January 2008

Identifying the fundamental units of bacterial diversity: A paradigm shift to incorporate ecology into bacterial systematics

A. F. Koeppel
*Wesleyan University*

E. B. Perry
*Wesleyan University*

J. Sikorski
*DSMZ*

A. Warner
*Wesleyan University*

D. M. Ward
*Montana State University - Bozeman*

*See next page for additional authors*

Follow this and additional works at: [https://wesscholar.wesleyan.edu/div3facpubs](https://wesscholar.wesleyan.edu/div3facpubs)

Part of the [Biodiversity Commons](https://wesscholar.wesleyan.edu/77199), [Ecology and Evolutionary Biology Commons](https://wesscholar.wesleyan.edu/77199), and the [Environmental Microbiology and Microbial Ecology Commons](https://wesscholar.wesleyan.edu/77199)

**Recommended Citation**

[https://wesscholar.wesleyan.edu/div3facpubs/198](https://wesscholar.wesleyan.edu/div3facpubs/198)

This Article is brought to you for free and open access by the Natural Sciences and Mathematics at WesScholar. It has been accepted for inclusion in Division III Faculty Publications by an authorized administrator of WesScholar. For more information, please contact nmealey@wesleyan.edu, jmlozanowski@wesleyan.edu.
Authors

This article is available at WesScholar: https://wesscholar.wesleyan.edu/div3facpubs/198
Identifying the fundamental units of bacterial diversity: A paradigm shift to incorporate ecology into bacterial systematics


Departments of *Biology and ‡Mathematics and Computer Science, Wesleyan University, Middletown, CT 06459; †Institute of Evolution, International Graduate Center of Evolution, University of Haifa, Haifa, Israel 31905; †Deutsche Sammlung von Mikroorganismen und Zellkulturen, GmbH, Mascheroder Weg 1 b, D-38124 Braunschweig, Germany; ‡Department of Land Resources and Environmental Sciences, Montana State University, Bozeman, MT 59717; †National Center for Agricultural Utilization Research, United States Department of Agriculture, Peoria, IL 61604; and **Infectious Diseases Laboratories, Institute of Medical and Veterinary Science, Frome Road, P.O. Box 14, Rundle Mall, Adelaide, South Australia 5000, Australia

Contributed by Evtair Nevo, December 28, 2007 (sent for review November 30, 2007)

The central questions of bacterial ecology and evolution require a method to consistently demarcate, from the vast and diverse set of bacterial cells within a natural community, the groups playing ecologically distinct roles (ecotypes). Because of a lack of theory-based guidelines, current methods in bacterial systematics fail to divide the bacterial domain of life into meaningful units of ecology and evolution. We introduce a sequence-based approach (“ecotype simulation”) to model the evolutionary dynamics of bacterial populations and to identify ecotypes within a natural community, focusing here on two Bacillus clades surveyed from the “Evolution Canyons” of Israel. This approach has identified multiple ecotypes within traditional species, with each predicted to be an ecologically distinct lineage; many such ecotypes were confirmed to be ecologically distinct, with specialization to different canyon slopes with different solar exposures. Ecotype simulation provides a long-needed natural foundation for microbial ecology and systematics.

