32. P. Nosil, T. H. Vines, D. J. Funk, *Evolution* **59**, 705 (2005).
33. H. D. Rundle, M. C. Whitlock, *Evolution* **55**, 198 (2001).
34. N. H. Martin, J. H. Willis, *Evolution* **61**, 68 (2007).
35. H. D. Bradshaw, D. W. Schemske, *Nature* **426**, 176 (2003).
36. C. Lexer, Z. Lai, L. H. Rieseberg, *New Phytol.* **161**, 225 (2004).
37. M. A. Beaumont, *Trends Ecol. Evol.* **20**, 435 (2005).
38. P. Nosil, D. J. Funk, D. Ortiz-Barrientos, *Mol. Ecol.* **18**, 375 (2009).
39. S. M. Rogers, L. Bernatchez, *Mol. Biol. Evol.* **24**, 1423 (2007).
40. S. Via, J. West, *Mol. Ecol.* **17**, 4334 (2008).
41. H. A. Orr, J. P. Masly, N. Phadnis, *J. Hered.* **98**, 103 (2007).
42. D. A. Levin, *Syst. Bot.* **28**, 5 (2003).
43. L. H. Rieseberg, J. H. Willis, *Science* **317**, 910 (2007).
44. D. C. Presgraves, *Trends Genet.* **24**, 336 (2008).
45. D. C. Presgraves, *Curr. Biol.* **17**, R125 (2007).
46. L. Fishman, J. H. Willis, *Evolution* **60**, 1372 (2006).
47. A. L. Case, J. H. Willis, *Evolution* **62**, 1026 (2008).
48. We thank M. Arnegard, R. Barrett, A. Case, H. Hoekstra, M. Noor, P. Nosil, S. Otto, T. Price, L. Rieseberg, S. Rogers, A. Schluter, S. Via, M. Whitlock, J. Willis, and a reviewer for assistance and comments. This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Canada Foundation for Innovation.

Supporting Online Material
www.sciencemag.org/cgi/content/full/323/5915/737/DC1
Tables S1 to S3
References

10.1126/science.1160006

REVIEW

The Bacterial Species Challenge: Making Sense of Genetic and Ecological Diversity

Christophe Fraser,[1]* Eric J. Alm,[2,3,4] Martin F. Polz,[2] Brian G. Spratt,[1] William P. Hanage[1]

The Bacteria and Archaea are the most genetically diverse superkingdoms of life, and techniques for exploring that diversity are only just becoming widespread. Taxonomists classify these organisms into species in much the same way as they classify eukaryotes, but differences in their biology—including horizontal gene transfer between distantly related taxa and variable rates of homologous recombination—mean that we still do not understand what a bacterial species is. This is not merely a semantic question; evolutionary theory should be able to explain why species exist at all levels of the tree of life, and we need to be able to define species for practical applications in industry, agriculture, and medicine. Recent studies have emphasized the need to combine genetic diversity and distinct ecology in an attempt to define species in a coherent and convincing fashion. The resulting data may help to discriminate among the many theories of prokaryotic species that have been produced to date.

The species debate in microbiology is not only about a human desire to catalog bacterial diversity in a consistent manner, but is also a fundamental argument because of what it reveals about our ignorance of how evolutionary forces form, shape, and extinguish bacterial genetic lineages, of the mechanisms of differentiation between subpopulations sharing common descent, and of the process of adaptation to new niches and changing environments. Animal species are defined by their morphological and behavioral traits and by their ability or inability to interbreed, but such categories cannot easily be applied to the Bacteria or Archaea (or indeed to many eukaryotic microbes). Instead, taxonomists have been forced to rely on biochemical tests and limited morphological characteristics for this purpose. Naturally, biochemical characters have been selected for the convenience of taxonomists; they reflect only a tiny subset of those characters that allow bacteria to use different resources in the environment, and only capture a small fraction of the true diversity in this superkingdom of life. More recently, molecular methods [particularly DNA-DNA hybridization and ribosomal RNA (rRNA) sequencing] have helped to define species, but these methods have serious limitations and cannot reliably assign a large collection of similar strains to species (e.g., rRNA sequences are too conserved to resolve similar species). rRNA sequence surveys have, however, revealed the extraordinary variety of microbial life, much of it uncultured (*1*). Beyond this, taxa too similar to be distinguished and circumscribed by rRNA sequences have revealed further diversity through multilocus sequence analysis (MLSA) (*2*) and metagenomic studies (*1*), and this diversity needs to be explained by theory. Thus, practical difficulties, lack of theory, and observations of vast amounts of as yet unclassified microbial diversity have all fueled the controversy of how one defines a bacterial species (*3*–*8*).

[1]Department of Infectious Disease Epidemiology, Imperial College London, London W2 1PG, UK. [2]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [3]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [4]Broad Institute of MIT and Harvard University, Cambridge, MA 02139, USA.

*To whom correspondence should be addressed. E-mail: c.fraser@imperial.ac.uk

Genetic Clustering

Darwin commented that "all true classification is genealogical" [(*9*), p. 404]. Taxonomists have thus used sequence relatedness to define cutoff values that place two bacterial isolates into the same or different species. The overall genetic relatedness of isolates may be measured by the extent of DNA hybridization between them, and those that show 70% or more DNA hybridization are defined as the same species (*2*, *10*). Such cutoffs imply that sequences that cluster together with a certain amount of similarity must be from the same species, and moreover that this cutoff value is applicable to all groups of bacteria or archaea. Recent MLSA studies, which use the concatenated sequences of multiple housekeeping genes to discern clustering patterns among populations of closely related taxa, suggest that species defined by taxonomists in many cases correspond to well-resolved sequence clusters. However, these studies also show that there is no universal cutoff or descriptor of clusters that characterizes a species. Furthermore, inspection of the clusters does not always clearly reveal which level in the hierarchy is more fundamental than any other (Fig. 1) (*7*).

