is one of the most popular coalescent simulators, which allows to produce patterns of neutral SNPs under various settings of demography (Hudson 2002). It incorporates two major outcomes of meiotic recombination, that is, meiotic crossing-over and gene conversion.

Prokaryotes are unique in that they are haploids and do not undergo meiosis, and therefore their recombination mechanisms are quite different from that of meiotic recombination in eukaryotes. Nevertheless, the coalescent can work with prokaryotes with a relatively simple modification: Recombination is treated as an event analogous to meiotic gene conversion because a prokaryote's circular chromosome needs double "crossing-over" to exchange a DNA fragment. This modification can well explain the nature of prokaryotes' homologous recombination as we will explain below. The application of the coalescent theory to bacteria became particularly popular since McVean *et al.* developed the software **LDhat** (McVean *et al.* 2002) for estimating the recombination rate, which is a modified version of Hudosn's composite likelihood method (Hudson 2001). Because **LDhat** allows 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;

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

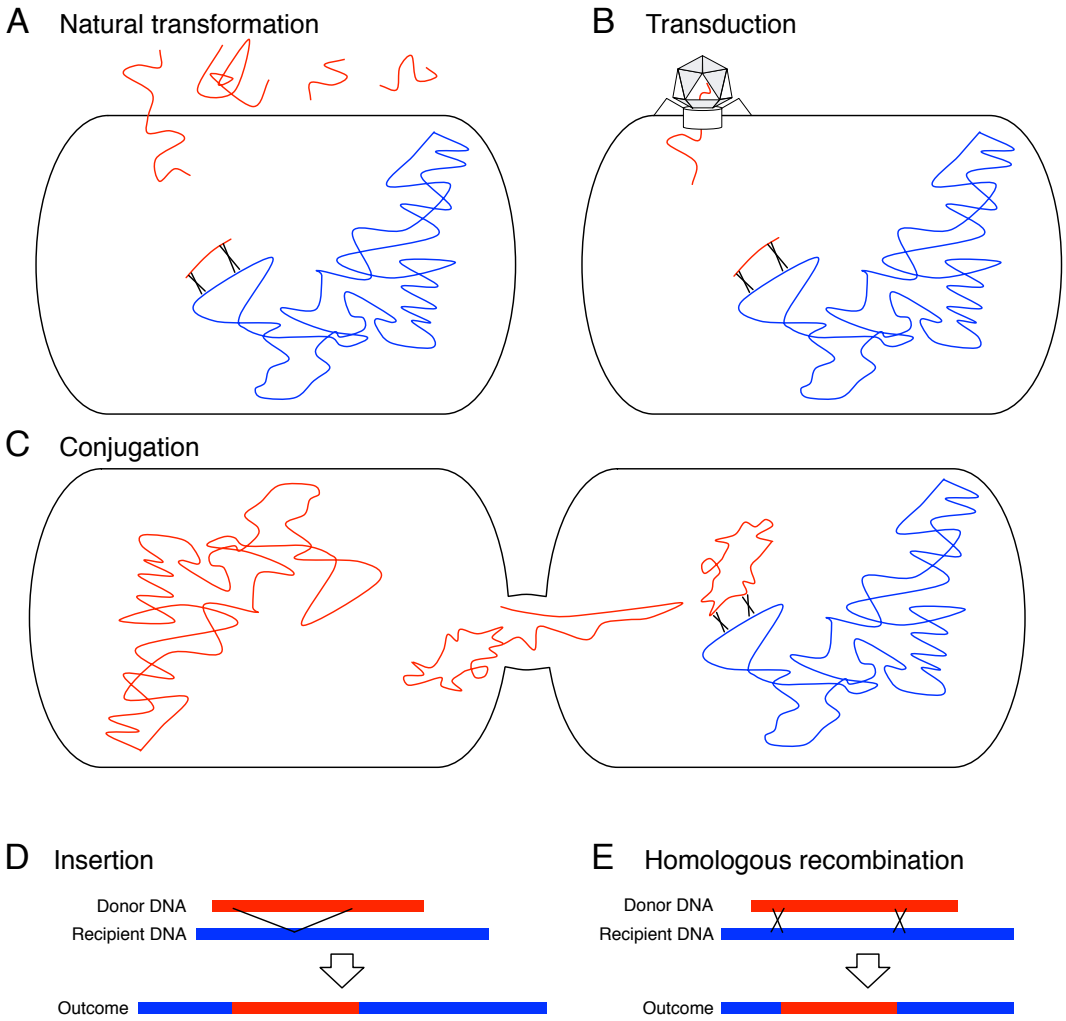

Figure 1 Three major mechanisms for prokaryotes to integrate external DNA: natural transformation (A), transduction (B), and conjugation (C). The host genome and external DNA are presented in blue and red, respectively. (D, E) Two outcomes of recombination via insertion (D) and homologous recombination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

Rosen *et al.* 2015).

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

Prokaryotes undergo recombination by incorporating DNA outside of the cell through three major mechanisms: natural transformation, transduction, and conjugation, as illustrated in Figure 1 (*e.g.,* Snyder *et al.* 2013). Natural transformation is a process involving direct uptake of a free extracellular DNA and the integration under natural bacterial growth conditions (Figure 1A). Transduction is a process in which bacterial DNA is introduced into the other bacteria through infection by a phage containing the DNA (Figure 1B). Conjugation is the transfer of DNA from one bacterial cell to another by the transfer functions of a self-transmissible DNA elements, frequently associated with plasmids (Figure 1C).

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

which markers are distributed through the genome to distinguish from those originating external source (*i.e.,* different species). In some bacterial species, such as *Neisseria gonorrhoeae* and *Haemophilus influenzae*, efficient natural transformation requires the presence of short sequence motifs (∼10bp), called as DNA uptake sequences (DUS) or uptake signal sequences (USS), that is interspersed among the genome, which may prevent the incoming DNA from different species (or strains) integrating into their genomes. Phage defense mechanisms may also work against incoming DNA from external source. The restriction-modification system is a common mechanism or degrading DNA that is not properly modified (*e.g.,* through DNA methylation), which have been identified in ∼90% of prokaryote species (Roberts *et al.* 2010). Clustered, regularly interspaced short palindromic repeat (CRISPR) loci and their associated proteins (Cas) are found in the genomes of ∼90% of archaea and ∼50% of eubacteria (Grissa *et al.* 2007; Rousseau *et al.* 2009). These sequences are separated by short sequences of DNA (23–50 bp) known as spacers, most of which exhibit homology to previously encountered phage or plasmid genomes, suggesting that these loci provide memory for the bacteria to prevent repeated incoming encounters.

