Chromosome Painting In Silico in a Bacterial Species Reveals Fine Population Structure

Koji Yahara,[1,2,3] Yoshikazu Furuta,[1,2] Kenshiro Oshima,[4] Masaru Yoshida,[5] Takeshi Azuma,[5] Masahira Hattori,[4] Ikuo Uchiyama,[6] and Ichizo Kobayashi*,[1,2]

[1]Department of Medical Genome Sciences, Graduate School of Frontier Sciences, University of Tokyo, Minato-ku, Tokyo, Japan

[2]Institute of Medical Science, University of Tokyo, Minato-ku, Tokyo, Japan

[3]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[4]Department of Computational Biology, Graduate School of Frontier Sciences, University of Tokyo, Kashiwa, Chiba, Japan

[5]Department of Gastroenterology, Graduate School of Medicine, Kobe University, Chuou-ku, Kobe, Hyogo, Japan

[6]Laboratory of Genome Informatics, National Institute for Basic Biology, Okazaki, Aichi, Japan

***Corresponding author:** E-mail: ikobaya@ims.u-tokyo.ac.jp.

Associate editor: James McInerney

Abstract

Identifying population structure forms an important basis for genetic and evolutionary studies. Most current methods to identify population structure have limitations in analyzing haplotypes and recombination across the genome. Recently, a method of chromosome painting in silico has been developed to overcome these shortcomings and has been applied to multiple human genome sequences. This method detects the genome-wide transfer of DNA sequence chunks through homologous recombination. Here, we apply it to the frequently recombining bacterial species *Helicobacter pylori* that has infected *Homo sapiens* since their birth in Africa and shows wide phylogeographic divergence. Multiple complete genome sequences were analyzed including sequences from Okinawa, Japan, that we recently sequenced. The newer method revealed a finer population structure than revealed by a previous method that examines only MLST housekeeping genes or a phylogenetic network analysis of the core genome. Novel subgroups were found in Europe, Amerind, and East Asia groups. Examination of genetic flux showed some singleton strains to be hybrids of subgroups and revealed evident signs of population admixture in Africa, Europe, and parts of Asia. We expect this approach to further our understanding of intraspecific bacterial evolution by revealing population structure at a finer scale.

***Key words:* fineSTRUCTURE, homologous recombination, phylogenetic network, human evolution, *Helicobacter pylori*.**

Introduction

Elucidation of population structure is an important basis for evolutionary studies in any species (Patterson et al. 2006; Robinson et al. 2010; Novembre and Ramachandran 2011) including studies such as the inference of demographic history and the examination of population differentiation and selection (Akey et al. 2004; Holsinger and Weir 2009; Henn et al. 2012). Identifying population structures and assigning individuals to the identified population categories (groups) are central issues to analyzing any genetic data set. They are, for example, an important basis for association studies (Devlin and Roeder 1999; Nakamura et al. 2005). Two of the most popular approaches for identifying population structures using genetic data are principal component analysis (PCA) (Menozzi et al. 1978) and STRUCTURE (Pritchard et al. 2000; Falush, Stephens et al. 2003). In addition to these, phylogenetic tree construction and BAPS (Corander et al. 2008) have also been used to describe population structure.

The increased availability of genome-wide sequence data poses challenges for current methods to identify population structure accurately and reliably in practical computational time. PCA can quickly process large data sets with hundreds of thousands of SNPs and thousands of samples (Patterson et al. 2006), and phylogenetic trees of core genomes can classify prokaryotic species (Snel et al. 2005; Zhi et al. 2012) and populations within a bacterial species (Kawai et al. 2011; Okoro et al. 2012). However, PCA and phylogenetic tree construction are not designed to infer the number of populations directly. Furthermore, in PCA and phylogenetic tree construction, correlation (linkage) between SNPs and their relative position is not taken into account. BAPS has a linkage model option (Corander and Tang 2007). STRUCTURE also has a linkage model option that accounts for recombination events (admixture) among populations and correlations among SNPs that arise in admixed populations (Falush, Stephens et al. 2003). But, STRUCTURE requires the number of populations (K) to be specified in advance and typically $K < 10$ for satisfactory convergence (Lawson et al. 2012).

As a solution to these problems, a new tool called fineSTRUCTURE was recently developed (Lawson et al. 2012). It is based on "chromosome painting" in silico, which infers recombination-derived "chunks" and reconstructs haplotypes on the chromosome of a "recipient" individual as a series of chunks from all other "donor" individuals in the sample (Lawson et al. 2012). The results are summarized into a "co-ancestry matrix," which contains the number of recombination events from each donor to each recipient individual. Using this matrix, fineSTRUCTURE conducts model-based clustering of hundreds or thousands of individuals. So far, fineSTRUCTURE has only been applied to humans; nevertheless, this application has already revealed subtle population structures in worldwide populations (Lawson et al. 2012).

In this study, we have applied this approach for the first time to a bacterial species. Knowledge of population structure has immediate applications to understanding bacteria's role in medicine, agriculture, and the environment (Achtman 2008; Downing et al. 2011; Argudin et al. 2013). Moreover, this knowledge has and will continue to directly affect bacterial molecular epidemiology and public health management among other fields (Robinson et al. 2010).

The population structures of bacterial species, however, are complex and often controversial due to uncertainty about the frequency and influence of recombination between lineages (Feil and Spratt 2001). Population structures of clonal bacteria with low rates of recombination can be inferred by phylogenetic trees (Hershberg et al. 2008; Morelli, Song, et al. 2010; Holt et al. 2012; Okoro et al. 2012), which has been by far the most common approach in bacteria (Snel et al. 2005; Zhi et al. 2012). However, the extent of recombination varies among bacterial species, some of which show much higher rates of recombination than others (Perez-Losada et al. 2006; Didelot and Maiden 2010). Bacterial recombination is mainly intraspecific but occasionally occurs between species (Hanage et al. 2009; Corander et al. 2011). The question of how to elucidate the population structures of such highly recombining bacterial species using genome-wide sequence data has been largely unexplored.

To address this question, we have focused on *Helicobacter pylori*, a stomach pathogen that infects over half of all humans and causes gastritis, ulcers, and cancer (Suerbaum and Josenhans 2007). *Helicobacter pylori* is commonly acquired in childhood and can persist in the stomach over the host's entire lifespan (Suerbaum and Josenhans 2007). *Helicobacter pylori* has been with *Homo sapiens* since its origin in Africa and shows geographic patterns of genetic diversity that parallel human diversity (Linz et al. 2007). Of the known bacteria, *H. pylori* is likely the most frequently recombining species (Doolittle and Zhaxybayeva 2009) with a greater effect of homologous recombination than mutation (Morelli, Didelot, et al. 2010; Kennemann et al. 2011) as well as signatures of homologous recombination throughout the genome (Yahara et al. 2012). It is worth exploring how chromosome painting and fineSTRUCTURE could elucidate the population structure of genomes subject to such frequent recombination events.

