Many bacteria are naturally competent, able to actively transport environmental DNA fragments across their cell envelope and into their cytoplasm. Because incoming DNA fragments can recombine with and replace homologous segments of the chromosome, competence provides cells with a potent mechanism of horizontal gene transfer as well as access to the nutrients in extracellular DNA. This review starts with an introductory overview of competence and continues with a detailed consideration of the DNA uptake specificity of competent proteobacteria in the Pasteurellaceae and Neisseriaceae. Species in these distantly related families exhibit strong preferences for genomic DNA from close relatives, a self-specificity arising from the combined effects of biases in the uptake machinery and genomic overrepresentation of the sequences this machinery prefers. Other competent species tested lack obvious uptake bias or uptake sequences, suggesting that strong convergent evolutionary forces have acted on these two families. Recent results show that uptake sequences have multiple “dialects,” with clades within each family preferring distinct sequence variants and having corresponding variants enriched in their genomes. Although the genomic consensus uptake sequences are 12 and 29 to 34 bp, uptake assays have found that only central cores of 3 to 4 bp, conserved across dialects, are crucial for uptake. The other bases, which differ between dialects, make weaker individual contributions but have important cooperative interactions. Together, these results make predictions about the mechanism of DNA uptake across the outer membrane, supporting a model for the evolutionary accumulation and stability of uptake sequences and suggesting that uptake biases may be more widespread than currently thought.

**THE MECHANISM AND FUNCTION OF NATURAL COMPETENCE**

How is DNA transported into the cell? Gram-negative bacteria take up DNA in two stages, customarily referred to as DNA uptake (across the outer membrane) and DNA translocation (across the inner membrane) (3). After DNA is bound to the cell (step A in Fig. 1), uptake occurs by retraction of cell surface fibers of the type IV pilus family (T4P), which pulls the DNA into the periplasm through secretin pores in the outer membrane (step B). In some species, the same pilus proteins form the long pili used for adhesion and twitching motility (7), but in others, the fibers are inferred to occur only as short competence-specific pseudopili that do not protrude detectably beyond the cell surface, like those used in type II secretion (8, 9). Gram-positive bacteria use similar T4P-related proteins to pull double-stranded DNA across their thick cell walls (1, 10).

Although the main components of the uptake machinery have been identified in multiple species, little is known about how they interact with DNA, and the detailed mechanics of DNA uptake are not well understood in any species (7). DNA uptake presents a particular challenge to Gram-negative bacteria, which must transport stiff and highly charged double-stranded DNA (dsDNA) across their outer membranes. Although uptake could in principle be initiated by threading one dsDNA end through the secretin pore, alongside or at the tip of the pilus/pseudopilus, this has not been demonstrated in any system. Instead, initiation likely occurs at internal sites on DNA fragments, as shown in Haemophilus influenzae, where closed circular plasmids are taken up into the periplasm as efficiently as linear DNAs (11). Although the power source for retraction of T4P is usually thought to be ATP-powered disassembly of the pilin subunits by the motor protein PilT (12), several competent species lack pilT homologs (13). The ability of cells to take up DNA fragments of >50 kb, much longer than the cell (14–16), creates another difficulty, since pulling these across the very narrow periplasmic space by pseudopilus retraction...
would require many cycles of pseudopilus elongation, DNA attachment, and retraction.

Once the DNA is in the periplasm (in Gram-negative bacteria) or at the cytoplasmic membrane (in Gram-positive bacteria), one strand is translocated across the membrane into the cytoplasm, with its 3′ end leading (step C). Although Gram-positive bacteria use a cell surface nuclease to cut DNA into smaller fragments before this step (17), Gram-negative bacteria do not. The strand that is not translocated is degraded at the membrane surface to nucleotides, which can then be dephosphorylated and taken up by nucleoside transporters (18). In all competent species, translocation uses a conserved membrane pore encoded by rec2 or comEC. If sequence similarity permits, the translocated strand may then replace a chromosomal strand by homologous recombination (step D); otherwise, it is degraded and its nucleotide subunits are recycled.

**What are the consequences of DNA uptake?** Because DNA has both biochemical and informational properties, the effects of DNA uptake depend on the nutritional needs of the cell, the presence of DNA damage, the ability of incoming DNA to recombine with chromosomal DNA, and the effects of this recombination on fitness, as summarized by Fig. 2. The most immediate consequence is nutritional (Fig. 2A). DNA is an excellent source of the deoxyribonucleotides needed for replication of the cell’s own genome, and de novo nucleotide synthesis is expensive both in terms of energy and in terms of the molecular constituents. All cells take up preformed bases and nucleotides where possible, and soil and sediment species often secrete nucleases that allow them to use preformed bases and nucleotides where possible, and soil and sediment species often secrete nucleases that allow them to use preformed bases and nucleotides where possible. All cells take up preformed bases and nucleotides where possible, and soil and sediment species often secrete nucleases that allow them to use preformed bases and nucleotides where possible (19). However, uptake of intact DNA is a more efficient way to obtain nucleotides, both because it limits losses due to diffusion and because it avoids the need for nucleoside repurification after uptake (18). Most incoming DNA is degraded even when it is identical to that of the chromosome, and even recombined DNA reduces the cell’s need for nucleotides (20, 21).