In addition, when these mechanisms do not work perfectly, there is another round of screening process to prevent integration of external DNA through homologous recombination (Majewski 2001). For example, recombination requires near identical regions (*e.g.,* monitored by RecA mediated homology search), (Shen and Huang 1986; Majewski and Cohan 1998) so that external DNA has less chance to be integrated. In addition, it is also pointed out that the mismatch repair system is effective in preventing recombination between highly mismatched sequences (Claverys and Lacks 1986; Majewski 2001; Overballe-Petersen *et al.* 2013). Thus, there are a number of molecular mechanisms to prevent integrating external DNA to the host genome. Nevertheless, it has been repeatedly demonstrated that prokaryote genomes undergo recombination, not only within the same species but also with different species (reviewed in Majewski 2001).

There are two possible outcomes of recombination as shown in Figures 1D and E (see Lawrence 2013, for a review). One is that the incorporated DNA is inserted into the genome (Fig. 1D), and the other is that the incorporate DNA is exchanged with its homologous part of the genome if any (Fig. 1E). The former is known as horizontal gene transfer or lateral gene transfer, and its evolutionary role is emphasized when a novel gene is acquired and contributes to adaptation (Ochman *et al.* 2000; Dobrindt *et al.* 2004; Fraser *et al.* 2009; Polz *et al.* 2013), although the frequency and importance of such illegitimate recombination is under debate (de Vries *et al.* 2001; Shapiro *et al.* 2012). The latter is known as homologous recombination, and it usually involves DNAs from the same species because the near-identity requirement of the RecA mediated homology search criteria is easily satisfied, whereas it is also possible that DNA from different species is integrated as long as it retains some homology. Homologous recombination between different species sometimes remains unique patterns of SNPs, from which we can search for their footprints in the sequence data (reviewed in Awadalla 2003; Didelot and Maiden 2010; Azad and Lawrence 2012; Nakhleh 2013).

The focus of this article is the latter, homologous recombination. Considering the mechanism of homologous recombination involving double crossing-over, the outcome is similar to meiotic gene conversion. Therefore, as mentioned earlier, the standard coalescent has been commonly applied for analyzing patterns of SNPs in bacteria with a simple modification in the setting; the rate of crossing-over is set to zero, so that all recombination events (*i.e.,* homologous recombination) are treated as if they are meiotic gene conversion. This application should be reasonable as long as the donor of homologous recombination is always another individual in the same species. However, it is well known that homologous recombination occasionally involves DNA from other species, and this is the case that the standard coalescent cannot handle. The purpose of this work is to develop the theoretical framework of the coalescent for prokaryotes, which allows homologous recombination both within and between species. We also developed a simulation software named **msPro**, which will be available at http://www.sendou.soken.ac.jp/esb/innan/InnanLab/).

Theoretical Framework

Overview: Consider a sample of prokaryote DNA sequences with length L bp from n haploids, and trace their ancestral lineages backward in time. Figure 2A shows an example of an ancestral recombination graph under the standard coalescent, in which all recombination is assumed to be homologous recombination within the same species (Hudson 1983a; Griffiths and Marjoram 1996). Under this setting, the process is analogous to meiotic gene conversion in the standard coalescent 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. 2A), and homologous recombination separates the lineage into two (*e.g.,* event 1, and 2 in Fig. 2A). For example, event 1 in Fig. 2A is a homologous recombination, in which a short fragment (presented by a gray box) is integrated into the recipient genome, so that the ancestral lineage is separated into two; one for the recipient genome and the other is for the integrated fragment. Then, following the standard treatment, we further trace their ancestral lineages until the lineages of all sampled chromosome merge to their MRCA (Most Recent Common Ancestor), which is referred to as MRCA_{all} in this article. It should be noted that with the presence of recombination, different parts of the region have different histories, so that MRCA_{all} cannot be identical across the region; different subregions chopped

by recombination should have their specific MRCA_{all}. For example, MRCA_{all} for the black region appears at time T_6, while that for the gray region is at time T_5 and for the other white regions are at T_4 (Fig. 2A). The ancestral recombination graph has all historical information for the entire regions as illustrated in Fig. 2A. With this ancestral recombination graph, a pattern of SNPs can be simulated by randomly distributing point mutations on the graph. Thus, the standard coalescent treatment works for prokaryotes with homologous recombination within species (McVean *et al.* 2002; Awadalla 2003).

The problem is when DNA from other species is integrated by homologous recombination. Figure 2B illustrates such a situation, in which event 2 is assumed to be a homologous recombination event from external source (*i.e.,* integration of DNA from other species), which is presented in a red box. In this case, the ancestral lineage of the transferred DNA originates from outside of the focal species, so that it is not involved in the coalescent process of the focal species before time T_2. This is the situation that the standard coalescent cannot handle. We here propose a simple solution to this problem: Tracing the ancestral lineage of 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 (MRCA_{all}) does not apply to such a region that experienced homologous recombination from external source. The direct donor of the external DNA is called MRCA_{ext}, most recent common ancestor from external source, and MRCA of the rest is referred to as MRCA_{int}, most recent common ancestor of internal lineages. In event 2 in Figure 2B, while the gray and white regions that are not involved in the integration of external DNA can be traced back to MRCA_{all} (at T_5 and T_4, respectively), we may stop tracing the ancestral lineage of the red region at $> T_2$, and the origin of this region is treated as a MRCA_{ext}. Thus, when a region experienced a homologous recombination from external source, the sampled sequences have two kinds of origins, one is MRCA_{ext} at T_2 as the origin of the red part (shown by a red box with a star in Figure 2B) and the other is MRCA_{int} at T_3 as the origin of the rest, shown by a black box with a yellow star in Figure 2B.

We here consider how to simulate a pattern of SNPs in such a region that experienced homologous recombination with external source. Note that provided the mechanism of homologous recombination, we assume a reasonable level of sequence identity between the external DNA and the focal species. This means that the external lineage should eventually coalesce with the common ancestor of the focal species (on the time scale of species-divergence). However, it is very difficult to know the probability distribution of the time to such eventual common ancestor, which could be far older than the MRCA of the focal species.

Alternatively, we develop an ad-hoc treatment that does not require any unknown ancient demographic history up to species divergence. The idea of our treatment is based on a number of empirical demonstrations that the rate of successful integration of external DNA heavily depends on the nucleotide divergence between the transferred fragment and the recipient sequence; the rate decays almost exponentially with increasing divergence as demonstrated by many authors (Albritton *et al.* 1984; Roberts and Cohan 1993; Vulić *et al.* 1997; Zahrt and Maloy 1997; Lorenz and Sikorski 2000; Majewski *et al.* 2000). See below for details.