Previously, by STRUCTURE analysis using 7 housekeeping (MLST) genes (*atpA*, *efp*, *mutY*, *ppa*, *trpC*, *urel*, and *yphC*), phylogeographic population structures of *H. pylori* were identified: hpEurope, hpSahul, hpEastAsia, hpAsia2, hpNEAfrica, hpAfrica1, and hpAfrica2 (Falush, Wirth et al. 2003; Linz et al. 2007; Moodley et al. 2009, 2012). The hpEastAsia population is known to have three subpopulations, hspAmerind, hspEAsia, and hspMaori (Moodley et al. 2009). Of these subpopulations, the complete genome sequences of hspEAsia strains were recently obtained (Kawai et al. 2011). In addition to these, this study included two newly sequenced strains from Okinawa, Japan. Okinawans and mainland Japanese are known to be genetically differentiated (Hammer and Horai 1995; Yamaguchi-Kabata et al. 2008). If the *H. pylori* strains from Okinawa are also genetically differentiated, these strains may form a regional subgroup useful for deepening our understanding of *H. pylori* population structure in East Asia.

Using chromosome painting and fineSTRUCTURE, we analyzed the complete genome sequences of *H. pylori* strains from various parts of the world. Our analysis, based on elucidation of genetic flux through homologous recombination between subgroups, has revealed the population structure of *H. pylori* at a finer scale than achieved by previous methods.

Results

Chromosome Painting In Silico of *H. pylori* Complete Genomes

To elucidate the population structure of *H. pylori*, we applied the "chromosome painting" algorithm accounting for linkage information to complete genome sequences of the two aforementioned Okinawa strains and 27 public *H. pylori* strains. Each genome (CDSs of all one-to-one orthologous genes in the core genomic regions, to be precise) was reconstructed using chunks of DNA donated by other individual genomes. The result is visualized in figure 1. Chunk donors are colored according to donor subgroups (the suffix "sg_" and the consecutive number [e.g., "_sg1"] is used to name each subgroup), which we will explain in the next section. The median length of a chunk is 14 bp (interquartile range: 5–39 bp). The distribution of chunk sizes is shown in supplementary figure S1, Supplementary Material online.

Based on the inference of recombination-derived chunks and their donors across the genomes, the chromosome painting algorithm calculates the expected number of chunks imported from a donor to a recipient genome and then summarizes these values into a matrix ("co-ancestry matrix"). The matrix for this study is visualized as a heat map (fig. 2*a*).

Population Structure at a Finer Scale

Based on the co-ancestry matrix visualized as a heat map, individual strains were assigned to subgroups by the fineSTRUCTURE clustering algorithm (fig. 2*a*). Hereafter, we use "subgroup" to designate a cluster, and each subgroup is named by adding the suffix "sg_" and the consecutive number (e.g., "Europe_sg1"). Unlike STRUCTURE, this algorithm can infer the number of clusters (K) and partition the strains into K subgroups with indistinguishable genetic ancestry. Based on likelihood of the co-ancestry matrix, the inference is performed by a Bayesian MCMC (Markov chain Monte Carlo) approach that explores the space of possible partitions by using an algorithm for proposing new partitions (Lawson et al. 2012).

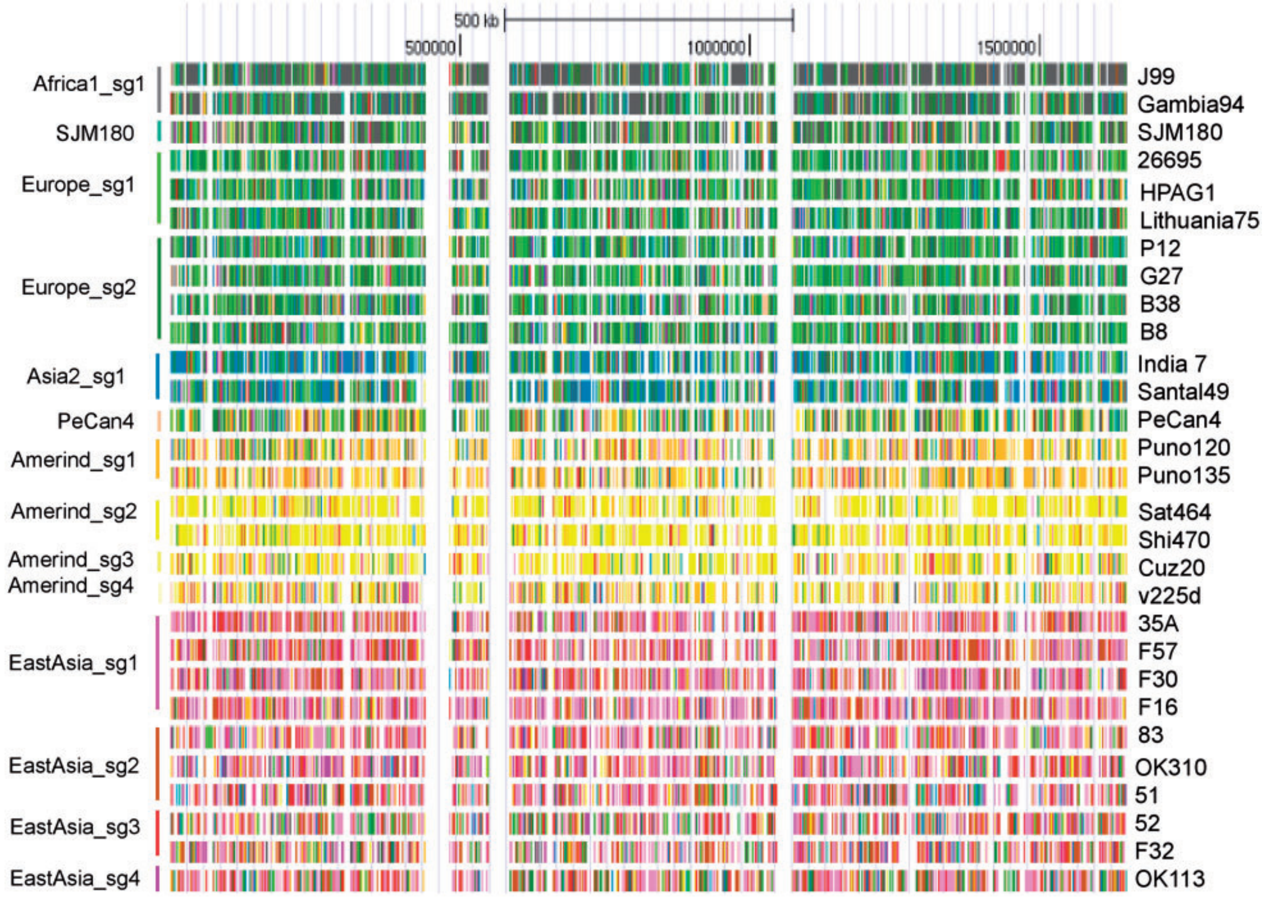

Fig. 1. Chromosome painting in silico. Each lane indicates the chromosome of a strain shown on the right. The strains are classified by fineSTRUCTURE into subgroups labeled by colors (table 1 and fig. 2) on the left. A color along the chromosome indicates the subgroup that donated a chunk of SNPs through homologous recombination. All genomic positions are transformed to those of a reference strain (26695).

The results of the model-based clustering are represented as a tree to the left of the co-ancestry matrix (fig. 2). The names of individual strains are replaced with those of the identified subgroups (table 1). We compared the results with those from the traditional STRUCTURE algorithm that is limited to seven housekeeping genes (also shown in table 1).

