Biological Species Are Universal across Life's Domains

Louis-Marie Bobay* and Howard Ochman

Department of Integrative Biology, University of Texas at Austin

*Corresponding author: E-mail: lbobay@utexas.edu.

Accepted: February 7, 2017

Abstract

Delineation of species is fundamental to organizing and understanding biological diversity. The most widely applied criterion for distinguishing species is the Biological Species Concept (BSC), which defines species as groups of interbreeding individuals that remain reproductively isolated from other such groups. The BSC has broad appeal; however, many organisms, most notably asexual lineages, cannot be classified according to the BSC. Despite their exclusively asexual mode of reproduction, Bacteria and Archaea can transfer and exchange genes though homologous recombination. Here we show that barriers to homologous gene exchange define biological species in prokaryotes with the same efficacy as in sexual eukaryotes. By analyzing the impact of recombination on the polymorphisms in thousands of genome sequences, we find that over half of named bacterial species undergo continuous recombination among sequenced constituents, indicative of true biological species. However, nearly a quarter of named bacterial species show sharp discontinuities and comprise multiple biological species. These interruptions of gene flow are not a simple function of genome identity, indicating that bacterial speciation does not uniformly proceed by the gradual divergence of genome sequences. The same genomic approach based on recombinant polymorphisms retrieves known species boundaries in sexually reproducing eukaryotes. Thus, a single biological species definition based on gene flow, once thought to be limited only to sexually reproducing organisms, is applicable to all cellular lifeforms.

Key words: speciation, recombination, Biological Species Concept, gene flow, asexuality.

Introduction

The most widely applied criterion for delineating species is the Biological Species Concept (BSC), which defines species as groups of interbreeding individuals that remain reproductively isolated from other such groups (Mayr 1942). Because it is based on genetic features, that is, the capacity for gene flow, rather than typological features, the BSC has broad appeal, particularly in defining plant and animal species. However, many contemporary organisms, most notably asexual and parthenogenetic lineages, defy classification by the BSC.

Similarly, the BSC cannot be applied to bacteria, the most ancient and widespread Domain of life—or can it? Bacteria, despite their exclusively asexual mode of reproduction, have long been known to transfer genes among themselves, presenting the possibility that barriers to gene exchange might also provide a metric for categorizing bacterial species. Strictly clonal bacteria, like asexual animals, are not amenable to classification by the BSC. But if gene exchange is limited to restricted sets of lineages—a process tantamount to reproductive isolation—then bacterial species might be distinguished in the same robust manner as animal species.

In bacteria, gene exchange can take on two forms: homologous recombination, which involves the replacement of resident genes with alleles from another lineage, and lateral gene transfer, which introduces new genes, leaving the sequence of resident genes intact. When quantifying species- or population-level diversity in bacteria, we are interested solely in recombination by homologous exchange. Because recombination has made a measurable impact on the allelic variation in the majority of bacterial species (Vos and Didelot 2009), we sought to determine whether barriers to recombination define species in bacteria, as they do in sexual eukaryotes. It has previously been argued that the interruption of gene flow is the critical component of bacterial speciation (Rayssiguier, et al. 1989; Dykhuizen and Green 1991; Ochman et al. 2005; Hanage et al. 2006; Fraser et al. 2007, 2009; Lawrence and Retchless 2009; Cadillot-Quiroz, et al. 2012; Shapiro et al. 2012, 2016), but the lack of genomic information for all but a few select species, and the absence of fast and efficient methods to quantify gene flow in such large data sets, have severely limited its application to bacteria as a whole.

In this study, we used a genome-based approach for detecting the patterns and interruption of gene flow among strains classified to a named bacterial species. In many cases, there were strong indications of true biological species comprising strains that were connected through gene exchange. However, there were often clear signals of strains that were genetically isolated and must be considered as separate biological species. The interruption of gene flow is not dependent on the degree of genetic divergence, discouraging the application of sequence identity thresholds to define bacterial species. The same approach readily distinguishes known species boundaries in sexually reproducing organisms and works equally well in defining species of Archaea, demonstrating that a single biological species definition is applicable to all cellular lifeforms.

Materials and Methods

Species Sampling and Defining Core Genomes

We downloaded the complete set of genomes of each bacterial and archaeal species (according to the named species designations at the NCBI website [ftp.ncbi.nlm.nih.gov/genomes/] on June 2015) represented by at least 20 genomes, resulting in a total of 20,690 genomes for 105 taxonomically defined species. For each named species, the protein sequences of all pairs of strains were compared with Usearch Global (Edgar 2010), and orthologs were defined as reciprocal best hits having at least 70% sequence identity and 80% length conservation. Gene families were curated to remove potential paralogs: each gene family was considered as part of the core genome if present exactly once in every strain within a species. When assembling core genomes, several genomes were manually excluded from the analysis when very few homologs could be identified, suggesting low quality or annotation issues. In species that comprise large numbers of sequenced strains ($n \geq 200$), performing all pairwise comparisons becomes computationally time-consuming, so core genomes were assembled in multiple steps: initially, core genomes were built for subgroups of up to 100 strains. The composite core genome was based on one genome from each subgroup such that those gene families that were present in every subgroup were assigned to the species' core genome. The core genes present in each sequenced representative of a species were aligned with MAFFT v7 (Katoh and Standley 2013) and merged into a single concatenate, from which we inferred pairwise distances D using RAxML v7 under a GTR + Γ model (Stamatakis 2006). In many species, the sequenced strains were identical (or very nearly so), so we randomly excluded strains whose core genome sequences were highly similar ($D < 0.00005$) to remove redundancies from the data set.

Several species comprised a very large number of strains (up to 4,221 genomes per species), and our quality-filtering procedures sometimes resulted in a very small number of genes constituting the core genome of a species due to the cumulative effect of large numbers of sequences and uneven assembly and annotation qualities. Due to these factors, the core genomes of several species (e.g. *Escherichia coli*, *Salmonella enterica*, *Vibrio parahaemolyticus*) were rebuilt and distance matrices recalculated, using only nonredundant strains, as determined based on the initial core genomes. Information about the core genomes of each species is given in supplementary table S1, Supplementary Material online. Additionally, after removal of redundant strains from the data, the numbers of strains representing a given species could be greatly reduced from the original number of available genomes. From the original set of 105 named species, only those with at least 15 genomes remaining after removing redundant strains were retained, leaving 91 bacterial and two archaeal species for subsequent analyses. The included and excluded strains are listed in supplementary table S2, Supplementary Material online, and the downloaded and processed genomes are listed in supplementary material S1, Supplementary Material online.