As an example, Fig. 1A shows the relationships among multiple isolates of three closely related streptococcal species. *Streptococcus pneumoniae* is a major human pathogen, *S. mitis* is a commensal bacteria with a history of taxonomic uncertainty (*11*), and *S. pseudopneumoniae* is a recently described organism of uncertain status that nonetheless corresponds to a distinct cluster in these data (*12*). There are striking differences in the amount of sequence diversity observed within homologous housekeeping genes in these named species, ranging from 1.2% for *S. pneumoniae* to 3.0% for *S. pseudopneumoniae* and up to 5.0% for *S. mitis*. The distance between two randomly selected *S. mitis* genotypes is similar to the average distance between *S. pneumoniae* and *S. pseudopneumoniae* genotypes (5.1%) (*2*). This implies that the use of a fixed level of sequence divergence for differentiating species would tend to either rejoin *S. pneumoniae* and *S. pseudopneumoniae*, or break up *S. mitis* so that nearly every isolate was a species of its own. This is clearly unsatisfactory.

Habitats and Ecological Differentiation

A clear natural criterion to identify clusters of evolutionary importance, which we might want to call species, is to find ecological features that distinguish them from close relatives. Among pathogens, the ability to cause a distinctive disease has historically been used to define species,

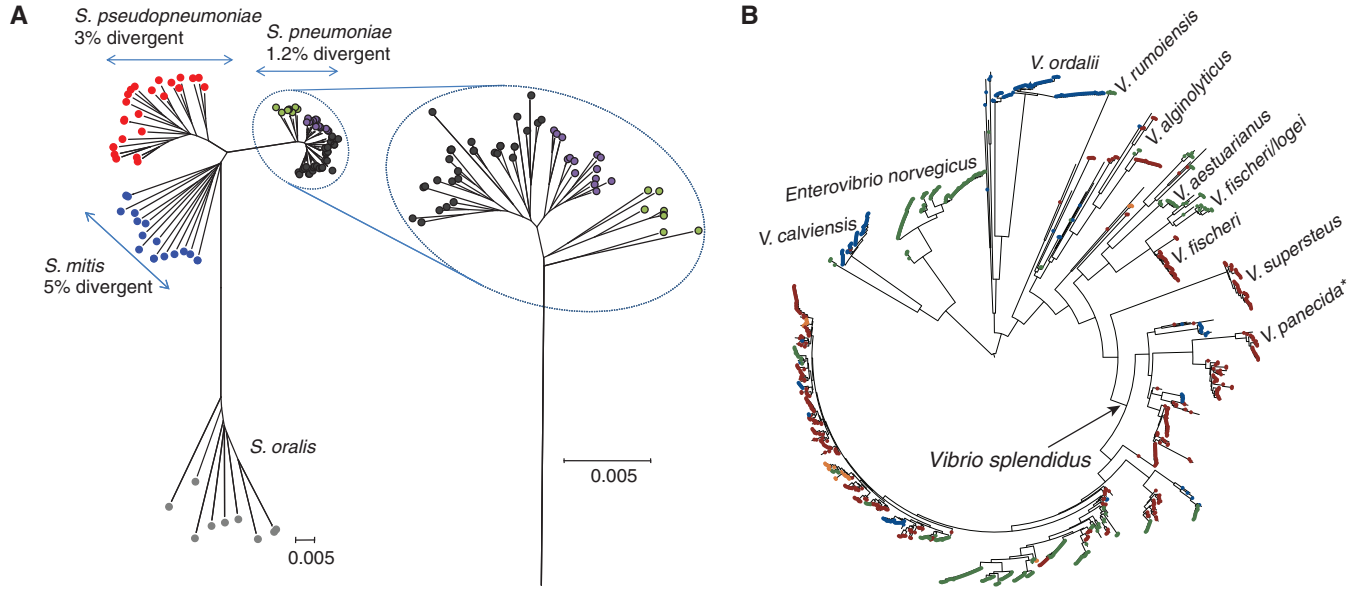

Fig. 1. Multilocus sequence analysis of closely related species. (**A**) Radial minimum evolution tree constructed using MEGA4, showing clusters among 97 isolates of four *Streptococcus* species identified as indicated. The tree was built using concatenates of six housekeeping loci, resulting in a total of 2751 positions in the final data set (*2*). Distances were calculated as the percentage of variant nucleotide sites. The mean distance within the clusters, calculated by MEGA4, is shown. To the right, the pneumococcal cluster is shown at larger scale, and putative subclusters are indicated in dark gray, purple, and green. (**B**) Ecological associations of Vibrionaceae sequence clusters (*13*). Habitats (colored dots) were estimated as differential distributions of groups of closely related strains among samples (size fractions enriched in different environmental resources). Clusters associated with named species are evident, and in most cases species show a clear predilection for one of the habitats. The exception is *V. splendidus*, which breaks up into many closely related ecological populations. Asterisk denotes that trees based on additional loci indicate that the placement of *V. panecida* within *V. splendidus* may be an artifact of horizontal gene transfer at the Hsp60 locus.