The results for Africa1 and Asia2 were the same for both methods. The European strains identified by STRUCTURE (hpEurope) were divided by fineSTRUCTURE into two subgroups and two singletons (SJM180, PeCan4). We will examine the two singleton strains later (in the "Hybrid genomes" section). The hpEastAsia strains (hspEAsia and hspAmerind strains) showed the clearest difference: fineSTRUCTURE divided hspAmerind into four subgroups including two singletons (we named them Amerind_sg1, Amerind_sg2, Amerind_sg3, and Amerind_sg4) and hspEAsia into four subgroups including one singleton (we named them as EastAsia_sg1, EastAsia_sg2, EastAsia_sg3, and EastAsia_sg4).

Comparison with Phylogenetic and No-Linkage Methods

From the same genomic data, we also constructed a phylogenetic network (fig. 3). Within each clade, the strains are colored according to the population assignment by fineSTRUCTURE. The result appeared globally consistent with the fineSTRUCTURE population assignments. The clades of East Asia, Amerind, Asia2, Europe, and Africa1 are seemingly polytomous, suggesting rampant recombination.

We also constructed a traditional phylogenetic tree (supplementary fig. S2, Supplementary Material online). The tree topology was also nearly consistent with the fineSTRUCTURE population assignments. However, the bifurcating branches in East Asia are short and contain lower bootstrap values, suggesting that the inference is not robust. This likely reflects a difference between the two methods with respect to whether they account for the occurrence of homologous recombination.

We also conducted the no-linkage approach of chromosome painting and fineSTRUCTURE treating markers as independent as PCA (Lawson et al. 2012), which ignores linkage and corresponds to assuming that recombination rate between any pair of markers is infinite. The result of the population assignment (supplementary fig. S3, Supplementary Material online) was also almost the same as that by the linkage approach.

Hybrid Genomes

In the phylogenetic network and tree (fig. 3 and supplementary fig. S2, Supplementary Material online), there is a clear

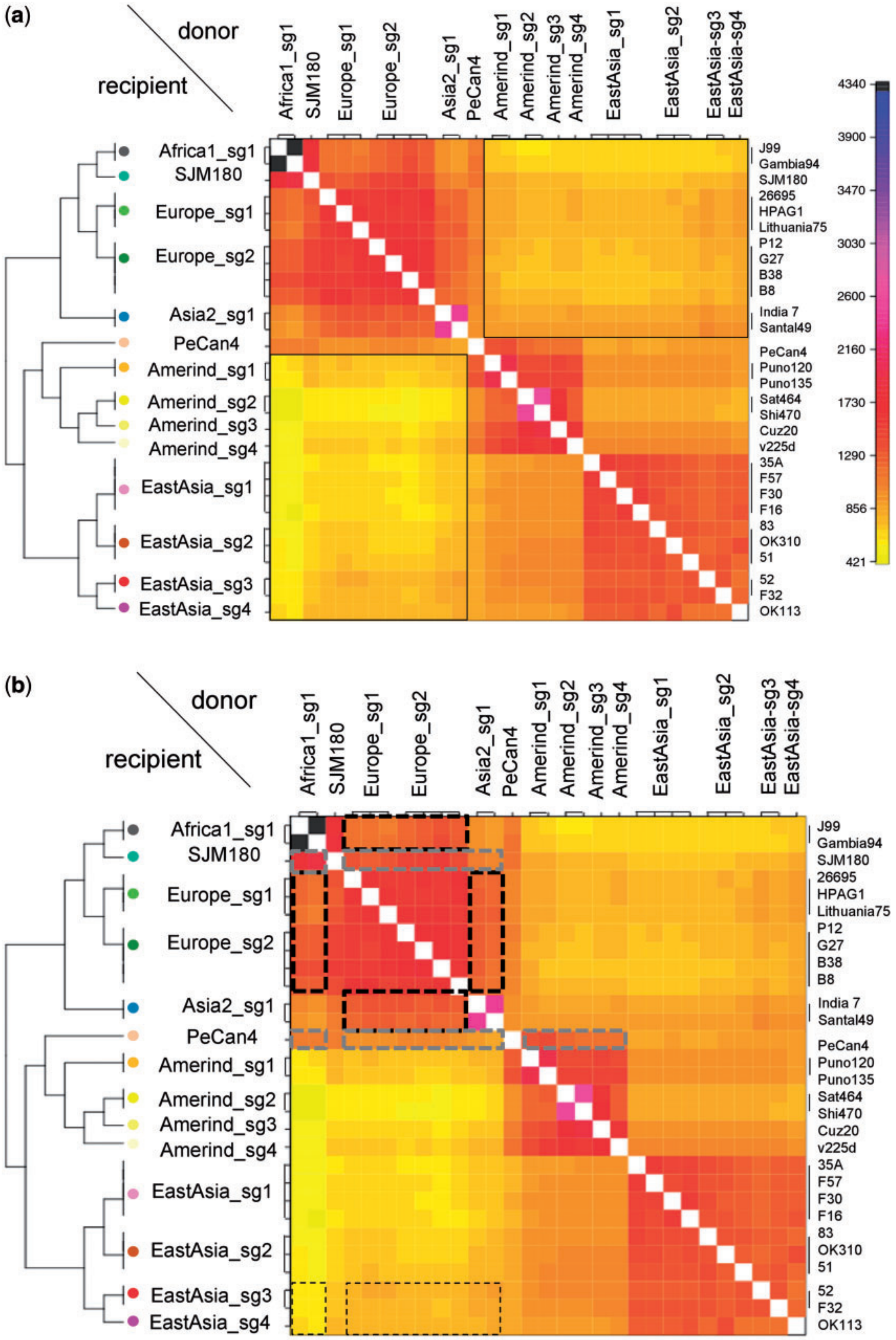

FIG. 2. Co-ancestry matrix with population structure and genetic flux. The color of each cell of the matrix indicates the expected number of chunks imported from a donor genome (column) to a recipient genome (row). The name of each strain is indicated on the right. (*a*) Population assignments and genetic flux. The tree in the left shows clustering for assignment of the listed population subgroups. The two black-lined boxes indicate asymmetry in genetic flux between EastAsia/Amerind and the other subgroups. (*b*) Hybrid strains and admixed subgroups. Two singleton strains, SJM180 and PeCan4, are hybrids as indicated by the gray dashed boxes. Signs of population admixture in Africa1, Europe, and Asia2 are indicated by bold black dashed boxes, whereas those in EastAsia_sg3 and EastAsia_sg4 are indicated by thin black dashed boxes.

Table 1. Comparison of Population Assignment.

Strain	fineSTRUCTURE (Linkage Model)	STRUCTURE
J99	Africa1_sg1	hpAfrica1
Gambia94		
SJM180	singleton (hybrid)	hpEurope
26695	Europe_sg1	
HPAG1		
Lithuania75		
P12	Europe_sg2	
G27		
B38		
B8		
India 7	Asia2_sg1	hpAsia2
Santal49 (SNT49)		
PeCan4	singleton (hybrid)	hpEurope
Puno120	Amerind_sg1	hpEastAsia (hspAmerind)
Puno135		
Sat464	Amerind_sg2	
Shi470		
Cuz20	singleton (Amerind_sg3)	
v225d	singleton (Amerind_sg4)	
35A	EastAsia_sg1	hpEastAsia (hspEAsia)
F57		
F30		
F16		
83	EastAsia_sg2	
OK310		
51		
52	EastAsia_sg3	
F32		
OK113	singleton (EastAsia_sg4)	

Note.—“sg” is abbreviated from “subgroup.”

