Genome-wide analysis of DNA uptake across the outer membrane of naturally competent *Haemophilus influenzae*

Marcelo Mora[1], Joshua Chang Mell[2], Garth D. Ehrlich[2,3], Rachel L. Ehrlich[2] and Rosemary J. Redfield[1¶*]

¶* Lead Contact and Corresponding Author: redfield@zoology.ubc.ca.

AUTHOR AFFILIATIONS:

1. Department of Zoology, University of British Columbia, Vancouver, BC, Canada V6T 1Z3.
2. Department of Microbiology & Immunology; Center for Genomic Sciences, Institute of Molecular Medicine and Infectious Disease; Drexel University College of Medicine, 12 Philadelphia PA, USA 19102.
3. Department of Otolaryngology – Head and Neck Surgery, Drexel University College of Medicine, 12 Philadelphia PA, USA 19102.

SUMMARY

The genomes of naturally competent Pasteurellaceae and Neisseriaceae have many short uptake sequences (USS), which allow them to distinguish self-DNA from foreign DNA. To fully characterize this preference we developed genome-wide maps of DNA uptake using both a

sequence-based computational model and genomic DNA that had been sequenced after uptake by and recovery from competent *Haemophilus influenzae* cells. When DNA fragments were shorter than the average USS spacing of ~1000 bp, sharp peaks of uptake were centered at USS and separated by valleys with 1000-fold lower uptake. Long DNA fragments (1.5-17 kb) gave much less variation, with 90% of positions having uptake within two-fold of the mean. All detectable uptake biases arose from sequences that fit the USS uptake motif. Simulated competition predicted that, in its respiratory tract environment, *H. influenzae* will efficiently take up its own DNA even when human DNA is present in 100-fold excess.

INTRODUCTION

Many bacteria are naturally competent, able to actively bind DNA fragments at the cell surface and pull them into the cytoplasm, where the incoming fragments may contribute nucleotides to cellular pools or recombine with homologous genomic sequences (Lorenz and Wackernagel, 1994). The genetic exchange associated with this latter process contributes to adaptation and is known to have promoted resistance to antibiotics (Bae et al., 2014) and increased strains' intracellular invasiveness (Mell et al., 2016) and vaccine resistance (Kress-Bennett et al., 2016; Straume et al., 2015). Thus, understanding how different genomic regions evolve via natural transformation processes could be used to predict the spread of pathogenic traits.

Most naturally competent bacteria that have been tested take up DNA regardless of sequence, but species in two families, the Pasteurellaceae and the Neisseriaceae, exhibit strong preferences for DNA containing short sequence motifs (Chen and Dubnau, 2004). Because these motifs have become highly enriched in the corresponding genomes, these biases

43 effectively limit uptake to DNA from close relatives with the same uptake specificity (Dougherty
44 et al., 1979; Scocca et al., 1974). The distribution of the preferred sequences around the
45 chromosome is uneven (Smith et al., 1995), which may cause different genes to experience
46 quite different rates of genetic exchange.

47 Most steps in the natural transformation process are highly conserved among transformable
48 species (Chen and Dubnau, 2004). In the Pasteurellaceae, the Neisseriaceae and most other
49 Gram-negative bacteria, DNA uptake is initiated by binding of a type IV pilus uptake machine to
50 double-stranded DNA (dsDNA) at the cell surface. The DNA-binding protein has not been
51 identified in *H. influenzae*, but in *Neisseria* it is a minor pilin-type protein that forms part of the
52 pilus (Cehovin et al 2013). DNA binding is followed by retraction of the pilus, which pulls the
53 DNA across the outer membrane into the periplasm. Because circular DNA is taken up as
54 efficiently as linear DNA, uptake is thought to begin internally on DNA fragments rather than at
55 their ends (Barany et al., 1983). Thus, it is likely that the stiff dsDNA molecule is transiently
56 kinked (folded sharply back on itself) at the site of initiation to allow it to pass through the
57 narrow secretin pore of the uptake machinery. Forces generated by the retraction of the type
58 IV pilus are thought to be responsible for this kinking, which might be facilitated by strand
59 separation at the AT-tracts (Danner et al., 1982). Once a loop of the DNA is inside the
60 periplasm, a ratchet process controlled by the periplasmic protein ComEA is thought to pull the
61 rest of the DNA through the outer membrane (Hepp and Maier, 2016; Salzer et al., 2016).
62 Below we use 'DNA uptake' to refer to the combined binding and membrane-transport steps
63 that move DNA from the extracellular environment across the outer membrane into the
64 periplasm. In Gram-positive bacteria similar machinery acts to pull the DNA through the thick

