DOI: 10.1534/genetics.110.119438

Bacterial DNA Uptake Sequences Can Accumulate by Molecular Drive Alone

H. Maughan,[1] L. A. Wilson[2] and R. J. Redfield[3]

Department of Zoology, University of British Columbia, Vancouver, British Columbia V6T 3Z4, Canada

Manuscript received May 30, 2010
Accepted for publication July 1, 2010

ABSTRACT

Uptake signal sequences are DNA motifs that promote DNA uptake by competent bacteria in the family Pasteurellaceae and the genus Neisseria. The genomes of these bacteria contain many copies of their canonical uptake sequence (often >100-fold overrepresentation), so the bias of the uptake machinery causes cells to prefer DNA derived from close relatives over DNA from other sources. However, the molecular and evolutionary forces responsible for the abundance of uptake sequences in these genomes are not well understood, and their presence is not easily explained by any of the current models of the evolution of competence. Here we describe use of a computer simulation model to thoroughly evaluate the simplest explanation for uptake sequences, that they accumulate in genomes by a form of molecular drive generated by biased DNA uptake and evolutionarily neutral (*i.e.*, unselected) recombination. In parallel we used an unbiased search algorithm to characterize genomic uptake sequences and DNA uptake assays to refine the *Haemophilus influenzae* uptake specificity. These analyses showed that biased uptake and neutral recombination are sufficient to drive uptake sequences to high densities, with the spacings, stabilities, and strong consensuses typical of uptake sequences in real genomes. This result greatly simplifies testing of hypotheses about the benefits of DNA uptake, because it explains how genomes could have passively accumulated sequences matching the bias of their uptake machineries.

MANY bacteria are able to take up DNA fragments from their environment, a genetically specified trait called natural competence (Chen and Dubnau 2004; Johnsborg *et al.* 2007; Maughan *et al.* 2008). Many other species have homologs of competence genes, suggesting that although they are not competent under laboratory conditions, they may be competent under natural conditions (Claverys and Martin 2003; Kovacs *et al.* 2009). Such a widespread trait must be beneficial but the evolutionary function of DNA uptake remains controversial. Cells can use the nucleotides released by degradation of both incoming DNA and any strands displaced by its recombination, thus reducing demands on their nucleotide salvage and biosynthesis pathways (Redfield 1993; Palchevskiy and Finkel 2009). Cells may also benefit if recombination of the incoming DNA provides templates for DNA repair or introduces beneficial mutations, but may suffer if recombination introduces damage or harmful mutations (Redfield 1988; Michod *et al.* 2008).

Although most bacteria that have been tested show no obvious preferences for specific DNA sources or sequences, bacteria in the family Pasteurellaceae and the genus Neisseria strongly prefer DNA fragments from close relatives. Two factors are responsible: First, the DNA uptake machineries of these bacteria are strongly biased toward certain short DNA sequence motifs. Second, the genomes of these bacteria contain hundreds of occurrences of the preferred sequences. The Pasteurellacean motif is called the uptake signal sequence (USS); its Neisseria counterpart is called the DNA uptake sequence (DUS). All Neisseria genomes contain the same consensus DUS [core GCCGTCTGAA (Treangen *et al.* 2008)], but divergence in the Pasteurellaceae has produced two subclades, one of species sharing the canonical *Haemophilus influenzae* 9-bp USS (*Hin*-USS core AAGTGCGGT) and the other of species with a variant USS that differs at three core positions (*Apl*-USS core: ACAAGCGGT) and has a longer flanking consensus (Redfield *et al.* 2006). Uptake sequence biases are strong but not absolute; for example, replacing the *Hin*-USS with the *Apl*-USS reduces *H. influenzae* DNA uptake only 10-fold (Redfield *et al.* 2006) and DNA from *Escherichia coli* is taken up in the absence of competing *H. influenzae* DNA (Goodgal and Mitchell 1984).

Supporting information is available online at http://www.genetics.org/cgi/content/full/genetics.110.119438/DC1.

[1]*Present address:* Department of Cell and Systems Biology, University of Toronto, Toronto, ON M5S 3G5, Canada.

[2]*Present address:* MRC Toxicology Unit, Hodgkin Bldg., P.O. Box 138, University of Leicester, Leicester LE1 9HN, United Kingdom.

[3]*Corresponding author:* Life Sciences Centre (Zoology), 2350 Health Sciences Mall, University of British Columbia, Vancouver, BC V6T 3Z4, Canada. E-mail: redfield@zoology.ubc.ca

Most studies of the distribution and consensus of uptake sequences in genomes have examined only those occurrences that perfectly match the above core DUS and USS sequences. Here we call these perfect matches "core-consensus" (cc) uptake sequences. These cc-uptake sequences occupy ~1% of their genomes; they are equally frequent in the plus and minus orientations of the genome sequence but are underrepresented in coding sequences, with the noncoding 14% and 20% of their respective genomes containing 35% of cc-USSs and 65% of cc-DUSs (Smith *et al.* 1995). Although many of these intergenic cc-DUSs and cc-USSs occur in inverted-repeat pairs that function as terminators (Kingsford *et al.* 2007), most uptake sequences are too far apart or in genes or other locations where termination does not occur. Within coding regions uptake sequences occur more often in weakly conserved genes, in weakly conserved parts of genes, and in reading frames that encode common tripeptides (Findlay and Redfield 2009), all of which are consistent with selection acting mainly to eliminate mutations that improve uptake from genome regions constrained by coding or other functions.

Analyses that focus on cc-uptake sequences effectively treat uptake sequences as replicative elements (Smith *et al.* 1995; Redfield *et al.* 2006; Ambur *et al.* 2007; Treangen *et al.* 2008). However, USS and DUS are known to originate *in situ* by normal mutational processes, mainly point mutations, and to spread between genomes mainly by homologous recombination (Redfield *et al.* 2006; Treangen *et al.* 2008). As they are not replicating elements, why are they up to 1000-fold more common in their genomes than expected for unselected sequences (*e.g.*, *H. influenzae*, 1471 cc-USS *vs.* 8 expected by chance; *N. gonorrheae*, 1892 cc-DUS *vs.* 2 expected by chance)?

The explanation for this abundance must lie with the specificity of the DNA uptake system, because the strong correspondence between the uptake bias and the uptake sequences in the genome is far too improbable to be a coincidence. However, uptake specificity is not easily accommodated by either of the hypothesized functions of DNA uptake. If bacteria take up DNA to get benefits from homologous genetic recombination, the combination of uptake bias and uptake sequences might serve as a mate-choice adaptation that maximizes these benefits by excluding foreign DNAs (Treangen *et al.* 2008). Although this explanation is appealing, it requires simultaneous evolution of bias in the uptake machinery and of genomic sequences matching this bias. Another problem is that the genomic sequences can be "selected" only after the cell carrying them is dead. On the other hand, if bacteria instead take up DNA as a source of nutrients, all DNAs should be equally useful, so uptake bias and uptake sequences would likely reduce rather than increase this benefit. Although the sequence bias could be explained as a consequence of mechanistic constraints on DNA uptake, this would not account for the high density of the preferred sequences in the genome.

However, both hypotheses may be simplified by a process called molecular drive, under which uptake sequences would gradually accumulate over evolutionary time as a direct consequence of biased uptake and recombination (Danner *et al.* 1980; Bakkali *et al.* 2004; Bakkali 2007), without any need for natural selection. This drive is proposed to work as follows: First, random mutation continuously creates variation in DNA sequences that affects their probability of uptake, and random cell death allows DNA fragments containing preferred variants to be taken up by other cells. Second, repeated recombination of such preferred DNA sequences with their chromosomal homologs gradually increases their abundance in the genomes of competent cells' descendants. Thus mutations that create matches to the bias of the uptake machinery are horizontally transmitted to other members of the same species more often than other mutations. Because some recombination may be inevitable even if DNA's main benefit is nutritional, molecular drive could account for uptake sequence accumulation under both hypotheses, leaving only the biased uptake process to be explained by natural selection for either genetic variation or nutrients.

Although drive is plausible, its ability to account for the observed properties of genomic uptake sequences has never been evaluated. To do this, we developed a realistic computer simulation model that includes only the processes thought to generate molecular drive. Below we first use this model to identify the conditions that determine whether uptake sequences will accumulate and then compare the properties of these simulated uptake sequences to those of the uptake sequences in the *N. meningitidis* and *H. influenzae* genomes. In parallel we use unbiased motif searches to better characterize genomic uptake sequences and DNA uptake assays to refine the *H. influenzae* uptake specificity.

MATERIALS AND METHODS

The model: Our goal was to find out whether uptake sequences resembling those in real genomes would evolve when a genome was influenced only by mutation, biased DNA uptake, and neutral recombination. The model, illustrated schematically in Figure 1, thus was designed to simulate evolution of uptake sequences in a single bacterial lineage that is subject to millions of generations of random mutation plus homologous recombination with DNA fragments derived from close relatives. Uptake bias is simulated by preferentially choosing the fragments that best match an uptake sequence motif. Although real uptake biases are likely to have arisen gradually over millions of years, for simplicity the model assumes that DNA uptake has a strong sequence bias that remains constant.

Each simulation run begins with one "focal" genome whose sequence is either provided as an input file or created as a random DNA sequence of specified length and base composition. The focal genome is then subject to many thousands of cycles of mutation and recombination. Because real biological

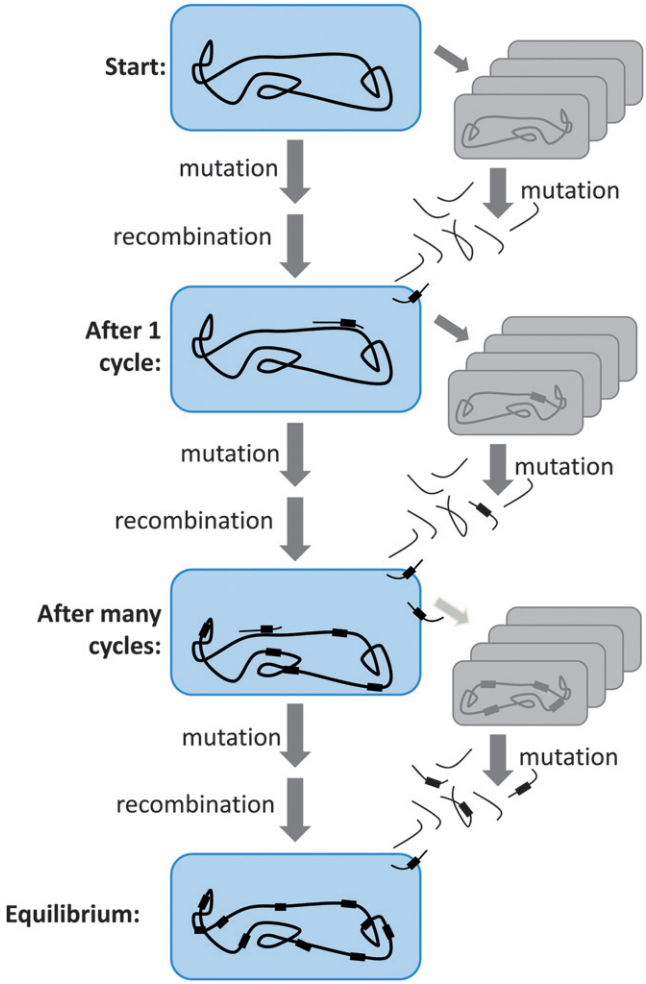


FIGURE 1.—Schematic drawing of the computer simulation model of molecular drive. Blue cells contain the evolving focal genome, and gray cells are its relatives whose deaths provide DNA fragments for uptake.

mutation rates are very low ($\sim 10^{-8}$–10^{-9}), each cycle is best considered as representing many generations of bacterial evolution (see DISCUSSION). In each cycle, transformation occurs when segments of the focal genome are replaced by homologous DNA fragments derived from related genomes. To accomplish this, candidate fragments are chosen randomly from each cycle's focal genome, mutated to simulate divergence, and examined with a sliding window for the best match to an uptake sequence motif specified by a position frequency matrix. Each fragment's score determines its probability of replacing the corresponding segment of the (now also mutated) focal genome. The user specifies the mutation rate of the focal genome, the mean divergence of the fragments, and the parameters affecting recombination—the number and length of DNA fragments recombined in each cycle, the matrix that defines the sequence preference of the uptake system, and how the matrix is used to score the fragments. All steps are simulated stochastically, so each run gives independent results.