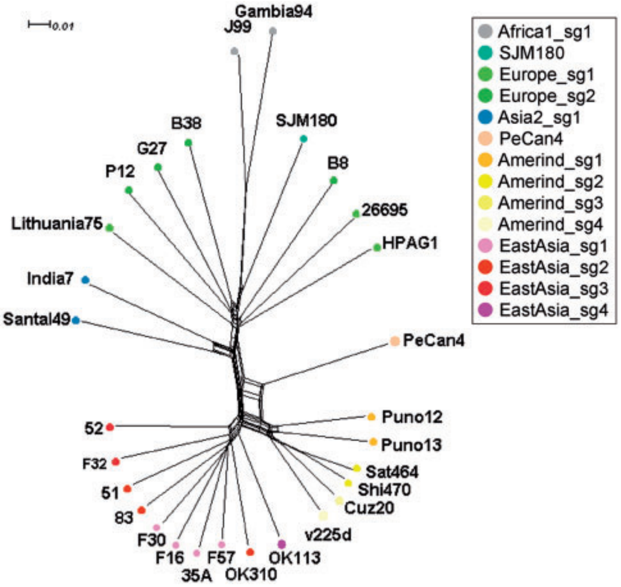

Fig. 3. Phylogenetic network. The colors indicate subgroups identified by fineSTRUCTURE (as in fig. 2 and table 1). Scale bar indicates substitutions per nucleic site.

separation between the Western (Europe and Africa) strains and the East Asian strains, a result that is also seen in the co-ancestry matrix (fig. 2*a*).

An exception to this separation is PeCan4. In the phylogenetic network (fig. 3), the PeCan4 genome appears to be a hybrid between Amerind genomes and Western genomes. The co-ancestry matrix (fig. 2) indicates that PeCan4 is a hybrid strain in that it has received a considerable number of chunks from the Amerind, Africa1_sg1, and European strains. The co-ancestry matrix (fig. 2) also indicates that SJM180 is a hybrid between Africa1_sg1 and European subgroups. This interpretation is consistent with the phylogenetic network (fig. 3). Thus, comparison of the co-ancestry matrix with the phylogenetic network reveals novel genome characteristics of these two strains, demonstrating the power of chromosome painting.

Signatures of Admixture Events

The boxes in figure 2*b* indicate evident signs of admixture. Here, we use the term “admixture” when there is a flow of chunks (more generally, “genetic flux”) between (sub)groups. It seems that the two European subgroups were evidently admixed with Africa1 and Asia2 populations, which is consistent with previous reports using MLST data (Falush, Wirth et al. 2003; Moodley et al. 2012). Conversely, Africa1 and Asia2 were evidently admixed with European populations, which is also consistent with a previous report (Linz et al. 2007). Interestingly, signs of admixture can also be seen in EastAsia_sg3 and EastAsia_sg4 (OK113 from Okinawa) subgroups—29–31% of the chunks were estimated to be imported from European and Asia2 populations, which is significantly higher than imported into the other subgroups in East Asia and Amerind ($P < 0.005$, Wilcoxon’s rank sum test). Although this seems to be a result of relatively recent admixture, there is currently no way to infer the date of admixture based on the co-ancestry matrix.

Asymmetry in Genetic Flux

In general, export from country A to country B might be larger than import from country B to country A. The co-ancestry matrix revealed such asymmetries in the genetic flux between two populations. Figure 2*a* indicate East Asia and Amerind populations imported a smaller number of chunks than they exported ($P < 10^{-15}$, Wilcoxon’s rank sum test) to external populations (African, European, and Asia2 populations). We will discuss its interpretation later (“Evolution of Amerind/ East Asia *H. pylori*” in Discussion).

Global View of Genetic Fluxes among Subgroups

For the subgroups identified by fineSTRUCTURE, we visualized the extents of genetic fluxes. Using the co-ancestry matrix, the proportion of chunks copied from a donor to a

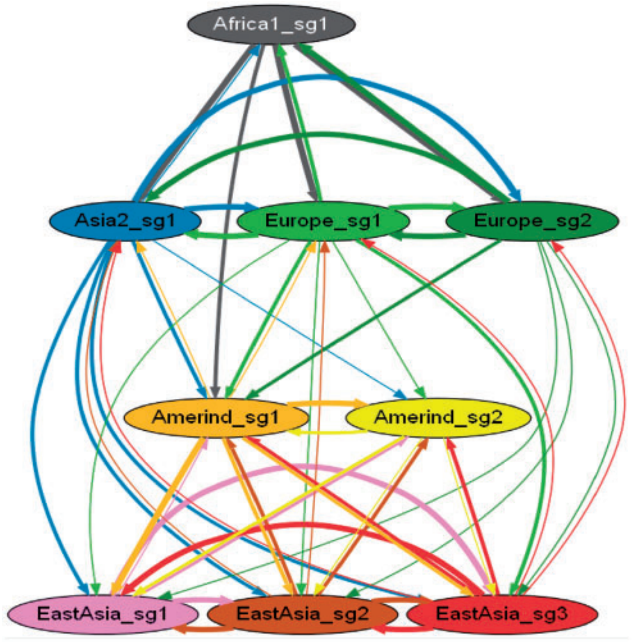

FIG. 4. Genetic fluxes between subgroups. Width of an arrow (in three grades) indicates the extent of flux. Arrows representing a small flux were omitted for clarity as explained in the text. Color of an arrow indicates the donor.

recipient subgroup on average was calculated (supplementary table S1, Supplementary Material online). The result is shown in figure 4. Arrows with the average proportion of chunks copied from a donor to a recipient subgroup less than 5.4% (first quartile of the values of all arrows) were omitted. Singletons were not included in the figures for simplicity. The figure illustrates that genetic flux into the admixed subgroups described earlier (Africa1_sg1, Europe_sg1, Europe_sg2, Asia2_sg1, and East Asia_sg3) is relatively large. Genetic flux into other subgroups can also be seen, indicating a complex network of genetic flux among *H. pylori* populations.

Discussion

Comparison of Methods for Population Structure Inference

We were the first group to apply “chromosome painting” in silico and fineSTRUCTURE to organisms other than humans in this study. We chose the bacterium *H. pylori* because humans and *H. pylori* are both highly recombining and share, to some extent, a phylogeographically differentiated population structure (Moodley and Linz 2009).

A recent study in humans demonstrated that chromosome painting followed by fineSTRUCTURE analysis is able to capture a more subtle, recent population structure compared with structures generated from STRUCTURE analysis or PCA (Lawson et al. 2012). The advantage of chromosome painting and fineSTRUCTURE became evident when accounting for linkage information, an advantage that is supported in a recent review comparing various algorithms for population identification (Lawson and Falush 2012). In this study, the analysis of chromosome painting and fineSTRUCTURE indeed revealed a population structure with more subgroups and singletons than were revealed by previous methods. Meanwhile, the benefit of accounting for linkage information was not clear because almost the same population structure was inferred by the no-linkage approach, which ignores linkage information as PCA. This is probably because the linkage is very weak in *H. pylori* because of its high recombination. Under such a condition with a large number of markers without linkage, fineSTRUCTURE and PCA utilize similar information (Lawson et al. 2012). However, we think there are at least three other major advantages to using the new method.

First and most importantly, chromosome painting and fineSTRUCTURE can elucidate the extent and direction of genetic flux between subgroups as shown in figure 2. This is impossible for the phylogenetic trees (supplementary fig. S2, Supplementary Material online), which do not take genetic flux between subgroups into account. The neighbor-net phylogenetic networks (fig. 3) can also visualize genetic flux between subgroups. However, their complex signatures are hard to interpret, and their output gives no information about the direction of genetic flux. Meanwhile, the new method is useful in visualizing the extent and direction of genetic flux and has disentangled the complex network interconnected by rampant recombination events.