Bacillus | Evolution Canyon | ecotype | periodic selection | species concept

To fully understand any community’s ecology, we need to identify its ecologically distinct populations and to determine their mutual interactions, because these are the units that contribute uniquely to community assembly, function, and dynamics (1). In the case of a bacterial community, identifying the ecologically distinct members is a particularly formidable task. This is due to the enormous number of bacterial species and ecological roles played within a typical community (2), our inability to cultivate more than a small fraction of these species for study within the laboratory (3), and our inability to predict the genes responsible for ecological divergence, owing to the role of horizontal genetic transfer in bacterial adaptation (4). A DNA sequence-based approach can help overcome these challenges, because the bacteria falling into sequence clusters for a given gene can correspond to ecologically distinct populations, even for genes not related to the adaptive divergence between populations (5). However, the ecological interpretation of a sequence-based phylogeny is not straightforward. Any phylogeny contains a hierarchy of subclusters within clusters, and it is generally not clear which level of sequence cluster corresponds to ecologically distinct populations. Also, a sequence-based phylogeny is complicated when factors, such as geographical isolation, genetic drift, plasmid gain and loss, or rapid speciation, result in a failure of correspondence between sequence divergence and ecological divergence (4, 6). Currently, bacterial systematics employs universal thresholds of molecular divergence values to help demarcate species (7–10), but there is no theoretical basis for identifying the thresholds that yield ecologically distinct populations, nor is there evidence to suggest that a single sequence-identity cut-off value appropriately demarcates the fundamental units of bacterial ecology and evolution (4, 6, 11). Indeed, the traditional approaches of bacterial systematics have led to species that are enormously diverse in their genome content, physiology, and ecology (4, 12).

Here, we propose and test a conceptual framework, based on the evolutionary dynamics of bacterial populations, to estimate the number of ecologically distinct populations within a given clade (a group of organisms sharing a common ancestor) and to identify the members of each such population. We present an algorithm, “ecotype simulation” (ES) (4, 13, 14), which models an ecotype as an ecologically distinct group whose diversity is limited by a force of cohesion, usually the genome-wide purging of diversity known as periodic selection but also genetic drift (4). Periodic selection occurs when a new adaptive mutant arises within an asexual or rarely sexual ecotype, and natural selection causes the mutant and its nearly clonal descendants to replace all competing variants within the ecotype (4, 15) [see supporting information (SI)]. A new ecotype is founded when an adaptive mutation (or a recombination event) allows a variant to invade a new ecological niche. Owing to ecological differences between ecotypes, a periodic selection event within one ecotype does not extinguish the diversity within other ecotypes (4). So defined, bacterial ecotypes have the quintessential properties of species recognized by many systematists outside of microbiology: They are ecologically distinct groups belonging to genetically cohesive and irreversibly separate evolutionary lineages, and they are each invented only once (4, 6, 16).

We have applied ES to examine two clades whose ecological diversity and habitats have been intensively studied. We surveyed multilocus diversity among strains of Bacillus simplex and the Bacillus subtilis–Bacillus licheniformis clade isolated from the “Evolution Canyons” of Israel, which are arid, east–west-running canyons providing three major habitat zones—north- and south-facing slopes with extremely different levels of insolation and extremes of hydration


The authors declare no conflict of interest.

Ffreely available online through the PNAS open access option.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. EU305743–EU306135, EU304829–EU304976, EF026354–EF026744, and EF015305–EF015395).

*To whom correspondence may be addressed. E-mail: nevo@research.haifa.ac.il or fcohan@wesleyan.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0712205105/DC1.

© 2008 by The National Academy of Sciences of the USA

2504–2509 | PNAS | February 19, 2008 | vol. 105 | no. 7

www.pnas.org/cgi/doi/10.1073/pnas.0712205105
...Clade Sequence Diversity. We surveyed sequence diversity at three protein-coding genes within \( B. \ simplex \) isolated from Evolution Canyons I and II near Haifa, Israel, and within the \( B. \ subtilis \)-\( B. \ licheniformis \) clade from Evolution Canyon III in the Negev Desert. We graphically characterized the sequence diversity for a clade by plotting the number of bins (or sequence clusters) required to encompass the sample of sequences from the clade, as increasingly stringent sequence identity criteria were used to define these clusters (21, 22) (Fig. 1). Complete linkage clustering (23) was used to bin the sequences into clusters with different levels of minimum pairwise identity (Fig. 1). The “clade sequence diversity” is the pattern of the number of bins over different sequence identity criteria for binning (Fig. 1). Because sequence divergence increases with the time since divergence, this pattern is a way of describing the evolutionary history of cladogenesis (splitting of lineages) within a clade (22).