The scoring matrices used are provided in Table 1. The matrix stringencies are provided to facilitate comparisons between them. The stringency of each matrix position is calculated as the ratio of the mean of its three low values to its highest value and the total stringency of each matrix as the product or sum of the individual stringencies. Matrices A–C specify the same generic uptake sequence (US) but with different stringencies, whereas DUS and USS matrices are derived from motif searches of the *N. meningitidis* and *H. influenzae* genomes. For each run of the simulation, the matrix is used both to score individual fragments in each cycle (to calculate recombination probabilities) and to score the focal genome (so the user can track the accumulation of uptake sequences as the run progresses). For fragment scoring, the score at each window position is either the product (stringent bias, "multiplicative" runs) or the sum (relaxed bias, "additive" runs) of the individual matrix-specified nucleotide preferences, and the fragment score is the single best score. For genome scoring, the score at each window position is always the product of the individual matrix-specified nucleotide preferences, and the genome score is the sum of the scores at all window positions, standardized by genome length. As the known uptake systems have strong biases, we present data primarily from runs using multiplicative scoring. To facilitate comparisons of evolved genomes from runs scored with matrix B or C, the final genomes were rescored using matrix A.

Simulation runs were typically continued until the genome score reached an equilibrium where integration of preferred uptake sequences by recombination was balanced by loss of these sequences through random mutation. The number of cycles needed to reach equilibrium depended on the parameter values and ranged from 5000 to 1,000,000 cycles. To unambiguously identify these equilibria, pairs of runs with identical parameters were initiated with focal genomes of either (a) a random sequence or (b) a random sequence seeded at random positions with five generic 10-bp uptake sequences per kilobase. The convergent endpoints of such pairs of runs were treated as equilibria (see Figure 2). For very time-consuming runs (*e.g.*, those with very low mutation or recombination rates), random-sequence and seeded runs were continued until their mean scores of all cycles were within twofold of each other, and then the average of the two runs was taken as the equilibrium.

Although equilibrium scores (per kilobase of sequence) did not depend on genome length (see RESULTS), the equilibrium scores of simulations with very short genomes showed substantial random fluctuations around the mean score, which reduced the value of their fast run times. At the other extreme, runs with biologically typical genome lengths (≥1 Mb) had stable equilibrium properties but took weeks or months to reach equilibrium. As a compromise most simulations used genome lengths of 20 or 200 kb; these runs typically took several hours to several weeks to reach equilibrium. The model was implemented in Perl and run on Macintosh computers and the MOA server at Dalhousie University; the code is available from the authors.

Sequence analysis methods: The *H. influenzae* Rd (NC000907) and *N. meningitidis* MC58 (NC003112) genome sequences were obtained from the J. Craig Venter Institute (http://www.jcvi.org/). The Linux version of the Gibbs motif sampler (THOMPSON *et al.* 2003, 2005) was run on the Westgrid computer facility (www.westgrid.ca); the Mac version was run on Mac desktops and laptops. To include uptake sequences in the reverse orientation, reverse-complement sequences were concatenated to the forward strand sequence before the combined sequences were searched for a single motif. A segmentation mask was used as a prior for *H. influenzae* genome searches, to specify internal spacing corresponding to that of known USSs. The mask specified a 10-bp motif followed by two 6-bp motifs separated by 1 and 6 bp, as follows: ++++++++++x++++++-xxxx-++++++, where "+'s" specify positions whose consensus is to be included in the motif, and "-" and "x" specify positions that are optional or not included, respectively. Some searches of gene

TABLE 1

Position-weight matrices used in simulations

Position	A	T	C	G	Stringency[a]
Matrix A (total multiplicative stringency 10^{-100}; total additive stringency 0.1)					
1	10	0.1	0.1	0.1	0.01
2	10	0.1	0.1	0.1	0.01
3	10	0.1	0.1	0.1	0.01
4	0.1	0.1	0.1	10	0.01
5	0.1	10	0.1	0.1	0.01
6	0.1	0.1	0.1	10	0.01
7	0.1	0.1	10	0.1	0.01
8	0.1	0.1	0.1	10	0.01
9	0.1	0.1	0.1	10	0.01
10	0.1	10	0.1	0.1	0.01
Matrix B (total multiplicative stringency 10^{-10})					
1	1	0.1	0.1	0.1	0.1
2	1	0.1	0.1	0.1	0.1
3	1	0.1	0.1	0.1	0.1
4	0.1	0.1	0.1	1	0.1
5	0.1	1	0.1	0.1	0.1
6	0.1	0.1	0.1	1	0.1
7	0.1	0.1	1	0.1	0.1
8	0.1	0.1	0.1	1	0.1
9	0.1	0.1	0.1	1	0.1
10	0.1	1	0.1	0.1	0.1
Matrix C (total multiplicative stringency 1.05×10^{-4})					
1	0.25	0.1	0.1	0.1	0.4
2	0.25	0.1	0.1	0.1	0.4
3	0.25	0.1	0.1	0.1	0.4
4	0.1	0.1	0.1	0.25	0.4
5	0.1	0.25	0.1	0.1	0.4
6	0.1	0.1	0.1	0.25	0.4
7	0.1	0.1	0.25	0.1	0.4
8	0.1	0.1	0.1	0.25	0.4
9	0.1	0.1	0.1	0.25	0.4
10	0.1	0.25	0.1	0.1	0.4
DUS matrix (total multiplicative stringency 3.2×10^{-20})					
1	0.831	0.054	0.054	0.060	0.068
2	0.031	0.742	0.047	0.181	0.116
3	0.066	0.007	0.000	0.927	0.026
4	0.000	0.061	0.939	0.000	0.022
5	0.015	0.044	0.941	0.000	0.021
6	0.007	0.000	0.000	0.993	0.002
7	0.005	0.882	0.113	0.001	0.045
8	0.000	0.027	0.973	0.000	0.009
9	0.006	0.869	0.098	0.027	0.050
10	0.034	0.006	0.004	0.956	0.015
11	0.935	0.005	0.060	0.000	0.023
12	0.945	0.021	0.016	0.019	0.020
USS matrix (total multiplicative stringency 5.2×10^{-22})					
1	0.727	0.113	0.093	0.067	0.125
2	0.915	0.04	0.019	0.025	0.031
3	0.912	0.04	0.025	0.023	0.032
4	0.068	0.039	0.023	0.87	0.050
5	0.077	0.863	0.033	0.026	0.053
6	0.044	0.039	0.018	0.898	0.037
7	0.031	0.035	0.915	0.019	0.031
8	0.046	0.033	0.022	0.899	0.037
9	0.045	0.033	0.024	0.898	0.038

(*continued*)

TABLE 1

(Continued)

Position	A	T	C	G	Stringency[a]
10	0.063	0.827	0.056	0.054	0.070
12	0.623	0.098	0.052	0.226	0.200
13	0.6	0.213	0.06	0.127	0.222
14	0.493	0.441	0.029	0.037	0.345
15	0.384	0.546	0.036	0.034	0.278
16	0.234	0.634	0.098	0.033	0.192
17	0.18	0.656	0.109	0.055	0.175
24	0.433	0.134	0.072	0.36	0.435
25	0.306	0.512	0.117	0.065	0.323
26	0.277	0.666	0.032	0.025	0.167
27	0.146	0.766	0.049	0.04	0.102
28	0.122	0.689	0.13	0.059	0.152
29	0.232	0.547	0.122	0.1	0.278

[a] Stringencies for individual matrix positions are calculated as the maximum value for that position divided by the mean of the three other values.

and intergenic sequences were also initialized using a prior file containing the base frequencies previously identified by a whole-genome search.

Most whole-genome searches used as "expected" number of occurrences 1.5 times the number of perfect matches to the standard DUS or USS core sequence (10 or 9 bp, respectively). The 1.5 value was arbitrarily chosen but should be conservative; some recent mutations will not yet have been subject to drive, including those that improve or worsen uptake. Since searches consistently returned substantially more than this number, the occurrences were ranked by their assigned scores and only the expected number were used to construct the logos shown in Figure 6 and supporting information, Figure S1. However, all of the identified occurrences were used for covariation analysis with MatrixPlot (www.cbs.dtu.dk/services/MatrixPlot) (Gorodkin *et al.* 1999), with the base composition settings appropriate to the sequence being searched. The random expectations for the spacing analyses in Figure 7 were determined by analyzing the spacings of random positions in 10 genomes of the same sizes and uptake sequence densities as the real *N. meningitidis* and *H. influenzae* genomes.

DNA uptake assays: DNA uptake was assayed in strain RR554, a *H. influenzae* KW20 derivative carrying pHKrec (Barcak *et al.* 1989), with competent and noncompetent *H. influenzae* KW20 as controls. Competence genes in strain RR554 are constitutively expressed due to overproduction of the positive regulator Sxy from the plasmid (Williams *et al.* 1994). Cells were cultured in brain–heart infusion supplemented with hemin and NAD (sBHI) as described (Poje and Redfield 2003a). RR554 and negative control KW20 cells were frozen on reaching a density of 10^9 cfu/ml in sBHI. All cells were stored as 1-ml aliquots frozen in 16% glycerol at −80°. Competence of each batch was confirmed by transforming thawed cells with chromosomal DNA of the multiply antibiotic resistant MAP7 *H. influenzae* strain (Poje and Redfield 2003b) and testing for novobiocin resistance.

Both strands of a 30-bp double-stranded DNA fragment containing the most frequent base at each position of the Smith *et al.* (1995) consensus were synthesized, annealed, and cloned by blunt-end ligation into the *Sma*I site of plasmid pGEM-7Zf(−) (Promega, Madison, WI) to generate plasmid pUSS-C. The same 30 bases in randomized order were synthesized, annealed, and cloned to give the negative control plasmid pUSS-R. Plasmids with mutant USS sequences were then generated from pUSS-C by overlap extension site-directed mutagenesis (Ho *et al.* 1989). Each mutagenized PCR product, containing a USS variant plus 193 bp of flanking plasmid sequence, was cloned into pCR2.1-TOPO (Invitrogen, Carlsbad, CA). The inserts of pUSS-C and pUSS-R were similarly amplified and subcloned, without mutagenesis. The sequences of all inserts were verified by sequencing.

The 222-bp plasmid inserts were amplified by PCR, and 60 ng was internally labeled with [α-^{33}P]dATP using the same primers and the Klenow fragment of *Escherichia coli* DNA polymerase I, in a reaction containing 62.5 μM d(CGT)TP, 8 μM forward and reverse primer, and 0.25 μM [α-^{33}P]dATP. Klenow (2 units) was added after a 3-min denaturation step at 94°, and the reaction was incubated for 20 min at 37°. To ensure complete copying of the DNA the reaction was continued for 30 min with additional dATP (60 μM). The DNA was purified using Sigma (St. Louis) PCR clean-up columns and eluted into 50 μl 10 mM Tris–HCl (pH 8.0) to a final concentration of 5–10 ng/μl and a specific activity of 2–6 × 10^4 cpm/ng.