Quantifying Gene Flow

Gene flow was estimated from the relative frequency of homoplasic polymorphisms (h) to nonhomoplasic polymorphisms (m) in the concatenate of the core genome (or a portion thereof). A homoplasic polymorphism is one whose evolution is incompatible with introduction by a single mutation inherited vertically from a shared ancestor. Homoplasies result either from recombination events or from convergent mutations occurring at the identical location in different strains. Since most species have a relatively short evolutionary history and harbor limited amounts of polymorphism, we reasoned that convergent mutations account for a very small proportion of homoplasies and that assessments of homoplasic polymorphisms would directly reveal the extent of recombination (gene flow) among strains. The aim is to identify groups of strains that exchange genes, not to infer precise rates of recombination, and comparisons of homoplasic polymorphisms offered a computationally tractable means of assessing gene flow from very large numbers of genomes. (Note that we estimated the contribution of convergent mutations to homoplasic polymorphisms in each species using procedures described in the following section.)

The large amount of data and the extensive resampling requires a method for identifying homoplasies that can be broadly applied to the 718,300 subsamplings specified by the analysis. Homoplasic polymorphisms were inferred from incongruences in the genomic distances of the corresponding strains (improved from Bobay et al. 2015) as follows: First, for each species, the matrix of maximum-likelihood distances, D, as defined earlier, was used to infer the pairwise distances between the core genomes for each pair of strains. Each

polymorphic site can have up to four alleles (i.e. one for each nucleotide), and in the simplest configuration, biallelic sites are composed of a major N_0 (most frequent) allele and a minor N_1 (least frequent) allele. (Major and minor alleles were assigned indiscriminately when at equal frequencies.) Each minor allele was considered homoplasic when $max\,(D_{N1N1}) > min\,(D_{N0N1})$, where $max\,(D_{N1N1})$ represents the highest genome-wide distance between the strains containing the minor alleles, and $min\,(D_{N0N1})$ represents the smallest genome-wide distance between the strains displaying the minor allele N_1 and the strains possessing the major allele N_0. For sites with three or four alleles, the one at highest frequency was defined as major and all others as minor. In such cases, each minor allele was considered homoplasic or nonhomoplasic by comparison of the genome-wide distances of the strains harboring the minor allele to the genome-wide distances of the strains displaying the major allele.

The extent of recombination among strains in a designated set of genomes was assessed by a resampling procedure. For each species, we randomly sampled 100 nonredundant combinations of strains for different numbers of genomes (from 4 to *n*–2, with *n* being the total number of strains analyzed for each species) (fig. 1*A*). This resulted in a total of 100*(*n*–5) combinations that were sampled randomly for each species, ranging from 1,000 to 36,300 combinations of strains for the species considered. For each combination of strains for a given species, the ratio of homoplasic to nonhomoplasic alleles (*h*/*m*) was inferred for the same 10-kb fragment of the core-genome concatenate. The ratio *h*/*m* was also computed for the entire core-genome concatenate for the complete set of strains of a species (table S1, Supplementary Material online).

Frequency of Convergent Mutations

The relative contributions of convergent mutations and recombination as the source of homoplasies vary among species depending on the amount of polymorphism, the number of strains considered, the G + C content of the genome, and the divergence and phylogenetic relationships among strains. We estimated the expected frequency of convergent mutations in each species by simulating sequence evolution without recombination. From the distance matrices computed on the core genome of each species, we built distance trees with BIONJ (Gascuel 1997) on which sequence evolution (without recombination) was simulated using Seq-Gen under a GTR model (Rambaut and Grassly 1997). The nucleotide composition of each species, estimated as the genome-wide average of all strains, was maintained during the simulation.

Because methods of phylogenetic reconstruction interpret all internal recombination events as occurring by convergent mutations, the branch lengths in the resulting trees overestimate the number of mutations in each alignment in cases where there is recombination. This becomes an issue when simulating the sequence evolution based on the parameters of the tree, producing simulated sequences with higher than actual levels of nucleotide diversity. To correct for this, the branch lengths of the tree of each species were rescaled using different coefficients, and we systematically inferred the number of polymorphic sites on the simulated alignments. We then selected the simulated alignments that displayed a polymorphism rate closest to that of the corresponding core genome.

Each of the simulated alignments for a species matched the number of strains, the alignment length, the amount of polymorphism, the tree topology and the nucleotide composition of the core genome of each species. The ratio of homoplasic to nonhomoplasic polymorphisms (*h*/*m*) was then calculated for each simulated alignment, but in this case *h* denotes homoplasies occurring only by convergent mutations, since we allow no recombination. We then applied the same resampling procedure described earlier on the simulated alignments, enabling us to directly compare the observed to the expected numbers of homoplasies attributable to convergent mutations (supplementary fig. S4, Supplementary Material online).

Exclusion Criterion

We developed an exclusion criterion for identifying the strain or strains responsible for a sharp interruption of gene flow with the rest of the species, as evident in the graphical representations. Identifying a single "sexually isolated" strain is straightforward; but since *h*/*m* ratios were calculated on randomly sampled subsets of strains, this task becomes more difficult when multiple strains do not participate in gene exchange with the rest of the species. A reduction of gene flow between some members of a species does not provide strong evidence of sexual isolation, since such variation in rates of genetic exchange is expected to occur within a species, so we developed a stringent exclusion criterion for eliminating a strain as a member of a biological species. We reasoned that if one or more strains display a strong reduction in gene flow with the rest of the strains, the distribution of the *h*/*m* ratios of the different subsampled combinations (that might or might not contain the sexually isolated strains) should be bimodal (or multimodal) and that the sexually isolated strain(s) should be unambiguously segregated into the lowest mode. Therefore, we examined the distribution of the *h*/*m* ratios in subsets of strains to identify those that do not participate in genetic exchange and should not be considered as part of the biological species comprising the other strains (supplementary fig. S2*A*, Supplementary Material online). For this analysis, we considered subsets of 10 or more strains, since the subsets of smaller size are consistently biased toward lower *h*/*m* values, as observed in the graphical representations (supplementary fig. S1, Supplementary Material online). We then partitioned subsets into two categories (modes) based on their *h*/*m* values using the *k*-means method with the Hartigan and Wong algorithm implemented