As typically seen in clade sequence diversity patterns (21, 24), there is an approximately log-linear increase in the number of bins with increasing sequence identity, which has been interpreted as revealing a constant net rate of ecotype formation (22), and there is a flair of increased diversity at an inflection point (in this case \( \approx 98\% \) identity), previously interpreted as reflecting the ephemeral sequence divergence within popula-

![Fig. 1. Observed and modeled clade sequence diversity patterns. Sequences for a gene (or a concatenation of genes) were binned by complete linkage clustering (23) into clusters with different levels of minimum pairwise identity. ES analysis of each clade yielded two solutions, one with low drift and one with high drift. The high-drift solutions require population sizes that are unrealistically low for these taxa (SI), so the model curves are based on the low-drift solutions. For taxa with low population sizes and/or extremely high recombination rates, high-drift solutions (with little or no periodic selection) may be the most appropriate. The individual points for each model are means based on 1,000 replications of the low-drift solution. (A) Diversity among 116 \( B. \ simplex \) isolates from Evolution Canyons I and II based on a concatenation of \( \text{gapA}, \ rpoB, \) and \( \text{uvrA} \), with recombinant organisms removed. (B) Diversity among 73 isolates within the \( B. \ licheniformis – B. \ subtilis \) clade, primarily from Evolution Canyon III, as based on a concatenation of \( \text{gapA}, \ \text{gyrA}, \) and \( \text{rpoB}, \) with recombinant organisms removed.

![Fig. 2. The ecotype simulation algorithm. The algorithm simulates the evolutionary history of the \( \sigma \) organisms sampled from nature, under different quartets of values for the net rate of ecotype formation (EF), the rates of periodic selection (PS) and drift (D), and the number of ecotypes in the sample. In the coalescence approach taken (36), the algorithm considers only the lineages that are directly ancestral to the \( \sigma \) sampled organisms (represented by black circles). These focal lineages are represented by solid lines; the many contemporary lineages not sampled from each ecotype are indicated by light dashed lines (\( E_1, \ E_2, \) and \( E_3 \)); the lineages extinguished by past PS and D are represented by bold short-dashed lines and long-dashed lines, respectively, with each extinction represented by a square. The program begins with a “backward” simulation that stochastically produces a phylogenetic representation of the history of the community, establishing nodes of coalescence of lineages (due to PS, EF, or D; indicated by gray circles) and time between nodes (\( t_1, t_2, \) etc.); this phylogeny is then taken as a scaffold for the forward simulation. The purpose of the forward simulation is to produce mutual nucleotide substitutions throughout the history of the clade, according to the phylogenetic scaffold. To begin a simulation, a set of \( \sigma \) contemporary organisms (representing the \( \sigma \) organisms sampled from nature) are distributed randomly (according to the canonical lognormal distribution) among \( n \) ecotypes (here, \( \sigma \approx 14 \) and \( n = 3 \)). Working backward from the \( \sigma \) organisms in the present, the processes of EF, PS, and D occur stochastically in time according to their respective rates (\( \Omega, \alpha, \) and \( d \)). For each such event, one or more lineages coalesce into a single ancestral lineage, as described in SI. Note that in the backward-looking view of the coalescence formulation, each PS appears as a coalescence event, in which all lineages after the PS coalesce into the survivor of the PS event. Likewise, each D event appears in this backward-looking view as the coalescence of a pair of lineages within an ecotype (e.g., two contemporary lineages coalesce into lineage \( D_1 \) to reflect the increased representation of lineage \( D_1 \) after the random loss by drift of lineage \( D_2 \)). Because \( \Omega \) is the net rate of EF events, taking into account extinction, we include in the simulation only those EF events resulting in ecotypes that survive into our contemporary sample. The backward phase of the simulation ends when all of the branches have coalesced into a single node; this represents the most recent common ancestor of all of the sampled organisms. Then the forward simulation begins when a sequence (of the same length as the observed sequence data) is assigned to this most recent common ancestor. Nucleotide substitutions then occur stochastically, going forward in time, between each pair of nodes in the phylogeny derived from the backward simulation, according to the time between the events determining the nodes. This generates a matrix of pairwise sequence divergence between all \( \sigma \) contemporary organisms for a simulation replicate, from which a clade sequence diversity curve is calculated; the simulated clade sequence diversity curve is then compared against the observed clade sequence diversity curve (see SI).
number of ecotypes, \( n \), in the sample of sequences. The ES
analysis evaluates different quartets of parameter values for their
likelihood of yielding an evolutionary history consistent with the
observed clade sequence diversity (Fig. 1). Thus, ES quantifies the
ecological diversity within a community (as the number of
ecotypes sampled, \( n \)) by analyzing the community’s evolutionary
history. ES also estimates the rates of net ecotype formation and
periodic selection, allowing for future tests of how a clade’s
ecological and life history characteristics might determine its
evolutionary rates (see Fig. 2 and SI). The ecotype simulation
software is available at http://fcohan.web.wesleyan.edu/ecosim.