65 cell wall (Chen and Dubnau, 2004). The next step in natural transformation is translocation of
66 the DNA out of the periplasm and into the cytoplasm. Only the 3'-leading strand remains intact,
67 passing through an inner membrane pore encoded by the *rec2*/*comEC* gene, while the other
68 strand is degraded in the periplasm and its nucleotides are dephosphorylated and imported as
69 nucleosides (Pifer and Smith, 1985). Circular DNA molecules are efficiently taken up across the
70 outer membrane but remain the periplasm because they lack free ends (Pifer and Smith, 1985).
71 As the single strand enters the cytoplasm it undergoes limited exonucleolytic degradation
72 before being complexed with cellular proteins. If sequence similarity permits the strand may
73 then recombine with homologous chromosomal sequences; otherwise the strand is degraded
74 to its constituent nucleotides (de Vries et al., 2001). Both linkage and sequencing studies
75 indicate that fragments much longer than the cell are readily taken up (Goodgal, 1982; Mell
76 and Redfield, 2014), as are fragments as short as 200 bp, although a lower limit has not been
77 established (Mell 2012, Maughan 2009). However, recombination of short fragments is limited
78 because they are usually degraded by cytoplasmic nucleases before they can recombine (Pifer
79 and Smith, 1985). Uptake speed has been estimated at 500-1000 bp/sec, with transformation
80 essentially complete by 15 min (Deich and Smith 1980).

81 **Direct measures of DNA uptake bias:** Uptake-competition experiments in the Pasteurellacean
82 *Haemophilus influenzae* and in *Neisseria gonorrhoeae* showed that genetically marked 'self-
83 derived' DNA competes for uptake with unmarked self-derived DNA but not with DNA from
84 unrelated sources (Dougherty et al., 1979; Scocca et al., 1974). Subsequent DNA uptake
85 experiments using cloned radiolabeled DNA fragments found that these self-preferences are
86 caused by the uptake machineries' strong biases for short sequence motifs, called uptake signal

87 sequences (USS) in *H. influenzae* and DNA uptake sequences (DUS) in *Neisseria* species (Sisco
88 and Smith, 1979) (Davidsen et al., 2004). Sequence comparisons and site-directed mutagenesis
89 initially identified the *H. influenzae* motif as an 11 bp sequence with a strong contribution by
90 flanking AT-rich sequences (Danner et al., 1980, 1982), and later genome sequencing identified
91 1465 occurrences of a 9 bp USS core in *H. influenzae* and 1892 occurrences of an unrelated 10
92 bp DUS motif in *N. meningitidis* (Smith et al., 1995, 1999). Subsequent motif search analyses by
93 Maughan *et al.* (Maughan et al., 2010) expanded this number to 2206 USSs in the genome of
94 the standard *H. influenzae* lab strain Rd, sharing the motif shown in Figure 1A. These genomic
95 analyses were later complemented by direct uptake experiments using mutated and
96 degenerate USS variants(Maughan, Wilson and Redfield, 2010; Mell, Hall and Redfield, 2012)..
97 Mell *et al.* used mutagenesis and sequencing of pools of degenerate USS-containing fragments
98 that had been recovered after uptake to identify the contribution to uptake of each USS
99 position, and found that the central GCGG bases are crucial for uptake, with much smaller
100 contributions made by the flanking bases and AT-rich segments. The motif in Figure 1B shows
101 the contribution of each base considered independently. Interaction effects between bases of
102 the AT-tracts and the core were also found to make important contributions to uptake, but only
103 a few of these have been directly measured. Although characterization of the unrelated
104 Neisseriacean DUS has not reached this level of detail (Mathis and Scocca, 1982), the two
105 uptake systems shares many features, apparently convergently evolved, including the presence
106 within each family of lineages with slightly different preferred motifs (Frye et al., 2013; Redfield
107 et al., 2006).

108 **Evolution of uptake sequences in the genome:** Alignment of homologous genomic regions
109 from different Pasteurellaceae species showed that USS evolve by point mutations (Redfield et
110 al., 2006); *i.e.* they are not inserted elements. As a mechanism for this evolution, (Danner et al.,
111 1980) proposed that the combination of uptake bias and genomic recombination creates an
112 evolutionary pressure that causes the preferred uptake sequences to accumulate throughout
113 the genome, with their numbers limited by their eventual interference with gene function.
114 Consistent with this, uptake sequences in both *Haemophilus* and *Neisseria* are
115 underrepresented in newly acquired segments, in rRNA genes, and in coding sequences,
116 especially those with strong functional constraints (Findlay and Redfield, 2009; Smith et al.,
117 1999). Modeling by Maughan *et al.* (Maughan et al., 2010) confirmed that this molecular drive
118 process could produce uptake sequence distributions like those of real genomes, with no need
119 for any fitness benefit from either the uptake sequences or the recombination they promote.
120 Thus, the presence of biased DNA uptake machinery may be sufficient in itself to explain the
121 abundance of uptake sequences. Such biases may be solely a consequence of direct selection
122 on the DNA uptake machinery for more effective DNA binding or may have been reinforced by
123 indirect selection for preferential uptake of conspecific DNA. These processes might be
124 especially important in respiratory tracts and other mucosal environments where
125 Pasteurellaceae and Neisseriaceae species mainly occur (Man et al., 2017). These environments
126 contain abundant host DNA, and transformation can only occur if the released bacterial DNA
127 competes successfully for binding to the uptake machinery (Lethem et al., 1990; Shak et al.,
128 1990).