but pathogens constitute only a minute fraction of overall bacterial diversity. Mapping of bacterial diversity onto environmental resources indicates that closely related groups of bacteria can be ecologically divergent. For example, fine-scale resource partitioning has been observed among coastal *Vibrio* populations coexisting in the water column (*13*). Partitioning was discovered because strains were collected from distinct, ecologically informative samples, and the phylogenetic structure of the ecologically differentiated populations was superimposed on their habitats. Habitats were defined using an empirical modeling approach. This analysis revealed high levels of specialization for some populations (e.g., *V. ordalii* is only found as single free-swimming cells), whereas others are more generalist (Fig. 1B) and can colonize a wide variety of surfaces, including organic particles and zooplankton in the water column (*13*). Most of the predicted *Vibrio* populations are deeply divergent from each other, and in many cases are congruent with named species; however, *V. splendidus* is a notable exception and splits into numerous closely related groups with distinct ecological preferences, presumably indicating recent ecological radiation from a sympatric ancestral population (*13*). Thus, genetic clusters that correlate with ecology can be discerned.

What do the genetic data tell us about mechanisms of population differentiation and the evolutionary history of the microbes in question? That bacteria are organized into genetic clusters is not, per se, a very interesting observation; many or most models of a population reproducing with a small amount of mutation will eventually produce populations consisting of clusters of related organisms, irrespective of the details of the evolutionary forces or ecological differentiation. A more substantial observation is that there is very little neutral diversity in many populations of microbes, from which we may infer some features of the selective landscape. Neutral diversity is the amount of polymorphism that is evident in noncoding regions or results in synonymous substitutions. One common measure of neutral diversity is the effective population size N_e, defined as the size of a population evolving in the absence of selection that would generate as much neutral diversity as is actually observed. Estimates of N_e for bacteria range from 10^5 to 10^9 (*14*–*18*). To put this into context, the numbers of *Vibrio* cells per cubic meter of seawater in temperate coastal regions range from 10^8 to 10^9 (*19*), which suggests vast census population sizes ($>10^{20}$). This observation—a mismatch of many orders of magnitude between effective population size and census population size (true of most bacteria studied to date)—was originally used to counter claims of neutrality and instead argue that all genetic variation was adaptive (*20*, *21*). However, there are several different mechanisms that can explain this mismatch (Fig. 2).

Whatever mechanisms are driving the differentiation of bacteria into clusters, they must restrict the accumulation of neutral diversity. The first proposed mechanism was based on artificial selection experiments with bacteria grown for extended periods under stable conditions in chemostats, which showed repeated selective sweeps in which the whole genome was thought to hitchhike to fixation along with an advantageous mutation (periodic selection) (*22*). Selective sweeps can purge almost all genetic diversity in the population and thus constitute a candidate mechanism for reducing neutral variation (*23*).

Niches and Ecotypes

To extend this model, one can consider multiple ecological niches characterized by the selective advantages they confer to specific genes. This is the ecotype model, where genes adapted to specific niches cause selective sweeps within those niches but not in other niches. In this way the population will undergo adaptation and differentiation while maintaining relatively low levels of neutral diversity, as selective sweeps confined to each ecotype regularly purge the population of any diversity that might have accumulated (Fig. 2A). Crucially, what neutral diversity we do observe is predicted to be associated with adaptive traits. The ability of such selective sweeps to limit the effective population size has been recognized for some time (*17*, *23*), and this model has been substantially developed by Cohan and colleagues (*4*, *16*, *24*). Because it links patterns of genetic differentiation with adaptation, and makes

reference to the unifying biological principles of selection and niche partitioning, the ecotype has rightly become popular as a framework within which to discuss bacterial evolution, speciation, and ecology.

The ecotype model (*4*, *16*, *24*) predicts that common ancestry will be preserved among bacterial populations within niches (which should be monophyletic), and thus predicts that ecotypes are coherent self-contained gene pools. As a result, it has been suggested that ecotypes should be considered as putative or actual species, depending on the level of genetic differentiation from the ancestral population. This model therefore has the advantage of providing a mechanistic understanding of the evolutionary processes, as well as an organizing principle for classifying species, that is based on experimental observations of bacterial populations.

However, these observations of repeated selective sweeps were made in chemostats, whereas natural environments are markedly unstable and diverse. How would one detect the presence of selective sweeps in natural bacterial populations? The most conclusive examples come not from bacteria but from RNA viruses, which mutate at much higher rates than DNA-based life forms. It has been established from sequences collected over many years that the population structure of the human influenza virus is predominantly driven by repeated selective sweeps (*25*) and that the resulting effective population size N_e (<100) is very much smaller than observed for bacteria. The use of longitudinal ecological and genetic data to distinguish between competing models of evolution has a long pedigree in eukaryotic biology (*26*). On the basis of these analogies, any inference of a population structure driven by selective sweeps would require good longitudinal data from natural bacterial populations, as well as observations of episodic crashes in diversity causally associated with genetic changes and not associated with changes in ecological covariates.