The ES of the history of the concatenated gene sequence in
*B. simplex* estimated the presence of 13 putative ecotypes within
this named species in Evolution Canyons I and II (Table 1). The
ES of the *B. subtilis–B. licheniformis* clade estimated 17 putative
ecotypes. The more rapidly evolving gene *gyrA* estimated some-
what more ecotypes than the other genes in the *B. subtilis–B.
licheniformis* clade. Also, the most rapidly evolving gene among
the three used for *B. simplex*, *rpoB*, estimated considerably more
ecotypes than the other genes, but the estimate based on the
concatenation was not affected greatly by the outlier gene.

We next extended the ES approach to identify the individual
ecotypes within each clade, with the ultimate aim of testing
whether each putative ecotype actually corresponds to an
ecologically distinct population. Our approach for ecotype
demarcation was to find the most inclusive clades that are each
consistent with being a single ecotype, as explained in SI. This
is a conservative approach that tends to yield fewer demarcated
ecotypes than indicated by the parameter estimates of Table 1
(see SI). Accordingly, nine and 13 putative ecotypes were
identified in the *B. simplex* and *B. subtilis–B. licheniformis*
clades’ concatenated gene sets, respectively (Fig. 3). For *B. simplex*,
the individual genes yielded the same demarcations as the concate-
nation, except for several cases in which *rpoB* split a putative
ecotype into two or more ecotypes; in one case, *gapA* split a
putative ecotype into two. Likewise, the individual gene analyses
of the *B. subtilis–B. licheniformis* clade gave the same demara-
cations as the concatenation, except in two cases where putative
ecotypes were split by *gyrA*. In general, analyses using rapidly
evolving genes were more likely to discern very closely related
ecotypes than were analyses using more slowly evolving genes. Future analyses that involve sequences of
many genes, perhaps even whole genomes, should focus on a
subset of genes with a history of rapid evolution and infrequent recombination (25).

In cases where a strain had been eliminated from the concate-
nation analysis because of recombination (Fig. 3), the strain was
classified into an ecotype based on single-gene ES analysis of the
two genes that did not recombine.