129 The goal of the present study was to measure DNA uptake at every position in the *H. influenzae*
130 genome and to use this uptake data to characterize the effects of USS and identify any other
131 factors affecting uptake. To prepare a framework for interpreting uptake biases, we developed
132 a computational model that predicted the effect of uptake sequences on DNA uptake across
133 the *H. influenzae* genome. We did not attempt to build a model that accurately simulated the
134 actual events of DNA uptake, since too little is known about these. Instead, the initial version of
135 this model was 'naïve' in that its parameters and settings were based only on previously
136 published information. Discrepancies between the model's predictions and the observed
137 uptake were then used to identify features of uptake that were poorly predicted. Hypothesized
138 biological explanations for these discrepancies then guided changes to the model, and the
139 effect of each change on the discrepancy was used to confirm or refute the hypothesis. The
140 most important product of this recursive analysis was not the model itself, but the improved
141 understanding of factors affecting DNA uptake across the genome. These in turn increased the
142 understanding of the genomic distribution of recombination and the effects of competition
143 with DNA from the host or other microbiota.

144

145 **RESULTS**

146 **A computational model of DNA uptake**

147 As a framework for interpreting DNA uptake data we developed a simulation model of USS-
148 dependent DNA uptake. It takes as input the locations and strengths of USSs in the DNA whose
149 uptake is to be simulated, the fragment-size distribution of this DNA, and binding and uptake

150 functions that specify how uptake probability depends on USS presence and strength. The
151 output is the expected relative uptake of every position in the genome.

152 Development of the model was guided by basic principles of sequence-specific protein-DNA
153 interactions (Halford and Marko, 2004; Rohs et al., 2010). The first step in these interactions is
154 thought to be a random encounter between a DNA fragment and the binding site of the
155 protein, usually at a DNA position that does not contain the protein's preferred sequence. This
156 non-specific binding dramatically increases the probability that the protein will subsequently
157 encounter any preferred sequence, either by sliding along the DNA or by transient dissociation
158 and reassociation, leading to specific binding between DNA and protein. In the case of the USS
159 this specific binding then enables uptake of the DNA fragment across the cell's outer
160 membrane.

161 The model did not explicitly simulate the first step, non-specific binding, since this is expected
162 to be equally probable for all DNA positions. The specific binding and DNA uptake steps were
163 separately modeled since they are expected to depend on the properties of the DNA uptake
164 machinery and on the length and sequence of the DNA fragment. Although in real cells both
165 steps may depend on the quality of the USS, for simplicity the initial version of the model
166 assumed that specific binding required only a threshold similarity to the USS consensus and
167 that the subsequent probability of uptake depended on the strength of this similarity.

168 Simulating these steps required first specifying the genomic sequences that should be treated
169 as USS. This was not straightforward because genomes contain many USS variants that differ in
170 how well they promote DNA uptake (Findlay and Redfield, 2009; Maughan and Redfield, 2009).
171 Our strategy was to score the information content (in bits) of every genome position, using the

172 Position-Specific Scoring Matrix (PSSM) from Mell *et al.*'s degenerate-sequence uptake
173 experiment (Mell et al., 2012) (Table S1), and to use overrepresentation of high-scoring
174 sequences as the USS criterion. We scored every genome position in the three *H. influenzae*
175 strains used for the experiments described below (Table 1), and in four randomly generated
176 sequences with the same length and base composition (Figure S1 shows the score
177 distributions). In the *H. influenzae* genomes, overrepresentation of high-scoring sequences was
178 detectable above a score of 7.0 bits and became dramatic above 10.0 bits, where the numbers
179 of high-scoring positions increased in *H. influenzae* genomes but became vanishingly small in
180 the random-sequence controls, (see inset in Figure S1). Since the slight overrepresentation of
181 scores between 7 and 10 bits was hypothesized to be not a direct effect of DNA uptake but an
182 indirect consequence of mutational degeneration of high scoring USSs, the model initially used
183 a USS cut-off score of 10 bits ('USS_{10}', n=1941 in strain 86-028NP). This was later reduced to a
184 less stringent 9.5 bits after weak uptake effects had been examined ('$USS_{9.5}$', n=2248 in strain
185 86-028NP).

186 The computational model used these USS scores to predict DNA uptake for every position in the
187 genome, summing the contributions of binding and uptake probabilities from DNA fragments of
188 different sizes (Figure 2A gives an overview). Each fragment under consideration was first
189 checked for the locations of any USS_{10}s, and the maximum fragment length (20 kb) and the
190 mean length of USS-free segments (*mean_gap*) were used to calculate the probability that a
191 DNA receptor protein initially encountering a random location in the fragment would then
192 encounter and bind specifically to a USS_{10} rather than disassociating from the fragment: *p_bind*
193 *= 1 – mean_gap/20000*). Fragments with no USS were initially assigned a baseline *p_bind* of 0.1.