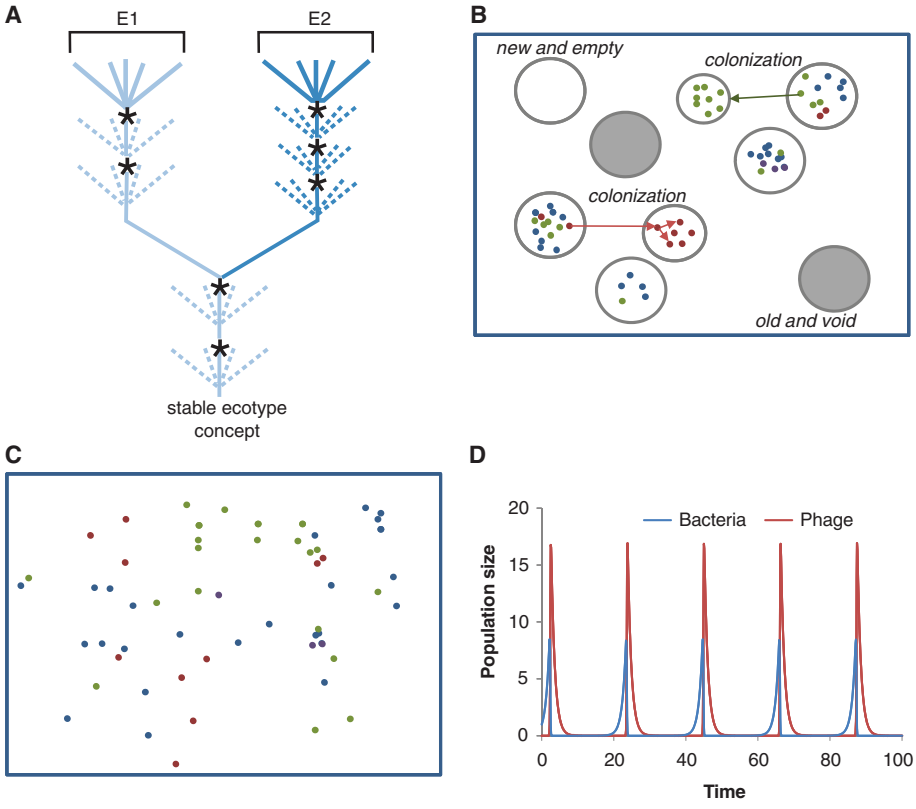

Fig. 2. Different models of microbial evolution that lead to low values of N_e. (**A**) The ecotype model of bacterial population differentiation. The tree shows a single bacterial lineage that differentiates into two sublineages (E1 and E2) that differ in some aspect of their ecology. Periodic selection (a selective sweep) occurs at the points marked by asterisks and eliminates almost all of the diversity that has arisen since the last episode of periodic selection, which is shown by the dashed branches (diversity purged by periodic selection) or solid branches (existing diversity) on the tree. As the two populations are ecologically distinct (i.e., ecotypes), periodic selection in one sublineage does not influence diversity in the other sublineage and vice versa. Each ecotype can therefore diverge to become separate species. Reproduced from (*24*) with permission. (**B**) A metapopulation. Patches of varying size (gray circles) are vacant (empty) or may be colonized by a single genotype randomly acquired from another patch. Strains may diversify within a patch (as shown by different colors representing distinct genotypes), which may colonize empty patches as described above. A characteristic of this sort of metapopulation is patch turnover, in which patches occasionally become unable to support colonization and their inhabitants are removed (solid gray circles). (**C**) A neutral model with small population size. Different genotypes (different colors) arise by mutation or recombination and increase or decrease in the population by random drift. For some purposes, this simple model is an adequate effective description of the more complex processes represented in (A), (B), and (D), and of other more complex evolutionary models not described in this review. (**D**) Predator-prey dynamics and population bottlenecks. Regular population bottlenecks can drastically shrink the effective population size. In this case, bacteria-phage predator-prey dynamics are simulated with a classical Lotka-Volterra model, which can generate oscillations in population size of any amplitude. Population sizes and time axes are in arbitrary units for illustrative purposes only.

Bottlenecks, Metapopulations, and Local Extinctions

The essential element of the ecotype model with respect to limiting neutral diversity is not niche adaptation per se, but rather the effective bottleneck caused by the replacement of the whole population by descendants from a single individual and the resulting extinction of all other lineages (Fig. 2A). Other mechanisms that induce or involve regular population bottlenecks will also restrict neutral diversity. Metapopulation structure, in which the population is divided into patches and where individuals disperse between patches, can generate very low effective population sizes if patches turn over (i.e., if patches are only intermittently able to support bacterial growth, and if a small number of bacteria are dispersed to colonize empty patches) (Fig. 2B) (*27*). This structure well describes the situation for parasites, which can colonize a host but are then forced to move on because the host develops immunity or dies (*17*). It also describes any situation where bacteria use a limited resource intensively for short bursts, followed by dispersal to new resource patches (e.g., colonization of organic particles in seawater by *Vibrio* populations). This metapopulation model is fundamentally different from the ecotype model because it does not predict an association between neutral diversity and adaptive traits.

The relevance of the metapopulation model to the species question is that, although highly idealized and simplified, it may capture some of the effects of complexity and instability of actual ecosystems on population structure. Selective sweeps are predicted to be inevitable in simple, stable environments but not in complex metapopulations [a point partly addressed in (*28*)].

A metapopulation may evolve, differentiate, and adapt without global selective sweeps. Diversity lost by a local selective sweep in one patch may be rescued and reintroduced from other patches. The ecotype model, with its predicted monophyletic relationship between niche and genotype, may therefore not be an appropriate model of speciation in complex ecosystems.