The genetic consequences of DNA uptake are less predictable. First, they depend on the sequence of the DNA being sufficiently similar to DNA in the cell’s genome that it can replace a genomic strand by homologous recombination, catalyzed by the ubiquitous RecA protein. If this replaced segment includes a position with DNA damage (Fig. 2B), the incoming strand could provide a template for DNA repair (22). However, unless the genome is heavily damaged, most recombination occurs at undamaged positions. When the incoming DNA is from cells of the same clonal population (Fig. 2C), recombination does not change the cells’ genotypes unless the donor or recipient DNA contains newly arisen mutations. A recent study comparing wild-type and competence-deficient strains of *Streptococcus pneumoniae* found that competence reduced the fixation of new mutations, suggesting that transformation eliminates new mutations from recipients more often than it introduces them from donors (23).

Natural populations may often be mixtures of strains, and transformation then creates new and possibly beneficial combinations of variant alleles and loci (Fig. 2D to F). DNA from one of the common noncompetent strains (see below) (24) may also replace functional competence genes with nonfunctional alleles (24). To the extent that DNA comes from relatives that have died due to deleterious mutations, recombination reduces fitness more often than it increases it (24, 25), but it may also provide cells with locally beneficial alleles if they are invading an established population (26).

Transformation is not limited to simple sequence variants; large insertions and deletions transform moderately well if flanked by sequences of chromosomal homology (Fig. 2E). Homology at only one end of a heterologous segment can be sufficient to promote recombination (Fig. 2F; line thicknesses indicate probable chance of different outcomes) (27), but “illegitimate” recombination with nonhomologous DNA is extremely rare (28). Although transformation’s dependence on sequence homology makes it inefficient at introducing novel genes into a species, this is balanced by the high efficiency with which it can spread genes through populations once they have been introduced by such homology-independent processes as specialized transduction or transposition.

Factors that evolved to protect cells against genetic parasites can also limit transformation. Lin et al. (29) found that recombination tracts in *H. pylori* often terminated at restriction sites where the donor DNA was unmethylated. This is unlikely to be due to
action of recipient-specific enzymes in the cytoplasm, since incoming DNA is single-stranded and thus not a target for most restriction enzymes. However, restriction enzymes released by cell lysis could also cut donor DNAs before uptake, and these may be an important limitation on the extent of recombination tracts. Similarly, Bikard et al. (30) showed that a clustered regularly interspaced short palindromic repeat (CRISPR) element engineered to target capsule genes prevented a *S. pneumoniae* strain from acquiring these genes by transformation. However, the likelihood of CRISPR elements acquiring sequences that target functional accessory genes has not been evaluated.

**How is competence regulated?** Unlike other uptake systems regulated by substrate availability (e.g., the lac operon), no bacteria are known to use DNA availability to signal induction of the competence machinery; this may be because environmental DNA is ubiquitous in most bacterial environments, especially biofilms. Instead, in most well-studied species, competence is regulated by other environmental and biochemical cues (4). The exception is *Neisseria*, whose competence appears to be constitutive (31). In these species, genes encoding proteins needed for DNA uptake and translocation are typically coregulated with genes for cytoplasmic proteins. Some of the latter contribute to transformational recombination, but others have no obvious connection to DNA uptake or have no known function at all. Unfortunately, the frequent coregulation of other cellular functions with competence makes it difficult to confidently delineate “competence regulons.”

The simplest known regulatory system is that of *H. influenzae*, where cells respond to a lack of phosphotransferase system (PTS) sugars and purine precursors by inducing expression of 25 genes from 12 operons, under the control of the catabolite regulator cyclic AMP receptor protein (CRP) and its competence-specific cofactor Sxy (also called TioX) (32–36). The functions of the Sxy-regulated genes have been examined using knockout mutations, confirming that most play direct roles in DNA uptake (37). Similar regulons induced by Sxy and CRP are present in other *Pasteurellaceae* and in the related *Vibrionaceae* and *Enterobacteriaceae*. Work in *Vibrio sp.* has shown that competence induction also depends on quorum sensing, pyrimidine precursors, and the presence of chitin breakdown products generated from crustacean exoskeletons (38–42). Although *Escherichia coli*’s competence gene homologs enable cells to use DNA as a sole source of carbon, attempts to demonstrate natural competence in *E. coli* have not been successful (43, 44). Less is known about regulatory mechanisms in other Gram-negative species, since culture conditions that induce competence have not been linked to specific regulators (4, 45). Gram-positive competence regulation is complex, with contributions from overlapping layers of quorum sensing, nutritional signals, and other stress responses (46–49). The competence regulons are also much larger and include many genes whose relationship to competence is not evident (4, 50–52). The extreme variability of competence regulation contrasts with the strong conservation of the DNA uptake machinery and perhaps reflects various benefits of coregulating DNA uptake with other cellular responses.

**What is the function of natural competence?** Although the action of natural selection with respect to competence genes remains controversial, the immediate consequences of DNA uptake (Fig. 2) provide a framework for thinking about its possible evolutionary function. Competence and transformation have customarily been viewed as adaptations to promote homologous recombination and genetic diversification. However, although transformation can clearly have long-term evolutionary benefits, natural selection acting on individual competent cells cannot foresee these, and the immediate selective advantages of DNA uptake for the cell are less clear (26, 53–56). Most transformation events are expected to be neutral or deleterious (the latter especially if using DNA from selectively killed cells [25]), and more immediate and reliable benefits of DNA uptake arise from DNA repair and nucleotide acquisition (57). A more extreme view is that DNA uptake could also be in part an unselected consequence of the adhesion and motility activities of T4P (58), though the coordinated regulation of the T4P genes responsible for DNA uptake and the non-T4P genes responsible for translocation into the cytoplasm argues against this. In this context, competence-induced cytoplasmic proteins coregulated with T4P genes take on special importance—are they modulations of the cellular response to their immediate environment or specific adaptations to promote transformation?