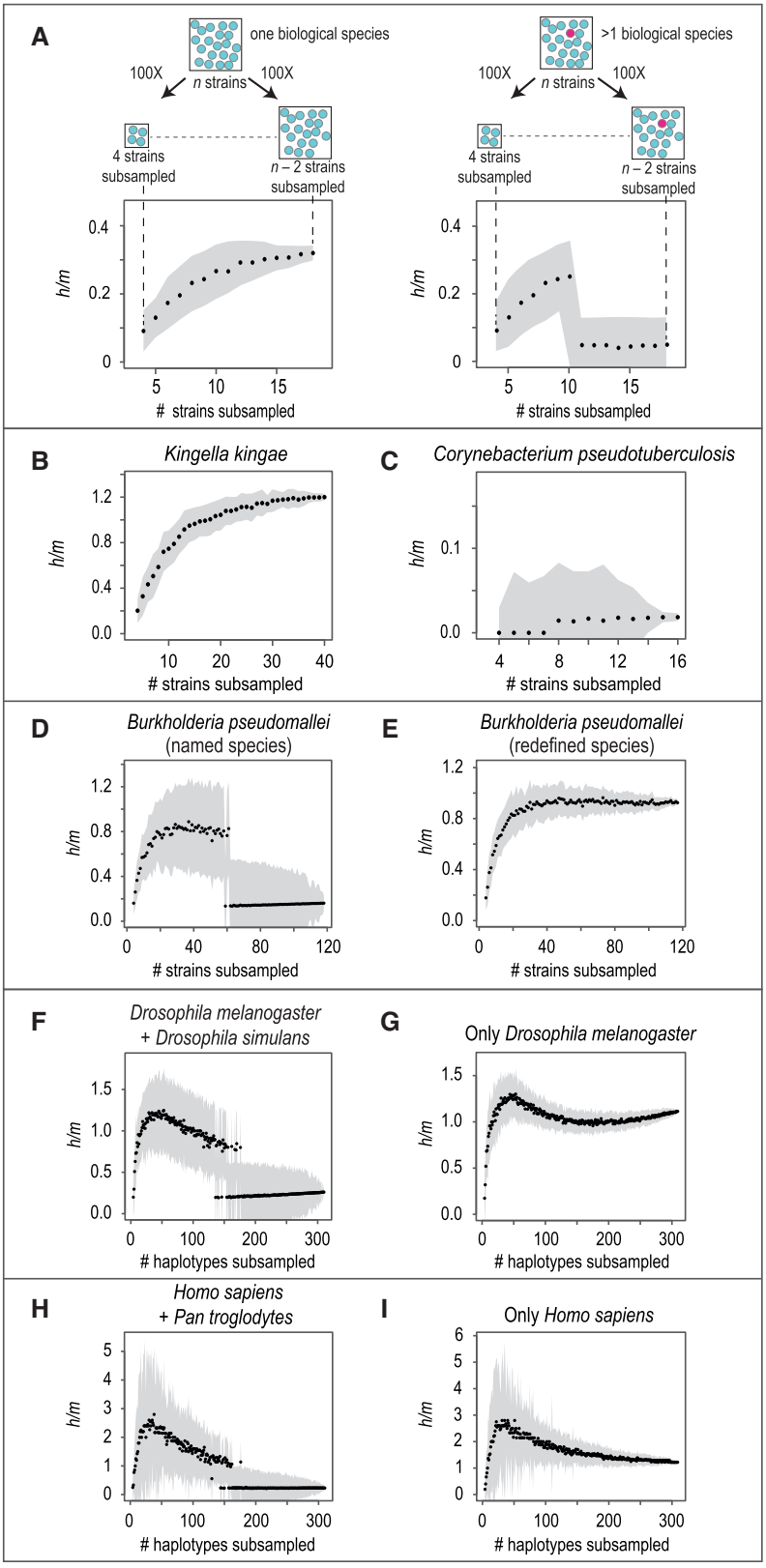

Fig. 1.—Recognizing biological species through genome analysis. *(A)* "Scheme used to test for gene flow". In each species or designated set of genomes composed of *n* strains (depicted as filled colored circles), nonredundant combinations of *i* strains (with *i* ranging from 4 to *n*–2 strains) were subsampled 100

in R (Hartigan and Wong 1979). Finally, we calculated the frequency at which each strain resided in subsets placed in each mode, asking whether a strain tended to be a member of subsets placed in the higher or lower distributions of *h/m* values (supplementary fig. S2*B*, Supplementary Material online).

To be considered "sexually isolated" (i.e. a member of a different biological species), strains had to be a member of subsets found exclusively in the lowest mode of *h/m* values and never in the highest mode. Additionally, subsets of strains containing the excluded strain(s) needed to display significantly lower *h/m* values than the rest of the subsets (*t*-test, $P < 0.00001$). A strain or a set of strains was excluded from the species when both conditions were met. The exclusion criterion was consistent with the sharp drop observed in the graphical representations of *h/m* ratios but also allowed the identification of several low-recombining strains that were not evident in the graphical representations (supplementary fig. S3, Supplementary Material online). Removal of these excluded strains redefined species boundaries, resulting in "biological species", whose constituents are linked by gene flow.

Simulations

We tested the sensitivity and robustness of our method through simulations that assessed the influence of several parameters on the recognition of sexually isolated strains. We simulated the evolution of a 10-kb circular sequence over 1,000, 5,000, 10,000, 15,000 or 20,000 generations. The original sequence corresponded to the first 10 kilobases of the *E. coli* K12 MG1655 genome, which was evolved *in silico* as in (Falush et al. 2006) under a constant population size of $N = 500$ sequences. Each generation was formed by randomly selecting, with replacement, 500 sequences from the preceding generation, and each sequence was subjected to random point mutations following a Poisson distribution of mean 0.1, corresponding to a mutation rate of 10^{-5} per generation per base pair. We implemented a mutational spectrum of *kappa* = 3 (i.e. transitions were three times more frequent than transversions) and maintained a constant nucleotide composition. Simulations were imposed different numbers of Poisson-distributed recombination events (either 5, 10, or 20 events per generation). Recombinant sequences, whose sizes were based on a normal distribution with a mean of 500 bp (±100 bp, SD), were selected at random from within a population, and replaced at the corresponding positions of a randomly selected recipient. Simulations were performed twice for each set of parameters in order to mimic the evolution of two sexually isolated populations evolving under the same conditions ($N = 500$ for each population). Additionally, we imposed different sampling biases for each pair of simulated populations: 50:50, 90:10 or 99:1 ratios after removal of near identical sequences with a total of 100 randomly selected sequences for each simulation.

After evolving populations under a given set of parameters, each pair was subjected to the same methodologies and graphical representations that we applied to actual genomes to detect interruptions of gene flow: Distance matrices were computed individually for each simulated population, nearly identical sequences were removed, the ratio of homoplasies to mutations (*h/m*) was calculated for different combinations of sequences following our resampling procedure, and sexually isolated strains were identified by our exclusion criterion.