### Table 1. Estimates of parameter values for the low-drift solutions of each clade with 95% confidence intervals in parentheses

<table>
<thead>
<tr>
<th>Clade</th>
<th>Gene</th>
<th>Net rate of ecotype formation (( \Omega ))</th>
<th>Rate of periodic selection (( \omega ))</th>
<th>Rate of drift (( d ))</th>
<th>Ecotypes, no. (( n ))</th>
<th>Ratio of ( r : n )</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>B. simplex</em></td>
<td>Concat.</td>
<td>0.084 (0.026, 0.18)</td>
<td>0.57901 (0.083, 4.19)</td>
<td>0 (0, 0.70)</td>
<td>13 (5, 28)</td>
<td>6.89</td>
</tr>
<tr>
<td></td>
<td><em>gapA</em></td>
<td>0.20 (0.042, 0.44)</td>
<td>3.56 (0.24, ∞)</td>
<td>0 (0, ∞)</td>
<td>11 (3, 40)</td>
<td>17.8</td>
</tr>
<tr>
<td></td>
<td><em>rpoB</em></td>
<td>0.43 (0.19, 0.93)</td>
<td>54.16 (1.17, ∞)</td>
<td>0 (0, ∞)</td>
<td>34 (9, 79)</td>
<td>126</td>
</tr>
<tr>
<td><em>B. subtilis–B. licheniformis</em></td>
<td>Concat.</td>
<td>0.028 (0.013, 0.041)</td>
<td>0.92 (0.14, 4.55)</td>
<td>0 (0, 0.70)</td>
<td>17 (9, 27)</td>
<td>32.9</td>
</tr>
<tr>
<td></td>
<td><em>gapA</em></td>
<td>0.083 (0.038, 0.18)</td>
<td>2.10 (0.44, 31.33)</td>
<td>0 (0, 5.12)</td>
<td>10 (6, 20)</td>
<td>25.3</td>
</tr>
<tr>
<td></td>
<td><em>gyrA</em></td>
<td>0.049 (0.036, 0.073)</td>
<td>2.54 (0.40, 27.66)</td>
<td>0 (0, 1.43)</td>
<td>20 (15, 30)</td>
<td>51.8</td>
</tr>
<tr>
<td></td>
<td><em>rpoB</em></td>
<td>0.10 (0.049, 0.16)</td>
<td>8.47 (1.34, ∞)</td>
<td>0 (0, 39.68)</td>
<td>12 (8, 19)</td>
<td>84.7</td>
</tr>
</tbody>
</table>

See SI for high-drift solutions. The confidence intervals were determined following the method of Felsenstein (see SI) and are indicated in parentheses. The number of ecotypes estimated here tends to be greater than the number of ecotypes demarcated in Fig. 3, because ecotype demarcation is performed conservatively—a clade is deemed an ecotype if the confidence interval of \( n \) includes 1, even if the maximum likelihood estimate for \( n \) is greater than 1 (SI). Concat., concatenation of three genes.

**Ecological Distinctness of Putative Ecotypes.** We next investigated whether the ecotypes hypothesized by ES are ecologically distinc-
t by making use of the topography of the Evolution Canyons.
As part of the Evolution Canyon paradigm developed by E. Neu-
and colleagues (18), the canyons have three major habitats, a
south-facing slope (SFS) with high solar insolation (and other
covarying physical and chemical parameters, and/or interactions
with other organisms adapted to this habitat), a north-facing
slope (NFS) with less insolation, and a canyon bottom with greater
access to water (18, 26).

In *B. simplex*, the nine putative ecotypes we identified were
significantly heterogeneous in their associations with the two
major habitats from which *B. simplex* was isolated (\( \chi^2 = 82.16; 
P < 0.0001; 8 \text{ df} \) (Fig. 3A). The clade at the top of Fig. 3A contains two putative ecotypes, one found primarily on the SFS
[putative ecotype 1 (PE1)] and the other on the NFS (PE2). Note also that PE3 and PE4 represent well supported clades found
nearly entirely on the NFS. The clade containing PE5–PE9
contains what appear to be specialists to the SFS (PE5 and PE9),
one ecotype specialized to the NFS (PE6), and one ecotype
(PE7) that is abundant on both slopes but in higher frequency on
the NFS, with physiological adaptation to the SFS (20).