**The Phylogenetic Distribution of Natural Competence**

Estimates of the distribution of natural competence come from three main approaches, each with specific limitations: (i) direct experimental assays, (ii) inferences from population genetics, and (iii) inferences from the presence of competence genes in sequenced genomes. All suggest that naturally competent species are widely distributed throughout the bacterial tree (Table 1) (5, 59).

**Direct assays.** Experimental demonstrations of natural competence are limited to only a few dozen species scattered across the

---

**TABLE 1** Distribution of competence and related traits

<table>
<thead>
<tr>
<th>Bacterial group(s)</th>
<th>Comp&lt;sup&gt;a&lt;/sup&gt;</th>
<th>r/m &gt; 1&lt;sup&gt;b&lt;/sup&gt;</th>
<th>T4P&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Rec2&lt;sup&gt;d&lt;/sup&gt;</th>
<th>DprA&lt;sup&gt;e&lt;/sup&gt;</th>
<th>Self&lt;sup&gt;f&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gammaproteobacteria</td>
<td>+</td>
<td>+/+/-</td>
<td>+</td>
<td>+</td>
<td>+/+/-</td>
<td>+/-</td>
</tr>
<tr>
<td>Betaproteobacteria</td>
<td>+</td>
<td>+/+/-</td>
<td>+</td>
<td>+</td>
<td>+/+/-</td>
<td>+/-</td>
</tr>
<tr>
<td>Alphaproteobacteria</td>
<td>+</td>
<td>+/+/-</td>
<td>-</td>
<td>+</td>
<td>+/+/-</td>
<td>+/-</td>
</tr>
<tr>
<td>Epsilonproteobacteria</td>
<td>+</td>
<td>+/+/-</td>
<td>+</td>
<td>+</td>
<td>+/+/-</td>
<td>+/-</td>
</tr>
<tr>
<td>Deltaproteobacteria</td>
<td>+</td>
<td>+/+/-</td>
<td>+</td>
<td>+</td>
<td>+/+/-</td>
<td>+/-</td>
</tr>
<tr>
<td>Bacillodes/Chlorobi</td>
<td>+</td>
<td>+/+/-</td>
<td>-</td>
<td>+</td>
<td>+/+/-</td>
<td>+/-</td>
</tr>
<tr>
<td>Spirochaetes</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+/+/-</td>
<td>+/-</td>
</tr>
<tr>
<td>Chlamydiae/Planctomycetes</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+/+/-</td>
<td>+/-</td>
</tr>
<tr>
<td>Firmicutes</td>
<td>+</td>
<td>+/+/-</td>
<td>+</td>
<td>+</td>
<td>+/+/-</td>
<td>+/-</td>
</tr>
<tr>
<td>Cyanobacteria</td>
<td>+</td>
<td>+/+/-</td>
<td>+</td>
<td>+</td>
<td>+/+/-</td>
<td>+/-</td>
</tr>
<tr>
<td>Chloroflexi</td>
<td>-</td>
<td>ND</td>
<td>-</td>
<td>+</td>
<td>+/+/-</td>
<td>+/-</td>
</tr>
<tr>
<td>Actinobacteria</td>
<td>+</td>
<td>+/+/-</td>
<td>-</td>
<td>+</td>
<td>+/+/-</td>
<td>+/-</td>
</tr>
<tr>
<td>Aquificae</td>
<td>-</td>
<td>ND</td>
<td>-</td>
<td>+</td>
<td>+/+/-</td>
<td>+/-</td>
</tr>
<tr>
<td>Thermotogae</td>
<td>+</td>
<td>+/+/-</td>
<td>-</td>
<td>+</td>
<td>+/+/-</td>
<td>+/-</td>
</tr>
<tr>
<td>Deinococcus/Thermus</td>
<td>+</td>
<td>+/+/-</td>
<td>-</td>
<td>+</td>
<td>+/+/-</td>
<td>+/-</td>
</tr>
</tbody>
</table>

<sup>a</sup> The bacterial group nomenclature follows Wu and Eisen (130).

<sup>b</sup> One or more species are known to be naturally competent (Comp) (+), or none are known to be naturally competent (–), following references 5 and 127.

<sup>c</sup> One or more species have an inferred recombination-to-mutation rate ratio (r/m) that is greater than 1 (+) or less than 1 (–), group has species with r/m <1 and r/m >1 (+/+), or species in group were not evaluated (ND) (data from reference 66).

<sup>d</sup> T4P machinery is present (+) or is not present (–) for one or more species, as reported by Pelicic (128).

<sup>e</sup> Rec2 and DprA homologs were found in one or more species of all groups by searches of the NCBI protein database.

<sup>f</sup> One or more species show self-specific DNA uptake (+) or no self-specificity (–), or no species were tested (ND) (references in text).
bacterial tree (5, 59); negative results are rarely reported (Table 1, column 2). Assays measuring genetic transformation are highly sensitive, but they can be done only in species where a selectable genetic marker is available (typically an antibiotic-resistance allele) and fail to discover competence in species where DNA uptake rarely leads to recombination. Fewer species have been directly tested for the ability to actively take up DNA; these assays typically use radiolabeled DNA and, although technically straightforward, are relatively insensitive (60).