The second advantage is its computational efficiency. STRUCTURE can also handle hundreds of thousands of SNPs as seen in human population genetic studies (Jakobsson et al. 2008). However, chromosome painting combined with fineSTRUCTURE is much faster than STRUCTURE. Furthermore, the computation is easily parallelizable and can therefore be applied to hundreds of genome sequences.

The third advantage is that chromosome painting is applicable even when sample size is small as in this study. In contrast, STRUCTURE has difficulty in population assignment for small sample sizes (Rosenberg et al. 2002; Yang et al. 2005). It requires at least 15–20 individuals per hypothesized population to achieve accurate clustering (Rosenberg et al. 2001) due to its assumptions of Hardy–Weinberg equilibrium and linkage equilibrium within each population. In this study, we applied STRUCTURE to the seven housekeeping (MLST) genes of more than 1,000 strains, but we were unable to apply STRUCTURE to the data of genome-wide SNPs of the 29 strains. Similarly, we were unable to analyze the data by another popular program BAPS because it is also based on assumptions of Hardy–Weinberg equilibrium and linkage equilibrium within each population, which becomes problematic with small sample sizes.

Genetic Flux between Subgroups

As mentioned earlier, an essential advantage of our approach is the ability to determine the extent and direction of genetic flux between identified population subgroups. Genetic flux among populations through homologous recombination has begun to be examined at the genomic scale. A recent study (Didelot et al. 2012) inferred the extent of genetic flux between four phylogroups in *Escherichia coli* by using ClonalOrigin (Didelot et al. 2010). ClonalOrigin assumes a clonal genealogy with some additional edges representing recombination. The clonal genealogy can be inferred by ClonalFrame (Didelot and Falush 2007b), which has been used in another related study (Didelot et al. 2011) that quantified the flux between five lineages in *Salmonella*. According to the developers however, this method is not appropriate if the recombination rate is high enough to obscure clonal structure, as is the case with *H. pylori* (Didelot and Falush 2007a). Therefore in this study, we used chromosome painting and fineSTRUCTURE instead.

A recent study developed another method to detect homologous recombinant segments imported from external populations (Marttinen et al. 2012). The application of this method to 241 genomes of *Streptococcus pneumoniae* demonstrated its ability to handle hundreds of genome sequence. Unlike chromosome painting however, it does not model recombination events between the observed sequences.

Another recent study reconstructed recombination events between lineages of *Chlamydia trachomatis* (Harris et al. 2012). The study succeeded in inferring donor branches by combining ancestral sequence reconstruction and a test statistics for detecting recombination (Croucher et al. 2011) to identify genomic regions in which the node sequence was likely derived from a distant branch on the phylogenetic tree. However, this method reconstructs a phylogenetic tree after removing the effects of recombination. Such a removal would make phylogenetic tree construction impossible for a species like *H. pylori*, where recombination occurs throughout the genome. Chromosome painting and fineSTRUCTURE are a more appropriate choice for such a highly recombining organism because they do not depend on reconstruction of a phylogenetic tree.

Evolution of Amerind/East Asia *H. pylori*

In the analysis by chromosome painting and fineSTRUCTURE, Amerind and East Asian *H. pylori* were interesting because more subgroups were identified compared to the other groups of *H. pylori*. Additionally, these two subgroups imported a significantly smaller number of chunks than they exported to external populations. Subgroups in the East Asian group (EastAsia_sg3 and EastAsia_sg4) also showed stronger signatures of population admixture. In the following paragraphs, we discuss relationships between these findings and previous studies on *H. pylori* and their host *Homo sapiens*.

A previous study on Amerind showed a distinctive *H. pylori* population in the Peruvian Amazon that had differentiated from Peruvian, Spainish, and Japanese strains (Kersulyte et al. 2010). The *H. pylori* strain Shi470 included in this study was sampled from the Peruvian Amazon, and phylogenetic analysis using a single locus identified no additional subgroups (Kersulyte et al. 2010). Our analysis was able to distinguish as many as four distinct subgroups (Amering_sg1 through Amerind_sg4) as well as a hybrid (PeCan4) for the Amerind area.

With regard to East Asia, we have recently obtained the complete genome sequences of four Japanese strains and examined the genomic characteristics of *H. pylori* in East Asia (Kawai et al. 2011). This previous study did not detect any subgroups and did not examine any sign of demographic events or genetic flux between subgroups. This study is the first to report these characteristics of *H. pylori* in East Asia.

Studies on human evolution in East Asia have been actively conducted using mitochondrial DNA, Y chromosomes, and genome-wide SNPs (Hammer and Horai 1995; Horai et al. 1996; Yamaguchi-Kabata et al. 2008; Ding et al. 2011; Peng and Zhang 2011). The ancestral Japanese human populations likely originated from two major migration events from the Asian continent (“the dual-origin hypothesis”), resulting in the Jomon (proposed to be a direct ancestor of Okinawa people) and Yayoi populations (Hanihara 1991; Hammer and Horai 1995; Horai et al. 1996; Hammer et al. 2006). Genetic studies support the idea that modern mainland Japanese derived from an admixture of the Yayoi and Jomon people. Among the East Asian subgroups of *H. pylori* identified by the present study, EastAsia_sg1 is purely (4/4) from Japan (excluding Okinawa), EastAsia_sg2 is from Japan, Korea, and Okinawa, EastAsia_sg3 is from Japan and Korea, and EastAsia_sg4 is from Okinawa. Parts of EastAsia_sg3 and EastAsia_sg4 showed stronger signatures of population admixture. Compared with these subgroups (EastAsia_sg3 and EastAsia_sg4), the signature of population admixture was weaker in the purely Japanese subgroup, EastAsia_sg1. These differences may reflect historical human population movements and admixture during the formation of modern human populations in East Asia. However, we currently have no additional evidence to connect EastAsia_sg1 with the Yayoi people or EastAsia_sg3 and EastAsia_sg4 groups with the Jomon people.

Additionally, it is interesting to consider what mechanism caused the asymmetric genetic flux in the Amerind and East Asian groups. The observation that they imported a smaller number of chunks than they exported to the external populations (African, European, and Asia2 populations) could indicate a less efficient genetic mechanism to import foreign DNA, or a different selective environment which had suppressed such imports in the past. The former possibility could be tested directly by experimentation, and the latter could be examined by simulating histories of populations with different scenarios of selection and comparing patterns of genetic flux in co-ancestry matrices.

Influence of Sample Structure and Concluding Remarks

This study used 29 complete genome sequences of *H. pylori*. The number of available genome sequences of *H. pylori* has been limited so far, and the genomes used in this study do not represent a systematic sample of the species' diversity. Sampling bias will have strong effect on inference of population structure and admixtures.

For example, genomes of the hpSahul population (in New Guinea and Australia) (Moodley et al. 2009) are not available and thus were not included in this study. They split from Asian populations of *H. pylori* before the Asian populations split into hpAsia2 (Central Asia) and hpEastAsia. Most likely, inclusion of the hpSahul genomes would reveal an additional subgroup(s) with additional information about genetic flux. Some chunks in the Amerind and East Asia genomes would probably be inferred to have derived from this unexamined region. Such results might affect the currently observed asymmetric genetic flux in the Amerind and East Asian groups. It should be noted that inference of admixture will sensibly depend on bias of donor genomes in the data set. To avoid this problem, we need to sample a broad range of genomes in a way that reflects the diversity of the species.