In each uptake assay, 1-ml aliquots of frozen cells were thawed, pelleted by centrifugation at 16,000 × *g* for 2 min at room temperature, and then resuspended by vortexing in an equal volume of the competence-induction medium MIV (Poje and Redfield 2003b). Aliquots of 0.5 ml were then added to 10–20 ng of labeled DNA in 2-ml tubes prewarmed to 37°, and the contents briefly mixed by vortexing. Cells and DNA were incubated for 15 min at 37° and then washed three times by centrifugation as above, with each supernatant retained. The final cell pellet was resuspended in 100 μl MIV, and the radioactivity (cpm) of each supernatant and the pellet was measured. The fraction taken up was calculated as the ratio of pellet activity to total activity (pellet plus all supernatants). Uptake of each USS variant was measured at least three times for competent KW20 and strain RR554.

RESULTS

Effects of model parameters: To identify the fundamental factors affecting uptake sequence accumulation the model was first run with a generic matrix that specified a strong uptake bias. This matrix (matrix A) specifies a 10-bp uptake sequence motif with each

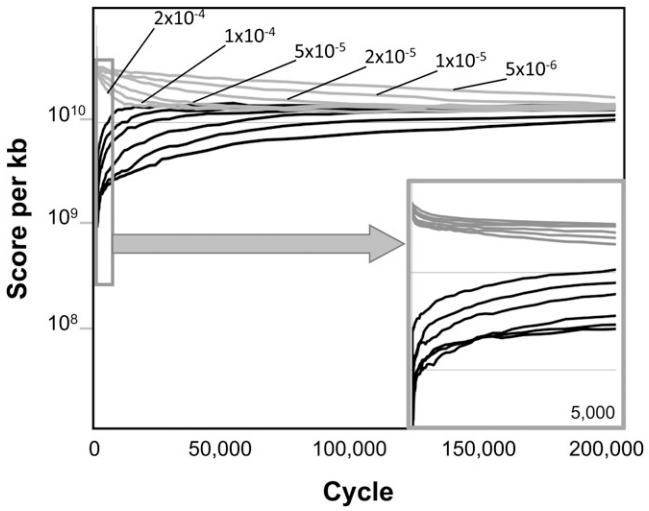


FIGURE 2.—Effect of mutation rate on uptake sequence accumulation. All runs had 200-kb genomes and used matrix A to choose 1000 × 100-bp fragments (100-fold divergent) for recombination in each cycle. Solid lines, runs initiated with random-sequence genomes (mean score 2.8×10^7); shaded lines, runs initiated with random-sequence genomes seeded with five USs per kilobase. Mutation rates are indicated above each upper line. Inset, expansion of the boxed area of the main graph.

position's preferred base making a 100-fold higher contribution to the fragment's uptake score than the other three bases. Beginning with this strong bias enabled us to investigate the other factors affecting uptake sequence accumulation before examining the effect of changing the bias. With this stringent matrix, genome score was a good measure of the density of perfectly matched uptake sequences. Although the sequence this matrix specifies is a 10-bp version of the cc-USS, it was always used with genome base composition set at 50% so this analysis applies equally to any 10-bp uptake sequence.

Effect of mutation rate: We began with a systematic analysis of 200-kb genomes evolving with mutation rates between 2×10^{-4} and 5×10^{-6}/bp per cycle. Although the number of model cycles (and thus the computer time) needed to reach equilibrium depended strongly on the mutation rate, the genome score at the equilibrium did not. Figure 2 shows that all runs had equilibrium scores of $\sim 1.4 \times 10^{10}$/kb (~1.4 uptake sequences per kilobase). Finding that the equilibrium score was independent of mutation rate allowed otherwise-slow simulations of large genomes to be speeded up by using high mutation rates, thus obtaining more stable estimates of equilibrium genome properties without excessive run times.

Effect of recombination parameters: The amount of recombination each genome received in each simulation cycle depended on both the number of fragments that recombined with the genome and the lengths of these fragments. Because little is known about the recombination rates of real genomes, we examined the effects of as wide a range of parameter values as possible. Figure 3A shows the equilibrium genome scores of a series of runs where the number of fragments recombined was varied while fragment length was held constant at 100 bp; between 0.0005 and 15 genome equivalents were recombined per cycle. Equilibrium score increased smoothly over a wide range, with an ~10-fold decrease in score for each 100-fold decrease in recombination; *i.e.*, uptake sequence accumulation is only moderately sensitive to recombination rate. The gradual leveling off seen for runs with very high levels of recombination (right side of Figure 3A) is expected because, when a position undergoes multiple recombination events in a single cycle of the model, only the last recombination event determines the sequence used for the next cycle. Figure 3B shows this "effective" recombination as a function of total recombination. At the other end of the recombination scale (left side of Figure 3A), when only a very small fraction of the genome was replaced in each cycle (one 100-bp fragment in a 200-kb genome), the equilibrium score was still ~100-fold higher than the typical scores of random-sequence genomes ($\sim 3.8 \times 10^6$/kb). This suggests that even very small amounts of recombination can have a significant impact on genome evolution if DNA uptake is biased. Because high-recombination fractions increased run time more than effective recombination, most simulations presented below recombined 0.5 genome equivalents in each cycle.

Fragment length: The default fragment size of 100 bp was initially chosen to maximize simulation speed, but Figure 3C shows that the length of the recombining fragments had a large effect on uptake sequence accumulation. In runs where fragment sizes were varied but the total amount of recombination was held constant at 0.5 genome equivalent (by covarying the numbers of recombining fragments), the equilibrium genome score decreased 30-fold in response to a 40-fold increase in fragment size. This is an expected consequence of choosing fragments by the score of their single best uptake sequence, because when the recombining fragments are longer, more of the fragment is hitchhiking along with the selected US. The effect of fragment length on the spacing of uptake sequences is considered in a later section.

Fragment divergence: Most simulations used fragments that had been mutated at rates 100-fold higher than the genomic mutation rate, to simulate their sharing a common ancestor with the focal genome an average of 50 cycles ago. This divergence is not expected to limit the efficiency of recombination; with genomic mutation rates of 10^{-4} or 10^{-5} such fragments would differ from the focal genome at no more than 1% of positions. Reducing the fragment divergence to 10-fold (5 cycles) reduced the equilibrium genome score by only about half, and eliminating it entirely reduced it only slightly

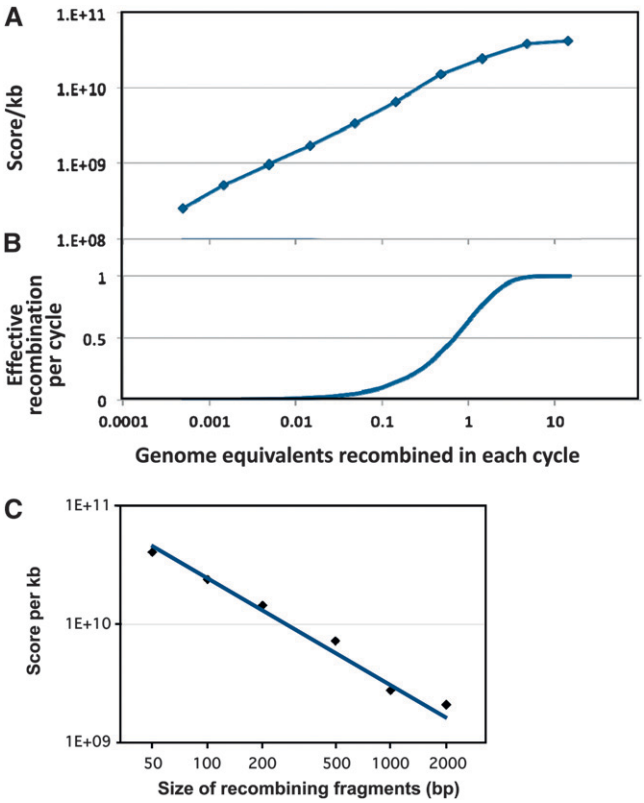


Figure 3.—Recombination effects. All runs had mutation rates of 1×10^{-4} and used matrix A to choose fragments (100-fold divergent) for recombination in each cycle. (A) Effect of amount of recombination on uptake sequence accumulation. Fragments were 100 bp, and genome size and number of recombining fragments were jointly varied to give the desired genome equivalents. (B) Effective recombination as a function of genome equivalents recombined. (C) Equilibrium genome score as a function of the size of the recombining fragments. The total amount of recombination was held constant at 50% by adjustments to the number of fragments recombined in each cycle.

more. This effect was independent of the amount of recombination over a 100-fold range.

Preference strength: We tested the effect of scoring fragments with a less stringent matrix, matrix B (Table 1). In this matrix, each position's preferred base is favored over the three other bases by only 10-fold (rather than by 100-fold as in matrix A). After the final genome scores from the matrix B runs were recalculated using matrix A (to allow comparison between final genome scores), use of the weaker matrix was seen to have caused only a 33% reduction in the final score per kilobase (1.0×10^{10} *vs.* 1.5×10^{10}, both averages of two replicate 200-kb runs). Runs with the even-weaker matrix C (2.5-fold base preference) gave much lower equilibrium scores (after rescoring with matrix A), only slightly higher than those of random sequences.

Additive and threshold scoring of fragments: In the simulations described above, the matrices were used multiplicatively to score fragments for recombination; that is, at each sliding-window position the scores for each of the 10 base positions were multiplied to give the score for that 10-bp sequence. However, the use of multiplication was quite arbitrary, as nothing is known about how the individual uptake sequence nucleotides interact in real bacteria to determine the probability that a fragment will be taken up. To examine alternatives we first ran similar simulations using additive scoring of fragments, where the scores for each of the 10 bases in the sliding window were instead added to give the score for that 10-bp sequence. (Note that the resulting genomes were still scored multiplicatively, so the outcomes can be directly compared to previous results.) In these runs uptake sequences did not accumulate even with matrix A (shaded line in Figure 4). We then tested a threshold scoring system, where the only fragments taken up were those that contained sequences matching at least a specified number of bases in the 10-bp US. The runs with a threshold of 10 required that each fragment contain a single perfectly matched uptake sequence. The solid lines in Figure 4 show that runs with thresholds of ≤6 gave no uptake sequence accumulation, but that higher thresholds were increasingly effective.

Properties of evolved genomes at equilibrium: *Proportions of perfect and mismatched sequences:* One notable attribute of the core-consensus uptake sequences previously examined in real genomes is the paucity of occurrences with one or two positions that do not match the consensus; for example, there are almost twice as many perfectly matched cc-USS sequences as singly mismatched ones, even though random mutations are 27 times more likely to convert the former to the latter than vice versa (Smith *et al.* 1995, 1999). Similar proportions were seen in the genomes evolved under matrix A (details in Table S1). The runs shown in Figure 2 gave equilibrium genomes that had, on average, 1.37 perfectly matched uptake sequences per kilobase, 0.14 singly mismatched ones, and 1.50 doubly mismatched ones (see Table S1). Genomes evolved with matrix B had only slightly more perfect consensus uptake sequences than singly mismatched ones, and those evolved with matrix C had only about one-quarter as many. Thus strong uptake biases could readily explain the proportions observed in real genomes.