A bigger problem for both assays is their dependence on prior induction of the postulated DNA uptake machinery; negative results are uninformative since a failure to observe uptake or transformation may simply mean that the appropriate inducing conditions have not been discovered. For example, Vibrio cholerae was not experimentally shown to be naturally competent until the inducing role of chitin breakdown products was discovered in 2005 (38). Another reason competence may often be missed is that levels of competence differ greatly even between isolates of the same species. For example, a survey of 34 H. influenzae isolates found 10^6-fold variation in the ability to be transformed, with many isolates being completely nontransformable (61, 62). Similar variation in transformability has been found in all other species that have been examined (see citations in reference 61). Thus, a species may be mistakenly thought to lack competence because only noncompetent isolates have been tested.

Inferences from population genetics. Less-direct methods can also be used to detect competence (Table 1, column 3). A number of studies have used population genetic models to estimate the historical frequency of homologous recombination from sequences of a standard set of seven housekeeping genes collected from many isolates of the same species (multilocus sequence typing [MLST]). One such study of pathogenic species found widely differing rates between taxonomic groups but consistently high rates among species known to be competent (63); this was confirmed for Neisseria species and H. influenzae by more recent studies (64, 65). A comprehensive survey of MLST data found high rates of historical recombination in many species not known to be competent (66), and in principle this might predict transformability.

Whole-genome approaches to detecting historical recombination are becoming increasingly practical as DNA sequencing costs decrease, and these reveal patterns of past recombination events that are not observable by MLST genotyping (67–69). However, these population-based approaches come with several caveats: a high historical recombination frequency may be due to horizontal gene transfer processes other than natural competence, and a competent species might show a low frequency due to scarcity of DNA from other strains, infrequent induction of competence, low recombination rates, or sampling of many nontransformable strains.

Inferences from gene distributions. In species with sequenced genomes, evidence of competence might also come from the presence of homologs of known competence genes (Table 1, columns 4 to 6). While investigations of this characteristic have met with some success, especially in relatives of well-studied species (33, 70), the results are not necessarily informative. The most striking counterexample is E. coli. Although it is not naturally transformable in the laboratory (despite extensive efforts) and its populations show little evidence of historical recombination, its genome contains an apparently complete set of competence genes inducible, like H. influenzae’s, by CRP and Sxy (71).

One reason the presence of competence gene homologs is not a reliable indicator of competence is that many genes required for DNA uptake have functions outside competence. In particular, the T4P machinery is widely distributed and strongly conserved at least partly because it also provides adhesion and twitching motility (8), and, conversely, the competent species Helicobacter pylori does not use T4P to take up DNA (72, 73). The taxonomic distributions of proteins that carry out homologous recombination are particularly uninformative, since these proteins have ubiquitous and conserved functions in DNA replication and repair (74).

Genes with the strongest cases for having functions specific to competence are those for Rec2/ComEC and DprA/Smf. Both proteins act after DNA has been transported across the outer membrane or through the cell wall. The Rec2/ComEC protein is required for translocation of single-stranded DNA into the cytoplasm; mutants are nontransformable. An alternative function for this protein has been recently identified in Listeria, in which ComEC (along with T4P genes) is important for infecting cells to escape host cell phagosomes (75). Competent cells defective in the cytoplasmic protein DprA take up and translocate DNA normally but produce few or no transformants; work in S. pneumoniae has shown that DprA/Smf is a recombination mediator protein, facilitating loading of RecA onto single-stranded DNA (76, 77). Although this suggests that DprA could function outside transformation, for example, by helping to stabilize stalled DNA replication forks, work in E. coli has failed to find a role in DNA repair (78). A GenBank search found 21,389 entries with dprA or smf homologs, 10,263 with rec2 or comEC homologs, and 3,058 with homologs of both genes, numbers far higher than the 75 verified competent species listed in a recent survey (5). Although the ubiquity of these proteins may indicate that many or even most bacteria are competent, it could also reflect as yet poorly understood cellular functions in a manner independent of DNA uptake.

Summary. Although naturally competent bacteria are common across the bacterial phylogeny, their sporadic distribution both within and above the species level implies frequent losses and/or gains of competence, which obscure its evolutionary origins (66, 79). The pleiotropic functions of the T4P machinery suggest that it might be repeatedly co-opted to function in DNA uptake, with the genes themselves maintained by selection for adhesion and motility functions. Another possibility is that competence is an ancient ancestral trait that is frequently lost but that the long-term consequence of lacking a genetic recombination system is an evolutionary dead end that ultimately leads to extinction (80).

THE MECHANISM AND DIVERSITY OF DNA UPTAKE SPECIFICITY

Convergent evolution of uptake specificity. Self-specific DNA uptake was first discovered in two competent Gram-negative species, H. influenzae and Neisseria gonorrhoeae. Although most competent bacteria that have been tested appear to take up any available DNA (Table 1, column 7) (81–86), competent cells of these species take up genomic DNA from their own species dramatically better than DNA of E. coli or other distant relatives (87, 88). Subsequent work showed that this self-specificity is not due to a preference for “self” DNA per se but to the genome of each species
being highly enriched for a sequence motif preferred by its own uptake machinery.

Before genome sequences became available, sequencing of short preferentially taken-up fragments identified the \textit{H. influenzae} uptake signal sequence (USS) as the 9-mer AAGTGCGGT, and the \textit{N. gonorrhoeae} DNA uptake sequence (DUS) as the 10-mer GCCGTCTGAA, with initial estimates of each of at least several hundred occurrences per genome and an uptake bias of at least 100-fold (89, 90). Later sequencing of both genomes found about one occurrence per kb and revealed the more nuanced motif models shown in Fig. 3A and B (6, 91). Such motif models found that the USS has two helically phased flanking segments of A/T bases, a feature absent from the DUS.