194 The probability that this specific binding led to uptake of the DNA fragment was then calculated
195 from the USS_{10} score (or the mean score if the fragment contained more than one USS_{10}), using
196 the function *p_uptake = 0.1 + (1 - 0.1)/(1 + exp(-5 * (score – 11))),* where *0.1* specifies the
197 baseline uptake of USS-free fragments and *-5* is an arbitrarily chosen coefficient specifying the
198 slope at the inflection point. The score of 11 bits specifying the inflection point of the function
199 was chosen because this score was the midpoint of the range of increasing USS overabundance
200 in the genomes (see Figure S1). When combined with the baseline *p_bind* of 0.1, the *p_uptake*
201 baseline assigned fragments with no USS_{10} a net uptake probability of 0.01.

202 Once the contributions of every size of fragment had been calculated for each position (Figure
203 2A), the model combined all the contributions, taking into account the frequency of each size in
204 the input DNA. The position-specific uptake predictions were then normalized to a genome-
205 wide mean uptake probability of 1.0. Except in simple test cases, for computational efficiency
206 DNA fragment lengths were specified as the median lengths of bins (10 bp bins for short
207 fragments, 200 bp bins for long fragments) rather than each being considered separately (e.g.
208 101-120 bp, 121-140 bp).

209 **Model results:** Figures 2B and 2C show examples of model predictions for simple situations.
210 Figure 2B shows the uptake predictions for an 800-bp simulated genome containing a single
211 USS with score 12.0 bits, considering three different input DNA fragment sizes (100, 200 and
212 300 bp). The peaks at the USS have straight sides, a basal width twice the length of the
213 fragments being taken up, and 31-bp flat tops arising from the model's requirement for a full-
214 length USS. When the DNA fragment sizes were evenly distributed between 25-300 bp in length
215 (Figure 2C), the peak had steep sides at its tops and gradually flattened at the base; maximum

216 width at the base equalled twice the maximum fragment length. In simulated mini-genomes
217 with more than one USS (Figures 2D, E and F), isolated peaks were only seen when the DNA
218 fragments being taken up were substantially shorter than the spacing of the USSs (Figure 2D),
219 and peaks disappeared entirely when the fragments were long enough that almost all
220 contained at least one USS (Figure 2F).

221 Figure 2G and H show the predicted uptake maps when this model analyzed a 50kb segment of
222 the *H. influenzae* 86-028NP genome, using the short-fragment and long-fragment length
223 distributions from the actual uptake experiments described below (Figures S2A and B), and
224 Figure 2.I shows the distribution of USSs over this segment. Because the 'short' DNA fragments
225 are shorter than the typical separation between USSs, the model predicts that uptake will be
226 restricted to sharp peaks at each USS. In contrast, uptake of long DNA fragments is predicted to
227 be much more uniform, since most of these will contain at least one USS.

228 **Generation of experimental DNA uptake data**

229 To obtain high-resolution measurements of actual DNA uptake, we sequenced *H. influenzae*
230 genomic DNA that had been taken up by and recovered from competent *H. influenzae* cells.
231 Competent cells of the standard laboratory strain Rd were first incubated with genomic DNA
232 preparations from strains 86-028NP and PittGG, whose core genomes are readily differentiated
233 from Rd (and each other) because they differ at ~3% of orthologous positions (Hogg et al.,
234 2007). To allow efficient recovery of the taken-up DNA, the Rd strain in which competence was
235 induced carried a *rec2* mutation that blocks translocation of taken-up DNA fragments, causing
236 the DNA to be trapped intact in the periplasm (16). The 86-028NP and PittGG genomic DNAs
237 were pre-sheared to give short (50-800 bp) and long (1.5-17 kb) DNA preparations (size

238 distributions are shown in Figure S2), and three replicate uptake experiments were done with
239 each DNA preparation. After 20 min incubation with competent cells, the taken-up DNA was
240 recovered from the cell periplasm using the cell-fractionation procedure of Kahn *et al.* (Barouki
241 and Smith, 1985; Kahn et al., 1983; Mell et al., 2012). Taken-up DNA samples were sequenced
242 along with samples of the input 86-028NP and PittGG DNAs and of the recipient Rd DNA. The
243 input and uptake reads were then aligned to the corresponding 86-028NP and PittGG reference
244 sequences and coverage at every position was calculated. Table S2 provides detailed
245 information about the four input samples, the twelve uptake samples, and the Rd sample.

246 **Removal of contaminating Rd DNA:** Preparations of DNA recovered from the periplasm after
247 uptake always included some contaminating DNA from the recipient Rd chromosome. The
248 divergence between the Rd and donor genomes allowed us to estimate the extent of this
249 contamination by competitively aligning the taken-up reads from each sample to an artificial
250 reference 'genome' consisting of both recipient and donor genomes as separate
251 'chromosomes'. Reads that uniquely aligned to only one chromosome could then be
252 unambiguously assigned to either donor (taken-up) or recipient (contamination). The resulting
253 estimates of Rd chromosomal contamination were between 3.2% and 19.3% of reads; sample-
254 specific values are listed in Table S2.