Choosing Between Models

It has proven difficult to discriminate between models of population differentiation that focus on ecotypes or metapopulations. For example, the ecotypic structure of a soil *Bacillus* has been modeled to predict a priori which sequence clusters were ecotypes, and hence which ones should be associated with specific ecological properties (*16*). Some clusters are associated with certain phenotypic traits, such as a propensity to grow on shady north-facing slopes or sunny south-facing slopes. However, this model fitted no better (and in fact slightly worse) than a version of the model with several subpopulations and diversity generated only by neutral drift. This version of the model was dismissed because of its association with a very low estimate of population size (*14*). However, estimates of effective population size N_e are often grossly disconnected from census population sizes. It has proven very challenging to find models that successfully explain low estimated values of N_e while providing better predictions than models based on simple neutral drift. The analysis of *Bacillus* partly did this by predicting more ecotypes in the model than were observed using established ecological criteria, a hypothesis that can be tested.

This problem of low power to detect selection (or, more accurately, to reject neutrality) is a very general problem in population genetics that does not negate the importance of adaptation in evolution, but rather suggests that more work is needed if we want model-based methods to discriminate among different biologically plausible explanations of genetic data. In Table 1 we propose a scheme for performing analyses that could be used to test, develop, and validate different competing models more systematically.

Homologous Recombination

One specific challenge to models that invoke ecotypic structure involves a feature of bacterial evolution—homologous recombination—that we have not yet discussed. Bacterial reproduction does not involve the obligate reassortment of genetic material observed in most higher organisms. However, recombination does occur in bacteria and archaea (*29*) and typically involves the replacement of a short piece of DNA with the homologous segment from another strain. Recombination becomes less probable with increasing sequence divergence between the donor and the recipient (*30*, *31*), which reduces but does not eliminate recombination between closely related species. Because of such interspecies recombination, any given isolate within a species is almost certain to contain at least some genetic material that is characteristic of other closely related species. Hence, whereas it was once thought that bacteria do not form species in the eukaryotic sense because they do not recombine at all (*32*), one current view is that they do not form species because they recombine too much (*5*).

In asexual clonal organisms, even in the absence of any selective pressure, clusters will spontaneously split into multiple lineages or "daughter" clusters (*15*). However, under certain circumstances recombination can prevent this, and we can hence divide the bacteria into "sexual" and "nonsexual" species. This effect, described at greater length elsewhere (*15*), is summarized in Fig. 3, which shows the rate at which two clusters diverge over time—that is, the increase in the mean genetic distance between them. If this becomes negative, then the two clusters will stop diverging and instead converge. The three examples shown in Fig. 3 differ only in the rate of homologous recombination between the clusters, all other parameters being held constant. As recombination increases, we see a distinction between a "clonal" organism in which clusters are predicted to diverge (the green line) and a "sexual" organism (the blue line) in which they are predicted to be held together by recombination. For "sexual" species, the divergence of clusters requires a process that reduces the rate of recombination between them—for example, a period of allopatry or ecological differentiation. The speciation point is the amount of divergence between clusters that needs to accumulate to prevent them from returning to a single cluster if the barriers to recombination are removed. A recent study hypothesized that two related *Campylobacter* species are currently undergoing this process of merging into a single species as a result of changes in their environment (*33*).

The above insights were reached using models based on the assumption that genetic variation is neutral. Although this is obviously not always an appropriate assumption, it is plausible that the number of loci explicitly involved in adaptive ecological differentiation will be small, and thus that in an unstable landscape, genomic barriers to recombination will depend more on the accumulation of differences at neutral loci than at adaptive loci. The models also assumed a homogeneous distribution of polymorphisms across

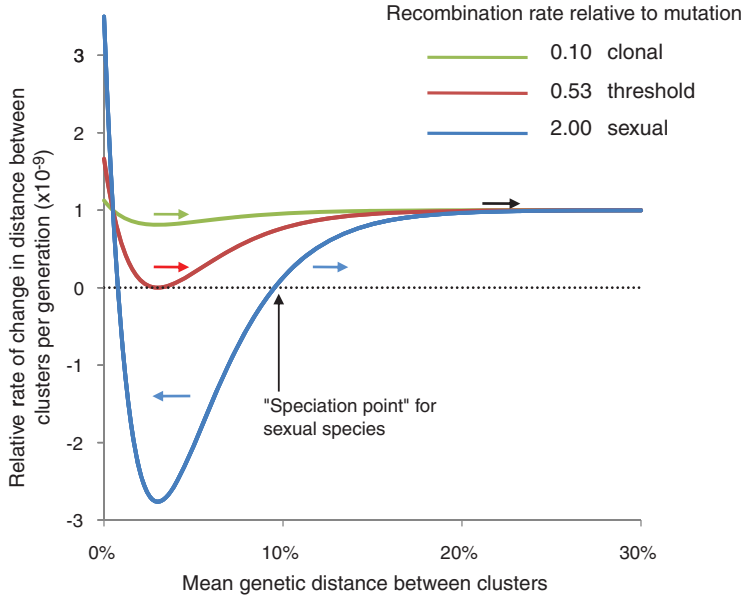

Fig. 3. The dynamics of cluster divergence. The figure summarizes some key results from (*15*) in a phase-space plot of the genetic dynamics of two populations, with recombination occurring between them at a rate that is varied for the three different simulations. The *y* axis shows the rate of change of genetic distance between the clusters as a function of the genetic distance itself (*x* axis). When the rate of change is positive, the populations will diverge genetically; when negative, they converge. The direction of change for each scenario is shown by arrows color-coded to each scenario. For low recombination rates, the populations are effectively clonal and always diverge (green line). As the recombination rate increases, the cohesive effects of recombination slow the rate of divergence, until a threshold is passed (red line) and the populations become effectively sexual in the sense that the populations no longer diverge. For recombination rates above this level, the fate of the two populations will depend on how genetically distinct they are at the outset. If they are within the "speciation point," then recombination will cause them to merge. If they are farther away than this "speciation point," they will continue to diverge from each other. These curves are derived using the model described in (*15*).

the genome, and violation of this may alter the tempo and mode of these processes (*34*, *35*).