Subsequent experimental work and genomic inference have found that uptake self-specificity is the norm throughout the \textit{Pasteurellaceae} and \textit{Neisseriaceae} families (80, 92, 93). Within the two families, the abundance of genomic uptake sequences suggests that many species are descended from recent competent ancestors even where their competence has not been demonstrated in laboratory cultures. This analysis cannot readily be extended beyond these families (using the presence of uptake-sequence-like repeats as indicators of competence), since motifs with properties similar to those of uptake sequences may occur for other reasons, and motifs whose properties are less dramatic may be difficult to detect without experimental data. The \textit{Pasteurellaceae} and \textit{Neisseriaceae} families are related only distantly, in the gamma and beta branches of the proteobacteria, respectively. Since the USS and DUS are distinct sequences and uptake sequences have not been identified in other groups, the self-specificities of the two families are likely to be of independent evolutionary origins. The only other bacterial species with demonstrated self-specificity are within the epsilonproteobacteria, \textit{Campylobacter} and \textit{Helicobacter}, but these have no strongly overrepresented genomic motif (94–96). The many similarities between the USS and DUS described below suggest that they are convergent products of the same evolutionary forces.

The genomic distribution of uptake sequences. In both \textit{H. influenzae} and \textit{Neisseria meningitidis} (and its close relative \textit{N. gonorrhoeae}), the density of uptake sequences is about 1/kb—more than 100-fold higher than expected by chance—and they collectively comprise \textasciitilde1\% of each genome (6). Both the USS and DUS are enriched in the relatively permissive regions of the genome (intergenic regions, poorly conserved genes, and poorly conserved parts of genes), presumably to minimize interference with protein-coding functions (97). In addition, both the USS and DUS in intergenic regions frequently occur as inverted repeats that act as \textit{rho}-independent transcriptional terminators (90, 98). Although DUS were reported to be enriched in “genome maintenance” genes, reanalysis of the data found no significant correlation (97, 99).

Uptake sequences are not insertions arising from a copy-paste or other duplicative mechanism. Instead, alignment of \textit{H. influenzae} USS-containing genes with homologs from relatives without USSs shows that they have evolved mainly by simple point mutations (80); DUSs in coding sequences show the same pattern (100). Once they have arisen at specific positions, uptake sequences are stable; many USSs are in homologous positions in \textit{H. influenzae} and its relative \textit{Pasteurella multocida} despite the hundreds of millions of years since their divergence (101). Comparison of DUS locations in three closely related species of \textit{Neisseria} also found strong conservation (100); more divergent genomes are now available for analysis (68).

Genes that have undergone horizontal gene transfer provide additional evidence of how uptake sequences accumulate. When genome sequences of multiple isolates of \textit{H. influenzae} and of \textit{Neisseria} species are compared, uptake sequences have a higher density in genes present in all isolates (core genes) than in genes found in only some isolates (accessory genes) (100, 102). This would be expected if uptake sequences were to accumulate slowly, since recently acquired accessory genes would not yet have accumulated them. Slow accumulation is also consistent with the absence of the USS from the Mu-like prophage recently acquired by
the Rd strain of *H. influenzae* (103). Independent estimates of the timing of accessory gene acquisition could permit accurate estimates of the rate of uptake sequence accumulation, which would in turn aid in interpreting other phenomena such as the high frequency of USs in the genome of *H. influenzae* phage HP1 (90).

Uptake sequences have also changed the frequencies of specific peptide signatures in proteomes. A total of 70% of *H. influenzae* USs and 35% of *N. meningitidis* USs are in open reading frames, where they encode specific tripeptides. Analysis of tripeptide frequencies reveals that these occur twice as often in the proteomes of their own species as in the proteomes of species lacking that uptake sequence (e.g., *E. coli*). This pattern can be explained by the combined effects on mutational divergence of selection for their coding function and of molecular drive arising from their effects on DNA uptake. Because many, often most, of these tripeptides are encoded by sequences that no longer match the uptake sequence consensus, they provide stark historical evidence of the power of biased DNA uptake to affect genome evolution and of the strong similarities of these evolutionary forces in the two families (97).

**Mechanism of uptake specificity.** Since only DNA fragments with uptake sequences become protected from experimentally added nucleases, uptake sequences must determine whether DNAs are transported across the outer membrane (104, 105). And because a single USS or DUS is sufficient to mobilize uptake of both short and long DNA fragments (106, 107), uptake specificity likely acts only at the initiation step, not continuously, during uptake of long molecules. The *H. influenzae* protein or proteins responsible for USS specificity have still not been identified, but the *N. meningitidis* ComP minor pilus protein has recently been shown to specifically bind USs (93, 108). *H. influenzae* lacks a ComP ortholog, but this role might instead be filled by a minor pilin encoded by the competence-regulated pilNOPQ operon, which is conserved across the Pasteurellaceae and Enterobacteriaceae (33).

We propose that strong binding of the uptake machinery to DNA is likely needed to overcome the physical difficulties posed by transport of dsDNA across the outer membrane pore, an obstacle not faced by Gram-positive bacteria. One difficulty is the 6-nm diameter of the secretin pore. Initiation of uptake from internal sites requires that DNA be sharply kinked back on itself at the point of initiation to pass through this pore, which is made difficult by DNA’s 50-nm persistence length. The secretin pore is also just wide enough to accommodate either the type 4 pilus (109) or the pair of DNA double helices on both sides of the initiation site, but not both, suggesting that the tip of the pilus/pseudopilus may lead the DNA through the pore. After initiation, tight DNA-protein interactions may also be required to transmit the extremely strong (up to 40 pN) DNA uptake forces that have been recorded during pilus retraction (7, 72, 110); we propose that such strong interactions are unlikely to be achieved without at least some degree of sequence specificity. In a later section, we discuss the implications of this for the many species that do not demonstrate self-specific uptake.