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

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

where $q = 1/\lambda$. This assumption is supported by empirical studies on transformation of many species including *Helicobacter pylori* (Lin *et al.* 2009), *Streptococcus pneumoniae* (Croucher *et al.* 2012), and *Haemophilus influenzae* (Mell *et al.* 2014). For mathematical convenience, we assume unidirectional elongation of conversion tract from 5' to 3', which has no quantitative effect on the pattern of SNPs. Given Equation 1, the rate that a region of L bp undergoes homologous recombination per generation is given by

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

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

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

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 undergo recombination when a lineage is traced back for a single generation.

It is interesting to note that Equation 2 does not include the probability that a recombination tract cover 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 we will explain in the next section (see below).

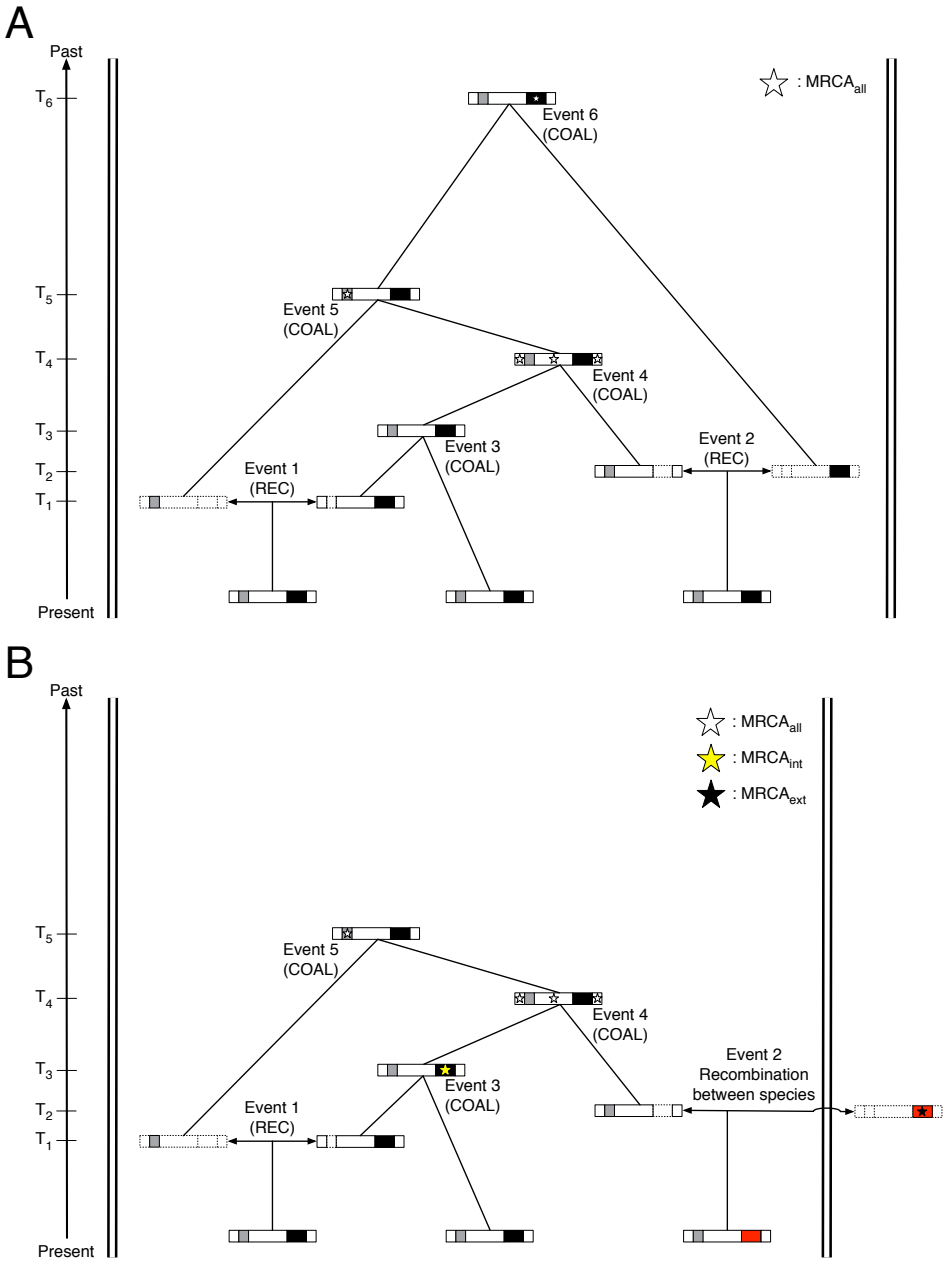

Figure 2 (A) Ancestral recombination graph with recombination within species alone. An example with sample size $n = 3$ is illustrated. The sampled three genomes are shown by long boxes, where regions with different histories are presented in different colors. The ancestral lineage is split into two 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 an black are transferred fragment 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.

It should be noted that there are three mechanisms for a cell to incorporate DNA, transformation, transduction and conjugation, through which homologous recombination can occur. They occur at different rates and typical lengths of integrated tracts should be different. Quite short fragments are usually integrated through transformation, while relatively large fragments may be involved in recombination through transduction and conjugation (Cohan 2002a). Therefore, it is biologically reasonable to model these processes separately, and this is what was done in earlier studies (see Maynard Smith 1994;

Hudson 1994).

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

More strictly, if we consider the three mechanisms separately, denote by $g_{\rm tf}$, $g_{\rm td}$, and $g_{\rm cj}$, respectively, the initiation rates of transformation-, transduction- and conjugation-oriented recombination per site per generation, and for each, let us assume that the tract length follows a geometric distribution (the mean lengths are $\lambda_{\rm tf}$, $\lambda_{\rm td}$, and $\lambda_{\rm cj}$ for the three mechanisms). Then, the total initiation rate per site is $g_{\rm total} = g_{\rm tf} + g_{\rm td} + g_{\rm cj}$, but the density distribution of tract length is not a simple geometric distribution with a single parameter, rather given by an average of three geometric distributions:

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

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

Thus, although it is mathematically correct to model the three mechanisms separately, there should be many cases where it is reasonable to use the simplified treatment. This is convenient to apply the coalescent theory to real polymorphism data for estimating the rate of homologous recombination, especially when the relative contributions of the three mechanisms are unknown. In this work, therefore, we employ the simplified treatment with a single rate of homologous recombination (Equation 2) with a single geometric distribution with parameter q (Equation 1), following previous theoretically studies (Falush *et al.* 2001; McVean *et al.* 2002; Fearnhead *et al.* 2005; Jolley *et al.* 2005; Didelot and Falush 2007).