Illegitimate Recombination and Gene Content Variation

Illegitimate recombination or gene acquisition is another unusual feature of bacteria. In this case, genes or clusters of genes are acquired that typically have no homolog(s) in the recipient strain. The importance of this phenomenon is evident in the clear and ubiquitous signature of such events in the growing body of genomic data. These are identified by differences in the characteristics of the acquired DNA and that of the host strain, for example, in base composition or codon usage; in most cases, the donor of the DNA in question is unknown. Gene acquisition leads to genomes being punctuated by stretches of foreign DNA. The largest of these (which may be many kilobases in length) were initially termed "pathogenicity islands," because the new functions encoded by the imports were often involved in virulence, but a better term is "genomic islands" as the phenomenon is far from limited to pathogens (*36*, *37*). Although it is hard to quantify the selective impact of importing any given gene(s) into a new background, the occasional ability to gain a new adaptation in this fashion—such as a new metabolic capability or a new mode of transmission for a pathogen—may be of enormous importance in terms of speciation.

Perhaps even more striking is the amount of variation in gene content revealed by multiple genomes from the same species, which implies that gene acquisition occurs at a surprisingly high frequency. It is now commonplace to speak of the "core" genome, which encodes fundamental functions shared by all members of a species (and, it should go without saying, other related species), onto which is bolted the "auxiliary" or "accessory" genome, composed of genes and operons that may or may not be present in all isolates. It seems likely that such auxiliary genes help to determine the specific ecological properties of the organism. For example, a group of related *Leptospirillum* has recently been hypothesized to adapt to different areas of an acid mine drainage system by shuffling of chromosome segments enriched in noncore genes (*38*, *39*). We should, however, be aware that changes in core genes may also lead to ecological differentiation, a phenomenon well documented in experimental studies of bacteria growing in structured environments (*40*).

Estimates vary, depending on the genomes that are available, but as little as 40% of genes may be present in all sequenced genomes of a named species (*41*). We may consider genes within a named species as being characteristic of different levels of ecological specificity, ranging from highly conserved core functions that are essential for growth in all environments to loci that are involved with adaptation to a specific habitat. Some narrow niche-specific genes may be distributed across species, being transferred between them by mobile elements. The evolutionary fate of such genes may hence be only loosely coupled with that of any particular species or strain in which they are found, and they are maintained through selection by the habitat to which each host strain is adapted. In the case of very mobile elements—for example, plasmids encoding resistance to antibiotics or heavy metals—the ecological specificity determined by these accessory loci may have no link to the sequence clusters we observe using housekeeping genes (Fig. 4).

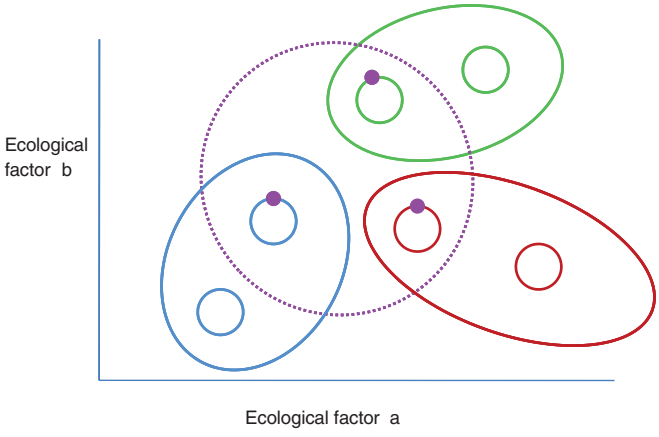

Fig. 4. Differences between core and auxiliary genes. This schematic illustrates the relationships between three species in "ecotype space," shown here in two dimensions, and a mobile gene common to all three. The areas occupied by the species are shown as solid lines in red, blue, and green. The part of the ecological space where the shared mobile gene is selected in each species is shown by a dashed purple line and overlaps all three species ranges. Examples of (circular) genomes from each species with and without the purple mobile element are also illustrated. Note that for each species, the locus is not selected for all isolates, and its evolutionary fate is uncoupled from that of each host species, because if one undergoes a selective sweep or goes extinct, the mobile gene may be reintroduced from one of the other species. Examples of such distributed loci include drug resistance determinants in pathogens (e.g., β-lactamase genes) and heavy metal resistance in environmental organisms. These genes may be transferred among strains and species by conjugative plasmids or other mobile elements (including transducing phage).

Identifying Mechanisms and Delineating Species

What do we want from bacterial species? Do we need theoretical consistency even at the expense of taxonomic practicality, incorporating both "clonal" and "sexual" populations into a single theoretical framework? One unifying theoretical concept is to consider species as the arena within which individuals are similar enough, or interbreed enough, that individual variant genes compete directly for reproductive success. Practical advances building on this or other theoretical concepts will only come when these are developed into explicit models and model-based algorithms that are tested and refined on a wide range of data. Alternatively, it may be sensible to suggest an ad hoc application of principles to different genera on the basis of their specific characteristics, including the extent of variation in gene content and recombination. In any case, no biologist would deny the importance of ecology to what we observe, but it may not be easy to incorporate it in a fashion that is convenient for taxonomists. Nonetheless, population geneticists may have little choice but to tackle the question of defining bacterial species or, at the very least, populations. Whether we are estimating effective population size from neutral diversity or choosing an appropriate set of strains to test for positive selection at a locus of interest, species definitions are implicit in much of the analytical toolkit of population genetics.

