Population Genomics of Early Events in the Ecological Differentiation of Bacteria

B. Jesse Shapiro *et al.*

Science **336**, 48 (2012);
DOI: 10.1126/science.1218198

Population Genomics of Early Events in the Ecological Differentiation of Bacteria

B. Jesse Shapiro,[1,2]* Jonathan Friedman,[1] Otto X. Cordero,[3] Sarah P. Preheim,[3] Sonia C. Timberlake,[4] Gitta Szabó,[3]† Martin F. Polz,[3]‡ Eric J. Alm[1,2,3,4]‡

Genetic exchange is common among bacteria, but its effect on population diversity during ecological differentiation remains controversial. A fundamental question is whether advantageous mutations lead to selection of clonal genomes or, as in sexual eukaryotes, sweep through populations on their own. Here, we show that in two recently diverged populations of ocean bacteria, ecological differentiation has occurred akin to a sexual mechanism: A few genome regions have swept through subpopulations in a habitat-specific manner, accompanied by gradual separation of gene pools as evidenced by increased habitat specificity of the most recent recombinations. These findings reconcile previous, seemingly contradictory empirical observations of the genetic structure of bacterial populations and point to a more unified process of differentiation in bacteria and sexual eukaryotes than previously thought.

How adaptive mutations spread through bacterial populations and trigger ecological differentiation has remained controversial. Although it is agreed that the key factor is the balance between recombination and positive selection, theory and observations are seemingly at odds. On one hand, evidence for genes spreading through populations independently via recombination ("gene-specific sweeps") is found in observations of environment-specific genes (*1*) and alleles (*2*), and reduced diversity at single loci amid high genomewide polymorphism (*3*, *4*). On the other hand, mathematical modeling suggests that empirically observed rates of homologous recombination should not be high enough to unlink a gene, which is under even moderate selection, from the rest of the genome (*5*, *6*). This recombination/selection balance, expressed most saliently by the ecotype theory, leads to a prediction that is actually observed but that is at odds with gene-specific sweeps [i.e., bacterial diversity is organized into ecologically differentiated clusters (*7*–*9*)]. The proposed mechanism involves cycles of neutral diversification punctuated by genomewide selective sweeps (*6*). Although the observations of environment-specific genes and locus-specific reduced diversity conflict with the ecotype model of selected clonal genomes, they do not explain why its prediction of coincident genetic and ecological clusters hold true, nor provide insights into the early genomic events accompanying adaptation. How to reconcile these seemingly contradictory empirical observations remains an open question.

Here, we test whether recombination is strong enough relative to selection to allow gene-specific rather than genomewide selective sweeps in natural microbial populations and explore the effect on population-level diversity. Using whole-genome sequences from two recently diverged *Vibrio* populations with clearly delineated habitat associations, we show that genome regions rather than whole genomes sweep through populations, triggering gradual, genomewide differentiation. Our proposed evolutionary scenario is based on three lines of evidence: (i) Most of the genetic divergence between ecological populations is restricted to a few genomic loci with low diversity within one or both of the populations, suggesting recent sweeps of confined regions of the genome. (ii) We show that only one of the two chromosomes constituting the genome has swept through part of one population. (iii) The most recent recombination events tend to be population specific but older events are not, reinforcing the notion that these populations are on independent evolutionary trajectories, which may ultimately lead to the formation of genotypic clusters with different ecology. Although such clusters have been interpreted as evidence for the ecotype model, our results suggest that they can arise even in populations that do not experience genomewide selective sweeps.

In a previous study, we noticed an instance of very recent ecological differentiation among two populations of *Vibrio cyclitrophicus* by their divergence in fast-evolving protein-coding genes and differential occurrence in the large (L) and small (S) size fractions of filtered seawater, suggesting association with different zoo- and phytoplankton or suspended organic particle types (*8*). This population structure was reproduced across independent samples taken in 2006 and 2009. We sequenced whole genomes from both populations (13 L and 7 S isolates, all obtained in 2006). As in other *Vibrionaceae*, these genomes consist of two chromosomes, each with a flexible and core component, defined as blocks of DNA not universally present in all isolates or shared by all, respectively. To estimate the extent and patterns of recombination among the isolates, we subdivided the core genome into blocks of DNA on the basis of their supporting different phylogenetic relationships among the 20 isolates (*10*). Overall, the ecological populations described here are among the most closely related (identical 16*S* and >99% average amino acid identity) studied with genomewide sequence data, making them an ideal test case for observing the early events involved in ecological differentiation.