As mentioned above, a homologous recombination event within prokaryote species is easily incorporated in the standard framework of the coalescent (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 with L bp, the process waits for either coalescent or recombination event, and the per-generation rate for the latter is given by Equation 2, while the rate of coalescence is given by $\binom{n}{2}/N$, where N is the population size and n is the number of lineages.

Note that this simple process holds in a single population, in which coalescence occurs randomly between any individuals in the population and so does recombination, but it is straightforward to incorporate population structure and demographic history into this framework as is done for eukaryote cases. The difference between eukaryote and prokaryote is the causes of population structure. In eukaryotes, limited migration between geographic barriers should be the major cause, and this also applies to prokaryotes although more complicated. For example, subpopulations of infectious species may form based on host individuals.

In addition, there are two major classes of isolation in prokaryote, ecological and genetic isolation. Ecological isolation is defined as a difference of niche that can reduce the rate of recombination between bacterial populations (Cohan 2002a,b). Physiological difference between donor and recipient would decrease the chance of recombination. For example, *Vibrio splendidus* lives in coastal bacterioplankton, exhibiting resource partitioning (specific season and/or free-living size fraction) among strains and phylogenetic divergence corresponding to each niche. It is suggested that ecological isolation is working as a barrier of DNA exchanges between niches (Hunt *et al.* 2008). Genetic isolation is defined as the establishment of mutation accumulation that prevents one strain from integrating foreign DNA of other strains (*e.g.,* Lawrence 2013). As described in the Introduction, the rate of successful integration of DNA of other strains depends on a number of self-recognition mechanisms, including short-specific sequences (*i.e.,* DUS or USS), restriction-modification systems, RecA-mediated homology search, and mismatch correction system. While both ecological and genetic isolation are often used in the context of homologous recombination with different species (or strains) involved in sexual isolation (Cohan 2002a; Fraser *et al.* 2009), these concept should work for homologous recombination within the same species but between different populations (or strains). Thus, there are many factors to cause isolation within the same species, which should heavily affect the pattern of the 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 (interspecific 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

$$h' = R^h_{\text{in}} + R^h_{\text{left}} + R^h_{\text{all}}, \tag{5}$$

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

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

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

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

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-

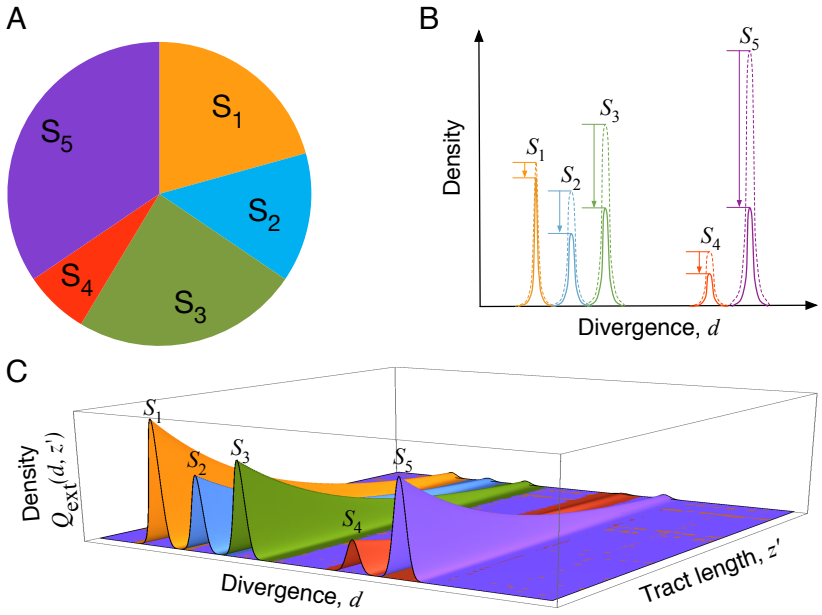

Figure 3 Illustrating a hypothetical environment where five different species (S_1 - S_5) are there around the focal species, S_0. (A) The proportion of the five species in the environment. (B) Density distribution of the divergence of environmental DNA from the focal species (dashed line) and the waited 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').

combination from external source (inter-specific recombination). The mutation rate $\theta = 0.01$ was fixed throughout this work. For recombination within the focal species (intra-specific recombination), the mean tract length was fixed to be $\lambda = 1000$ bp, and the rate (g) was changed. We first considered a relatively low recombination rate from external source ($2Nh = 0.00005$). We here used a simplified assumption to demonstrate the point, that is, the average divergence to external DNA was fixed to be 20% and the tract length followed a geometric distribution with a fixed mean ξ (*i.e.*, $Q_{\rm ext}(d = 0.2, z') = \xi^{-1}(1-\xi^{-1})^{z'-1}$). Figure 4 shows typical patterns of SNPs from the simulation results with $n = 10$, and $L = 5,000$. The positions of SNPs are presented by solid vertical lines along the simulated region. In Figure 4A, no recombination within species is assumed ($2Ng = 0$). One recombination event (607 bp) from external source occurred $t = 0.23N$ generations ago on the ancestral lineage of individuals 2, 5 and 10, and the positions of two breakpoints of the recombination event are shown by 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 θ). This region is referred to as Region 1 and boxed in red. Neighbor-jointing tree for this region is completely different form that for the other region: Individuals 2, 5 and 10 are highly diverged from the other seven individuals in Region 1.

It is thus obvious that the level of polymorphism increases as recombination events from external source increases. As shown in Figure 5A, the level of polymorphism increases with increasing the initiation rate (h), mean tract length (ξ) and divergence (d), where the amount of polymorphism is measured by π, the average number of nucleotide differences per site. This simulation result agrees with theoretical prediction that the expectation of π is given by a simple function of h, ξ, d:

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

where $\phi = 2Nh\xi d$. See Appendix for the derivation. It is obvious 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 flipped by recombination (see Appendix), as Figure 5B clearly demonstrates that π is given by a simple liner function of $h\xi d$.