255 The effects of this contamination were not expected to be uniform across each donor genome,
256 since segments of the 86-028NP and PittGG genomes with high divergence from or with no
257 close homolog in Rd would be free of contamination-derived reads. We used the competitive-
258 alignment described above to create contamination-corrected uptake coverages by discarding
259 all reads that could not be uniquely mapped to the donor genome; in addition to removing Rd

260 contamination this also removed reads from segments that are identical between the donor
261 and recipient strains ('double-mapping reads') and reads that mapped to repeats, such as the
262 six copies of the rRNA genes. For consistency, the same changes were applied to the input
263 samples although they did not experience any contamination. This correction removed an
264 average of 18.6% of reads (range 8.9%-28.3%), left some segments of the 86-028NP and PittGG
265 genomes with no coverage in all samples (2.3% and 2.1% respectively), and reduced coverage
266 adjacent to these segments. Figure S3 shows the locations of the missing data. Contamination
267 details for each sample are provided in Table S2, and the impacts are considered below. The
268 uptake analysis described below showed this correction to be effective.

269 **Uptake ratios:** To control for position-specific differences in sequencing efficiency,
270 contamination-corrected read coverage at each position in each uptake sample was divided by
271 read coverage at that position in the corresponding input sample (*e.g.* coverages of each 86-
272 028NP-short uptake sample were divided by 86-028NP-short input coverages). Normalizing the
273 mean of the three replicates to a genome-wide mean ratio of 1.0 then gave a mean 'uptake
274 ratio' measurement for each genome position for each DNA type. Finally each position's uptake
275 ratio was smoothed using a USS-length (31 bp) window. Figures 2.5 and 2.6 show the resulting
276 uptake ratio maps.

277 Figure 3A shows the short-fragment uptake ratio map for the first 50kb of the 86-028NP
278 genome; the ticks in Figure 3C indicate locations and scores of USS_{10}s. The pattern is strikingly
279 similar to that predicted for the same DNA segment by the model (Figure 2G). Sharp uptake
280 peaks are seen at USS_{10} positions, some separated by flat-bottomed valleys and others
281 overlapping. Figure S4A, B and C show similar analysis for the first 50 kb of strain PittGG's

282 genome, and Figure S5A and B show expanded maps for two examples of the 86-028NP peaks.
283 The full-genome maps of these uptake ratios are provided in Figure S4D (86-028NP) and S4G
284 (PittGG;) they display the consistency of the peak heights across each genome.

285 As expected, the long-fragment DNA samples (Figure 3B and Figure S4B, E and H) had much less
286 variation in uptake ratio than the short-fragment samples; 90% of positions had uptake ratios
287 within two-fold of the mean, and there were few high peaks or low valleys. The few extended
288 segments with low or no uptake coincided with large gaps between USS_{10}s. The largest gap was
289 in the 86-028NP segment between 95 and 145kb —the site of a genomic island that is absent
290 from the PittGG and Rd strains, has few USS and has high similarity to an *H. influenzae* plasmid
291 (Harrison et al., 2005). However, uptake ratios did exhibit substantial short-range variation not
292 predicted by the model, which is considered further below.

293 **Sensitivity is limited by low sequencing coverage:** At some genome positions our ability to
294 detect USS-dependent uptake biases and possible USS-independent biases in uptake coverage
295 was limited by low sequencing coverage. Although some of this low coverage arose from the
296 contamination-correction step described above, segments of low coverage were also seen in
297 the raw data, likely arising from biases in the library preparation and sequencing steps. Figure
298 S6 compares the raw coverage of short and long input samples for a 50 kb segment of the 86-
299 028NP genome, illustrating the strong variation in sequencing coverage that was both broadly
300 reproducible and sequence dependent. Low coverage had similar effects in all samples,
301 precluding calculation of uptake ratios where input coverage was zero and generating high
302 levels of stochastic variation where coverage was low (segments with no coverage and

303 positions with coverage below 20 reads are indicated in uptake ratio maps by gaps and grey
304 dots respectively).

305 We used the low uptake ratios of positions at least 1000 bp from the nearest $USS_{9.5}$ to assess
306 the effectiveness of the contamination correction described above, since these positions are
307 where contamination would have had its largest effects. This peak-separation distance was
308 chosen because a peak-shape analysis of uptake around 158 USS_{10}s that were separated by at
309 least 1200 bp from other USS_{10}s and had uptake ratios of at least 3.0 found that mean uptake
310 ratio had fallen to baseline at positions 600 bp from the USS (Figure S7). Figure 4 compares
311 uptake ratios between the valley positions that received the contamination correction
312 described above (those with a Rd homolog) and the positions where no correction was needed
313 (those with no Rd homolog). The two distributions had nearly identical medians (0.0022 and
314 0.0020 respectively), indicating that the correction was sufficient but not excessive. These low
315 values also confirm that the DNase I treatment and washing steps removed at least 99.99% of
316 the donor DNA that had not been taken up.

317 **Uptake ratios show no periodicity across the genome:** Bacterial genomes show periodicity for
318 several features related to DNA curvature and codon usage biases (Mrazek, 2010), so we
319 examined the distribution of uptake ratios across each genome by Fourier analysis, using the R
320 package TSA. The log-log plots in Figure S8 show that this found no strong influence of any
321 specific repeat period on either the variation in input-sample coverage (panels A-D) or the
322 variation in uptake ratios (panels E-H). Instead, to explain the observed variation, the analysis
323 needed to invoke small contributions from almost every possible repeat period.