Stability of uptake sequences positions: When real genome sequences of related species are aligned, many of their uptake sequences are found in homologous positions (Bakkali *et al.* 2004; Treangen *et al.* 2008). This could be because individual uptake sequences are stabilized over hundreds of millions of years by functional constraints that limit mutational degeneration of existing uptake sequences and fixation of new ones, or it could be because their mode of evolution makes uptake sequences intrinsically stable. We examined the stability of the simulated uptake sequences in a variety of equilibrium genomes. With 100-bp fragments and a mutation rate of 10^{-5}, 81% of the perfect uptake

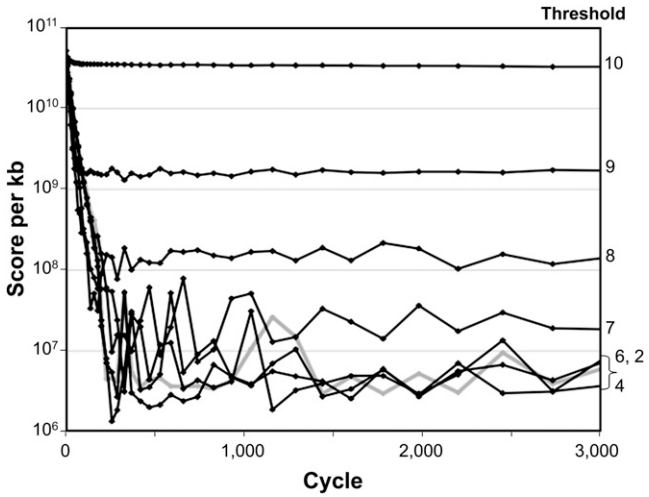


Figure 4.—Effect of alternative modes of scoring on uptake sequence accumulation. All runs began with 200-kb genomes randomly seeded with five USs per kilobase, with 1000 × 100-bp fragments (100-fold divergent) recombining in each cycle. The mutation rate was 1×10^{-4}. Shaded line: Fragments were scored additively using matrix A. Solid lines: Fragments were scored by requiring a minimum number of matches to the 10 bp of the US, with the threshold (number of matches required for uptake) indicated to the right of each line.

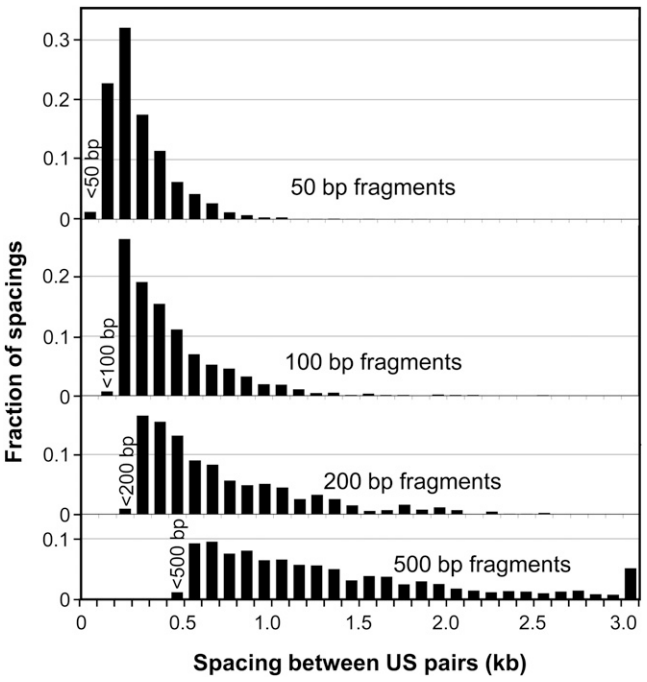


Figure 5.—Dependence of uptake sequence spacing on size of recombining fragments. All runs began with 200-kb random-sequence genomes and a mutation rate of 1×10^{-4}. Fragments were scored with matrix A. The sizes of the recombining fragments (100-fold divergent) are indicated on the graphs; the total length of DNA recombined in each cycle was 1.5 genome equivalents. The rightmost bar in the bottom panel represents all of the US pairs with spacing >3.0 kb.

sequences present after 40,000 cycles of evolution were still present after 60,000 additional cycles. The other uptake sequences had been lost and replaced by ones at new positions. Stability was slightly lower for runs using longer fragments: 75% with 200-bp fragments, 60% with 500-bp fragments, and 59% with 1000-bp fragments, presumably because of increased recombination of unselected flanking sequences, and was slightly higher when the positions of perfect and singly mismatched uptake sequences were combined. When the mutation rate was raised to 10^{-4}, uptake sequences were much less stable, with only 8% of sites present at 40,000 cycles still present at 100,000 cycles. Because real mutation rates are much lower than those used here, these results suggest that simple molecular drive may be sufficient to explain the long-term stability of uptake sequences.

Spacing of uptake sequences: Karlin *et al.* (1996) reported that perfect USS cores are more evenly spaced around their genomes than would be expected if they were randomly positioned, so we examined the center-to-center separations of the uptake sequences generated in our simulated genomes to find if drive alone generated nonrandom spacing. The histograms in Figure 5 show the equilibrium separations between uptake sequences evolved when the lengths of the recombining fragments were 50, 100, 200, or 500 bp. All showed an almost complete absence of separations smaller than the size of the recombining fragments and a corresponding excess of moderate separations. This result is nicely consistent with the similarity between the mean length of Neisseria recombination tracts and the mean spacing of DUSs (Treangen *et al.* 2008) and demonstrates that relatively even spacing need not reflect selection either for a chromosomal function or for more effective generation of variation.

Simulating uptake sequence evolution driven by DUS and USS matrices: The above analyses show that uptake sequences can be driven to high densities in simulated genomes by a simple model of biased uptake plus neutral recombination and that they have properties typical of uptake sequences in real genomes. However, these analyses all used matrices specifying a simple generic motif, with all 10 positions having the same degree of preference. But uptake sequences in real genomes suggest that real uptake biases are more complex. Measurements of variant uptake sequences have found that the uptake machinery is more sensitive to variation at some positions than at others (Ambur *et al.* 2007), and analyses of perfect and singly mismatched cc-DUSs and cc-USSs in real genomes have found different consensus strengths at different positions (Redfield *et al.* 2006; Ambur *et al.* 2007).

Motif-based DUS and USS data sets: Since previous analyses of genome sequences looked only at perfect and singly mismatched cc-DUS and cc-USS, they are likely to have overlooked some uptake-related variation in the genomes they analyzed. To develop DUS and USS

matrices and data sets that better capture the full variation, the *N. meningitidis* and *H. influenzae* genomes were analyzed with the Gibbs motif sampler (Thompson *et al.* 2003). This program uses a Bayesian sampling method to examine sequences for patterns, requiring as input only the length of the expected motif and an initiating expected number of occurrences of whatever motif might be discovered. Each analysis produces both an aligned list of the occurrences of the motif that was found and a matrix describing the variation at each position in the alignment.

Gibbs motif sampler searches of the forward and reverse strands of the *N. meningitidis* and *H. influenzae* genomes found motifs strongly resembling the DUS and USS consensuses previously identified from aligned cc-DUSs and cc-USSs. Consistent with the evidence for a strong consensus, most of the occurrences found by these searches had already been identified in searches for perfect and singly mismatched cc-DUSs and cc-USSs. For each genome we retained a set of sequences corresponding to 1.5 times the number of cc-DUSs or cc-USSs (sets of 2902 and 2206, respectively). Position-weight matrices of these sequences gave logos very similar to the perfect-consensus ones previously reported (Redfield *et al.* 2006; Ambur *et al.* 2007) (Figure 6); changing the number of sequences retained changed the strength of the consensus but not its sequence. For completeness we also carried out Gibbs motif sampler analyses of all the other available Neisseria and Pasteurellaceae genomes; logos of these are provided in Figure S2.

We also evaluated uptake sequences in functional subsets of the *N. meningitidis* and *H. influenzae* genomes. If uptake sequences play roles in DNA replication or scaffolding (Karlin *et al.* 1996), they might be expected to have different motifs in the genome strands replicated in leading and lagging directions, akin to the well-characterized strand-specific differences in base composition (Maliszewska-Tkaczyk *et al.* 2000). However, the motifs produced by Gibbs searches of leading and lagging strands were indistinguishable (data not shown), suggesting that any functional constraints on uptake sequences are independent of replication direction. The possible effect of direction of transcription was similarly investigated by separate Gibbs searches of coding sequences and their reverse complements (Francino and Ochman 2001). Only minor differences were seen between the two motifs (data not shown).

The Gibbs analyses produced position weight matrices for the DUS and USS motifs they found (see Table 1). These matrices had stringencies intermediate between those of the generic matrices A and B, and we used them to address the following questions: First, would simulations that used these genome-derived matrices maintain the numbers of uptake sequences already present in the concatenated intergenic sequences of the real *N. meningitidis* and *H. influenzae* genomes (378 and 215 kb, respectively)? Second, would simulation of molecular drive using these genome-derived matrices cause the corresponding uptake sequence to accumulate in random sequences?

Maintenance of DUS and USS in real intergenic sequences: Simulated evolution of the concatenated *N. meningitidis* intergenic sequences with the DUS matrix maintained high DUS densities. After 55,000 cycles the scores and DUS densities were stable and about two-thirds as high as in the real sequences [scores, 0.20/kb *vs.* 0.31 (note that the DUS and USS scores cannot be compared to those for the generic US matrix A, as the numbers of positions are different); densities, 224 12-mer DUSs *vs.* 331] (Table S1). All but 19 of the DUSs present after 55,000 cycles were at new positions, confirming that the simulation did not simply preserve existing DUSs. As a negative control, simple decay of DUSs was simulated using a "null" matrix where every base was weighted equally. Over 5000 cycles the score decayed from 1.77×10^{2} to 3.17×10^{-3}, a score typical of random sequences of the same length and base composition, confirming that the high score at 55,000 cycles was due to the DUS matrix.

The complexity of the USS matrix derived from the *H. influenzae* genome caused simulations using it to run so slowly that the final equilibrium was not attained. After 5000 cycles, the evolved intergenic sequence "genome" had a slightly lower score (0.95×10^{-8} *vs.* 1.2×10^{-8}/kb) and only about one-quarter the number of cc-USSs as the original intergenic sequences (51 *vs.* 211) (Table S1). However, all of these USSs were at new positions, again indicating active drive rather than persistence. Furthermore, comparison with the control simulation showed that this is very strong drive for USSs, as just 500 cycles using the null matrix produced a sequence with a score typical of a random sequence and containing no perfect cc-USSs and only 7 singly mismatched ones.

DUS and USS accumulation in random sequences: Both DUS and USS matrices also caused uptake sequences to accumulate to high levels in random-sequence genomes. Simulated evolution under the DUS matrix increased the genome score to 0.27/kb, comparable to that of the real intergenic sequences, with 0.875 perfectly matched 10-bp DUS cores and 0.15 singly mismatched cores per kilobase. Because the complexity of the USS matrix caused accumulation of USSs in a random sequence to be extremely slow, the simulation program was modified to allow the stringency of the bias to be reduced within a cycle if too few fragments had initially recombined. This reduced bias allowed less-well-matched sequences to promote recombination, and their accumulation in turn decreased the need for bias reduction in subsequent cycles. The output from this run was then fed into a run with normal (not self-reducing) bias. The result was an evolved genome with a score about twice as high as that of the real intergenic