In Figure 4B, a moderate level of recombination within species (intra-specific recombination) is introduced ($2Ng = 0.001$). Two external DNA fragments are integrated (Regions 2 and 3). In Region 2, a 643 bp of foreign DNA was integrated $t = 0.61N$ generations ago. It is important to notice that whereas individuals 2, 3, 7 and 9 have the entire fragment, only a part of the integrated fragment is observed in individuals 4, 5, 6, 8 and 10. This is due to intra-specific recombination that occurred after the integration; the integrated fragment was chopped into pieces and distributed into the population. By looking at the simulated ancestral recombination graph, we found three such intra-specific recombination

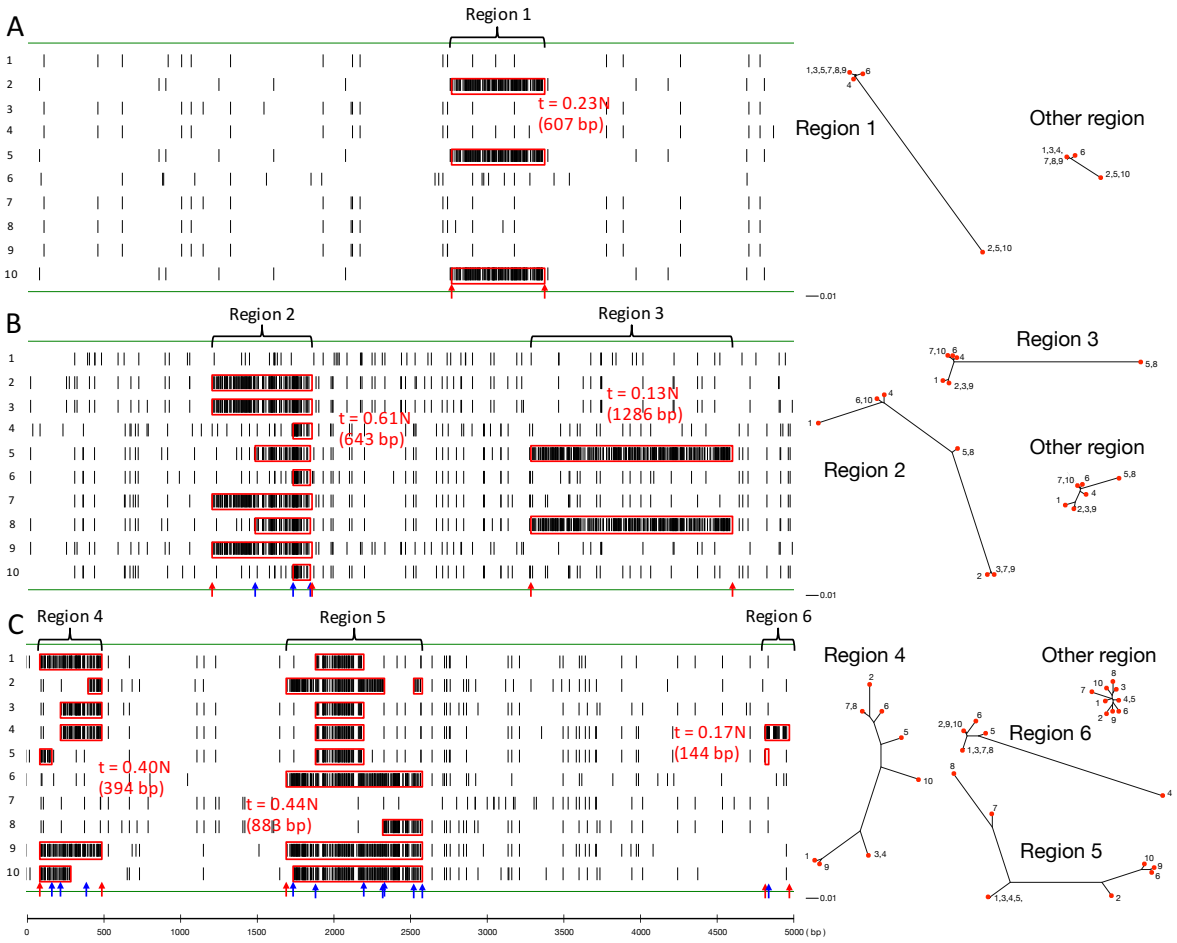

Figure 4 Typical patterns of SNPs with inter-specific recombination from external source with no intra-specific recombination (A; $2Ng = 0$), with a moderate level of intra-specific recombination (B; $2Ng = 0.001$), and with a high recombination rate (C; $2Ng = 0.005$). Vertical bars indicate the locations 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-jointing 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.

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

With even a higher intra-specific recombination rate ($2Ng = 0.005$) in Figure 4C, fragmentation is more enhanced. There are three regions that experienced recombination from external source (Regions 4, 5 and 6), and all of them involved intra-specific recombination. An intriguing pattern is seen in Region 5, where only a part of the integrated flagrant is observed in the sample. The recombination occurred $t = 0.44N$ generations ago. The actual length of the integrated foreign fragment was more than 883 bp, but none of the sampled ten individuals have the 5' breakpoint. This process can be well understood with the cartoon in Figure 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

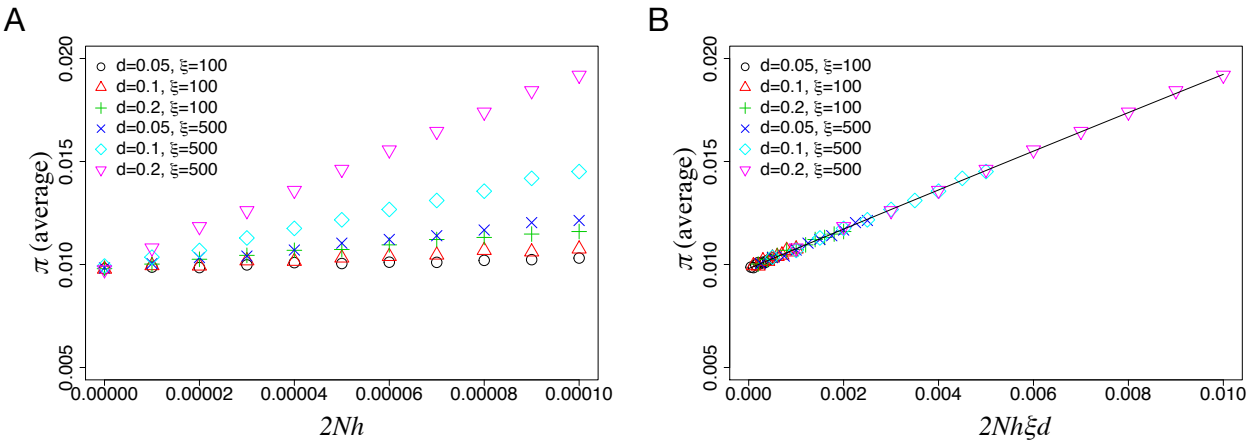

Figure 5 The effect of recombination from external source on the amount of polymorphism measure by π (A) π as a function of $2Nh$. (B) π is in a clear linear correlation with $2Nh\xi d$ (Equation 8). The averages π over 10,000 runs of simulations with $n = 15$ and $L = 10,000$ are shown.