Drosophila melanogaster *and* Homo sapiens *Data Sets*

We downloaded all haplotypes of *D. melanogaster* from the Drosophila Population Genomics Project (dpgp.org/) (Mackay et al. 2012; Pool et al. 2012) and generated an alignment of 311 autosomal haplotypes (including chromosomal arms 2R, 2L, 3R, and 3L) using VCFtools (Danecek et al. 2011). The exome of *D. simulans* v2.02 was downloaded from FlyBase (flybase.org/), and exons ≥100 bp were blasted against the autosomes of *D. melanogaster* using Blastn with a 90% identity score and *e*-value < 0.00001 (Altschul et al. 1997). We selected exons that best matched those of *D. melanogaster* over 95% of their length, allowing a maximal gap-opening of

Fig. 1.—Continued

times for each value of *i*. At each iteration, the *h/m* ratio was calculated for a randomly selected 10-kilobase fragment from the alignment of the core genome concatenate common to all strains. Within the bivariate plots, black dots are medians, and the grey-shaded region is the standard deviation, for the indicated number of subsampled combinations of strains. Left panel is the graphical representation observed when there are no barriers to gene flow among strains, and right panel depicts the discontinuity produced by inclusion of a strain that does not participate in gene exchange. *(B–D)* "Patterns of genetic exchange observed in taxonomically defined bacterial species". *Kingella kingae*, in which there is gene flow among the entire set of sequenced strains; *C. pseudotuberculosis*, in which there is too little gene flow to assess species status; *Buckholderia pseudomallei*, in which there is a sharp drop in *h/m* ratios, denoting the presence of a sexually isolated strain or strains. The complete set of graphical representations of gene flow for 93 taxonomically defined bacterial and archaeal species is presented in supplementary fig. S1, Supplementary Material online. *(E)* Graphical representation of *B. pseudomallei* after removal of the sexually isolated strain, showing that the remaining strains constitute a biological species. The complete set of graphs is presented for the redefined species in fig. S3 before and after species redefinition. *(F–I)* Verification of method for recognizing biological species using obligatory sexual animals. Graphical representation of gene flow in *Drosophila* when including 311 haplotypes of *D. melanogaster* and one haplotype of *D. simulans*, and when the analysis is restricted to the 311 haplotypes of *D. melanogaster*. Graphical representation of gene flow in *Homininae* when including 311 haplotypes of chromosome 21 in humans and one haplotype of chimpanzees (*P. troglodytes*), and when the analysis is restricted to the 311 chromosome 21 haplotypes of humans.

20% of the exon length on the chromosome. To prevent the inclusion of paralogs, we excluded all exons that significantly matched an overlapping region of the chromosome. Finally, the aligned portions of the exon and of the chromosome of each haplotype were extracted, aligned independently in MAFFT v7 (Katoh and Standley 2013) and merged into a single alignment. On account of the size of the concatenate (~5Mb), we restricted the analysis to a 1-Mb fragment.

We downloaded all the haplotypes of the chromosome 21 of *H. sapiens* from the 1000 Genome Project (ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/) and the chromosome 21 of the corresponding reference genome (ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/reference/phase2_reference_assembly_sequence/) (Auton et al. 2015). To match the sample size to that used for *D. melanogaster*, we restricted the analysis to 311 haplotypes, each from a different individual. The exome of chromosome 21 of *Pan troglodytes* (NC_006488) was downloaded from the NCBI (www.ncbi.nlm.nih.gov/genome/), and exon alignments to the human chromosome 21 sequences were generated as described above. Calculation of *h/m* values, the subsampling procedures and the estimation of convergent mutations follow procedures identical to those described earlier.

Results

Bacteria Can Be Classified into Biological Species

We examined over 20,000 genomes from 91 named bacterial species from 13 phyla with the goal of identifying individuals that are sexually isolated from the rest of the population. Our data set was limited to species, which, after removal of identical and nearly identical genomes, represented by at least 15 unique genome sequences. For each genome, the set of orthologs present in all representatives of a species (the core "genome") was extracted, merged into a single concatenate and aligned to the core genomes of all conspecifics. We quantified the extent of recombination among strains through examination of homoplasies—shared polymorphisms that did not arise by vertical inheritance from a single ancestor and likely arose through recombination. Assessment of homoplasies offers a computationally tractable approach for identifying individual recombinant sites in such large numbers of strains, and this procedure is amenable to the iterative subsampling procedure that we use to test species boundaries.

For each set of alignments, we inferred the frequencies of homoplasic polymorphisms (*h*) relative to nonhomoplasic polymorphisms (*m*) (see "Materials and Methods" section). We reasoned that if named species harbor strains that do not exchange DNA with the rest of the species, these strains should sharply reduce *h/m* ratios. To determine whether strains typed to a species show gaps in gene flow, whereby recombination occurs within some subsets of strains but not between them, we resampled sets of strains within a named species and calculated *h/m* for each set. We conducted a procedure that randomly sampled 100 nonredundant combinations of strains for different numbers of genomes (from 4 to *n*–2, with *n* being total number of strains analyzed for each species) (fig. 1*A*). The recalculation of *h/m* ratios for different numbers and combinations of strains allows identification of those strains whose inclusion or exclusion in subsampled combinations modifies *h/m* ratios. Our subsampling procedure involves a total of 100*(*n*–5) combinations per species, which resulted in a range of 1,000 to 36,300 subsampled sets for the 91 species considered. Calculating *h/m* for the >700,000 subsampled genome sets was accomplished by basing each estimate on a randomly selected 10-kilobase fragment from the concatenated alignment for each species. To verify that the selected fragment is representative of the genome as a whole, we computed the *h/m* ratio of the entire core-genome concatenate for all *n* strains in a species and confirmed that the selected fragment and the core genome converge to similar *h/m* ratios (supplementary table S1, Supplementary Material online).

For each species, gene flow is graphically represented as the function of *h/m* ratios calculated for the subsampled combinations of strains (fig. 1 and supplementary fig. S1, Supplementary Material online). Based on their pattern of genetic exchange among strains, species assorted into three categories: 1) Those species in which the different subsets of strains display *h/m* ratios that plateau as more genomes are analyzed (fig. 1*B*). This is the most common pattern—observed for 54/91 cases—and is indicative of relatively homogeneous gene flow among all strains within the named species. 2) Those species displaying *h/m* ratios with high standard deviations and showing a sharp drop when *h/m* ratios are calculated with larger subsets of strains (fig. 1*D*). The steep decline occurs because random subsets of larger size will contain a strain that does not participate in gene exchange, thereby reducing the *h/m* ratios. This pattern—observed for 21/91 cases—denotes the inclusion of strains that do not recombine with the others and indicates that the named species contains two or more reproductively isolated groups. 3) Those species displaying low *h/m* ratios among all subsets of strains, thereby preventing the assessment of gene flow based solely on these graphs (fig. 1*C*). This last category includes assemblages either of purely clonal lineages or of multiple biological species whose individual sample sizes are too low to detect recombination.