The ecological distinctness of the *B. simplex* ecotypes is further
confirmed by their habitat-related physiological properties. The
ecotypes associated with the hotter SFSs have greater growth rates
at a stressful high temperature than do the ecotypes associated with
NFS’s, but the differences disappear at optimal temperature (20)
(see SI). Also, the SFS-associated ecotypes constitutively produce
greater amounts of isomethyl-branched fatty acids, which are beneficial for heat tolerance, than NFS-associated
ecotypes (J.S., E.B., R. Kroppenstedt, and B. Tindall,
unpublished data). Resistance to UV-C radiation does not
appear to contribute to SFS adaptation (19). Other aspects of
ecotype divergence yet to be explored in this system are differences
in nutrient resources and interactions with other micro-
organisms (27).

Within the *B. subtilis–B. licheniformis* clade, the 13 hypothe-
sized ecotypes are heterogeneous in their associations with
the major habitats (\( \chi^2 = 84.25; P < 0.0001; 24 \text{ df} \) (Fig. 3B).
PE1 appears to be specialized to the SFS; PE2 and PE9 appear to
be specialized to the canyon bottom; PE5 (which includes the type
strain of *B. licheniformis*) appears to be specialized to the canyon bottom and NFS; and PE8 (which includes the type strain of *B.
atrophaeus*) appears to be specialized to the NFS. As seen also in
SFS-adapted ecotypes in *B. simplex*, the SFS-adapted PE1 of the
*B. subtilis–B. licheniformis* clade has heat-adapting isomethyl-
branched fatty acids in higher abundance than closely related
ecotypes not specialized to SFS (J.S., E.B.P., and F.M.C.,
unpublished data).

We do not currently have sufficient ecological data to discern
the specific factors that may contribute to the ecological distinctness of every putative ecotype. We anticipate that in future applications of ES, microbiologists will confirm the ecological distinctness of putative ecotypes through microhabitat distribution studies and comparisons of genome content and analyses of genome-wide gene expression and comprehensive metabolic phenotype. In general, we expect closely related ecotypes to differ in their adaptations to a great diversity of ecological dimensions, including different temperatures (14), photic zones (14), sources of inorganic nutrients (28), carbon sources used (29), and host organisms (30).

**Resolution of Ecotype Simulation.** Except in cases of compelling phenotypic differentiation (usually related to human disease), bacterial systematists have frequently used 16S rRNA gene divergence as a guide to species demarcation, in which divergence scores >3% between strains denote that they should be recognized as different species; recently, this value has been changed to 1% divergence (7). Although this guideline does not require clades <1% divergent to be subsumed within a single species, it provides no encouragement for systematists to discover ecologically significant diversity among such clades. We note that the 1% guideline would group together within a single species many of the most closely related ecotypes identified by ES, many of which have been confirmed to be ecologically distinct. This is particularly striking in the case of *B. simplex*—here, the 131 isolates comprising the nine ecotypes identified by ES (Fig. 3A) are absolutely identical in their 16S rRNA se-
quences (~1350 nt were sequenced) and are thus invisible to a 1% 16S rRNA divergence criterion for species demarcation. The putative ecotypes identified within the *B. subtilis–B. licheniformis* clade correspond more closely to the named species, except that four ecotypes (PE2–PE5) are within 0.72% 16S rRNA divergence of the *B. licheniformis* type strain and would most probably be recognized by bacterial systematics as members of that species (Fig. 3B). Also, PE11 and PE12 are within 0.40% divergence of *B. spizizenii* and would fit within that species. One newly discovered ecotype (PE1) is just outside the 1% divergence guideline from *B. licheniformis*. Thus, ES has discovered ecological diversity that would probably have been ignored within the current systematics of bacteria (as well as one ecotype that would have been recognized).