Genes, not genomes, sweep populations. Our first line of evidence favoring gene-specific rather than genomewide selective sweeps is that most of the differentiation between populations is restricted to a few small patches of the core genome. Ecological differentiation is supported by 725 "ecoSNPs" (single-nucleotide polymorphisms)—defined as dimorphic nucleotide positions with one variant present in all S strains and a different variant in all L strains—that cluster in a few discrete patches of the genome (11 in total, three of which contain >80% of ecoSNPs). By contrast, the rest of the genome is dominated by 28,744 SNPs, supporting phylogenetic intermingling of S and L strains (e.g., nucleotide C in 3 S and 6 L strains, G in 4 S and 7 L strains), therefore rejecting the ecological partition (Fig. 1 and figs. S1 and S2). Any signal of clonal ancestry has been obscured by homologous recombination, which affects equally genes of all functions, and is therefore likely not driven by selection [fig. S3 (*10*)], such that no single bifurcating tree relating the 20 strains adequately describes the evolution of more than 1% of the core genome (Fig. 1C). Such a pattern could have been produced either by an ancient genomewide selective sweep in one or both populations, followed by recombination between populations eroding the "clonal frame" down to a few regions, or by recent gene-specific selective sweeps centered on these few regions. The latter explanation is favored because most major ecoSNP clusters (three out of the four peaks in Fig. 1B) have significantly lower within-habitat diversity (in one or both habitats) than the chromosome-wide average. The exception is the

[1]Program in Computational and Systems Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [2]Broad Institute, Cambridge, MA 02142, USA. [3]Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA. [4]Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA.

*Present address: Center for Communicable Disease Dynamics, Harvard School of Public Health, Boston, MA 02115, and Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA 02138, USA.

†Present address: Department of Microbial Ecology, University of Vienna, Vienna, Austria.

‡To whom correspondence should be addressed. E-mail: ejalm@mit.edu (E.J.A.); mpolz@mit.edu (M.F.P.)

highly diverse *RTX/RpoS* locus, which may be under diversifying selection both within and between habitats. The low within-habitat diversity in the other three regions, which account for the majority of ecoSNPs, suggests that they arrived recently by recombination [likely from a distantly related population (*10*)] and swept through a population before accumulating much polymorphism.

Our second line of evidence shows that genomic fragments can sweep through populations in an ecology-specific manner without purging genomewide variation. In particular, a large fraction of chromosome II has swept through a subset of the S population, without affecting the diversity of chromosome I. As evidence for this, each chromosome has a distinct core phylogeny, with five of the seven S strains grouping together on chromosome II, but not chromosome I (Fig. 1). This "5-S" clade (grouping together strains 1F97, 1F111, 1F273, FF274, and FF160; blue branch in Fig. 1A and blue points in Fig. 1B) is supported by 796 SNPs: 790 on chromosome II and six on chromosome I—a >200-fold imbalance after normalizing by the 1.45 times as many SNPs per site on chromosome II. Chromosome II also strongly supports one phylogeny within the 5-S strains; SNPs inconsistent with this phylogeny are restricted almost entirely to chromosome I (figs. S4 and S5). The degree of support for the 5-S group on chromosome II suggests that a variant of this chromosome swept through these five S strains, independently of chromosome I. The sweep likely occurred recently, before the clear phylogenetic signal within the 5-S strains was disrupted by recombination. This signature of a long stretch of DNA (in this case, a chromosome) largely uninterrupted by recombination is a hallmark of recent positive selection in sexual eukaryotes (*11*), suggesting a selective sweep of chromosome II independently of the rest of the genome (chromosome I). The mobilization of genomic fragments on the size scale of chromosomes may also explain the hybrid genomes observed in novel pathogenic variants of *Vibrio vulnificus* (*12*).