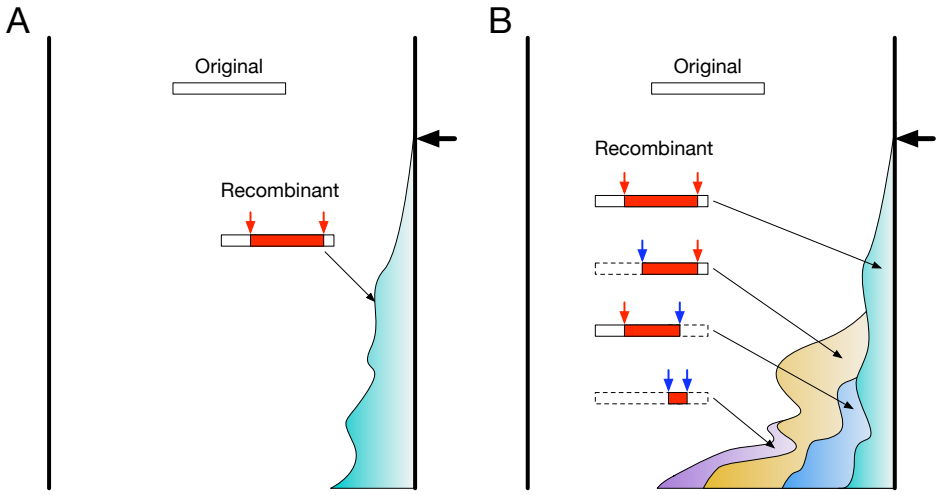

Figure 6 Cartoons of the typical behavior of population frequency of a foreign DNA, (A) without intra-specific recombination and (B) with intra-specific recombination. The time of the foreign DNA introduced into the population is denoted by a thick black arrow, producing a recombinant haplotype in which 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.

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.

Given this effect of inter-specific recombination, it is predicted 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

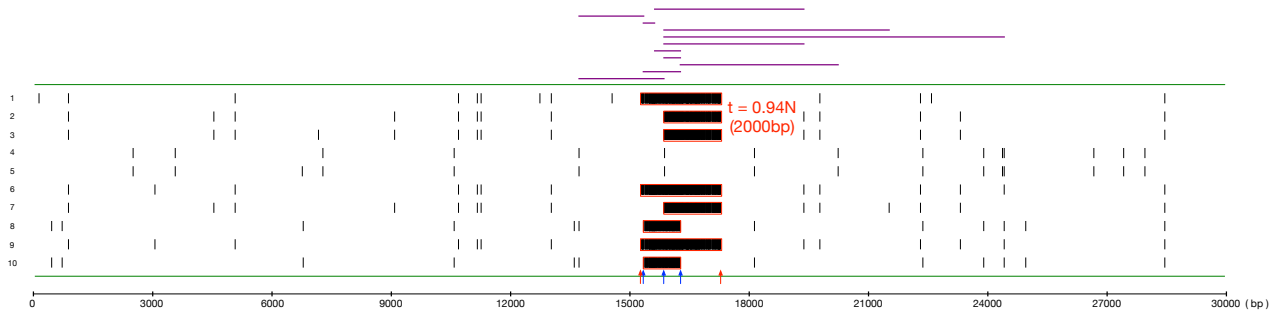

Figure 7 Application of **GENECONV** to a simulated region ($n = 10$, $L = 30,000$ bp), in which a 2000 bp of foreign DNA was integrated $0.94N$ generation ago (red boxed), followed by three additional intra-specific recombinations that fragmented the foreign DNA. The breakpoints of inter-specific recombination 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.

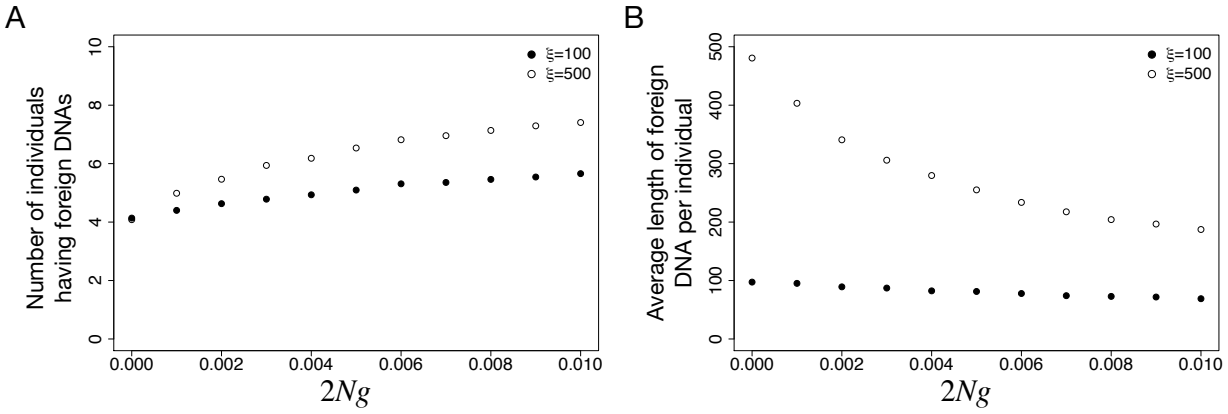

Figure 8 The effect of intra-specific recombination on (A) the number of individuals having foreign DNA and (B) average length of integrated foreign DNA, from 10,000 runs of simulations with $n = 15$ and $L = 10,000$.

Brown, T., X. Didelot, D. J. Wilson and N. De Maio, 2016 SimBac: simulation of whole bacterial genomes with homologous recombination. Microbial genomics 2.

Chen, I., and D. Dubnau, 2004 DNA uptake during bacterial transformation. Nat. Rev. Microbiol. 2: 241–9.

Claverys, J. P., and S. A. Lacks, 1986 Heteroduplex deoxyribonucleic acid base mismatch repair in bacteria. Microbiol. Rev. 50: 133–165.

Cohan, F. M., 2002a Sexual isolation and speciation in bacteria. Genetica 116: 359–370.

Cohan, F. M., 2002b What are bacterial species? Annu. Rev. Microbiol. 56: 457–87.

Cornejo, O. E., T. Lefébure, P. D. P. Bitar, P. Lang, V. P. Richards, *et al.*, 2013 Evolutionary and population genomics of the cavity causing bacteria *Streptococcus mutans*. Mol. Biol. Evol. 30: 881–93.