324 **Uptake of short-fragment 86-028NP DNA**

325 To investigate how USSs contribute to the DNA uptake process, our strategy was to first analyze
326 discrepancies between model predictions and observed uptake ratio peaks in the 86-028NP
327 short-fragment dataset, since these would reveal ways in which the simple assumptions
328 underlying the model mischaracterized the actual steps of DNA uptake. Model changes that
329 improved the predictions were considered to better reflect the true constraints on uptake of
330 short DNA fragments. The refined model's predictions were then compared to the real uptake
331 ratios for 86-028NP long-fragment DNA, allowing additional refinement, and finally to the
332 short- and long-fragment uptake ratios for PittGG DNA.

333 Figure 5 compares predicted and measured uptake (orange and blue lines respectively) of
334 short-fragment 86-028NP DNA for the first 50kb of the genome. The close correspondence
335 between the model's predictions and observed peak locations and shapes confirmed that
336 almost all of the variation in uptake was due to USSs. (The Pearson correlation coefficient over
337 all positions was 0.92). However, the more detailed analyses below identified components of
338 the model that could be improved.

339 First, the predicted low-uptake valleys were too high. This is not easily seen in Figure 5, but the
340 log-scale inset in Figure 3 and the analysis in Figure 4 show that most valley-bottom positions
341 had uptake ratios between 0.01 and 0.001, well below the model's predicted baseline uptake of
342 0.052 (dashed orange line in Figure 4).

343 Second, the score cutoff of 10.0 bits was too high. Analysis of score-subsets within the uptake-
344 valley dataset showed that 9.5 bits was a better cutoff. For the 78 positions with scores
345 between 9.5 and 10.0 bits, the correlation between score and uptake ratio was 0.27. Although

346 a similar correlation (0.27) was seen for the 136 positions with scores between 9.0 and 9.5 bits,
347 it was driven mainly by very small effects, and only 4 positions had uptake ratios higher than
348 0.1. For the 461 positions with USS scores between 8.5 and 9.5 bits the correlation was only
349 0.09.

350 Third, most of the predicted peaks were too high. Figure 6 complements Figure 4's analysis of
351 uptake valleys with an analysis of similarly isolated uptake peaks. The blue dots show the
352 uptake ratio at every $USS_{9.5}$ position in the genome that is at least 1000 bp from the nearest
353 USS_{10} (n=209), and the small orange dots show predicted uptake at the same positions. The lack
354 of scatter in the orange points confirms that the separation distance was sufficient to avoid
355 predicted effects of nearby USS; the least-squares difference between predicted and observed
356 uptake ratios at these 209 USS positions was 0.521.

357 In the original-version of the uptake model, the sigmoidal parameters of the *p_uptake* function
358 (baseline and location and slope of inflection point) were set using the frequencies of USS
359 scores in the 86-028NP genome; as expected, the same parameters were obtained when a
360 sigmoidal function was fit to the uptake points predicted by this model (orange line in Figure 6).
361 To improve the predictions we replaced these *p_uptake* parameters with those of a sigmoidal
362 function fitted to the real uptake data (blue line in Figure 6), changed the *p_uptake* baseline to
363 0.005, lowered the USS score cutoff from 10.0 bits to 9.5 bits, and ran the model again. The
364 purple points in Figure 6 show that, although peak-height predictions for low-scoring USS were
365 somewhat improved by the changed *p_uptake* function, those for USS with scores > 11.5 bits
366 became slightly worse; the least-squares difference for the 209 positions was modestly reduced
367 from 0.521 to 0.420.

368 We initially suspected that the over-prediction of peak heights might be due to overestimating
369 the proportion of very short fragments in the input DNA, but eliminating the contributions of
370 fragments shorter than 80 bp had very little effect on predicted peak heights. (This is likely
371 because these fragments contribute little DNA and rarely contain USS.) We then considered an
372 alternative explanation, that USS might be ineffective when they were very close to fragment
373 ends. The red dots in Figure 6 shows that requiring USS to be at least 50 bp from fragment ends
374 lowered predicted peak heights to the mean observed height while maintaining a low baseline
375 uptake. Distances of 30 bp and 70 bp were also tested, but gave peaks that were, respectively,
376 too high and not high enough. Incorporating this modification into the model barely changed
377 the least-squares difference of its predictions with the observed uptake ratios (0.93, up from
378 0.92), but dramatically reduced the least-squares difference from 0.420 to 0.052. This revised
379 model was used by the analyses described below.

380 The above analyses do not explain why many uptake ratio peaks were substantially higher or
381 lower than predicted by their USS score. The scatter is unlikely to be due to noise alone, since
382 more extreme smoothing of the uptake ratio data with windows as large as 150 bp improved
383 the correlation by only 0.001. Below we consider three other factors that might influence DNA
384 uptake: weak non-USS uptake biases, effects of interactions between bases at different
385 positions in the USS, and effects of DNA shape.