Emergent habitat-specific recombination. Our third line of evidence shows how, despite the lack of genomewide selective sweeps, tight genotypic clusters may eventually emerge as a result of preferential recombination within, rather than between, habitats. This is evident from quantification of recent recombination in the core genome, using three very recently diverged pairs of "sister strains"—1F175-1F53, 1F111-1F273, and ZF30-ZF207—that group together at nearly all SNPs in the genome (Fig. 1A). The grouping of such young sister pairs should only be broken by

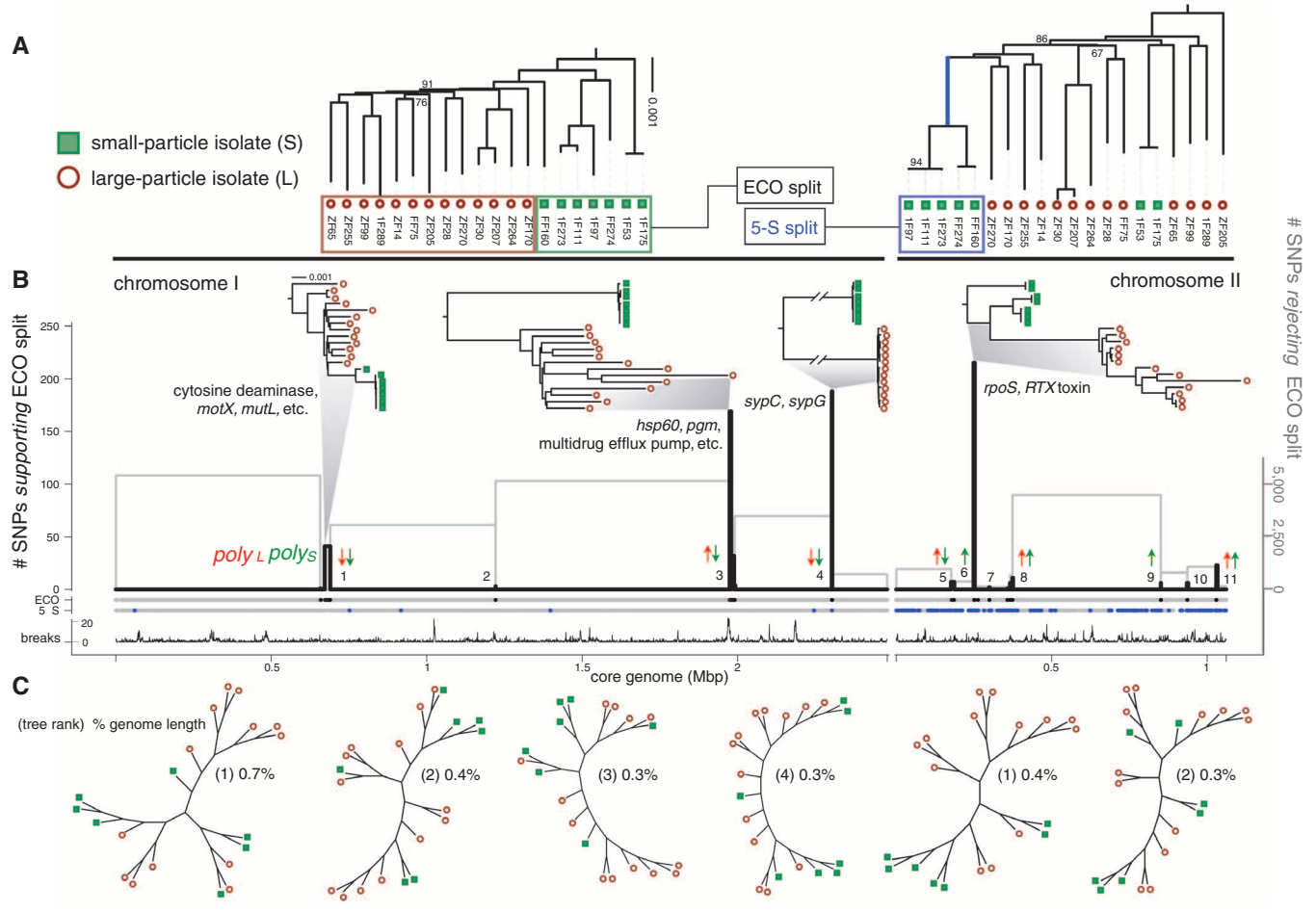

Fig. 1. Phylogeny follows ecology at just a few habitat-specific loci. (**A**) Maximum-likelihood (ML) *V. cyclitrophicus* phylogenies rooted by *V. splendidus* 12B01, based on core genome nucleotide sequence for chromosome I (left) and II (right). Scale is substitutions per site; all nodes have 100% bootstrap support unless indicated. (**B**) Genome regions with uninterrupted support for (black bars) or against (gray bars; note different scale) the ecological split of strains into distinct habitats (S/L). Bar height indicates the number of informative SNPs in each region. ECO-sup regions 1 to 11 are described in table S2; ML trees for four major regions are shown, rooted with 12B01; $poly_L$/$poly_S$ indicates regions with significantly higher (up