Recent work has found that only 3 to 4 “core” bases within the USS and DUS make strong contributions to DNA uptake, with the remaining bases in each uptake motif individually making smaller contributions to uptake specificity (91–93, 111, 112). In *H. influenzae*, changing the consensus USS within a DNA fragment dramatically reduces uptake if the change affects one of the four bases GCGG, while mutations at other USS positions have substantially smaller effects (Fig. 3C). Similarly, in *N. meningitidis* the 3 bases CTG dramatically reduce DNA uptake, while other bases make smaller individual contributions (Fig. 3D). These data suggest that the core bases of the USS and DUS make strong sequence-specific contacts that tightly bind DNA during the retraction that initiates uptake. Initial assays of DUS binding by ComP did not distinguish between core and noncore binding, since they compared a perfect 12-bp DUS to a control construct mismatched at every other base (108), but later assays found that mutations to the core had much larger impacts on binding than mutations to noncore bases, in a pattern closely matching the *in vivo* DNA uptake results shown in Fig. 3D (92, 93).

These data also raise the issue of why noncore bases show such strong signatures in the accumulated genomic uptake sequence motif, despite making apparently minor contributions to uptake (Fig. 3). A detailed characterization of the *H. influenzae* bias was accomplished using deep sequencing of DNA fragments containing a degenerate USS; a pool of DNA fragments were recovered from competent cells’ periplasm, and ~10^7 fragments of both this periplasmic pool and the input pool were sequenced and compared. This corroborated the previous analysis of singly mutated USS constructs, but because of the high degeneracy and high sequence yield, that study was also able to show the effects of interactions (“positional dependencies”) between noncore bases, with doubly mutated constructs having on average ~5-fold-less-efficient uptake than that predicted from singly mutated constructs (indicated by the “synergistic interactions” shown at the bottom of Fig. 3C) (112). Thus, noncore bases still make important contributions to DNA uptake; although their individual effects are small, their collective effects are substantial. Similar interaction effects are suggested by the decreased uptake seen for some multiply mismatched DUS variants (92), but no systematic survey has yet been undertaken. Such interaction effects could arise because noncore bases weakly but cooperatively bind the uptake machinery, but they could also arise from cooperative contributions to DNA bending or kinking.

**Subclade-specific uptake sequence dialects.** Uptake sequences have often been assumed to be species-specific mate recognition signals, but several lines of evidence show both sharing and divergence of uptake sequences across related species. Analysis of heterospecific transformation in *H. influenzae* found that DNAs from many other *Haemophilus* species are readily taken up although they are otherwise too divergent to produce recombinants (113). Sharing of the same uptake sequence across diverged species can also promote harmful uptake of foreign DNA. In *H. influenzae*, this has been shown to kill cells by inducing the SOS response, which activates a resident prophage (114, 115).

Later work analyzing eight *Pasteurellaceae* genomes identified two monophyletic subclades with divergent USS types (dialects), *Hin*-USS and *Apl*-USS; *Fig. 4* shows the number of each USS type in the genome of each species (80). The two USS dialects share the GCGG core sequence and the helically phase T tracts, but several other bases do not align and the *Apl*-USS extends beyond the second T tract. The densities of the 9-mer consensus *Hin*-USS and *Apl*-USS within each subclade range from 117 to 836 per Mb, and bacteria in each subclade exhibit a strong preference for their USS dialect. More recent work in the *Neisseriaceae* has identified eight distinct DUS dialects, with each dialect highly enriched in the genomes of subclades within the *Neisseriaceae* phylogenetic tree.
As previously seen for USS dialects, the densities of 12-mer consensus DUS of different dialects differed dramatically, ranging from 705 to 1,754 per Mb, and measurements of uptake biases and transformation frequencies for several species also indicate preferential uptake of the corresponding genomic consensus.

Variation between dialects in both families is found outside the core bases that are particularly crucial for DNA uptake; i.e., the core DUS bases CTG, like the USS core bases GCCG, are invariant between dialects. This suggests a remarkable similarity between the families in both the mechanism of DNA uptake and the coevolutionary constraints operating between uptake specificity and genomic accumulation of uptake sequences. We reason that changes to uptake specificity that affect binding to a crucial core base are highly constrained, since these would also drastically reduce DNA uptake. Changes in uptake specificity affecting binding to a noncore base would have comparatively weaker effects on DNA uptake, allowing for less-constrained coevolutionary changes of the specificity machinery and the uptake sequences accumulated in the genome. However, the strong specificities of the uptake machineries for the different dialects and the deep-
sequencing results from *H. influenzae* suggest that sequence-specific DNA-protein interactions are still important at noncore positions but that several different combinations of bases and protein residues can provide for this.

Exceptional species within each family may be particularly useful for better understanding of the coevolution of uptake specificity and uptake sequence accumulation: analysis of the Pasteurellaceae species *Gallibacterium anatis* found self-specificity, but the densities of *Hin*-USS and *Apl*-USS are substantially lower than for other species in the family (Fig. 4) (73, 116). *G. anatis* represents an outgroup of the two main subclades (117, 118), suggesting that this species would be an ideal system for investigation of the origins of uptake bias and uptake sequence accumulation. In contrast, *Haemophilus ducreyi* (Apl-USS subclade) has a substantially lower density of USS than other Pasteurellaceae, suggesting either that it has lost competence or that its uptake specificity is changing (Fig. 4). Notably, one of the Neisseriaceae genomes analyzed, *Simonsiella muelleri*, contained evidence of two distinct dialects, both its own AG-simDUS and the AG-kingDUS found in related *Kingella oralis*, in roughly equal proportions (Fig. 5)(92). It remains unclear whether this is the result of two distinct uptake biases in the same organism or whether one of the DUS dialects is a remnant from a time before the uptake bias changed. Further experimental and genomic study of these species is clearly in order.