Similarly, the inference of hybrid ancestry of the singleton strains can also be affected by the sampling bias. If a closely related strain were found and included in the analysis, it would be clustered with the previous singleton strains, possibly resulting in the disappearance of the evident signature of hybridization with other populations. We should keep this possibility in mind when interpreting the hybrid ancestry of singleton strains.

Therefore, sampling bias currently limits our ability to interpret the results of population structure and admixture. Questions such as how many subgroups exist in a region and to what extent they are admixed with other subgroups will be difficult to answer reliably without reducing this bias. By including more genome sequences in future analyses, we could reduce the sampling bias and interpret the results with greater confidence. Because the algorithms we used in this study can be applied to hundreds of genome sequences, it will be interesting and desirable to analyze more genome sequences that have been systematically sampled from various regions.

In summary, by making use of chromosome painting and fineSTRUCTURE algorithms on complete genome sequences of a bacterial species, we were able to reveal both population structure at a finer scale as well as the extent and direction of genetic flux between subgroups of a bacteria subject to frequent recombination events. This procedure will form a basis for applying the novel algorithms to large-scale population genomic data of other species.

Materials and Methods

Sequencing *H. pylori* Strains from Okinawa

Helicobacter pylori strains collected in Okinawa Prefectural Chubu Hospital, Uruma, Okinawa, Japan (OK107, OK130, OK139, OK144, OK155, OK160, OK180, OK181, OK185, OK187, OK204, and OK210 (Yamazaki et al. 2005); OK113, OK168, OK308, OK310 (Satomi et al. 2006); OK112, (Azuma et al. 2004); OK188 (this study) were used. OK188 was separated and cultivated from an epithelium biopsy tissue from a patient with duodenal ulcer. For strains carrying *cagA* genes, phylogenetic trees of *cagA* genes were drawn and OK107, OK113, OK130, OK139, OK144, OK155, OK160, OK180, OK181, OK185, OK187, OK204, OK210, OK308, OK310, and OK168 all carried *cagA* of the J-Western type (Truong et al. 2009; Furuta, Yahara, et al. 2011). Standard multilocus sequence typing was applied to all strains. Nucleotide sequence data of seven housekeeping (MLST) genes of other *H. pylori* strains were downloaded from the database pubMLST (http://pubmlst.org/, last accessed April 5, 2013). In the resulting neighbor-joining phylogenetic tree (supplementary fig. S4, Supplementary Material online), OK 113, OK310, OK188, OK130, OK139, OK204, and OK168 cluster with East Asian strains. Among them, OK 113 and OK310 were chosen for sequencing. OK112 and OK210 are similar to each other and cluster with European strains 26695, G27 and P12. OK107, OK144, OK155, OK160, OK180, OK181, OK185, OK187, and OK308 are very similar to one another and cluster with a European strain HPAG1.

From OK 113 and OK310, genomic DNA was isolated as described earlier (Kawai et al. 2011). In brief, cells were inoculated onto a fresh TSA-II plate from 20% glycerol stock and cultured at 37 °C for 3 days under micro-aerobic conditions (O_2, 5%; CO_2, 15%; N_2, 80%). Colonies were collected and transferred into 20 ml of Brucella broth culture medium with 10% fetal calf serum and cultured at 37 °C for 3 days under micro-aerobic conditions. Genomic DNA was extracted from culture pellets by the protease/phenol–chloroform method and eluted in 300 μl of TE buffer (10 mM–Tris HCl, 1 mM EDTA).

Genome sequences were determined by the whole-genome shotgun strategy using Sanger sequencing. Approximately 20 μg of genomic DNA was sheared using a HYDROSHEAR (Gene Machine). DNA fragments were fractionated by agarose gel electrophoresis and subcloned into the plasmid pTS1 vector (Nippon Gene) to construct shotgun libraries with an average insert size of 3 and 10 kb using the

3730xl sequencer (Applied Biosystems). Template DNA was prepared from an aliquot of bacterial culture by amplifying the inserted DNA of each clone using PCR.

We produced 19,200 and 17,664 reads of the genomes of strains 113 and 310 by sequencing both ends of the clones, giving 7.9- and 7.7-fold coverage, respectively. The sequencing reads were assembled with the Phred-Phrap-Consed program (Gordon et al. 2001) and gaps were closed by direct sequencing of either clones that spanned the gaps or PCR products amplified with oligonucleotide primers designed to anneal to each end of the neighboring contigs. Finally, a finished sequence with an error rate of less than 1 per 10,000 bases (QV - 40) was obtained. The obtained sequences were deposited in DDBJ with the following accession numbers: OK113, AP012600; OK310 chromosome, AP012601; and OK310 plasmid, AP012602.

Genome Sequences and Alignments of Their Genes

We used complete genome sequences of 29 *H. pylori* strains collected from various parts of the world. Names and accession numbers are as follows: 26695, NC_000915.1; J99, NC_000921.1; HPAG1, NC_008086.1 and NC_008087.1; Shi470, NC_010698.2; G27, NC_011333.1 and NC_011334.1; P12, NC_011498.1 and NC_011499.1; F57, DDBJ:AP011945; F32, DDBJ:AP011943 and AP011944; F30, DDBJ:AP011941 and AP011942; F16, DDBJ:AP011940; B38, NC_012973.1; 51, CP000012.1; 53bP001680.1; v225d, CP001582.1 and CP001583.1; B8, NC_014256.1 and NC_014257.1; SJM180, NC_014560.1; PeCan4, NC_014555.1 and NC_014556.1; Cuz20, CP002076.1; Sat464, CP002071.1 and CP002072.1; OK113, DDBJ: AP012600; OK310, DDBJ: AP012601 and AP012602; 35A, CP002096.1; 83, CP002605.1; Gambia94, CP002332.1 and CP002333.1; India7, CP002331.1; Lithuania75, CP002334.1 and CP002335.1; Puno120, CP002980.1 and CP002981.1; Puno135, CP002982.1; Santal49, CP002983.1 and CP002984.1.

An entire data set of one-to-one orthologous genes in the core genomic regions was prepared through ortholog clustering by CoreAligner (Uchiyama 2008), DomClust (Uchiyama 2006), and RECOG (http://mbgd.genome.ad.jp/RECOG/, last accessed April 5, 2013). Alignment of each orthologous gene was conducted by MAFFT (Katoh et al. 2005). SNP call of each orthologous gene was conducted by adegenet (Jombart 2008). We combined the SNPs while preserving information of SNP positions to prepare genome-wide haplotype data. The genome of strain 26695 was used as a reference to record and examine positions of SNPs.

Chromosome Painting In Silico and fineSTRUCTURE Analysis

"Chromosome painting" was applied to the genome-wide haplotype data by the linkage model implemented in ChromoPainter (Lawson et al. 2012) (version 0.02). We followed the instructions from the official web page (http://www.paintmychromosomes.com, last accessed April 5, 2013). We prepared a recombination map file by specifying the same recombination rate per-site per-generation for the SNPs based on previous estimates of recombination rate and generation time (Webb and Blaser 2002; Morelli, Didelot, et al. 2010). The results were visualized in UCSC (The University of California Santa Cruz) browser (Schneider et al. 2006) after filtering SNPs with uncertain estimates of their donor and chunks (a series of SNPs with the same expected donor) of more than 20 kb (because of sparse distribution of SNPs).