To better define breaks in gene flow within named species, and to provide statistical support for the different patterns that were charted graphically (fig. 1 and supplementary fig. S1, Supplementary Material online), we developed an exclusion criterion to identify those strains that when added to a sampled subset of strains caused disruptions in the distribution of *h/m* ratios as a result of their lack of gene exchange with other members of the species (see "Materials and Methods" section). This exclusion criterion was designed to be very

stringent, since members of true species can vary in their contributions to genetic exchange. For each of the 91 named bacterial species that we considered, we used the subsampling of strains and applied the *k*-means method to partition the subsets of strains into two groups based on the distribution of *h/m* values (supplementary fig. S2, Supplementary Material online). Strains were earmarked for exclusion from a named species when the following conditions were met: 1) a strain was found exclusively among subsets of strains with the lowest mode of the distribution of *h/m* values and 2) the inclusion of a strain in a subset significantly reduced the overall *h/m* ratio (see "*Materials and Methods*" section). Strains that fulfill these conditions can be viewed as 'sexually isolated' from the rest of the strains in the species. The exclusion criterion was met for at least one strain from 23 of the named species, and these species included all the species that displayed sharp drops in *h/m* ratios in the graphical representations (fig. 1*D* and supplementary fig. S3, Supplementary Material online). Rebuilding the graphical representations after excluding the sexually isolated strains yielded groups of constituent strains that are connected by gene flow (fig. 1*E* and supplementary fig. S3, Supplementary Material online), thereby redefining species boundaries in bacteria based on standards identical to those of the BSC.

Because homoplasies can be generated either by recombination or by convergent mutations, we performed a series of tests to quantify the proportion of homoplasies that are expected to originate from convergent mutations as opposed to recombination. For each species, we used Seq-Gen (Rambaut and Grassly 1997) to generate a set of alignments that retained all of the features of original data set with respect to strain number, concatenate size, base composition, level of polymorphism, and tree topology, but introduced no recombination (see "Materials and Methods" section). We subjected these alignments to the identical procedures used on the original alignments of concatenates to produce graphical representations of the frequencies of homoplasic and nonhomoplasic polymorphisms in subsampled combinations of strains. (Note that in these analyses, homoplasy/nonhomoplasy ratios are the same as *h/m* ratios but that homoplasies can only be introduced by convergent mutations). For the majority of species (77/91), convergent mutations produce very few homoplasies (supplementary fig. S4, Supplementary Material online), indicating that most homoplasies in the original data sets are introduced by recombination not convergent mutations. In contrast, frequencies of homoplasies in the simulated and original data sets are, as expected, similar in the 14 low-recombining species, indicating that we cannot use gene flow to delineate species boundaries in a minority of bacterial species (supplementary table S1, Supplementary Material online). Therefore, homoplasies offer the possibility to measure gene flow among bacterial strains, and gene flow, in turn, can serve as the benchmark for delineating biological species for the majority of bacteria.

Because a genomic delineation of species depends on the amount of polymorphism, the recombination rate and the sampling of analyzed strains, we tested the sensitivity and robustness of our approach with simulations under 45 sets of parameter conditions. These conditions included various degrees of divergence (accrued from 1,000 to 20,000 generations), recombination rates (5, 10, and 20 events per generation) and sampling biases (50:50, 90:10, and 99:1 mixtures of the two source populations); and for each pair of evolved populations, we applied our resampling procedure and exclusion criterion (see "Materials and Methods" section). These simulations ascertained the minimum number of generations and the amount of recombination necessary to identify populations that are sexually isolated based on the distribution of homoplasic polymorphisms. As expected, when the number of generations separating two populations was low (<5,000 generations), the recombination rate low (<10 events per population per generation) and there was equal sampling of the two populations, members of the two populations cannot be distinguished (supplementay fig. S5, Supplementary Material online). As populations diverge, recombination rates increase, and/or sampling is more skewed, discontinuities become evident in the simulated populations. But even at the lowest recombination rate examined (five events per population per generation), individual strains from the independently evolved populations were recognized after 10,000 generations. Under the tested parameters, our method fails at distinguishing multiple species when the two sexual populations are present at identical frequencies, yielding a signal resembling clonality (i.e. consistently low *h/m* ratios), but this situation is resolved when one population predominates (supplemetary fig. S5, Supplementary Material online). In summary, our approach is conservative and does not subdivide genetically cohesive populations; however, it lacks sensitivity when sexually isolated populations are present at similar frequencies in the sample.

A Single Approach Defines Biological Species across All Cellular Lifeforms

We next tested if the same approach, when applied to obligatory sexual organisms, is capable of recognizing species' boundaries, especially since multicellular eukaryotes can harbor low amounts of variation while obviously engaging in recombination each generation. We focused on two strictly sexual species—*D. melanogaster* and *H. sapiens*—for which large genomic data sets comparable in scope to those of bacteria are available. Polymorphism data for autosomes in *D. melanogaster* (n = 311) (Mackay et al. 2012; Pool et al. 2012) and for chromosome 21 in *H. sapiens* (n = 311) (Auton et al. 2015) were assembled and aligned for each species, and the corresponding sequences were obtained for the orthologous genes in their respective sister species (*D. simulans* and *P. troglodytes*). As before, *h/m* ratios

were computed on 10-kb fragments based on subsampling multiple combinations of haplotypes. Applying the same graphical representations and exclusion criterion as above clearly designate *D. simulans* and *P. troglodytes* as having an abrupt reduction of gene flow with their respective sister species (fig. *1F–I*). Next, we applied the same methodology to the two archaeal species (*Methanosarcina mazei* and *Sulfolobus islandicus*); and in the case of *M. mazei*, the sequenced strains convey a clear signal of one cohesive species (supplementarty fig. S1, Supplementary Material online). Thus, patterns of gene flow and reproductive isolation make it possible to define biological species in all Domains of life.