**A MODEL FOR THE EVOLUTION OF SELF-SPECIFICITY**

Molecular drive can explain the accumulation of uptake sequences in genomes. A satisfactory explanation of self-specificity must account for both the sequence bias of the DNA uptake machinery and the accumulation of its preferred sequences in the genome. The commonly assumed mate recognition function requires that both components evolve simultaneously by natural selection operating on their combined effects, but relaxing this assumption greatly simplifies the evolutionary steps. Here we first consider how—regardless of whether or not recombination is a selected function of competence—uptake bias causes a “molecular drive” leading to accumulation of uptake sequences in genomes, in the same way that biased gene conversion causes a molecular drive leading to allele fixation in sexual eukaryotes (91, 101, 119). We then consider how weak uptake biases could be amplified over time.

Consider an ancestral species that is competent but has few or no uptake sequences. This species is likely to have had some uptake bias, for reasons discussed above and below. Provided some fraction of the available DNA is conspecific (or otherwise able to recombine with the chromosome), the combination of random mutation, biased DNA uptake, and subsequent recombination will enrich the genomes of this species’ descendants with the preferred sequences, as follows (Fig. 6): at any position in the genome whose sequence is not already well matched to the uptake bias, random mutation will sometimes create variants that better fit this bias, and cell death will create a pool of environmental DNA that includes this variation. Competent cells will then preferentially take up those DNA variants that better match their bias, and recombination of these with their chromosomal homologs will transfer these preferences into the genome (91). The preferred sequences need not be functionally beneficial in any way; their accumulation will inevitably continue until it is limited by mutational degradation of uptake sequences and by selection against uptake sequences that conflict with genetic functions. In simulations, the mutation and transformation rates determine the final density of uptake sequences, the uptake bias determines their sequence distribution, the sizes of available DNA fragments determine their spacing, and natural selection determines their relationships to protein coding and other genome functions (91). Molecular drive is also predicted to prevent accumulation of uptake-reducing mutations once strong uptake sequences have arisen.

Notably, the molecular drive model does not preclude the possibility that competence is in part maintained for its genetic/informational consequences (26, 53–56); rather, it stipulates only that the existence of self-specificity does not necessarily mean that uptake sequences are there to screen for homologous DNA. Consistent with this, the only known specificity factor, ComP from *Neisseria*, is not an “add-on” filter for DUS-containing DNA but is an intrinsic part of the uptake mechanism, since knockouts have strong defects in DNA uptake (92, 93).

**Causes of uptake biases.** The evolutionary model described above begs the question of why DNA uptake would have a preexisting sequence bias, but uptake biases are likely to be more common than is suggested by the known occurrences of self-specificity. Sequence biases are typical of DNA-binding proteins in general, even those whose function is sequence independent, with tighter binding typically associated with stronger bias (120, 121).
As discussed above, the strong forces and topological deformations needed for DNA uptake, especially by Gram-negative bacteria, require that DNA be tightly bound to the uptake machinery. Thus, organisms with no reported self-specificity are predicted to nevertheless exhibit some degree of sequence bias in a manner independent of any enrichment in the genome.

**Development of self-specificity.** If the above conditions are met, the genome will gradually and inevitably become enriched for sequences preferred by the DNA uptake machinery. This enrichment then sets the stage for selection on the genes encoding this machinery, amplifying the bias to promote more-efficient uptake or better exclusion of harmful foreign DNAs. Such selection could occur regardless of whether the main benefit of the presence of the DNA is its nutrient content or its genetic information.

Why then do most tested groups have no self-specificity? Two factors are important here—whether preferred sequences recombine often enough to create a signature in the genome and whether this signature is strong enough to cause self-specific uptake. If conspecific DNA is rare in the local microenvironment, or if incoming DNA rarely recombines with the chromosome, even strong uptake biases would not give rise to a genomic signature. And even a strong signature does not cause strong self-specificity when it also occurs frequently in other DNAs; e.g., if cells had an absolute requirement for a simple motif such as AGTC, they would still take up many foreign DNAs as well as their own.

Is there something special about *Pasteurellaceae* and *Neisseriaceae* that predisposed them to evolve strong uptake biases and accumulate genomic uptake sequences? Species in both families predominantly live in respiratory tracts and other mucosal environments rich in host DNA, but other competent Gram-negative species lacking uptake sequences coexist with these (see, for example, reference 122).

It has been suggested that beneficial genetic recombination is optimized by quorum-sensing regulation as well as by self-specific uptake, since quorum sensing would ensure that cells become competent only when surrounded by other members of their species (4). However, the links between quorum sensing and competence are often incomplete or indirect. For example, only a subset of *V. cholerae*’s competence genes are regulated by a secreted autoinducer (39), and in *S. pneumoniae* and *Bacillus subtilis*, competence is part of a much larger set of genes and processes influenced by secreted autoinducers (39, 50, 51). Interpretation of this regulation is further complicated by uncertainties about whether secreted autoinducers exist primarily for sensing cell densities or for sensing the physical properties of the cell’s microenvironment (123). Further comparative studies into factors that limit DNA uptake to close relatives, such as uptake specificity and quorum sensing, are warranted.