For fineSTRUCTURE (version 0.02) (Lawson et al. 2012), both the burn-in and Markov chain Monte Carlo (MCMC) chain after the burn-in were run for 100,000 iterations. The thin interval was specified as 100. We performed the inference twice at the same parameter values and confirmed the population assignments.

Population Assignment by STRUCTURE Using Multilocus Sequence Typing Data

We also conducted population assignment by the "no admixture" model of the program STRUCTURE version 2.0 as described in previous studies (Linz et al. 2007; Moodley et al. 2009). We used nucleotide sequences of 7 housekeeping (MLST) genes of all *H. pylori* strains registered in the database pubMLST (http://pubmlst.org/, last accessed April 5, 2013) plus those of strains we sequenced in earlier work (Furuta, Kawai, et al. 2011; Kawai et al. 2011) and in this work. A burn-in and MCMC chain after burn-in were conducted for 10,000 and 20,000 iterations, respectively. We varied the parameter K (number of population) from 7 (Moodley et al. 2009) through 11, and used $K = 9$, which maximized likelihood. Similar to previous works (Linz et al. 2007; Moodley et al. 2009), we assigned each genome into a traditional group (e.g., hpEurope).

Construction of a Phylogenetic Network and Tree

A neighbor-net phylogenetic network and a neighbor-joining tree were also constructed from concatenated nucleotide sequence alignments of the genome-wide haplotype by SplitsTree4 (Huson and Bryant 2006) and by MEGA5 (Tamura et al. 2012), respectively.

Supplementary Material

Supplementary figures S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Daniel Lawson and Daniel Falush for personal discussion with us. The computational calculations were done at the Human Genome Center at the Institute of Medical Science at the University of Tokyo and at Max Plank Institute. This work was supported by JSPS Research Fellowships for Young Scientists to K.Y., the global COE (center of excellence) project "Genome Information Big Bang" from the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) of the Japanese government, "Grants-in-Aid for Scientific Research" (21370001, 25291080) from the Japan Society for the Promotion of Science, "Grant-in-Aid for Scientific Research on Innovative

Areas" (24113506, 24119503) from MEXT, and Grant in Promotion of Basic Research Activities for Innovative Biosciences (121205003001002100019) from Bio-oriented Technology Research Advance Institution to I.K.

References

Achtman M. 2008. Evolution, population structure, and phylogeography of genetically monomorphic bacterial pathogens. *Annu Rev Microbiol.* 62:53–70.

Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, Kruglyak L. 2004. Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol.* 2:e286.

Argudin MA, Argumosa V, Mendoza MC, Guerra B, Rodicio MR. 2013. Population structure and exotoxin gene content of methicillin-susceptible *Staphylococcus aureus* from Spanish healthy carriers. *Microb Pathog.* 54:26–33.

Azuma T, Yamazaki S, Yamakawa A, et al. (13 co-authors). 2004. Association between diversity in the Src homology 2 domain—containing tyrosine phosphatase binding site of *Helicobacter pylori* CagA protein and gastric atrophy and cancer. *J Infect Dis.* 189:820–827.

Corander J, Connor TR, O'Dwyer CA, Kroll JS, Hanage WP. 2011. Population structure in the *Neisseria*, and the biological significance of fuzzy species. *J R Soc Interface.* 9:1208–1215.

Corander J, Marttinen P, Siren J, Tang J. 2008. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics* 9:539.

Corander J, Tang J. 2007. Bayesian analysis of population structure based on linked molecular information. *Math Biosci.* 205:19–31.

Croucher NJ, Harris SR, Fraser C, et al. (24 co-authors). 2011. Rapid pneumococcal evolution in response to clinical interventions. *Science* 331:430–434.

Devlin B, Roeder K. 1999. Genomic control for association studies. *Biometrics* 55:997–1004.

Didelot X, Bowden R, Street T, et al. (12 co-authors). 2011. Recombination and population structure in *Salmonella enterica*. *PLoS Genet.* 7:e1002191.

Didelot X, Falush D. 2007a. ClonalFrame user guide. http://www.stats.ox.ac.uk/~didelot/files/clonalframe-userguide.pdf (last accessed April 5, 2013).

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

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 MC. 2010. Impact of recombination on bacterial evolution. *Trends Microbiol.* 18:315–322.

Didelot X, Meric G, Falush D, Darling AE. 2012. Impact of homologous and non-homologous recombination in the genomic evolution of *Escherichia coli*. *BMC Genomics* 13:256.

Ding Q, Wang C, Faria S, Li H. 2011. Mapping human genetic diversity on the Japanese archipelago. *Adv Anthropol.* 1:19–25.

Doolittle WF, Zhaxybayeva O. 2009. On the origin of prokaryotic species. *Genome Res.* 19:744–756.

Downing T, Imamura H, Decuypere S, et al. (20 co-authors). 2011. Whole genome sequencing of multiple *Leishmania donovani* clinical isolates provides insights into population structure and mechanisms of drug resistance. *Genome Res.* 21:2143–2156.

Falush D, Stephens M, Pritchard JK. 2003. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 164:1567–1587.

Falush D, Wirth T, Linz B, et al. (13 co-authors). 2003. Traces of human migrations in *Helicobacter pylori* populations. *Science* 299: 1582–1585.

Feil EJ, Spratt BG. 2001. Recombination and the population structures of bacterial pathogens. *Annu Rev Microbiol.* 55:561–590.

Furuta Y, Kawai M, Yahara K, et al. (12 co-authors). 2011. Birth and death of genes linked to chromosomal inversion. *Proc Natl Acad Sci U S A.* 108:1501–1506.

Furuta Y, Yahara K, Hatakeyama M, Kobayashi I. 2011. Evolution of cagA oncogene of *Helicobacter pylori* through recombination. *PLoS One* 6: e23499.

Gordon D, Desmarais C, Green P. 2001. Automated finishing with autofinish. *Genome Res.* 11:614–625.

Hammer MF, Horai S. 1995. Y chromosomal DNA variation and the peopling of Japan. *Am J Hum Genet.* 56:951–962.

Hammer MF, Karafet TM, Park H, Omoto K, Harihara S, Stoneking M, Horai S. 2006. Dual origins of the Japanese: common ground for hunter-gatherer and farmer Y chromosomes. *J Hum Genet.* 51: 47–58.

Hanage WP, Fraser C, Tang J, Connor TR, Corander J. 2009. Hyper-recombination, diversity, and antibiotic resistance in pneumococcus. *Science* 324:1454–1457.

Hanihara K. 1991. Dual structure model for the population history of the Japanese. *Japan Rev.* 2:1–33.

Harris SR, Clarke IN, Seth-Smith HM, et al. (24 co-authors). 2012. Whole-genome analysis of diverse *Chlamydia trachomatis* strains identifies phylogenetic relationships masked by current clinical typing. *Nat Genet.* 44:413–419, S1.

Henn BM, Botigue LR, Gravel S, et al. (12 co-authors). 2012. Genomic ancestry of North Africans supports back-to-Africa migrations. *PLoS Genet.* 8:e1002397.

Hershberg R, Lipatov M, Small PM, et al. (11 co-authors). 2008. High functional diversity in *Mycobacterium tuberculosis* driven by genetic drift and human demography. *PLoS Biol.* 6:e311.