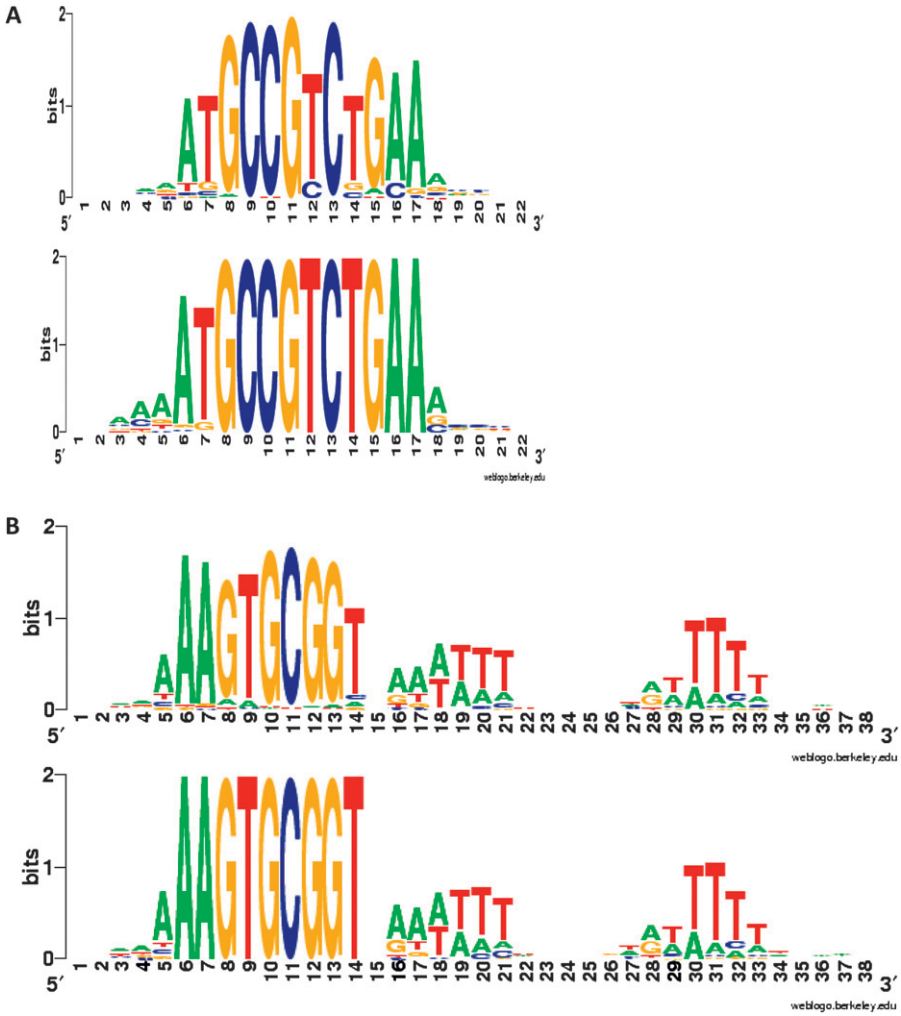


FIGURE 6.—Sequence logos of uptake sequence motifs derived from unbiased genome searches. The corresponding matrices are in Table 1. (A) *N. meningitidis* MC58. Top panel, 2902 aligned sequences; bottom panel, 1935 perfect matches to the cc-DUS (GCCGTCTGAA). (B) *H. influenzae* Rd KW20. Top panel, 2206 aligned sequences; bottom panel, 1471 perfect matches to the cc-USS (AAGTGCGGT).

sequence (1.2×10^{-6}/kb) and with 0.56 perfectly matched and 0.115 singly mismatched 9-bp USS cores per kilobase.

Spacing of DUS and USS in real and simulated genomes: As described above, uptake sequences in simulated genomes were rarely closer than the lengths of the recombining fragments. This caused the overall spacing to be relatively even, as has been reported for real cc-USSs in the *H. influenzae* genome (KARLIN *et al.* 1996) (TREANGEN *et al.*'s 2008 analysis of DUS spacing did not address its randomness). We used the genomic DUS and USS sets produced by the Gibbs motif sampler to reinvestigate the spacing of uptake sequences in real genomes, comparing their spacing distributions to random distributions with the same means. Figure 7, A and B, confirms that both DUS and USS have many uptake sequences in closely spaced pairs (centers within 30 bp; note that these are the centers of the full 12-bp DUS and 30-bp USS motifs, not of the cores). All but one of the 169 close-USS pairs and 97% of the 646 close-DUS pairs are in inverted repeat orientations, and many of these have been previously identified as potential transcription terminators (KINGSFORD *et al.* 2007). However, when these close pairs are each treated as a single occurrence, the distributions of uptake sequence positions are very similar to random distributions (black lines in Figure 7, A and B). Thus, although the locations of uptake sequence have clearly been influenced by both coding constraints and their frequent roles as transcription terminators, there is no evidence of selection for a chromosomal function requiring relatively even spacing.

Several factors could account for the difference between the lack of close neighbors in simulated genomes (Figure 5) and the excess of them in real genomes (Figure 7). The former effect will be blurred in real genomes, which will have undergone recombination with fragments of varying lengths. On the other hand, a major factor in real-genome spacing is the contribution of closely spaced and oppositely oriented uptake sequences to transcription termination. Selection against uptake sequences in coding regions may also play a role.

Laboratory measurements of DNA uptake specificity: One remaining discrepancy between real and simulated uptake sequences is that the fine structure of the real USS motif does not agree well with what little

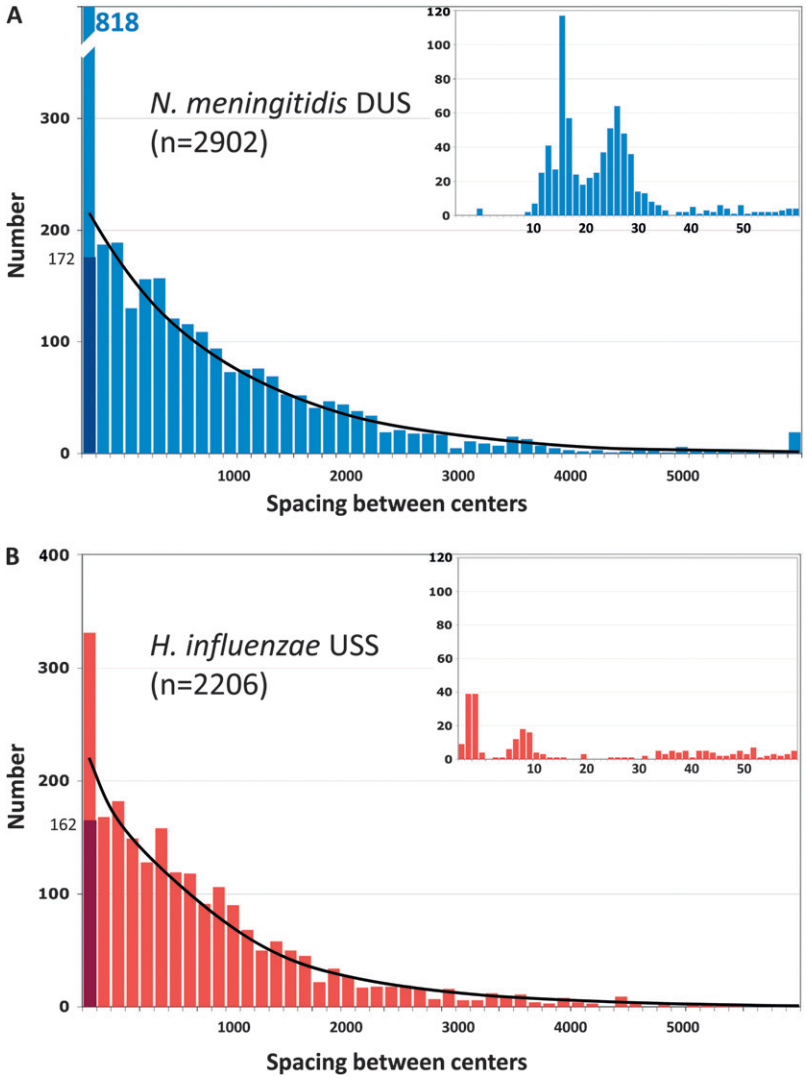


FIGURE 7.—Center-to-center separations of uptake sequences in real genomes. The uptake sequences were those identified by Gibbs searches. (A) A total of 2902 *N. meningitidis* DUSs; (B) 2206 *H. influenzae* USSs. The black line in each graph shows a random distribution of the same number of uptake sequences over the same length of DNA. The dark bar on the left in each graph is the number of separations within 100 bp when inverted-repeat pairs within 30 bp are treated as single occurrences.

is known of the actual *H. influenzae* uptake specificity. BAKKALI (2007) reported uptake of USS-containing DNA fragments with singly mismatched cores. Although uptake was dramatically reduced by some changes at positions with strong genome consensus, it was unchanged by others (BAKKALI 2007). A comparable analysis of *N. meningitidis* uptake specificity has not been done, but the limited information available (AMBUR *et al.* 2007) is also discrepant with the genomic motif, in that the 10-bp core appears to be less important for uptake than the AT bases that precede it (see Figure 7 in that article). We investigated three possible explanations for these discrepancies.

First, the confidence limits on the published *H. influenzae* DNA uptake data were very broad, so we repeated the uptake experiments using a slightly different system. The new results confirm both the previous results and the discrepancy; the combined uptake data are presented in Figure 8.

Second, we considered whether the position biases specified by the matrix might have disproportionate effects on the sequences that accumulate in the focal genome. To find out whether bases at uptake sequence positions only weakly favored by the matrix might nevertheless accumulate to a strong consensus the uptake sequences in simulated genomes that had evolved with the DUS and USS matrices were identified with the Gibbs motif sampler. The resulting logos were very similar to those of the DUS and USS data sets used to generate the matrices (data not shown), confirming that the uptake sequences that accumulate in a simulated genome accurately reflect the bias of the matrix used. Thus, if molecular drive due to biased uptake was the only evolutionary force, uptake sequences in the genome should reflect the biases of the uptake machinery.

The third explanation we considered was that uptake might depend not only on the individual base pairs in the motif, but also on specific interactions between these bases. Such interactions might contribute to sequence recognition or facilitate DNA deformation needed to initiate uptake and would cause different positions within the uptake sequence to coevolve. To

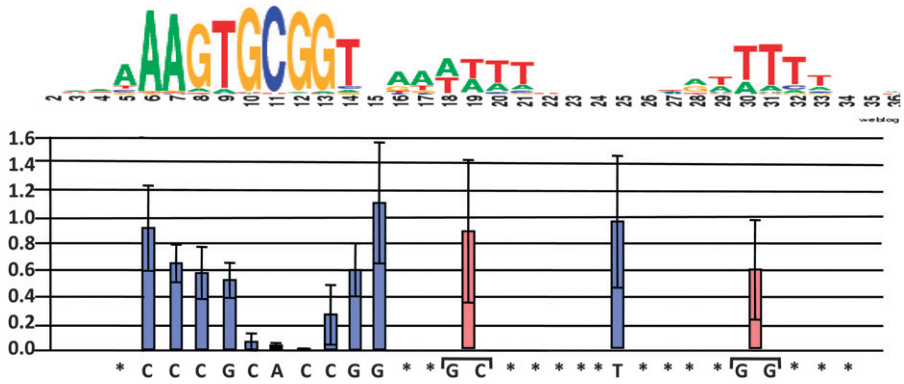


FIGURE 8.—DNA uptake assays. Mean uptake of thirteen 222-bp DNA fragments relative to uptake of the full-length perfect-consensus USS sequence is shown. Each fragment contains a USS sequence that differs from the consensus USS at one position (blue bars) or at a pair of positions (bracketed, red bars). The nucleotide that replaced the consensus base is shown for each position changed. Bars illustrate confidence values for each result. The logo above is for the 2206 genomic USSs identified by the Gibbs search of the *H. influenzae* genome.

test this we looked for evidence of covarying positions in the sets of USS and DUS sequences identified by the Gibbs searches, using MatrixPlot (GORODKIN *et al.* 1999). This covariation analysis used both whole-genome data sets and intergenic sequence data sets (because most closely spaced USS pairs had confounding overlaps, they were first removed from the *H. influenzae* intergenic data set). The analysis found no strong correlations in the uptake sequence cores and only very weak covariation between adjacent positions at the centers of the flanking AT-rich motifs (Figure S2), suggesting that the contributions of individual base pairs to uptake are largely independent. The lack of covariation also confirms that genomes do not contain subsets of USS or DUS motifs with differing sequences. The above three analyses thus leave unresolved the discrepancy between the measured *H. influenzae* uptake bias and its genomic USS motif.

DISCUSSION

We used computer simulations to test whether molecular drive can cause the DNA sequences preferred by a cell's DNA uptake machinery to gradually accumulate in its genome, testing wide ranges of values for the major parameters to compensate for the lack of empirical data. The results showed that molecular drive is very robust; accumulation of preferred sequences was independent of mutation rate and only moderately dependent on how the bias was applied, on the amount of recombination, and on the length, number, and extent of divergence of the recombining fragments. This result greatly simplifies our understanding of the evolution of uptake specificity in competent bacteria.