This conclusion is supported also by results from preliminary versions of ES, in which 11 putative ecotypes were discovered within the species *Legionella pneumophila*; some ecotypes were confirmed as distinct in their host ranges and in their gene expression patterns during infection (13). Similarly, two ecotypes were identified within subclade A of hot spring *Synechococcus*, one of which was associated with a specific temperature (65°C); also, two ecotypes were identified within subclade A′, each of which was associated with a different depth in the photic zone (14). Within each of the *Synechococcus* subclades, sequences were within 0.71% divergence in 16S rRNA sequence, and so the current framework of systematics would most probably not distinguish these ecotypes.

Thus, ES promises to be an effective way to discover ecological diversity. Many of the ecotypes hypothesized by ES and confirmed as ecologically distinct would fit within the species demarcated by traditional bacterial systematics approaches. ES has an important advantage over current methods in bacterial systematics in that it does not employ a universal threshold of molecular divergence that is arbitrarily defined and subjectively applied. In ES, analysis of the evolutionary history of a particular clade yields the appropriate criteria for demarcating ecotypes of that clade. The result is that ES can identify ecotypes that are not discerned by our current framework for bacterial systematics (Fig. 3 and Table 1).

That ecotypes can be found lumped within established species suggests that bacterial systematics is failing in its fundamental mission—to precisely provide the ecological properties of any organism that is classified to species (31). The ES approach promises to rectify this by offering a general theory-based approach for identifying a multiplicity of ecotypes per taxon, even before the ecological differences among putative ecotypes can be confirmed.

**Incorporating Ecology into Bacterial Systematics.** We propose a paradigm by which bacterial systematics may use ES to demarcate ecotypes, while taking into account a potential diversity of evolutionary models. The ES approach is most likely to reveal ecotypes under a “stable ecotype” model in which new ecotypes are formed only rarely and each ecotype endures many periodic selection events during its lifetime. Under these circumstances, there is time for accretion of sequence divergence between ecotypes, with recurrent purging of diversity within but not between ecotypes. The sequence clustering we observe in these systems is thus dominated by periodic selection, yielding a close correspondence between ecotypes and sequence clusters (4, 6). We note that, for each clade, ES has estimated the rate of periodic selection to be much greater than the net rate of ecotype formation—the condition promoting correspondence of ecotypes and sequence clusters (Table 1).

However, bacterial systematics must take into account that factors other than periodic selection may contribute to sequence clustering in certain lineages (4). To accommodate these additional factors, we suggest that long-standing ecotypes be demarcated as the smallest clades that show (i) a history of coexistence as separate ecologically distinct lineages, as inferred from ES (or an equivalent sequence-based approach) and supported as monophyletic groups by bootstrap or similar analysis, and (ii) a prognosis for future coexistence, as inferred from the ecological distinctness of the groups in nature (4). Ecotypes cannot be inferred by sequence clustering alone, because it is possible for one ecotype to fall into multiple sequence clusters when geography and genetic drift have been major factors in the history of the lineage (4) (see SI). We note also that showing ecological distinctness alone, even in nature, is not sufficient to infer that the groups will coexist as separate lineages, given the possibility of recurrent, plasmid-based evolution into a particular ecological niche (see SI). Also, ecological differences inferred from laboratory tests may not be relevant to coexistence in nature. Identification of ecotypes therefore requires both a sequence-based approach to formulate hypotheses about putative ecotypes and an ecological approach to confirm these hypotheses.

Ecotypes that are confirmed to have a history of coexistence as distinct lineages and a prognosis of future coexistence are the fundamental units of bacterial ecology and evolution (4, 6, 11). We recommend that they be recognized also as the fundamental units of diversity in bacterial systematics (32). We suggest that, when multiple ecotypes are discovered within the accepted phylogenetic range of an established species (e.g., with 1% divergence in 16S rRNA), the ecotypes should be recognized and named by adding an “ecovar” epithet to the species binomial, for example, by naming the ecotypes within *B. simplex*. When an ecotype is discovered that is outside the phylogenetic range of any established species, we propose that the ecotype should be given a new species name, for example, by naming PE1 of the *B. subtilis–B. licheniformis* clade as a separate species. This dual approach should enrich bacterial systematics with ecologically significant but previously ignored groups, while respecting the stability of taxon names.