Croucher, N. J., S. R. Harris, L. Barquist, J. Parkhill and S. D. Bentley, 2012 A high-resolution view of genome-wide pneumococcal transformation. PLoS Pathog. 8: e1002745.

de Vries, J., P. Meier and W. Wackernagel, 2001 The natural transformation of the soil bacteria *Pseudomonas stutzeri* and *Acinetobacter* sp. by transgenic plant DNA strictly depends on homologous sequences in the recipient cells. FEMS Microbiol. Lett. 195: 211–215.

De Maio, N., and D. J., Wilson, 2017 The bacterial sequential Markov coalescent. Genetics 206: 333–343.

Didelot, X., and D. Falush, 2007 Inference of bacterial microevolution using multilocus sequence data. Genetics 175: 1251–66.

Didelot, X., D. Lawson and D. Falush, 2009 SimMLST: simulation of multi-locus sequence typing data under a neutral model. Bioinformatics 25: 1442–4.

Didelot, X., and M. C. J. Maiden, 2010 Impact of recombination on bacterial evolution. Trends Microbiol. 18: 315–22.

Didelot, X., and D. J. Wilson, 2015 ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol 11: e1004041.

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

Donati, C., N. L. Hiller, H. Tettelin, A. Muzzi, N. J. Croucher, *et al.*, 2010 Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. Genome Biol. 11:R107.

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

Doolittle, W. F., and R. T. Papke, 2006 Genomics and the bacterial species problem. Genome biol. 7: e116.

Doroghazi, J. R., and D. H. Buckley, 2011 A model for the effect of homologous recombination on microbial diversification. Genome Biol. Evol. 3: 1349–56.

Earl, A. M., R. Losick and R. Kolter, 2008 Ecology and genomics of *Bacillus subtilis*. Trends Microbiol. 16: 269–75.

Falush, D., C. Kraft, N. S. Taylor, P. Correa, J. G. Fox, *et al.*, 2001 Recombination and mutation during long-term gastric colonization by *Helicobacter pylori*: estimates of clock

rates, recombination size, and minimal age. Proc. Natl. Acad. Sci. USA 98: 15056–15061.

Falush, D., M. Torpdahl, X. Didelot, D. F. Conrad, D. J. Wilson, *et al.*, 2006 Mismatch induced speciation in *salmonella*: model and data. Philos. Trans. R. Soc. B 361: 2045–53.

Fearnhead, P., 2006 Perfect simulation from nonneutral population genetic models: variable population size and population subdivision. Genetics 174: 1397–406.

Fearnhead, P., N. G. C. Smith, M. Barrigas, A. Fox and N. French, 2005 Analysis of recombination in *Campylobacter jejuni* from MLST population data. J. Mol. Evol. 61: 333–40.

François, O., M. G. B. Blum, M. Jakobsson and N. A. Rosenberg, 2008 Demographic history of european populations of *Arabidopsis thaliana*. PLoS Genet. 4: 1–15.

Fraser, C., W. P. Hanage and B. G. Spratt, 2007 Recombination and the nature of bacterial speciation. Science 315: 476–80.

Fraser, C., E. J. Alm, M. F. Polz, B. G. Spratt and W. P. Hanage, 2009 The bacterial species challenge: making sense of genetic and ecological diversity. Science 323: 741–746.

Fraser, C., W. P. Hanage and B. G. Spratt, 2005 Neutral microepidemic evolution of bacterial pathogens. Proc. Natl. Acad. Sci. USA 102: 1968–1973.

Griffiths, R. C., and P. Marjoram, 1996 Ancestral inference from samples of DNA sequences with recombination. J. Comput. Biol. 3: 479–502.

Grissa, I., G. Vergnaud and C. Pourcel, 2007 The CRISPRdb database and tools to display CRISPRs and to generate dictionaries of spacers and repeats. BMC Bioinformatics 8: 172.

Hartl, D. L., and A. G. Clark, 2007 *Principles of population genetics*. Sinauer Associates, Sunderland.

Haven, J., L. C. Vargas, E. F. Mongodin, V. Xue, Y. Hernandez, *et al.*, 2011 Pervasive recombination and sympatric genome diversification driven by frequency-dependent selection in *Borrelia burgdorferi*, the lyme disease bacterium. Genetics 189: 951–66.

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

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

Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–43 in *Oxford surveys in evolutionary biology*, edited by D. Futuyma and J. Antonovics. Volume 7. Oxford Univ Press, Oxford.

Hudson, R. R., 1994 Analytical results concerning linkage disequilibrium in models with genetic transformation and conjugation. J. evol. biol. 7: 535–548.

Hudson, R. R., 2001 Two-locus sampling distributions and their application. Genetics 159: 1805–1817.

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

Hunt, D. E., L. A. David, D. Gevers, S. P. Preheim, E. J. Alm, *et al.*, 2008 Resource partitioning and sympatric differentiation among closely related bacterioplankton. Science 320: 1081–5.

Jolley, K. A., D. J. Wilson, P. Kriz, G. McVean and M. C. J. Maiden, 2005 The influence of mutation, recombination, population history, and selection on patterns of genetic diversity in *Neisseria meningitidis*. Mol. Biol. Evol. 22: 562–9.

Kingman, J. F., 1982 On the genealogy of large populations. J.Appl. Probab. 19: 27–43.

Krause, D. J., X. Didelot, H. Cadillo-Quiroz and R. J. Whitaker, 2014 Recombination shapes genome architecture in an organism from the archaeal domain. Genome Biol. Evol. 6: 170–8.

Kreitman, M., 2000 Methods to detect selection in populations with applications to the human. Annu. Rev. Genom. Hum. Genet. 1: 539–559.

Krone, S. M., and C. Neuhauser, 1997 Ancestral processes with selection. Theor. popul. biol. 51: 210–237.

Lawrence, J. G., 2013 Gradual speciation: Further entangling the tree of life, pp. 243–262 in *Lateral Gene Transfer in Evolution*, edited by U. Gophna. Springer, New York.

Lin, E. A., X.-S. Zhang, S. M. Levine, S. R. Gill, D. Falush, *et al.*, 2009 Natural transformation of *Helicobacter pylori* involves the integration of short dna fragments interrupted by gaps of variable size. PLoS Pathog. 5: e1000337.

Linz, B., M. Schenker, P. Zhu and M. Achtman, 2000 Frequent interspecific genetic exchange between commensal neisseriae and neisseria meningitidis. Mol. Microbiol. 36: 1049–1058.