Holsinger KE, Weir BS. 2009. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nat Rev Genet.* 10:639–650.

Holt KE, Baker S, Weill FX, et al. (12 co-authors). 2012. *Shigella sonnei* genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nat Genet.* 44:1056–1059.

Horai S, Murayama K, Hayasaka K, Matsubayashi S, Hattori Y, Fucharoen G, Harihara S, Park KS, Omoto K, Pan IH. 1996. mtDNA polymorphism in East Asian Populations, with special reference to the peopling of Japan. *Am J Hum Genet.* 59:579–590.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. *Mol Biol Evol.* 23:254–267.

Jakobsson M, Scholz SW, Scheet P, et al. (12 co-authors). 2008. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451:998–1003.

Jombart T. 2008. Adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* 24:1403–1405.

Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* 33:511–518.

Kawai M, Furuta Y, Yahara K, et al. (12 co-authors). 2011. Evolution in an oncogenic bacterial species with extreme genome plasticity: *Helicobacter pylori* East Asian genomes. *BMC Microbiol.* 11:104.

Kennemann L, Didelot X, Aebischer T, et al. (12 co-authors). 2011. *Helicobacter pylori* genome evolution during human infection. *Proc Natl Acad Sci U S A.* 108:5033–5038.

Kersulyte D, Kalia A, Gilman RH, et al. (15 co-authors). 2010. *Helicobacter pylori* from Peruvian Amerindians: traces of human migrations in strains from remote Amazon, and genome sequence of an Amerind strain. *PLoS One* 5:e15076.

Lawson D, Falush D. 2012. Population identification using genetic data. *Annu Rev Genomics Hum Genet.* 13:337–361.

Lawson DJ, Hellenthal G, Myers S, Falush D. 2012. Inference of population structure using dense haplotype data. *PLoS Genet.* 8:e1002453.

Linz B, Balloux F, Moodley Y, et al. (12 co-authors). 2007. An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature* 445:915–918.

Marttinen P, Hanage WP, Croucher NJ, Connor TR, Harris SR, Bentley SD, Corander J. 2012. Detection of recombination events in bacterial genomes from large population samples. *Nucleic Acids Res.* 40:e6.

Menozzi P, Piazza A, Cavalli-Sforza L. 1978. Synthetic maps of human gene frequencies in Europeans. *Science* 201:786–792.

Moodley Y, Linz B. 2009. *Helicobacter pylori* sequences reflect past human migrations. In: de Reuse H, Bereswill S, editors. Microbial pathogenomics. Basel (Switzerland): Karger. p. 62–74.

Moodley Y, Linz B, Bond RP, et al. (12 co-authors). 2012. Age of the association between *Helicobacter pylori* and man. *PLoS Pathog.* 8: e1002693.

Moodley Y, Linz B, Yamaoka Y, et al. (15 co-authors). 2009. The peopling of the Pacific from a bacterial perspective. *Science* 323:527–530.

Morelli G, Didelot X, Kusecek B, Schwarz S, Bahlawane C, Falush D, Suerbaum S, Achtman M. 2010. Microevolution of *Helicobacter pylori* during prolonged infection of single hosts and within families. *PLoS Genet.* 6:e1001036.

Morelli G, Song Y, Mazzoni CJ, et al. (24 co-authors). 2010. *Yersinia pestis* genome sequencing identifies patterns of global phylogenetic diversity. *Nat Genet.* 42:1140–1143.

Nakamura T, Shoji A, Fujisawa H, Kamatani N. 2005. Cluster analysis and association study of structured multilocus genotype data. *J Hum Genet.* 50:53–61.

Novembre J, Ramachandran S. 2011. Perspectives on human population structure at the cusp of the sequencing era. *Annu Rev Genomics Hum Genet.* 12:245–274.

Okoro CK, Kingsley RA, Connor TR, et al. (21 co-authors). 2012. Intracontinental spread of human invasive *Salmonella typhimurium* pathovariants in sub-Saharan Africa. *Nat Genet.* 44:1225–1221.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet.* 2:e190.

Peng MS, Zhang YP. 2011. Inferring the population expansions in peopling of Japan. *PLoS One* 6:e21509.

Perez-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.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* 155:945–959.

Robinson DA, Feil EJ, Falush D. 2010. Bacterial population genetics in infectious disease. Hoboken (NJ): Wiley-Blackwell.

Rosenberg NA, Burke T, Elo K, et al. (12 co-authors). 2001. Empirical evaluation of genetic clustering methods using multilocus genotypes from 20 chicken breeds. *Genetics* 159:699–713.

Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW. 2002. Genetic structure of human populations. *Science* 298:2381–2385.

Satomi S, Yamakawa A, Matsunaga S, et al. (13 co-authors). 2006. Relationship between the diversity of the cagA gene of *Helicobacter pylori* and gastric cancer in Okinawa, Japan. *J Gastroenterol.* 41:668–673.

Schneider KL, Pollard KS, Baertsch R, Pohl A, Lowe TM. 2006. The UCSC archaeal genome browser. *Nucleic Acids Res.* 34:D407–D410.

Snel B, Huynen MA, Dutilh BE. 2005. Genome trees and the nature of genome evolution. *Annu Rev Microbiol.* 59:191–209.

Suerbaum S, Josenhans C. 2007. *Helicobacter pylori* evolution and phenotypic diversification in a changing host. *Nat Rev Microbiol.* 5: 441–452.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S. 2012. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol.* 28:2731–2739.

Truong BX, Mai VT, Tanaka H, et al. (11 co-authors). 2009. Diverse characteristics of the CagA gene of *Helicobacter pylori* strains collected from patients from southern Vietnam with gastric cancer and peptic ulcer. *J Clin Microbiol.* 47:4021–4028.

Uchiyama I. 2006. Hierarchical clustering algorithm for comprehensive orthologous-domain classification in multiple genomes. *Nucleic Acids Res.* 34:647–658.

Uchiyama I. 2008. Multiple genome alignment for identifying the core structure among moderately related microbial genomes. *BMC Genomics* 9:515.

Webb GF, Blaser MJ. 2002. Dynamics of bacterial phenotype selection in a colonized host. *Proc Natl Acad Sci U S A.* 99:3135–3140.

Yahara K, Kawai M, Furuta Y, et al. (12 co-authors). 2012. Genome-wide survey of mutual homologous recombination in a highly sexual bacterial species. *Genome Biol Evol.* 4:628–640.

Yamaguchi-Kabata Y, Nakazono K, Takahashi A, Saito S, Hosono N, Kubo M, Nakamura Y, Kamatani N. 2008. Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies. *Am J Hum Genet.* 83:445–456.

Yamazaki S, Yamakawa A, Okuda T, et al. (11 co-authors). 2005. Distinct diversity of vacA, cagA, and cagE genes of *Helicobacter pylori* associated with peptic ulcer in Japan. *J Clin Microbiol.* 43: 3906–3916.

Yang BZ, Zhao H, Kranzler HR, Gelernter J. 2005. Practical population group assignment with selected informative markers: characteristics and properties of Bayesian clustering via STRUCTURE. *Genet Epidemiol.* 28:302–312.

Zhi XY, Zhao W, Li WJ, Zhao GP. 2012. Prokaryotic systematics in the genomics era. *Antonie Van Leeuwenhoek* 101:21–34.