Distinguishing among mechanisms of population differentiation in bacteria ultimately comes down to testing the ability of different models to explain highly variable patterns within and between genetic-ecological clusters (Fig. 1). It is still unclear whether these patterns are maintained by gene flow or selection, and what the effect of population structure is. The joint distribution of genetic and ecological data can be used, as described above for *Vibrio* species (*13*), to define populations without making a strong theoretical commitment to either of these alternatives. One clear result from all of the studies discussed here is that the underlying theoretical questions concerning species will not be answered in the absence of more detailed genetic-environmental mapping. Moreover, some guidelines for the types of ecological studies that will be most informative are emerging. Most important, the ecological data collected must be relevant to the niche boundaries of the populations studied. And if genetic groups do not map exclusively onto sampling categories (as is likely to be the case), more complex statistical models will be needed to identify and describe the underlying niche structure. Longitudinal studies that measure the dynamics of ecological associations over time will also be helpful to determine how transient natural habitats are, and thus how likely bottlenecks are to result. Finally, whole-genome sequences from entire populations of environ-

Table 1. A proposed strategy for developing and validating models of bacterial evolution that might eventually be used to classify genetic diversity data and provide a firm foundation for a bacterial species concept.

1. Collect samples according to systematic ecological stratification. Focus on longitudinal studies, geographical studies, and measurement of physical and chemical gradients affecting bacterial growth. Consider biotic factors such as the presence of other competing bacteria or parasitic phage.
2. For each isolate, sequence as much as possible and affordable (16*S* rRNA, MLSA, auxiliary genes, full genomes, etc.).
3. Use empirical classification algorithms that use genetic and ecological data to jointly map isolates.
4. To guide model formulation, use population genetic tests on observed clusters, focusing on tests for selection, population structure, and gene flow.
5. Generate evolutionary models and simulate populations.
6. Test, then reject or adapt, evolutionary models according to agreement between simulations and real populations; if necessary, return to step 1.
7. For successful models, develop model-based methods for interpreting pure genetic data (without ecological covariates) and test on new data.
8. If one or more validated models emerge, use these to classify genetic data and to develop bacterial species concepts.

mental bacteria will be useful in dissecting the roles of the auxiliary and core genome in ecological differentiation. If after this process it emerges that some model or models are consistently validated for different study systems, these would inevitably form a good basis for identifying fundamental levels of clustering, or species.

In the foregoing we have emphasized ecotype and metapopulation models, but there are others that deserve consideration—notably the epidemic clonal model (*42*) and the impact of phage epidemics causing classic Lotka-Volterra boom-bust dynamics (*43*) illustrated in Fig. 2D—and it is possible, even likely, that more than one of these mechanisms may be relevant to any given problem in speciation and cluster formation. Distinguishing among these mechanisms is the bacterial species challenge (Table 1), described in 1991 by John Maynard Smith as follows: "Ecotypic structure, hitch-hiking, and localized recombination can explain the observed patterns of variation. The difficulty, of course, is that the model is sufficiently flexible to explain almost anything. To test the hypothesis of ecotypic structure, we need to know the distribution of electrophoretic types [i.e., genotypes] in different habitats" (*17*).

Much research on bacterial species to date has come from studies on pathogens, where the correct identification of species is crucial for accurate clinical diagnoses. However, for pathogens the identification of the multiple ecological niches within (for example) the nasopharynx or gut is difficult, and studies of the relationships between bacterial populations and ecology may be more fruitful for some environmental species where the categorization of niches is a more tractable enterprise. Hopefully, we will soon obtain richer data sets that map bacterial diversity onto ecology and provide a way to distinguish among various models of population differentiation and speciation, including those based on ecotypes or metapopulations.