arrows) or lower (down arrows) nucleotide diversity and density of segregating polymorphic sites within the L (red) or S (green) habitat, relative to the chromosome-wide average. Tracks below *x* axis are as follows. "ECO": locations of ECO-supporting (black points) and -rejecting (gray) SNPs. "5-S": SNPs supporting (blue points) or rejecting (gray) the 5-S branch. "Breaks": number of inferred recombination breakpoints per kb. (**C**) Tree topologies accounting for most genome length. Top four ranked unrooted topologies are shown for chromosome I, top 2 for chromosome II, and the percentage of the core genome accounted for (*10*).

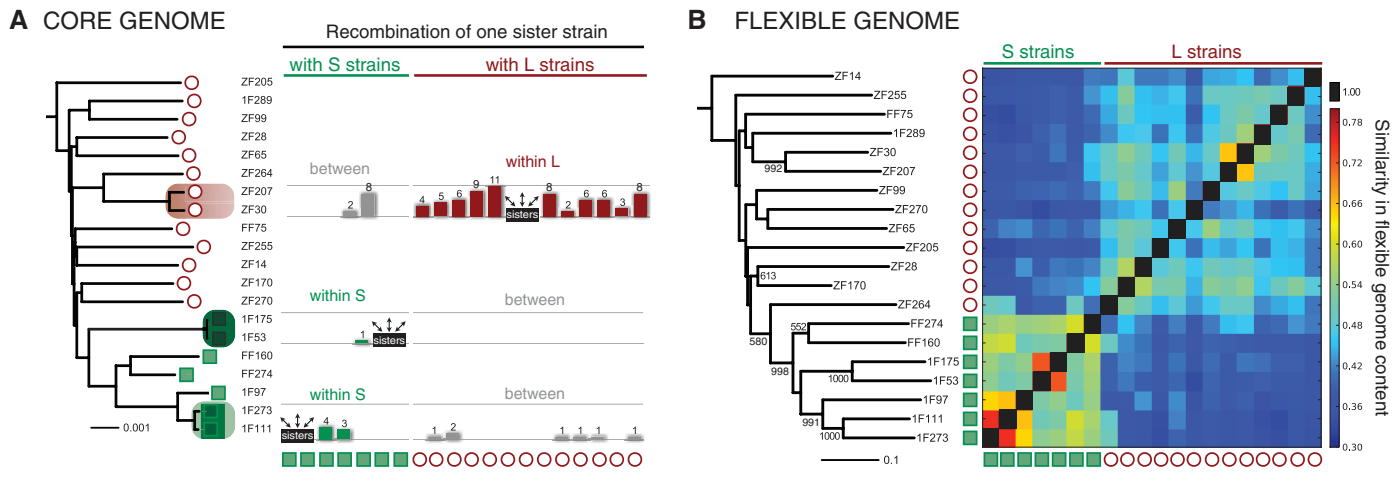

Fig. 2. Recent recombination is more common within than between habitats. (**A**) Genomewide ML phylogeny based on 3.54 Mb of aligned core genome, with sister strains highlighted in red or green. All nodes have 79 to 100 bootstrap support. Bar graphs show events (number of core genome blocks) that split up sisters by recombination between (gray bars) or within habitats (S, green; L, red). (**B**) Relative amount of shared flexible genomic blocks between strains. The neighbor-joining (NJ) tree (left) is a consensus across 1000 bootstrap resamplings of the flexible blocks. Only nodes with support >500 are shown. Scale bar: Bray-Curtis distance used to construct the NJ tree (*10*).