Four major questions remain. First, how did the ancestral state of nonspecific DNA uptake evolve into the strong uptake specificity seen in the Pasteurellaceae and Neisseria species? Second, have increases in uptake bias been driven by selection for more efficient uptake or for prescreening DNA fragments to exclude foreign sequences? Third, why would some bacteria have evolved uptake specificity (strong uptake bias for an abundant genomic motif) while most others have not? Fourth, has selection for genetic benefits of transformation affected the properties of uptake sequences?

To address the first question, consider the likely steps in the evolution of uptake specificity, starting with a naturally competent ancestral species. Although these bacteria would not have discriminated between DNAs from different sources, the ubiquity of sequence preferences in all well-characterized DNA-binding proteins predicts that their DNA uptake machinery might have already had a modest sequence bias. Such biases are especially likely for the uptake proteins that directly contact DNA, because specific contacts between nucleotides and amino acid residues will provide the tight binding needed to pull strongly on DNA (MAIER *et al.* 2004a,b; STINGL *et al.* 2010). Here we assume that any such ancestral biases were too slight to cause significant accumulation of the preferred sequences in the genome.

The first step toward uptake specificity would be a mutational change in a cell surface protein that modestly increased both its sequence bias and its DNA-binding affinity. This protein might have already been a component of the DNA uptake machinery, but it could also have had no previous role in uptake and only by chance acquired the ability to bind DNA. The increased bias need not have decreased uptake and may even have increased it, either because the higher affinity increased the effective DNA concentration at the cell surface or because it was accompanied by an improvement in the efficiency of the uptake machinery.

The now-preferred sequences would begin to accumulate in the genome if two conditions were met: (i) the bacteria lived in biofilms or other environments where a significant amount of the available DNA came from close relatives (conspecifics) and (ii) some of the DNA brought into the cells underwent homologous recombination with the chromosome. These conditions are likely to be widely met, as all bacteria that grow on surfaces or in poorly mixed environments encounter abundant DNA from close relatives, and all have the DNA repair and replication machinery responsible for

recombination. They are also necessary if competence is to have any other genetic consequences.

The accumulation of preferred sequences in this ancestral genome would then increase the benefits or reduce the cost of the initial uptake bias, because these sequences would now be more common in the available DNA. Genetic benefits of DNA uptake would be higher because more of the DNA taken up would be able to recombine. Nutrient effects could also be enhanced, both because of more efficient uptake and because the base composition of the incoming DNA would more often match that of the cell's own genome. Although nucleotides are needed primarily for synthesis of RNA, deoxynucleotides are most efficiently used for DNA, and matching of base composition would avoid the need for interconversion of deoxynucleotide bases by complex salvage pathways. Once the preferred sequences became more common in the available DNA, additional mutations that further strengthened the bias would become beneficial, and this increased bias would in turn cause further accumulation of preferred sequences. Although our simulations did not incorporate this gradual increase in bias strength, they showed that uptake sequences could also evolve under the more severe condition of a sudden imposition of a strong bias.

This scenario must also consider the extent to which accumulated uptake sequences would interfere with genome functions, especially coding. Our analysis of uptake sequences in the proteomes of *H. influenzae*, *Actinobacillus pleuropneumoniae*, and *N. meningitidis* (Findlay and Redfield 2009) found that these are accommodated in two ways. First, nonsilent mutations created uptake sequences at positions where the specific amino acid substitutions are well tolerated; this roughly doubled the frequencies of those tripeptides specified by each species' uptake sequence in different reading frames. Second, silent codon changes created uptake sequences at positions where these tripeptides were already present; the cost of using less-favored codons may be relatively small because uptake sequences are rare in highly conserved genes. However, the cost of eliminating mutations that create uptake sequences at other coding positions will still oppose the benefits of biased uptake.

Gradually, over many millions of generations, the feedback between uptake bias and molecular drive could produce the strong uptake specificity seen in the modern Pasteurellaceae and Neisseria. If substantial genetic benefits arise from genetic variation and/or recombinational repair (Davidsen *et al.* 2004; Michod *et al.* 2008; Treangen *et al.* 2008), increasing uptake bias might continue to be selected by its ability to prescreen DNA, preventing uptake of potentially harmful foreign genes. Uptake sequences would then also accumulate by "hitchhiking," being recombined along with the linked beneficial alleles whose uptake they have facilitated, as has been simulated in a hybrid model incorporating features of both molecular drive and beneficial recombination (Chu *et al.* 2006). If DNA provides mainly nutritional benefits, the strong uptake biases typical of the Pasteurellaceae and Neisseria might evolve only if conspecific DNA was usually available or if foreign DNAs contained sequences that fit the uptake bias well enough that they could be taken up when conspecific DNA was unavailable. Most importantly for evolutionary models, although the benefits of uptake specificity would depend on the accumulation of genomic uptake sequences, only the genes responsible for the sequence bias need be under selection.

How long might this accumulation take? The computer simulations give only a very rough guide to real evolutionary timescales, both because real generation times and recombination rates are not known and because the uptake sequences in bacterial genomes may not have reached equilibrium between drive and mutation. If cells in mucosal environments undergo about one division per day, each model cycle would represent 10^3–10^4 days ($\sim$10 years), and the equilibria shown in Figure 2 would take $\sim 2 \times 10^6$ years to achieve. However, shorter generation times could reduce this estimate by $\sim$10-fold and lower uptake or recombination rates could increase it substantially. Timescales of many millions of years are consistent with the persistence of USS at homologous sites in Pasteurellacean genomes (Bakkali *et al.* 2004), and Pasteurellacean uptake sequences predate the divergence of this family's two major clades (Redfield *et al.* 2006) hundreds of millions of years ago. The costs and benefits of sequence-biased uptake are unlikely to have remained constant over such long periods, being affected not only by the slow accumulation of the preferred sequences in the genome but also by more rapid changes in both external factors (sources and amounts of DNA in the environment) and internal factors [changes in uptake specificity, frequency of recombination, potential for genetic benefits (Redfield *et al.* 2006; Maughan and Redfield 2009)]. Thus the levels of uptake bias and uptake sequence abundance seen in modern bacteria need not represent either true stable equilibria or stages on the way to such equilibria, but may instead integrate the effects of varying selection over long evolutionary periods. Phylogenetic studies of homologous uptake sequences could shed light on the history of uptake sequences, but we also badly need data about DNA uptake and recombination by bacteria in their natural environments.

To dissect the components that contribute to the accumulation of uptake sequences by molecular drive, the simulation model could be modified in several ways:

i. The gradual increase of uptake bias discussed above could be simulated by beginning with a very weak bias and tracking the proportion of fragments scored that went on to recombine, raising the bias when this proportion passes a specified threshold.

ii. The length of the simulated recombining fragments was found to set the minimum spacing between genomic uptake sequences. Although this result is nicely consistent with the similarity between the mean length of Neisseria recombination tracts and the mean spacing of DUSs (Treangen *et al.* 2008), comparisons to real genomes would be improved by having the model use a range of fragment sizes within each cycle.
iii. Transforming DNA is known to undergo partial sequence degradation in the cytoplasm, which could allow closer spacing of uptake sequences because of shorter recombination tracts but could also reduce their accumulation because some are lost to degradation after promoting uptake. Simulations could clarify these potentially opposing effects.
iv. Most of the available uptake data are consistent with the model's assumption that the single best uptake sequence determines each fragment's probability of uptake (Ambur *et al.* 2007), but a model that integrates the effects of multiple uptake sequences on a fragment might also be tested.
v. Although most uptake sequences in bacterial genomes appear to be accommodated with little cost to fitness, the locations of uptake sequences in real genomes are clearly sensitive to coding and other functional constraints (Kingsford *et al.* 2007; Treangen *et al.* 2008; Findlay and Redfield 2009). Selection for coding could be simulated by designating 1-kb blocks of sequence to have reduced probabilities of recombination.

Major improvement to understanding of uptake sequence evolution will come not from improved models or more sequence analyses but from detailed molecular characterization of actual uptake biases in a wide range of competent species. The best available uptake data are for *H. influenzae*; they suggest substantial discrepancies between the genomic motif and measured position-specific uptake biases (see Figure 8). Such discrepancies are unlikely to be caused by selection for genetic benefits of recombination and may disappear when better uptake data become available. Alternatively, discrepancies may reflect intrinsic biases of postuptake events such as DNA processing, recombination, and mismatch repair, or selection for such cellular functions as transcription termination and protein coding.

The genes and proteins responsible for uptake specificities have not been identified. Their identities should shed light on how selection has acted; finding that bias is due to a prescreening protein extrinsic to the uptake machinery would suggest selection to exclude foreign DNA, whereas finding a mechanism-intrinsic protein would be more consistent with selection for an efficient mechanism. These two cases also make contrasting predictions about the effects of mutations that reduce uptake specificity on the amount of DNA taken up—reducing the bias of a prescreening protein should increase overall uptake, whereas reducing the bias of the mechanism should decrease it.

Why is uptake specificity known only for two bacterial groups? In some other competent bacteria, the conditions necessary for drive may not be met. For example, highly active cytoplasmic nucleases may limit recombination, or fragments from conspecifics may be only a small part of the local DNA pool (this likely applies to planktonic species). At the other extreme, prescreening DNA for uptake sequences would not increase genetic benefits if the DNA in the cells' microenvironment came only from conspecifics, as may be the case in some biofilm environments. DNA uptake may also have weaker mechanistic constraints in gram-positive bacteria, which transport single-stranded DNA only across a single membrane. Nevertheless, uptake biases may be more widespread than appreciated, as no attempts have been made to detect modest effects, and the possible role of DNA uptake biases in genome dinucleotide signatures has not been explored (Campbell *et al.* 1999; van Passel *et al.* 2006).

To find out whether selection for genetic benefits of transformation has affected the properties of uptake sequences (by hitchhiking or other processes), a new model will be needed, one that incorporates both drive and the selective effects of making new combinations of alleles. Such a model would be substantially more complex than ours, as selection can be simulated only in a population-based model capable of tracking multiple alleles of multiple loci. The model could be run with and without the kinds of recombination benefits previously identified by evolution-of-sex theory (Otto and Gerstein 2006), and the outcomes could then be compared to each other and to the properties of uptake sequences in real genomes.

We thank Bill Thompson for his extensive assistance with the Gibbs motif sampler and Jan Gorodkin and Dave Ardell for advice on using MatrixPlot. Steve Schaeffer provided assistance with correlation analysis and Mike Whitlock assistance with assessing randomness of distributions. We thank S. Highlander and the Sanger Centre for draft sequences of *Mannheimia haemolytica* and *Neisseria lactamica*. We also thank Becky Dinesen for technical assistance and Reece Burborough for programming. The probing questions of an anonymous reviewer greatly clarified the framework in which the results are interpreted. The Westgrid facility and the MOA cluster at Dalhousie University provided much-needed computer facilities, the latter supported by the Canada Foundation for Innovation grant "A Canadian platform for advanced comparative genomics"; Roman Baranowski and Rob Beiko provided invaluable advice on using these facilities. This work was supported by an operating grant to R.J.R. from the Canadian Institutes for Health Research and by fellowships to H.M. from National Institutes of Health and Killam Trust.

LITERATURE CITED

Ambur, O. H., S. A. Frye and T. Tonjum, 2007 New functional identity for the DNA uptake sequence in transformation and its presence in transcriptional terminators. J. Bacteriol. **189:** 2077–2085.

Bakkali, M., 2007 Genome dynamics of short oligonucleotides: the example of bacterial DNA uptake enhancing sequences. PLoS One **2:** e741.

Bakkali, M., T. Y. Chen, H. C. Lee and R. J. Redfield, 2004 Evolutionary stability of DNA uptake signal sequences in the *Pasteurellaceae*. Proc. Natl. Acad. Sci. USA **101:** 4513–4518.