By identifying taxa at the level of ecotypes, bacterial systematics will provide a long-needed, biologically meaningful, taxonomic grouping of microbial diversity, to the benefit of other biological disciplines such as ecology, evolution, biochemistry, genomics, epidemiology, and biotechnology (4). The identification of these groups will be a critical step forward in our venture to understand the myriad ecological interactions within a natural microbial community.

**Materials and Methods**

**Diversity Within *B. simplex***. The details of soil collection and isolation of *B. simplex* from Evolution Canyons I and II, each with a south-facing slope and a north-facing slope, are described in ref. 19. PCR and sequencing of the *gapA*, *rpoB*, *uvrA*, and 16S rRNA genes (GenBank accession nos. EU305743–EU306135) were as described in refs. 19, 20, and 33.

**Diversity Within the *B. subtilis–B. licheniformis* Clade.** Evolution Canyon III is located in the southern Negev desert, at Nahal Shaharut, a tributary of Nahal Hiyyon (lat 29°55’S, long 34°58’E); EC III has an SFS (~35° rise) and an NFS (~30° rise), separated by ~150 m at the bottom (26). The soil (from the top 1- to 3-cm layer, taken on March 25, 2003) was collected from three elevation stations each from the SFS and NFS habitats and from one collecting station at the canyon bottom, with three collecting sites per station. Strains from the *B. subtilis–B. licheniformis* clade were identified by metabolic tests (34) and confirmed by sequences of *rpoB*. Additional strains and species from this clade were obtained from National Center for Agricultural Utilization Research of the U.S. Department of Agriculture.

DNA extraction, PCR, and sequencing were as described in SI. The *gapA*, *gyrA*, *rpoB*, and 16S rRNA sequences are accessible as GenBank numbers EU304829–EU304903, EF026654–EF026744, EF015305–EF015395, and EU304904–EU304976, respectively.

**Preparation of Sequences for Ecotype Simulation.** Alignment positions with gaps or indeterminate nucleotides in any sequence were removed. We compensated for PCR and sequencing error by “correcting” a random subset of the
singleton nucleotides (i.e., occurring at a particular nucleotide position in only one sequence) equal in number to the expected number of PCR and sequencing errors (see SI). Each chosen singleton nucleotide was corrected to the singleton nucleotides (i.e., occurring at a particular nucleotide position in only one sequence) equal in number to the expected number of PCR and sequencing errors (see SI).

Accommodation for Recombination. The ES algorithm takes recombination into account, allowing that periodic selection may not be significant in some taxa with high recombination rates. ES allows for periodic selection and/or genetic drift, to constrain within-ecotype sequence diversity.

The ES algorithm does not take into account that recombination between ecotypes can introduce sequence diversity into an ecotype at a gene locus being surveyed. Therefore, we identified strains that underwent recombination and the particular gene involved in each recombination event, using the “majority rules” rationale (35): In cases where a strain had changed its phylogeny and the particular gene involved in each recombination event, using the taxa with high recombination rates. ES allows for periodic selection and/or genetic drift, to constrain within-ecotype sequence diversity.


ACKNOWLEDGMENTS. We thank Mary Bateson for suggesting improvements in earlier manuscripts. This work was funded by National Science Foundation Frontiers in Biological Research Program Award EF-0328698 (to D.M.W. and F.C.) and the National Aeronautics and Space Administration Exobiology Program Award NAG5–8824 (to D.M.W. and F.M.C.), a grant from the Ancell–Teicher Research Foundation for Genetics and Molecular Evolution (to E.N.), Deutsche Forschungsgemeinschaft Grant 1352/1–1 (to J.S.), and research grants from Wesleyan University (to D.K. and F.M.C.).