386 Weak uptake biases could arise from either low-scoring USS or non-USS sequence factors. Since
387 weak biases would only be detectable in genome segments that lack strong USSs, we searched
388 for them using a far-from-USS_{10} dataset containing only DNA segments whose ends were at
389 least 0.6kb from the closest USS_{10} and that had input coverage of at least 20 reads. This dataset

390 contained 513 segments where weak uptake effects could in principle be detected (22% of the
391 genome); their mean uptake ratio over the 428678 positions was 0.0102. In these segments,
392 the only positions with distinct uptake ratio peaks >0.2 were 9 weak USSs with scores between
393 9.5 and 9.9 bits (an example is shown in panel C of Figure S5). Since this analysis did not find
394 any non-USS positions giving uptake higher than 0.2, it suggests that other sequence factors do
395 not detectably promote uptake in the absence of a USS.

396 The degenerate-USS analysis of Mell *et al.* (Mell et al., 2012) found that pairwise interactions
397 between the AT-tract bases and core bases of a USS made substantial contributions to uptake
398 of a 200 bp synthetic fragment. To evaluate the effects of these interactions in predicting
399 uptake of genomic DNA, the USS scores used by our revised model were raised or lowered in
400 proportion to the interaction effects reported in Figure 6 of Mell *et al*. Figure 7 shows that, for
401 the same 209 isolated $USS_{9.5}$s used above (Figure 6), this adjustment had little effect on high-
402 scoring USS but further reduced the scores of low-scoring USS. The data points are coloured by
403 their uptake ratios, showing that weak USS giving unusually high or low uptake were equally
404 likely to have their scores reduced. However, these scoring changes had no effect on the short-
405 fragment uptake predictions of the revised model (both Pearson correlations 0.93), probably
406 because (i) the scores of strong USS were not significantly changed and (ii) the scores of weak
407 USS already gave near-baseline uptake.

408 **DNA shape effects:** Most of the interactions in Figure 7 were between bases separated by 10
409 bp or more, so we complemented that analysis with analysis of DNA shape, which reflects both
410 pairwise and more complex interactions over a 5 bp range. The major shape features that can
411 be predicted from DNA sequence are the minor groove width, the propeller twist between two

412 paired bases, the helix twist between one base pair and the next, and the roll of one base pair
413 relative to the next. In Figure 8, the wide grey line reproduced in each panel shows these
414 feature predictions for the consensus USS. The USS inner core (orange shading) has a relatively
415 wide minor groove and high propeller twist, which would facilitate sequence recognition by
416 proteins (Rohs et al., 2009). To the left of this and in both AT-tracts (yellow shading) the minor
417 groove is narrow with low propeller twist and negative roll, predicting that these segments are
418 both rigid and slightly bent.

419 To see if shape features affect uptake in ways that were not captured by scores alone, we
420 compared the shape features of subsets of the 209 isolated USS with similar scores but
421 different uptake ratios. Panels A-D of Figure 8 compare the shape features of very weak USS
422 ($USS_{9.5\text{-}10}$) whose uptake ratios were either low (<0.6, blue lines) or high (>2.0, orange lines).
423 Similarly, panels E-H and I-L show the same comparisons for USSs with low and moderate
424 scores ($USS_{10\text{-}10.5}$ and $USS_{10.5\text{-}11}$ respectively). USSs with scores higher than 11 bits were not
425 analyzed since they did not exhibit enough uptake variation to reveal correlations between
426 uptake and DNA shape.

427 In all three score subsets the inner-core shape features were very similar for low-uptake and
428 high-uptake subsets (blue and orange lines), probably because this sequence perfectly matches
429 the consensus in 206 of the 209 $USS_{9.5}$. (The other three had otherwise-perfect core sequences
430 and good AT tracts but had only baseline uptake ratios.) However, the AT-tract shapes had
431 marked differences, with low-uptake USSs having no distinctive shape features and high-uptake
432 USSs resembling the USS consensus shape. This suggests that the predicted rigidity and slight
433 bend caused by interactions within the consensus AT tracts facilitate DNA uptake.

434 **Uptake of fragments with more than one USS:** Many genomic USSs are sufficiently close that
435 they will co-occur even on short DNA fragments; 23% of 86-028NP USS_{10}s are within 100 bp of
436 another USS_{10}, and 17% are within 30 bp (Figure S9A**)**. Fragments with multiple USS might be
437 expected to have relatively high uptake, since they provide more targets to which the uptake
438 machinery receptor could bind, but only one of the two previous studies in *Neisseria* found this
439 effect (Ambur et al., 2007; Goodman and Scocca, 1991).

440 Visual examination of uptake ratio maps around the 230 pairs of 86-028NP USS_{10}s within 100
441 bp found only single peaks; Figure S9B shows that the midpoints of these USS pairs (red and
442 green points) have very similar uptake ratios to those of the 209 isolated USSs (pale blue
443 points). Figure S5 panels E and F show examples of peaks at USS pairs separated by 69 bp and
444 230 bp.