Barcak, G. J., J. F. Tomb, C. S. Laufer and H. O. Smith, 1989 Two *Haemophilus influenzae* Rd genes that complement the recA-like mutation rec-1. J. Bacteriol. **171:** 2451–2457.

Campbell, A., J. Mrazek and S. Karlin, 1999 Genome signature comparisons among prokaryote, plasmid, and mitochondrial DNA. Proc. Natl. Acad. Sci. USA **96:** 9184–9189.

Chen, I., and D. Dubnau, 2004 DNA uptake during bacterial transformation. Nat. Rev. Microbiol. **2:** 241–249.

Chu, D., J. Rowe and H. C. Lee, 2006 Evaluation of the current models for the evolution of bacterial DNA uptake signal sequences. J. Theor. Biol. **238:** 157–166.

Claverys, J. P., and B. Martin, 2003 Bacterial "competence" genes: Signatures of active transformation, or only remnants? Trends Microbiol. **11:** 161–165.

Danner, D. B., R. A. Deich, K. L. Sisco and H. O. Smith, 1980 An eleven-base-pair sequence determines the specificity of DNA uptake in *Haemophilus* transformation. Gene **11:** 311–318.

Davidsen, T., E. A. Rodland, K. Lagesen, E. Seeberg, T. Rognes *et al.*, 2004 Biased distribution of DNA uptake sequences towards genome maintenance genes. Nucleic Acids Res. **32:** 1050–1058.

Findlay, W. A., and R. J. Redfield, 2009 Coevolution of uptake sequences and bacterial proteomes. Genome Biol. Evol. **1:** 45–55.

Francino, M. P., and H. Ochman, 2001 Deamination as the basis of strand-asymmetric evolution in transcribed *Escherichia coli* sequences. Mol. Biol. Evol. **18:** 1147–1150.

Goodgal, S. H., and M. Mitchell, 1984 Uptake of heterologous DNA by *Haemophilus influenzae*. J. Bacteriol. **157:** 785–788.

Gorodkin, J., H. H. Staerfeldt, O. Lund and S. Brunak, 1999 MatrixPlot: visualizing sequence constraints. Bioinformatics **15:** 769–770.

Ho, S. N., H. D. Hunt, R. M. Horton, J. K. Pullen and L. R. Pease, 1989 Site-directed mutagenesis by overlap extension using the polymerase chain reaction. Gene **77:** 51–59.

Johnsborg, O., V. Eldholm and L. S. Havarstein, 2007 Natural genetic transformation: Prevalence, mechanisms and function. Res. Microbiol. **158:** 767–778.

Karlin, S., J. Mrazek and A. M. Campbell, 1996 Frequent oligonucleotides and peptides of the *Haemophilus influenzae* genome. Nucleic Acids Res. **24:** 4263–4272.

Kingsford, C. L., K. Ayanbule and S. L. Salzberg, 2007 Rapid, accurate, computational discovery of Rho-independent transcription terminators illuminates their relationship to DNA uptake. Genome Biol. **8:** R22.

Kovacs, A. T., W. K. Smits, A. M. Mironczuk and O. P. Kuipers, 2009 Ubiquitous late competence genes in *Bacillus* species indicate the presence of functional DNA uptake machineries. Environ. Microbiol. **11:** 1911–1922.

Maier, B., I. Chen, D. Dubnau and M. P. Sheetz, 2004a DNA transport into *Bacillus subtilis* requires proton motive force to generate large molecular forces. Nat. Struct. Mol. Biol. **11:** 643–649.

Maier, B., M. Koomey and M. P. Sheetz, 2004b A force-dependent switch reverses type IV pilus retraction. Proc. Natl. Acad. Sci. USA **101:** 10961–10966.

Maliszewska-Tkaczyk, M., P. Jonczyk, M. Bialoskorska, R. M. Schaaper and I. J. Fijalkowska, 2000 SOS mutator activity: unequal mutagenesis on leading and lagging strands. Proc. Natl. Acad. Sci. USA **97:** 12678–12683.

Maughan, H., and R. J. Redfield, 2009 Extensive variation in natural competence in *Haemophilus influenzae*. Evolution **63:** 1852–1866.

Maughan, H., S. Sinha, L. Wilson and R. J. Redfield, 2008 Competence, DNA uptake and transformation in the Pasteurellaceae, pp.79–98 in *Pasteurellaceae: Biology, Genomics and Molecular Aspects*, edited by P. Kuhnert and H. Christensen. Caister Academic Press, Norfolk, UK.

Michod, R. E., H. Bernstein and A. M. Nedelcu, 2008 Adaptive value of sex in microbial pathogens. Infect. Genet. Evol. **8:** 267–285.

Otto, S. P., and A. C. Gerstein, 2006 Why have sex? The population genetics of sex and recombination. Biochem. Soc. Trans. **34:** 519–522.

Palchevskiy, V., and S. E. Finkel, 2009 A role for single-stranded exonucleases in the use of DNA as a nutrient. J. Bacteriol. **191:** 3712–3716.

Poje, G., and R. J. Redfield, 2003a General methods for culturing *Haemophilus influenzae*. Methods Mol. Med. **71:** 51–56.

Poje, G., and R. J. Redfield, 2003b Transformation of *Haemophilus influenzae*. Methods Mol. Med. **71:** 57–70.

Redfield, R. J., 1988 Evolution of bacterial transformation: Is sex with dead cells ever better than no sex at all? Genetics **119:** 213–221.

Redfield, R. J., 1993 Genes for breakfast: the have-your-cake-and-eat-it-too of bacterial transformation. J. Hered. **84:** 400–404.

Redfield, R. J., W. A. Findlay, J. Bosse, J. S. Kroll, A. D. Cameron *et al.*, 2006 Evolution of competence and DNA uptake specificity in the *Pasteurellaceae*. BMC Evol. Biol. **6:** 82.

Smith, H. O., J. F. Tomb, B. A. Dougherty, R. D. Fleischmann and J. C. Venter, 1995 Frequency and distribution of DNA uptake signal sequences in the *Haemophilus influenzae* Rd genome. Science **269:** 538–540.

Smith, H. O., M. L. Gwinn and S. L. Salzberg, 1999 DNA uptake signal sequences in naturally transformable bacteria. Res. Microbiol. **150:** 603–616.

Stingl, K., S. Muller, G. Scheidgen-Kleyboldt, M. Clausen and B. Maier, 2010 Composite system mediates two-step DNA uptake into *Helicobacter pylori*. Proc. Natl. Acad. Sci. USA **107:** 1184–1189.

Thompson, W., E. C. Rouchka and C. E. Lawrence, 2003 Gibbs Recursive Sampler: finding transcription factor binding sites. Nucleic Acids Res. **31:** 3580–3585.

Thompson, W., L. A. McCue and C. E. Lawrence, 2005 Using the Gibbs motif sampler to find conserved domains in DNA and protein sequences, in *Current Protocols in Bioinformatics*, Unit 2.8, Wiley-InterScience, New York.

Treangen, T. J., O. H. Ambur, T. Tonjum and E. P. Rocha, 2008 The impact of the neisserial DNA uptake sequences on genome evolution and stability. Genome Biol. **9:** R60.

van Passel, M. W., E. E. Kuramae, A. C. Luyf, A. Bart and T. Boekhout, 2006 The reach of the genome signature in prokaryotes. BMC Evol. Biol. **6:** 84.

Williams, P. M., L. A. Bannister and R. J. Redfield, 1994 The *Haemophilus influenzae* sxy-1 mutation is in a newly identified gene essential for competence. J. Bacteriol. **176:** 6789–6794.

Communicating editor: J. Lawrence

GENETICS

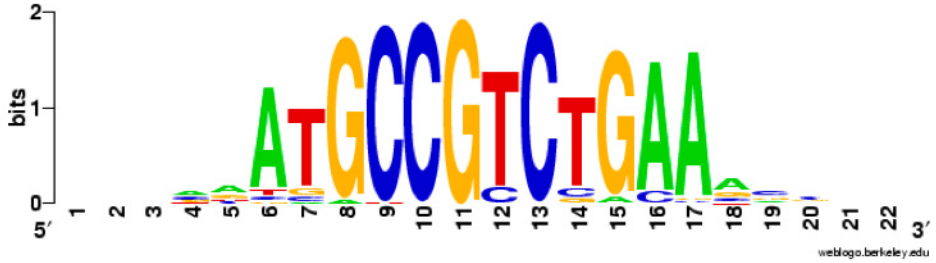
Supporting Information
http://www.genetics.org/cgi/content/full/genetics.110.119438/DC1

Bacterial DNA Uptake Sequences Can Accumulate by Molecular Drive Alone

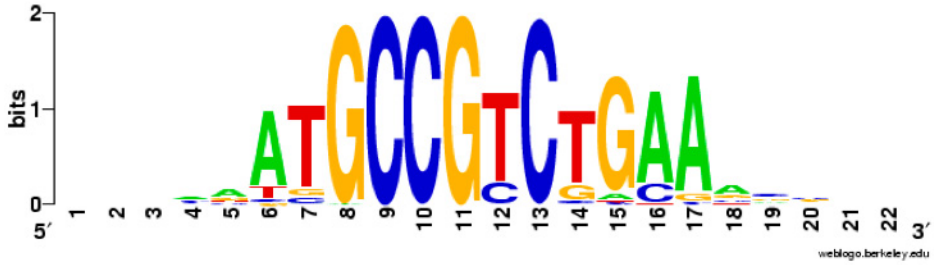
H. Maughan, L. A. Wilson and R. J. Redfield

DOI: 10.1534/genetics.110.119438

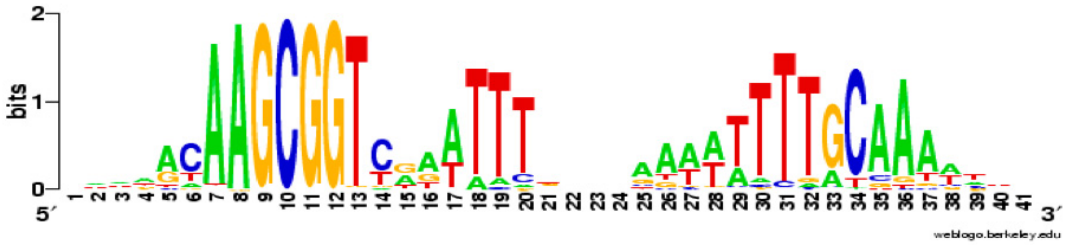
Neisseria gonorrhoeae (n=2944)



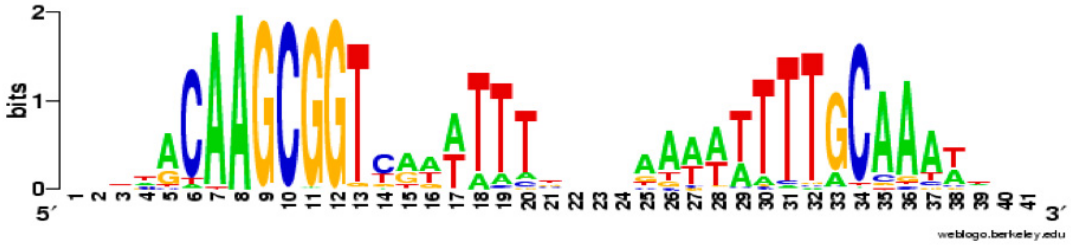
Neisseria lactamica (n=3388)



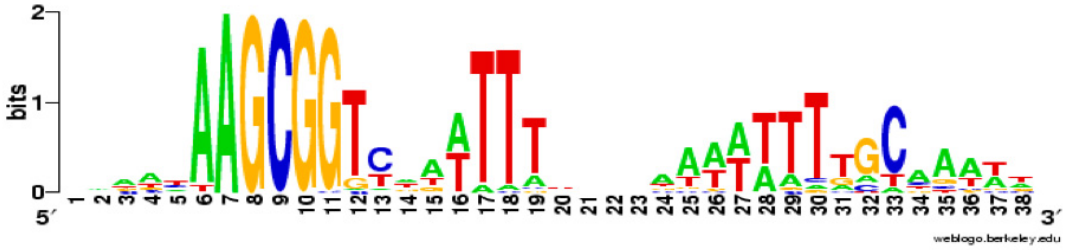
Actinobacillus pleuropneumoniae (n=1114)