References and Notes

1. J. Handelsman, *Microbiol. Mol. Biol. Rev.* **68**, 669 (2004).
2. W. P. Hanage, C. Fraser, B. G. Spratt, *Philos. Trans. R. Soc. London Ser. B* **361**, 1917 (2006).
3. M. Achtman, M. Wagner, *Nat. Rev. Microbiol.* **6**, 431 (2008).
4. F. M. Cohan, *Annu. Rev. Microbiol.* **56**, 457 (2002).
5. W. F. Doolittle, R. T. Papke, *Genome Biol.* **7**, 116 (2006).
6. D. Gevers *et al.*, *Nat. Rev. Microbiol.* **3**, 733 (2005).
7. M. F. Polz, D. E. Hunt, S. P. Preheim, D. M. Weinreich, *Philos. Trans. R. Soc. London Ser. B* **361**, 2009 (2006).
8. D. B. Rusch *et al.*, *PLoS Biol.* **5**, e77 (2007).
9. C. Darwin, *The Origin of Species* (Penguin Classics, London, 1985).
10. E. Stackebrandt *et al.*, *Int. J. Syst. Evol. Microbiol.* **52**, 1043 (2002).
11. R. Facklam, *Clin. Microbiol. Rev.* **15**, 613 (2002).
12. J. C. Arbique *et al.*, *J. Clin. Microbiol.* **42**, 4686 (2004).
13. D. E. Hunt *et al.*, *Science* **320**, 1081 (2008).
14. See supplementary information in (*16*) for a statistical comparison of the ecotype model with an effective neutral model and an implicit estimate of N_e.
15. C. Fraser, W. P. Hanage, B. G. Spratt, *Science* **315**, 476 (2007).
16. A. Koeppel *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 2504 (2008).
17. J. Maynard Smith, *Proc. R. Soc. London Ser. B* **245**, 37 (1991).
18. H. Ochman, A. C. Wilson, in Escherichia coli *and* Salmonella typhimurium*: Cellular and Molecular Biology*, F. C. Neidhart, Ed. (ASM Press, Washington, DC, 1987), pp. 1649–1654.
19. J. R. Thompson *et al.*, *Appl. Environ. Microbiol.* **70**, 4103 (2004).
20. M. Kimura, *Trends Biochem. Sci.* **1**, N152 (1976).
21. R. Milkman, *Trends Biochem. Sci.* **1**, N152 (1976).
22. K. C. Atwood, L. K. Schneider, F. J. Ryan, *Proc. Natl. Acad. Sci. U.S.A.* **37**, 146 (1951).
23. B. R. Levin, *Genetics* **99**, 1 (1981).
24. F. M. Cohan, E. B. Perry, *Curr. Biol.* **17**, R373 (2007).
25. A. Rambaut *et al.*, *Nature* **453**, 615 (2008).
26. R. A. Fisher, E. B. Ford, *Heredity* **1**, 143 (1947).
27. M. Slatkin, *Theor. Popul. Biol.* **12**, 253 (1977).
28. J. Majewski, F. M. Cohan, *Genetics* **152**, 1459 (1999).
29. E. J. Feil, B. G. Spratt, *Annu. Rev. Microbiol.* **55**, 561 (2001).
30. J. Majewski, F. M. Cohan, *Genetics* **153**, 1525 (1999).
31. J. Majewski, P. Zawadzki, P. Pickerill, F. M. Cohan, C. G. Dowson, *J. Bacteriol.* **182**, 1016 (2000).
32. S. T. Cowan, in *Microbial Classification, 12th Symposium of the Society for General Microbiology*, G. C. Ainsworth, P. H. A. Sneath, Eds. (Cambridge Univ. Press, Cambridge, 1962), pp. 433–455.
33. S. K. Sheppard, N. D. McCarthy, D. Falush, M. C. J. Maiden, *Science* **320**, 237 (2008).
34. A. C. Retchless, J. G. Lawrence, *Science* **317**, 1093 (2007).
35. K. Vetsigian, N. Goldenfeld, *Proc. Natl. Acad. Sci. U.S.A.* **102**, 7332 (2005).
36. A. Tuanyok *et al.*, *BMC Genomics* **9**, 566 (2008).
37. U. Dobrindt, B. Hochhut, U. Hentschel, J. Hacker, *Nat. Rev. Microbiol.* **2**, 414 (2004).
38. P. Wilmes, S. L. Simmons, V. J. Denef, J. F. Banfield, *FEMS Microbiol. Rev.* **33**, 109 (2009).
39. V. J. Denef *et al.*, *Environ. Microbiol.* **11**, 313 (2008).
40. E. Bantinaki *et al.*, *Genetics* **176**, 441 (2007).
41. R. A. Welch *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 17020 (2002).
42. J. Maynard Smith, N. H. Smith, M. O'Rourke, B. G. Spratt, *Proc. Natl. Acad. Sci. U.S.A.* **90**, 4384 (1993).
43. K. H. Hoffmann *et al.*, *FEMS Microbiol. Lett.* **273**, 224 (2007).

44. We thank T. Connor and S. Deeny for useful discussions. Supported by University Research Fellowships from the Royal Society (C.F. and W.P.H.), a program grant from the Wellcome Trust (B.G.S.), grants from the U.S. Department of Energy Genomes to Life program (M.F.P. and E.J.A.), and the NSF/National Institute of Environmental Health Sciences Woods Hole Centre for Oceans and Human Health, the NSF Biological Oceanography Program, and the Moore Foundation (M.F.P.).

10.1126/science.1159388

REVIEW

Is Genetic Evolution Predictable?

David L. Stern[1]* and Virginie Orgogozo[2]*

Ever since the integration of Mendelian genetics into evolutionary biology in the early 20th century, evolutionary geneticists have for the most part treated genes and mutations as generic entities. However, recent observations indicate that all genes are not equal in the eyes of evolution. Evolutionarily relevant mutations tend to accumulate in hotspot genes and at specific positions within genes. Genetic evolution is constrained by gene function, the structure of genetic networks, and population biology. The genetic basis of evolution may be predictable to some extent, and further understanding of this predictability requires incorporation of the specific functions and characteristics of genes into evolutionary theory.

One hundred and fifty years ago, Charles Darwin and Alfred Russell Wallace proposed that biological diversity results from natural selection acting on heritable variation in populations. Both Darwin and Wallace recognized the importance of heritable variation to evolutionary theory, but neither man knew the true cause of inheritance. Early in the 20th cen-

The Bacterial Species Challenge: Making Sense of Genetic and Ecological Diversity

Christophe Fraser, Eric J. Alm, Martin F. Polz, Brian G. Spratt and William P. Hanage

Science **323** (5915), 741-746.
DOI: 10.1126/science.1159388