Identity Thresholds Do Not Delineate True Biological Species

Bacterial species have traditionally been defined as assemblages of strains that possess certain diagnostic properties or that attain a prescribed level of sequence identity, for example, ≥97% 16S rRNA identity or ≥94% average nucleotide identity of shared genes (Konstantinidis and Tiedje 2005; Richter and Rossello-Mora 2009). As already shown (fig. 1*D* and supplementary fig. S3, Supplementary Material online), numerous named bacterial species that were delineated by DNA identity thresholds actually comprise multiple biological species. Since recombination rates are expected to decrease as nucleotide sequences diverge due to the action of the mismatch repair systems (Matic et al. 1995; Rayssiguier et al. 1989; Shen and Huang 1986), we first asked the extent to which sequences of genes constituting the core genome can diverge but still recombine, and then, whether the extent of sequence divergence is related to the amount of gene flow.

The most divergent members in the biological species that we recognized usually average no more than 5% difference in the nucleotide sequences of their core genes; however, there is a long tail to the distribution of DNA identity values (fig. 2). At the extreme is *Prochlorococcus marinus*, in which the most divergent strains of the same biological species share only 72% average nucleotide sequence identity. It is possible that in such species, the most divergent members are connected only through a continuum of conspecific strains that are able to recombine [analogous to eukaryotic "ring species" (Cain 1954)]. However, even the most divergent members of *P. marinus*, uniquely share many homoplasic polymorphisms, implying that they still exchange genes, possibly due to the loss of certain recombination and repair genes during genome reduction (Rocap et al. 2003; Dufresne et al. 2005).

Gene Flow between Near and Distant Relatives

Given the range of sequence diversity within some bacterial species, we questioned the relationship between nucleotide divergence and rates of recombination by comparing the *h/m* ratio to the average sequence identity for each of the subsampled combination of strains of each species. Correlation

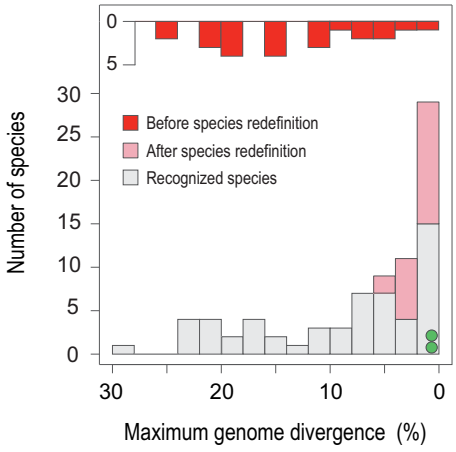

Fig. 2.—Maximum sequence divergence within species of bacteria. Shown are the average nucleotide sequence identity values for orthologs shared by the two maximally divergent strains within each biological species of bacteria (grey and pink bars). The inset shows these values for those named species (red bars) that were subsequently redefined into biological species (pink bars) based on gene flow (n = 23). Within strictly clonal species (n = 14; not included in figure), maximally divergent strains ranged from 73 to > 99% nucleotide sequence identity (95% on average). Green dots indicate the degree of sequence identity between two maximally divergent haplotypes of *D. melanogaster* and of *H. sapiens*.

coefficients were calculated independently for each species and for all subsampled combinations having the same number of strains (supplementary fig. S6, Supplementary Material online). Across species, there is no consistent trend between degree of sequence divergence and the amount of gene flow (fig. 3): several species display a clear positive correlation between sequence identity and *h/m* ratio, indicating that as strains diverge, recombination becomes more restricted; but as many species, such as *Pseudomonas putida*, show a negative correlation, such that strains preferentially exchange DNA with more distantly related conspecifics. Additionally, *Campylobacter coli* exhibits one of the strongest negative associations between sequence identity and gene flow, and this species was previously reported as fusing with its congener *Campylobacter jejuni* (Sheppard et al. 2008). That *C. coli* recently absorbed a substantial fraction of alleles from *C. jejuni* confirms its tendency toward recombining with distant relatives and shows that hybridization events that disrupt species boundaries can occur in bacteria as well as eukaryotes.

Discussion

The ability to obtain large numbers of genome sequences allows application of the same biologically relevant criterion used in eukaryotes—that is, gene flow—to define and differentiate microbial species. To date, most named bacterial species that have been analyzed by multilocus sequence typing