the most recent recombination events identifiable in our sample, involving one of the sister strains as a donor or acceptor. We quantified such events by counting core genome blocks inconsistent with phylogenetic pairing of sister strains (*10*). Out of 93 such blocks (Fig. 2A), 76 resulted from one sister strain pairing with another strain from the same habitat. This is significantly more within-habitat recombination than expected under a model with random recombination across habitats [$P < 1 \times 10^{-5}$ (*10*)]. The excess within-habitat recombination was detectable in both S ($P = 0.03$) and L ($P < 1 \times 10^{-5}$) populations considered separately and is robust to variation in our assumptions about the relative S:L population sizes (*10*). By contrast, the pairing of more anciently diverged S strains, FF160 to FF274, is more often broken up by recombination with L (222 blocks) than with S strains (8 blocks) ($P < 1 \times 10^{-5}$), perhaps owing to the higher abundance of L strains in the past (e.g., if the ancestral, undifferentiated population was L-associated). This finding suggests that the trend toward the habitat-specific gene flow that we identified has emerged relatively recently.

The preference for within-habitat recombination is also apparent in the flexible genome. This component of the genome changes so rapidly that even the two most closely related genomes in our study (1F175 and 1F53), differing by only 66 substitutions in 3.54 Mb of core genome, each contain about 4500 base pairs of unique DNA (fig. S6). The flexible genome tree also has a topology that differs from that of the core (Fig. 2), suggesting that the flexible genome is shaped largely by horizontal transfer (integrase-mediated and illegitimate recombination), with limited clonal descent. The separate grouping of S and L strains (Fig. 2B; 99.8% bootstrap support), when clustered by the proportion of shared flexible DNA (Fig. 2B), indicates that preferential recombination occurs within habitats. Compared with a model of random recombination among habitats, there is significantly more habitat-specific sharing of flexible blocks than expected by chance [$P < 5.5 \times 10^{-58}$ (*10*)] (table S1). All seven S strains—not just the 5-S strains hypothesized to have undergone a selective sweep on chromosome II—share a relatively high fraction of their flexible DNA on this chromosome (fig. S7). Therefore, flexible genome turnover is sufficiently rapid that flexible DNA does not hitchhike with selective sweeps for very long. Rather, high turnover, with a clear bias toward within-habitat sharing of DNA, maintains distinct but dynamic and habitat-specific gene pools.

Functions of ecologically differentiated genes. The revelation that there is a suite of habitat-specific genes and alleles has shed light on the selective pressures associated with specialization to different microhabitats in the ocean [tables S1 and S2 (*10*)]. The *RTX* locus and *syp* operon exhibit both allelic variation (core) and gene content variation (flexible). Several *syp* genes, present in all L but absent from S genomes, and their upstream regulator *sypG*, present in different allelic variants between habitats, are involved in biofilm formation and host colonization (*13*). RTX proteins are important virulence factors in pathogens (*14*) and may play a role in interactions with different hosts. The stress-response sigma factor RpoS encoded in the core genome near the *RTX* locus, has been shown to mediate a trade-off between stress tolerance and nutritional specialization in environmental *Escherichia coli* isolates (*15*). Finally, genes responsible for the biosynthesis of mannose-sensitive hemagglutinin (MSHA), many of which are unique to L flexible genomes, promote adherence to chitin (*16*) and zooplankton exoskeletons (*17*). Together, this evidence suggests that ecological specialization, possibly through differential host association, can be achieved by fine-tuning genes in a few key functional pathways.

A model for ecological differentiation in bacteria. Our observations can be generalized with a model predicting independent evolutionary trajectories for nascent populations triggered by gene-specific sweeps (Fig. 3). The mosaic genomes that we observed, with different genome blocks supporting different phylogenies, suggest a frequently recombining, ecologically uniform ancestral population (Fig. 3B, early time points). The recent acquisition of habitat-specific flexible genes and core alleles likely initiated specialization to different hosts or habitats, leading to decreased gene flow between populations. The populations that we studied are in a very early stage of ecological specialization, with little genetic divergence between them. However, if the trend toward greater within-population recombination can be extrapolated into the future [as might indeed be expected given that recombination drops log-linearly with sequence divergence (*18*–*22*)], they will eventually form distinct genetic clusters, potentially indistinguishable from those predicted by (and often taken as evidence for) the ecotype model (Fig. 3A). Genetic isolation by preferential recombination has been suggested previously (*23*), and this trend might be enhanced if homologous recombination between populations is reduced in the vicinity of acquired habitat-specific genes (*24*). Thus, a mechanism of gene-centered sweeps may eventually lead to a pattern characteristic of genomewide sweeps. In this way, our study of the very early stages of ecological specialization has provided a simple resolution to seemingly conflicting empirical observations.