Mannheimia haemolytica (n=1460)

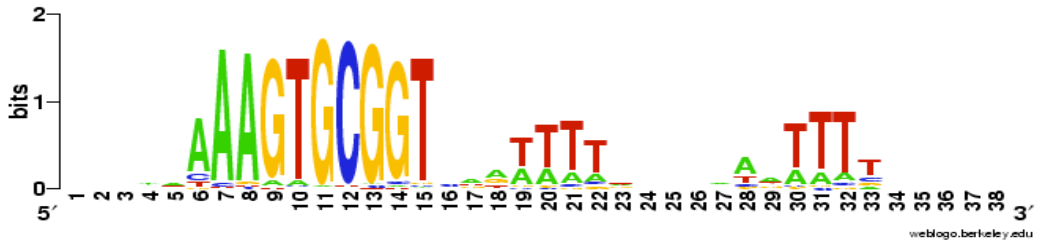


Haemophilus ducreyi (n=298)

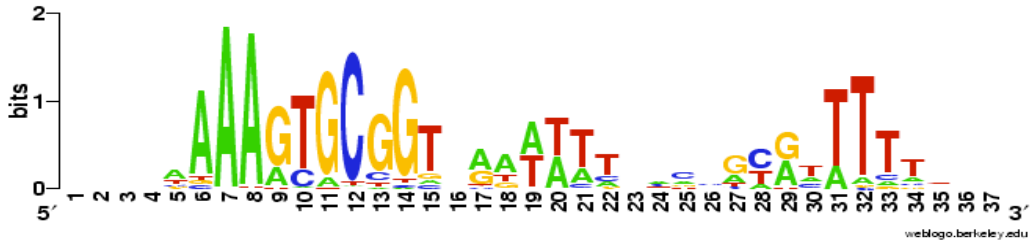


Mannheimia succiniciproducens (n=2228)

Pasteurella multocida (n=1390)



Actinobacillus actinomycetemcomitans (n=2640)



Haemophilus somnus (n=1824)

Actinobacillus succinogenes (n=2536)

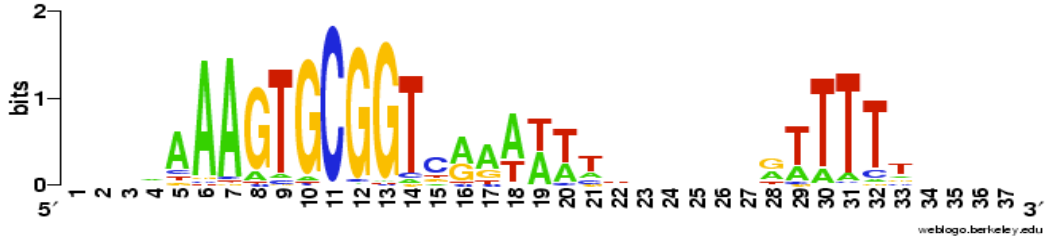


FIGURE S1.—Sequence logos of uptake sequence motifs derived from unbiased genome searches. The numbers in brackets are the numbers of aligned sequences used for each logo. These were chosen from the larger number of sequences identified by a Gibbs Motif Sampler search of each genome, by first ranking the sequences by Gibbs score and then retaining a number equal to 1.5 times the number of cc-USS or cc-DUS present in the corresponding genome.

Accession numbers: *N. gonorrhoeae*: NC002946; *A. pleuropneumoniae*: NC009053; *H. ducreyi*: NC002940; *M. succiniproducens*: NC006300; *P. multocida*: NC002663; *A. actinomycetemcomitans*: NC013416; *H. somnus*: NC008309; *A. succinogenes*: NC009655. The draft sequence of *N. lactamica* was obtained from www.sanger.ac.uk/Projects/N_lactamica/ . The draft sequence of *M. haemolitica* was obtained from Sarah Highlander.

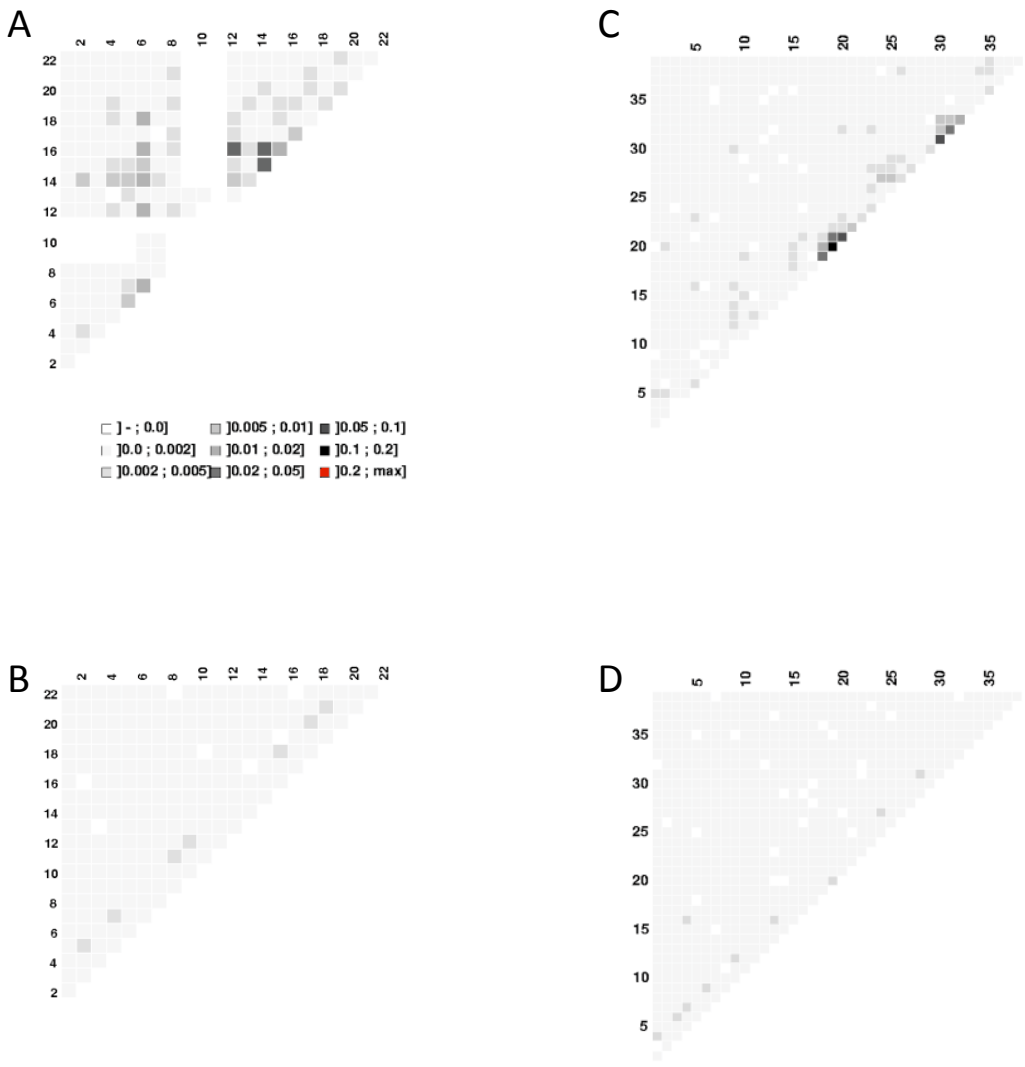


FIGURE S2.—Matrixplot analysis of correlations between residues at different uptake sequence positions. A. *N. meningitidis* DUSs; B. random *N. meningitidis* segments; C. *H. influenzae* USSs; D. 2200 random *H. influenzae* segments. The colour scale shown in part A also applies to parts B, C, and D.

TABLE S1
Uptake sequences in evolved genomes

Fig. 2 data

Run pair	μ-genome	Total perfect	Total oneoff	perf/kb	1off/kb	Score	Score/kb
XA6&XB6	0.0002	289	27	1.45	0.14	2.96E+12	1.46E+10
XC6&XD6	0.0001	272	24	1.36	0.12	2.77E+12	1.42E+10
XE6&XF6	0.00005	265	21	1.33	0.11	2.67E+12	1.32E+10
XG6&XH6	0.00002	274	34	1.37	0.17	2.78E+12	1.39E+10
XI6&XJ6	0.00001	260	25	1.30	0.13	2.64E+12	1.32E+10
XK6&XL6	0.000005	280	32	1.40	0.16	3.08E+12	1.54E+10

Fig. 3A data

Run pair	frac recomb	Total perfect	Total oneoff	perfect/kb	1off/kb	Score	Score/kb
ZK&ZU	0.0005	13	10	0.07	0.05	0.0005	2.51E+08
ZJ&ZT	0.0015	6	5	0.03	0.03	0.0015	5.11E+08
ZI&ZS	0.005	3	0	0.02	0.00	0.005	9.60E+08
ZH&ZR	0.015	7	3	0.04	0.02	0.015	1.70E+09
ZG&ZQ	0.05	11	4	0.06	0.02	0.05	3.37E+09
ZF&ZP	0.15	25	12	0.13	0.06	0.15	6.60E+09
ZE&ZO	0.5	59	6	0.30	0.03	0.5	1.51E+10
ZD&ZN	1.5	96	11	0.48	0.06	1.5	2.45E+10
ZC&ZM	5	143	5	0.72	0.03	5	3.86E+10
ZB&ZL	15	152	4	0.76	0.02	15	4.23E+10

Fig. 3C data

Run	frag size	Total perfect	Total oneoff	perfect/kb	1off/kb	Score	Score/kb
CB3	50	818	34	4.09	0.17	8.27E+12	4.14E+10
CC3	100	478	27	2.39	0.14	4.81E+12	2.41E+10
CD1-3	200	287	24	1.44	0.12	2.91E+12	1.46E+10
CE1-8	500	144	11	0.72	0.06	1.43E+12	7.15E+09
TL4	1000	56	8	0.28	0.04	5.40E+11	2.70E+09
TW3	2000	53	12	0.27	0.06	3.10E+11	1.55E+09

Fig. 4 data

Run	Threshold	Total perfect	Total oneoff	Total twooff	perfect/kb	1off/kb	2off/kb	Score	Score/kb
PA7	0	1	4	77	0.01	0.02	0.39	7.34E+08	3.67E+06
PD7	2	0	2	82	0.00	0.01	0.41	8.47E+08	4.24E+06
PF7	4	0	3	82	0.00	0.02	0.41	1.34E+09	6.70E+06
PH7	6	0	9	88	0.00	0.05	0.44	1.33E+09	6.65E+06
PI7	7	1	23	216	0.01	0.12	1.08	3.09E+09	1.55E+07
PJ7	8	4	82	1062	0.02	0.41	5.31	3.24E+10	1.62E+08
PL7	9	16	717	54	0.08	3.59	0.27	3.19E+11	1.60E+09
PN7	10	523	9	79	2.62	0.05	0.40	5.24E+12	2.62E+10

DUS/USS data

Run	Input seq.	Genome	Matrix	Total cc-US	Total oneoff	perfect/kb	1off/kb	Score	Score/kb
VI	Nme intergen	380 kb	Nme2902	224	41	0.59	0.11	7.68E+01	2.02E-01
VC	Hin intergen	220 kb	Hin2206	51	24	0.23	0.11	2.09E-06	9.50E-09
RW5&RW6	Random 200kb	200 kb	Nme2902	159	27	0.88	0.15	5.50E+01	2.70E-01
RV1-10	Random 200kb	200 kb	Hin2206	112	23	0.56	0.12	4.54E-06	2.27E-08