Lorenz, M. G., and J. Sikorski, 2000 The potential for intraspecific horizontal gene exchange by natural genetic transformation: sexual isolation among genomovars of *Pseudomonas stutzeri*. Microbiology 146: 3081–3090.

Lorenz, M. G., and W. Wackernagel, 1994 Bacterial gene transfer by natural genetic transformation in the environment. Microbiol. Rev. 58: 563–602.

Majewski, J., 2001 Sexual isolation in bacteria. FEMS Microbiol. Lett. 199: 161–169.

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

Majewski, J., P. Zawadzki, P. Pickerill, F. M. Cohan and C. G. Dowson, 2000 Barriers to genetic exchange between bacterial species: *Streptococcus pneumoniae* transformation. J. Bacteriol. 182: 1016–1023.

Marjoram, P., J. Molitor, V. Plagnol and S. Tavaré, 2003 Markov chain monte carlo without likelihoods. Proceedings of the National Academy of Sciences 100: 15324–15328.

Marraffini, L. A., and E. J. Sontheimer, 2010 CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. Nat. Rev. Genet. 11: 181–90.

Maynard Smith, J., 1994 Estimating the minimum rate of genetic transformation in bacteria. J. Evol. Biol. 7: 525–534.

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

Mell, J. C., J. Y. Lee, M. Firme, S. Sinha and R. J. Redfield, 2014 Extensive cotransformation of natural variation into chromosomes of naturally competent *Haemophilus influenzae*. G3 (Bethesda) 4: 717–31.

Nakhleh, L., 2013 Computational approaches to species phylogeny inference and gene tree reconciliation. Trends Ecol. Evol. 28: 719–728.

Nell, S., D. Eibach, V. Montano, A. Maady, A. Nkwescheu, *et al.*, 2013 Recent acquisition of *Helicobacter pylori* by baka pygmies. PLoS Genet. 9: e1003775.

Neuhauser, C., and S. M. Krone, 1997 The genealogy of samples in models with selection. Genetics 145: 519–534.

Nordborg, M., 2001 Coalescent theory in *Handbook of statistical genetics*, edited by D. J. BALDING, M. Bishop, and C. Cannings. Wiley-Blackwell, Chichester, UK.

Nordborg, M., and H. Innan, 2002 Molecular population genetics. Curr. Opin. Plant Biol. 5: 69–73.

Ochman, H., J. G. Lawrence and E. A. Groisman, 2000 Lateral gene transfer and the nature of bacterial innovation. Nature 405: 299–304.

Overballe-Petersen, S., K. Harms, L. A. A. Orlando, J. V. M. Mayar, S. Rasmussen, *et al.*, 2013 Bacterial natural transformation by highly fragmented and damaged DNA. Proc. Natl. Acad. Sci. USA 110: 19860–19865.

Pepperell, C., V. H. Hoeppner, M. Lipatov, W. Wobeser, G. K. Schoolnik, *et al.*, 2010 Bacterial genetic signatures of human social phenomena among *M. tuberculosis* from an Aboriginal Canadian population. Mol. Biol. Evol. 27: 427–40.

Pérez-Losada, M., E. B. Browne, A. Madsen, T. Wirth, R. P. Viscidi, *et al.*, 2006 Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. Infect. Genet. Evol. 6: 97–112.

Polz, M. F., E. J. Alm and W. P. Hanage, 2013 Horizontal gene transfer and the evolution of bacterial and archaeal population structure. Trends Genet. 29: 170–175.

Roberts, M., and F. Cohan, 1993 The effect of DNA sequence divergence on sexual isolation in bacillus. Genetics 134: 401–408.

Roberts, R. J., T. Vincze, J. Posfai and D. Macelis, 2010 REBASE–a database for DNA restriction and modification: enzymes, genes and genomes. Nucleic Acids Res. 38: D234–236.

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

Rosenberg, N. A., and M. Nordborg, 2002 Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. Nat. Rev. Genet. 3: 380–90.

Rousseau, C., M. Gonnet, M. Le Romancer and J. Nicolas, 2009 CRISPI: a CRISPR interactive database. Bioinformatics 25: 3317–8.

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

Shapiro, B. J., 2014 Signatures of natural selection and ecological differentiation in microbial genomes, pp. 339–359 in *Ecological Genomics*. Springer, New York.

Shapiro, B. J., J. Friedman, O. X. Cordero, S. P. Preheim, S. C. Timberlake, *et al.*, 2012 Population genomics of early events in the ecological differentiation of bacteria. Science 336: 48–51.

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

Snyder, L., J. E. Peters, T. M. Henkin and W. Champness, 2013 *Molecular genetics of bacteria*. ASM Press, Washington, DC.

Sousa, V., and J. Hey, 2013 Understanding the origin of species with genome-scale data: modelling gene flow. Nat. Rev. Genet. 14: 404–14.

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

Takuno, S., T. Kado, R. P. Sugino, L. Nakhleh and H. Innan, 2012 Population genomics in bacteria: A case study of *Staphylococcus aureus*. Mol. Biol. Evol. 29: 797–809.

Thomas, C. M., and K. M. Nielsen, 2005 Mechanisms of, and barriers to, horizontal gene transfer between bacteria. Nat. Rev. Microbiol. 3: 711–21.

Thomas, J. C., P. A. Godfrey, M. Feldgarden and D. A. Robinson, 2012 Candidate targets of balancing selection in the genome of *Staphylococcus aureus*. Mol. Biol. Evol. 29: 1175–86.

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

Vasu, K., and V. Nagaraja, 2013 Diverse functions of restriction-modification systems in addition to cellular defense. Microbiol. Mol. Biol. Rev. 77: 53–72.

Vulić, M., F. Dionisio, F. Taddei and M. Radman, 1997 Molecular keys to speciation: DNA polymorphism and the control of genetic exchange in enterobacteria. Proc. Natl. Acad. Sci. USA 94: 9763–9767.

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

Wirth, T., D. Falush, R. Lan, F. Colles, P. Mensa, *et al.*, 2006 Sex and virulence in escherichia coli: an evolutionary perspective. Mol. Microbiol. 60: 1136–51.

Wiuf, C., 2001 Recombination in human mitochondrial DNA? Genetics 159: 749–756.

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

Zahrt, T. C., and S. Maloy, 1997 Barriers to recombination between closely related bacteria: MutS and RecBCD inhibit recombination between *Salmonella typhimurium* and *Salmonella typhi*. Proc. Natl. Acad. Sci. USA 94: 9786–9791.

Zawadzki, P., and F. M. Cohan, 1995 The size and continuity of dna segments integrated in bacillus transformation. Genetics 141: 1231–1243.

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,

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

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,

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

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.