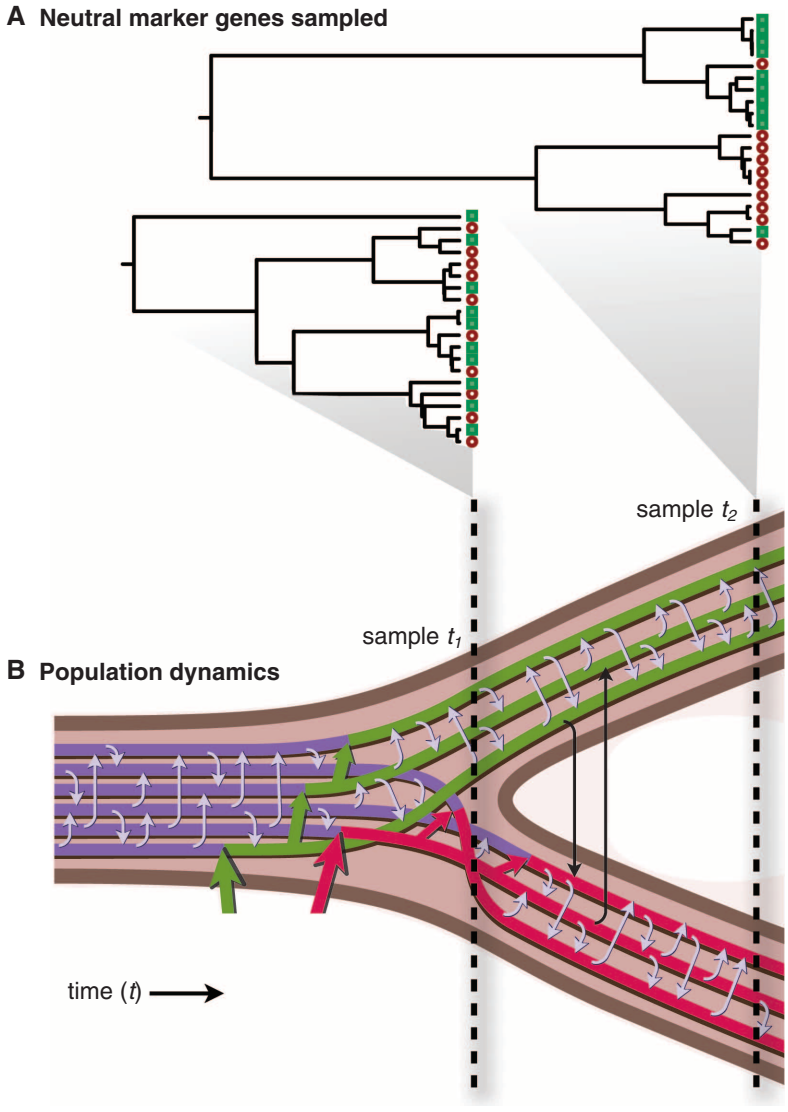

Fig. 3. Ecological differentiation in recombining microbial populations. (**A**) Example genealogy of neutral marker genes sampled from the population(s) at different times. (**B**) Underlying model of ecological differentiation. Thin gray or black arrows represent recombination within or between ecologically associated populations. Thick colored arrows represent acquisition of adaptive alleles for red or green habitats.

Outlook. Our findings of ecological differentiation driven by gene-specific rather than genomewide selective sweeps, followed by gradual emergence of barriers to gene flow, leave open three major questions for future investigation: What mechanisms (aside from unrealistically high recombination rates) are responsible for preventing genomewide selective sweeps (e.g., negative frequency-dependent selection by viruses and protozoa), how often and by what mechanism are entire chromosomes mobilized, and what are the barriers to gene flow between sympatric ecological populations (e.g., reduced encounter rates or some form of assortative mating)? No matter how marked the decline in gene flow between ecological populations, they will always remain open to uptake of DNA from other populations, thus remaining fundamentally different from biological species of sexual eukaryotes (*2*). Yet notably, the process of ecological differentiation that we have inferred for these ocean bacteria is similar to that in models of sympatric speciation by habitat-specific allelic sweeps in sexual eukaryotes (*25*, *26*). Despite differences in how adaptive alleles are acquired, our results suggest that how they spread within populations may follow a more uniform process in both prokaryotes and eukaryotes than previously imagined.