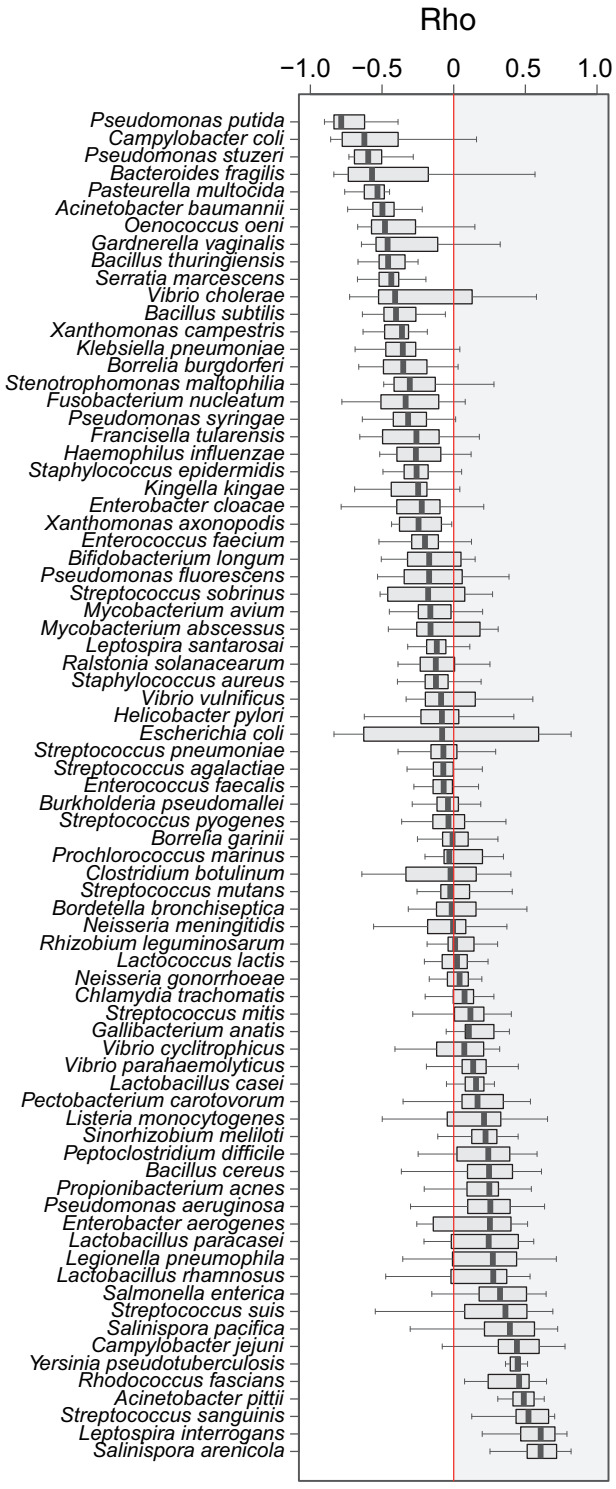

Fig. 3.—Differential effects of sequence divergence on gene flow in biological species of bacteria. For each species, the *h*/*m* ratio for each of the subsampled combination of strains was compared to the average genome-wide sequence identity (supplementary fig. S6, Supplementary Material online). Spearman correlation coefficients Rho were computed independently for all nonredundant combinations of strains (ranging from 4 to *n*–2 strains) each subsampled 100 times and whose distributions are displayed as box-and-whiskers (first and third quartiles, 1.5 interquartile range) plots for each species. Positive values of Rho indicate a positive correlation between nucleotide sequence identity and gene flow (such that genetic exchange occurs between more similar strains), whereas negative values indicate that genetic exchange occurs preferentially among divergent strains.

(MLST) show evidence of gene exchange, albeit some at very low frequencies (Perez-Losada et al. 2006; Vos and Didelot 2009; Hanage 2016). Therefore, it is imaginable that all assemblages of related bacteria—even those with exceedingly low rates of gene flow—are amenable to species-level classification based on the BSC provided the availability of a sufficient number of genome sequences. For example, even in *Staphylococcus aureus*, which stands near the extreme in its level of clonality based on MLST (Vos and Didelot 2009), we were able to delineate a true biological species given the current availability of several hundred genome sequences. And for the 14 named species that defied classification (or reclassification) by our approach, it is likely that additional genome sequences will provide the recombinant polymorphisms necessary for their classification, since many (e.g. *Brucella suis*, *Corynebacterium pseudotuberculosis*, *Coxiella burnetti, Mycobacterium africanum, Xanthomans citri*) were at or near the lower limit of the number of genomes that we considered. The only bacteria for which there is no indication of gene exchange are strictly maternally transmitted endosymbionts, such as *Buchnera* (Moran et al. 2008), but there are currently too few closely related genomes for any of these endosymbionts to perform a species-level analysis.

Given the speed at which new genome sequences are being produced, the processing and classification of bacteria to species presents a colossal and dynamic endeavor. In this analysis, we started with the naïve assumption that strains and species in the NCBI RefSeq database were classified correctly. We found that no bacterial genome was completely aberrant, such that its sequence was extremely divergent from other members of the species; however, some were misclassified at the level of genus. For example, our approach distinguished one strain of *Vibrio cholerae* from others classified to the same species, and more careful examination revealed that this strain was more closely related to *Vibrio mimicus* than to *V. cholerae*. This case of misclassification reinforces the importance of redefining species boundaries prior to conducting computational analyses.

With regard to the applicability of the BSC, prokaryotes do not differ all that much from eukaryotes. Although only a small proportion (<1%) of animals are parthenogenetic, it has been estimated that 20–30% of plants self-fertilize or reproduce vegetatively and are not suited for classification by the BSC (Stebbins 1963; White 1978). We find that among bacteria, only a minor fraction (<15%) undergoes too little gene flow to be assigned to species based on the

BSC, so it appears that all groups—bacteria, eukaryotes and archaea alike—are similar in that they comprise some small percentage of nonrecombining lineages.

Although the concept of microbial species has been controversial, we show that it is possible to delineate species by assessing gene flow from population genomic data. This approach does not abrogate the advantages of other concepts and methodologies, since it will often be necessary to identify an organism prior to access to full genome sequences. Furthermore, our method identifies the interruption of gene flow, and this suggests—but does not guarantee—that different populations have lost the capacity to exchange DNA. In that genome sequencing might soon supplant other methods for categorizing organisms, our approach to species recognition, although currently only practicable in small-genomed lineages, will become increasingly applicable to other groups of organisms. Based on our results and the rapidity with which sequencing technology is advancing, it might soon become unnecessary to classify bacteria based on sequence identity thresholds, which are easily but inappropriately applied (Krause and Whitaker 2015), or on unique metabolic capabilities, which can be both impractical to test on noncultivable organisms and variable among members of species.

We note that many bacterial and even some eukaryotic populations do not represent true biological species *stricto sensu*, because they occasionally exchange DNA with other species through horizontal gene transfer. Nevertheless, external imports from distant sources are rare relative to the amount of recombination between conspecifics, making it possible to define species based the occurrence of homologous gene exchange. Additionally, a BSC-based approach towards defining microbial species helps in building a framework for population genetics studies, which rely on the assumption that populations are panmictic and that alleles can freely circulate.

Recognition of biological species by our genome-based method is currently problematic when the number of available genomes is low or when samples contain multiple biological species present at similar frequencies. Thus, it is possible that some of the named species with small sample sizes and relatively low recombination rates, such as *P. marinus* and *S. islandicus*, actually comprise multiple biological species, as has been suggested in other studies (Cadillot-Quiroz et al. 2012; Kashtan et al. 2014). Despite the few cases where additional genomes might assist in defining species, most of the species analyzed displayed *h/m* ratios that plateaued at relatively high values, suggesting that they each represent a single biological species.

The sensitivity of our genome-based method for recognizing biological species might also be improved by resampling additional assemblages of genomes. In this study, we did not systematically increase the sampling for all species, since it would have substantially increased the computational time required for the analysis, and it is not possible to predict *a priori* what or how many combinations of strains need to be sampled in order to identify all the biological species in a given set of genomes. But instead of evaluating all genomes assigned to a named species, our resampling procedure could possibly gain sensitivity by first limiting analyses to particular clades or geographic groups (e.g. using synapomorphic SNPs or ecoSNPs; Shapiro et al. 2012), which would then allow testing for species boundaries within phylogenetically or ecologically relevant sets of strains.

What does the delineation of bacterial species by the BSC tell us about the process of speciation in bacteria? Most early models of bacterial speciation did not consider gene flow as a major component of the speciation process (Cohan 2001; Shapiro et al. 2016). Despite evidence that strains from different bacterial species display highly reduced levels of recombination (Rayssiguier et al. 1989), simulations of the role of sequence divergence in the disruption of gene flow suggested that the gradual reduction of gene flow that might result from sequence divergence, in the absence of selection, was unlikely to drive speciation (Fraser et al. 2009; Hanage et al. 2006). We, too, find little support for the view that bacteria gradually diverge to a point where they are no longer members of the same species in that we detected no systematic relationship between higher levels of sequence divergence and the reduction of recombination.