**OUTSTANDING ISSUES**

The finding that subclades of the *Pasteurellaceae* and *Neisseriaceae* have distinct dialects is consistent with mutational divergence of specificity components of the uptake machinery as the subclades diverged. The resulting changes in uptake bias would then lead to corresponding changes in the population of uptake sequences in the genome by the same combination of random mutation, biased uptake, and recombination that caused their original accumulation and maintenance. Within the *Hin-USS* subclade, the positions of uptake sequence have remained quite stable (101), but their stability over the deeper evolutionary time separating the two subclades is unknown. A more systematic examination of synteny between uptake sequences of distinct dialects in both families is clearly in order.

The divergence of uptake specificity within each family could also give insights into the molecular basis of uptake specificity. Do changes in specific amino acid residues in components of the uptake machinery cooccur with changes in specificity? For example, if ComP interacts with dialect-specific bases in the DUS, mutations at specific surface residues in its positively charged channel could be responsible for shifts in specificity and the subsequent turnover of DUS dialects observed in the different *Neisseriaceae* species. Recent results show that single mutations to noncore bases in DUS have only weak effects on ComP binding (93), suggesting either that ComP does not directly interact with these positions or that effects are detected only when multiple noncore bases are changed. Experiments with ComP protein purified from species with different dialects will help sort this out. A comparative approach could also potentially also identify and validate candidates for the still unknown genes responsible for uptake specificity in the *Pasteurellaceae*.

The drive model predicts that the uptake sequences in modern genomes will have lower divergence rates than the parts of the genome that are under otherwise similar constraints, a pattern we have observed in preliminary studies of *H. influenzae* genomes (J. C. Mell and R. J. Redfield, unpublished data). This pattern could be informative about the function of competence, since molecular drive predicts reduced variation only at uptake sequences whereas selection for a mate-choice function predicts that reduced variation would also affect flanking sequences containing beneficial alleles. Since only a subset of genes in the USS-poor species *H. ducreyi* and *G. anatis* have nearby uptake sequences, similar population genetic analysis results at these loci could reveal biases in recombination to these loci. Finally, since species with particular dialects often contain many uptake sequences matching other dialects (Fig. 4 and 5), cooccurrence could cause genes close to these to exhibit a higher probability of cross-clade recombination. These and other population-based studies are becoming increasingly feasible as the number of species with many sequenced isolates increases.

We argue above that uptake biases are expected to be the norm rather than the exception among competent bacteria, which predicts that species with no reported self-specificity nevertheless would still have some degree of uptake bias. Although identifying weak biases was previously impractical, this can now be done using a modification of the deep-sequencing approach used to characterize the uptake bias of *H. influenzae* (112), in which periplasmic DNA taken up by competent cells was purified away from their chromosomes. The best candidates for these tests are the competent Gram-negative species demonstrated to lack self-specificity, *Acinetobacter baysii*, *Thermus thermophilus*, *Vibrio cholerae*, and *Pseudomonas stutzeri* (82–84), but the well-studied competent models in the Gram-positive bacteria *Bacillus subtilis* and *Streptococcus pneumoniae* should also be examined (85, 86). Noting that uptake biases are still present even in the absence of self-specificity would confirm that intrinsic sequence constraints play important roles in DNA uptake.

The epsilonproteobacteria *Helicobacter pylori* and *Campylobacter jejuni* have been shown to have uptake specificity, but the mechanism is unclear since no uptake sequences have been identified in their genomes (95, 96). Although this may simply reflect
lower enrichment or less-specific biases that are not as readily detected as those of the Pasteurellaceae and Neisseriaceae, these species could also require specific DNA modifications for efficient uptake. Glucosylation of DNA has been shown to affect DNA binding by B. subtilis (124), and ethylation of specific USP positions can either reduce or enhance their uptake (89). Since Helicobacter and Campylobacter use a type IV secretion system for DNA uptake instead of T4P (114), insights into uptake specificity in this independently evolved system will provide powerful evidence of the forces responsible for uptake specificity.

That there is a lack of consensus on the evolutionary function of natural competence is not because we lack evidence of how natural selection acts on genetic recombination. Many microbiologists are probably unaware of the extensive body of theoretical and experimental work on the evolution of genes causing sexual (meiotic) recombination in eukaryotes (reviewed in reference 125). This work has shown that such genes and alleles are not favored by natural selection except in restricted circumstances, and even in these circumstances selection for recombination is readily swamped by molecular drive forces (126). Lack of selection for recombination is not inconsistent with the strongly beneficial recombination events observed in bacterial genomes, since rare beneficial genotypes will be preserved while harmful recombinants are swept under the rug of evolutionary history.

ACKNOWLEDGMENTS

This work was supported by grants to R.J.R. from the Canadian Institutes for Health Research and the Natural Sciences and Engineering Research Council of Canada and by a postdoctoral fellowship to J.C.M. from the National Institutes of Health.

We thank three anonymous reviewers for their detailed and helpful comments.

REFERENCES

30. Cameron ADS, Redfield RJ. 2006. Non-canonical CRP sites control competence regulators in Escherichia coli and many other gamma-


58. Stingl K, Muller S, Scheidgen-Kleyboldt G, Clausen M, Maier B. 2010. Continuous culture populations of Bacillus subtilis show a dynamic but stable population and evolutionary dynam-
Minireview


97. Mell JC, Hall IM, Redfield RJ. 2012. Defining the DNA uptake speci-