References and Notes

1. M. L. Coleman, S. W. Chisholm, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 18634 (2010).
2. R. T. Papke *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **104**, 14092 (2007).
3. D. S. Guttman, D. E. Dykhuizen, *Genetics* **138**, 993 (1994).
4. V. J. Denef *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 2383 (2010).
5. B. J. Shapiro, L. A. David, J. Friedman, E. J. Alm, *Trends Microbiol.* **17**, 196 (2009).
6. F. M. Cohan, E. B. Perry, *Curr. Biol.* **17**, R373 (2007).
7. A. Koeppel *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 2504 (2008).
8. D. E. Hunt *et al.*, *Science* **320**, 1081 (2008).
9. S. P. Preheim, S. Timberlake, M. F. Polz, *Appl. Environ. Microbiol.* **77**, 7195 (2011).
10. Materials and methods are available as supplementary materials on *Science* Online.
11. P. C. Sabeti *et al.*, *Science* **312**, 1614 (2006).
12. N. Bisharat *et al.*, *Emerg. Infect. Dis.* **11**, 30 (2005).
13. K. L. Visick, *Mol. Microbiol.* **74**, 782 (2009).
14. K. J. F. Satchell, *Annu. Rev. Microbiol.* **65**, 71 (2011).
15. T. King, A. Ishihama, A. Kori, T. Ferenci, *J. Bacteriol.* **186**, 5614 (2004).
16. K. L. Meibom *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 2524 (2004).
17. D. A. Chiavelli, J. W. Marsh, R. K. Taylor, *Appl. Environ. Microbiol.* **67**, 3220 (2001).
18. J. Majewski, *FEMS Microbiol. Lett.* **199**, 161 (2001).
19. D. Falush *et al.*, *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **361**, 2045 (2006).
20. C. Fraser, W. P. Hanage, B. G. Spratt, *Science* **315**, 476 (2007).
21. J. M. Eppley, G. W. Tyson, W. M. Getz, J. F. Banfield, *Genetics* **177**, 407 (2007).
22. V. J. Denef, R. S. Mueller, J. F. Banfield, *ISME J.* **4**, 599 (2010).
23. D. E. Dykhuizen, L. Green, *J. Bacteriol.* **173**, 7257 (1991).
24. J. G. Lawrence, *Theor. Popul. Biol.* **61**, 449 (2002).
25. T. Turner, M. Hahn, S. Nuzhdin, *PLoS Biol.* **3**, e285 (2005).
26. D. E. Neafsey *et al.*, *Science* **330**, 514 (2010).

Acknowledgments: We thank E. DeLong, S. W. Chisholm, J. Wakeley, P. Sabeti, W. Hanage, D. Neafsey, and M. Coleman for valuable suggestions and comments, and X. Didelot and P. Marttinen for help with software. Funding for this work was provided by NSF grant DEB-0918333 (to E.J.A. and M.F.P.), the NSF-supported Woods Hole Center for Oceans and Human Health (COOH), and grants from the Gordon and Betty Moore Foundation and the Department of Energy Genomes to Life program (M.F.P.). Funding for genome sequencing was provided by the Moore Foundation and the Broad Institute's SPARC program. Computational resources were provided by NSF grant 0821391. Support was provided by a Canada Graduate Scholarship from the Natural Sciences and Engineering Research Council of Canada and a postdoctoral fellowship from the Harvard MIDAS Center for Communicable Disease Dynamics (B.J.S.); a Merck-MIT fellowship (J.F.); the Netherlands Organisation for Scientific Research (O.X.C.); and the Rosztoczy Foundation (G.S.). Whole genomes sequences have been deposited at the DNA Data Bank of Japan, European Molecular Biology Laboratory, and GenBank under accessions AHTI00000000, AICZ00000000, and AIDA00000000-AIDS00000000 (table S6).

Supplementary Materials

www.sciencemag.org/cgi/content/full/336/6077/48/DC1
Materials and Methods
Figs. S1 to S14
Tables S1 to S6
References (*27*–*59*)

20 December 2011; accepted 2 March 2012
10.1126/science.1218198