Although our approach was designed to identify which, if any, lineages in a named species should be excluded based on genetic criteria, a related problem is whether lineages currently assigned to different named species actually constitute a single biological species. For example, the species boundaries in the genus *Neisseria* are considered "fuzzy" due to the presence of strains that possess sequences typical of multiple species (Hanage 2013; Hanage et al. 2005). We tested whether strains from these two species can be discriminated by applying the same graphical representations and exclusion criterion to the core genome assembled from the 153 sequenced strains of *Neisseria meningitidis* and one randomly selected strain of *Neisseria gonorrhoeae*. Despite a potential for hybridization that might blur the species boundaries, the inclusion of *N. gonorrhoeae* in the *N. meningitidis* data set introduces an abrupt reduction of gene flow (supplementary fig. S7, Supplementary Material online), indicating that even in highly promiscuous bacterial genera, it is possible to resolve true biological species.

That species can be universally defined based on gene flow implies that many of the same factors are operating in the process of speciation across all lifeforms. Differences in genomic properties (such as ploidy, recombination frequencies, and reproduction, and rates of gene acquisition) and demographic parameters (such as population sizes, geographic distribution, and rates of migration) will impact the pace at which microbes speciate relative to sexual organisms. However, the application of a single genomic-based BSC criterion to delineate species makes it possible to define species and study speciation under a similar framework across the tree of life.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Author Contributions

L.M.B. designed the study, performed the analyses and wrote the manuscript. H.O. designed the study and wrote the manuscript.

Acknowledgments

We thank Sarah McCarthy for assistance in data processing, Kim Hammond for help in the preparation of figures, and Nancy Moran for comments on the article. This work was supported through funding by National Institutes of Health awards R01GM101209 and R01GM108657.

Literature Cited

Altschul SF, et al. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389–3402.

Auton A, et al. 2015. A global reference for human genetic variation. Nature 526:68–74.

Bobay LM, Traverse CC, Ochman H. 2015. Impermanence of bacterial clones. Proc Natl Acad Sci U S A. 112:8893–8900.

Cadillot-Quiroz H, et al. 2012. Patterns of gene flow define species of thermophilic Archaea. PLoS Biol. 10:e1001265.

Cain AJ. 1954. Animal species and their evolution. London: Hutchinson House.

Cohan FM. 2001. Bacterial species and speciation. Syst Biol. 50:513–524.

Danecek P, et al. 2011. The variant call format and VCFtools. Bioinformatics 27:2156–2158.

Dufresne A, Garczarek L, Partensky F. 2005. Accelerated evolution associated with genome reduction in a free-living prokaryote. Genome Biol. 6:R14.

Dykhuizen DE, Green L. 1991. Recombination in *Escherichia coli* and the definition of biological species. J Bacteriol 173:7257–7268.

Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460–2461.

Falush D, et al. 2006. Mismatch induced speciation in Salmonella: model and data. Philos Trans R Soc Lond B Biol Sci. 361:2045–2053.

Fraser C, Alm EJ, Polz MF, Spratt BG, Hanage WP. 2009. The bacterial species challenge: making sense of genetic and ecological diversity. Science 323:741–746.

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

Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. Mol Biol Evol. 14:685–695.

Hanage WP. 2013. Fuzzy species revisited. BMC Biol. 11:41.

Hanage WP. 2016. Not so simple after all: bacteria, their population genetics, and recombination. Cold Spring Harb Perspect Biol. 8(7):a018069.

Hanage WP, Fraser C, Spratt BG. 2005. Fuzzy species among recombinogenic bacteria. BMC Biol. 3:6.

Hanage WP, Spratt BG, Turner KM, Fraser C. 2006. Modelling bacterial speciation. Philos Trans R Soc Lond B Biol Sci. 361:2039–2044.

Hartigan JA, Wong MA. 1979. A K-means clustering algorithm. Appl. Stat. 28:100–108.

Kashtan N, et al. 2014. Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. Science 344:416–420.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 30:772–780.

Konstantinidis KT, Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A. 102:2567–2572.

Krause DJ, Whitaker RJ. 2015. Inferring speciation processes from patterns of natural variation in microbial genomes. Syst Biol. 64:926–935.

Lawrence JG, Retchless AC. 2009. The interplay of homologous recombination and horizontal gene transfer in bacterial speciation. Methods Mol Biol. 532:29–53.

Mackay TF, et al. 2012. The *Drosophila melanogaster* Genetic Reference Panel. Nature 482:173–178.

Matic I, Rayssiguier C, Radman M. 1995. Interspecies gene exchange in bacteria: the role of SOS and mismatch repair systems in evolution of species. Cell 80:507–515. PMC][7859291.

Mayr E. 1942. Systematics and the origin od species. New York: Columbia University Press.

Moran NA, McCutcheon JP, Nakabachi A. 2008. Genomics and evolution of heritable bacterial symbionts. Annu Rev Genet. 42:165–190.

Ochman H, Lerat E, Daubin V. 2005. Examining bacterial species under the specter of gene transfer and exchange. Proc Natl Acad Sci U S A. 102(Suppl. 1):6595–6599.

Perez-Losada M, et al. 2006. Population genetics of microbial pathogens estimated from multilocus sequence typing (MLST) data. Infect Genet Evol. 6:97–112.

Pool JE, et al. 2012. Population Genomics of sub-saharan *Drosophila melanogaster*: African diversity and non-African admixture. PLoS Genet. 8:e1003080.

Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. Comput Appl Biosci. 13:235–238.

Rayssiguier C, Thaler DS, Radman M. 1989. The barrier to recombination between *Escherichia coli* and *Salmonella typhimurium* is disrupted in mismatch-repair mutants. Nature 342:396–401.

Richter M, Rossello-Mora R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A. 106:19126–19131.

Rocap G, et al. 2003. Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature 424:1042–1047.

Shapiro BJ, et al. 2012. Population genomics of early events in the ecological differentiation of bacteria. Science 336:48–51.

Shapiro BJ, Leducq JB, Mallet J. 2016. What Is Speciation?. PLoS Genet. 12:e1005860.

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

Sheppard SK, McCarthy ND, Falush D, Maiden MC. 2008. Convergence of *Campylobacter* species: implications for bacterial evolution. Science 320:237–239.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–2690.

Stebbins GL. 1963. Perspectives—I. Am Sci. 51:362–370.

Vos M, Didelot X. 2009. A comparison of homologous recombination rates in bacteria and archaea. isme J. 3:199–208.

White MJ. 1978. Modes of speciation. San Francisco (CA): W.H. Freeman and Co.

Associate editor: Rachel Whitaker