445 A special subclass of USS pairs consists of oppositely oriented pairs that are so close that they
446 overlap; these can form RNA hairpins and are usually located at the ends of genes where they
447 act as transcriptional terminators (Kingsford et al., 2007; Smith et al., 1995, 1999). Figure S9A
448 shows that the 86-028NP genome has 109 USS_{10} pairs whose centers are within 14 bp: 69 in the
449 -/+ orientation, all 0-3 bp apart, and 40 in the +/- orientation, all 10-14 bp apart. The green
450 points in Figure S9B shows that midpoint uptake ratios of these pairs were very similar to those
451 of other pairs or at isolated USS_{10}s with similar scores.

452 Overall, these results indicate that the presence of two USS_{10}s within 100 bp does not
453 detectably increase the probability of the receptor binding to a USS, a result consistent with
454 that of Ambur *et al*. for pairs of closely spaced DUS in *Neisseria meningitidis* (Ambur et al.,

455 2007). Because individual fragments were not tracked, we cannot make any conclusions about
456 uptake of fragments with more widely separated USS.

457 **Uptake of long-fragment 86-028NP DNA**

458 Figure 9A compares the revised model's predictions for long-fragment 86-028NP DNA with the
459 uptake ratios observed over the same 50kb genome segment as in Figure 3B. In contrast to
460 predictions for short-fragment uptake (correlation of 0.93), it seriously underpredicted the
461 variation in long-fragment uptake ratios (correlation of 0.60). Although 51% of 86-028NP
462 positions had long-fragment uptake ratios lower than 0.8 or higher than 1.2, only 13% had
463 predicted uptake outside these limits.

464 Because long-fragment uptake ratios lacked the dramatic peaks and valleys seen for short-
465 fragments, stochastic noise arising at the regions of low sequencing coverage described earlier
466 was expected to play a larger role. To estimate the magnitude of this effect, we compared the
467 effects of adding different amounts of artificially generated noise to simulated (noise-free)
468 uptake data (Figure S10A shows examples of noise-free and noise-added coverages). Figure
469 S10B shows that, as expected, the correlation between noisy and noise-free data worsened as
470 the arbitrary level of noise increased for both short-fragment (blue) and long-fragment (red)
471 simulations, and that increasing noise had a much stronger effect on the long-fragment
472 simulations. For short DNA fragments, simulations with noise levels of 2 and 2.5 gave
473 correlations of 0.94 and 0.92, very close to the 0.93 correlation between the revised model and
474 the real data. (dashed blue line in Fig. S9B) For long DNA fragments with the same noise levels,
475 more noise was needed, with a 3.0 multiplier giving a correlation of 0.56, slightly lower than the
476 observed correlation of 0.60 (dashed red line in Fig. S9B). This confirms that the disparities

477 between measured uptake ratios and USS-based predictions were at least partially due to noise
478 in the sequencing data. However, these correlations are still 21% and 10% higher than the best
479 correlation obtained between the revised model's predictions and real data.

480 One notable feature of the long-fragment uptake ratio maps in Figure 3B and 9A is the presence
481 of occasional spikes of unusually high uptake, *e.g.* at position 18,540 (examples are shown in
482 panels G and H of Figure S5). These spikes were much narrower than expected for true uptake
483 biases acting on long DNA fragments, so they were likely due to stochastic differences between
484 input and uptake coverage in regions of low coverage, not to true differences in uptake.
485 Consistent with this explanation, 70.4% of positions with uptake ratios greater than 2.0 had
486 coverage less than 100 reads, compared to only 8% of positions with more typical uptake ratios
487 between 0.5 and 2. Table S3 shows analysis of the distribution of uptake extremes at positions
488 with different levels of sequencing coverage, and panels I and J of Figure S5 show the low
489 coverage at the example spikes. However, this explanation did not apply to most positions with
490 very low uptake, which were in broad segments with few or weak USS, not in narrow spikes at
491 low coverage positions, and thus likely reflect genuinely low uptake.

492 **Prediction of uptake of PittGG DNA**

493 Since the uptake model was refined using uptake data for DNA of strain 86-028NP, the revised
494 model's predictions were assessed using uptake data for DNA of a different strain, PittGG,
495 which differs from 86-028NP DNA by SNPs and indels affecting about 11% of its genome. Figure
496 S11 compares the uptake predictions with the observed PittGG uptake ratios. For short-
497 fragment data the peak heights and valley depths were similar to those for 86-028NP, as was
498 the Pearson correlation between predicted and observed uptake (0.92). Although the model

499 predicted similar long-fragment variation for 86-028NP and PittGG, the PittGG experimental
500 data showed more extreme variation, and the correlation was only 0.41, substantially worse
501 than the 0.60 obtained for 86-028NP. As with the 86-028NP data, much of this discrepancy may
502 be due to noise in regions of low sequencing coverage. However this may not fully explain
503 PittGG's lower correlation, because input samples of the two strains had similar frequencies of
504 low-coverage positions (Table 2.2).

505 **Predicted competition with other DNAs in the human respiratory tract**

506 *H. influenzae*'s natural environment is the human respiratory tract, where its DNA must
507 compete for uptake with DNAs from other bacteria, and with host-derived DNA whose
508 concentration in respiratory mucus of healthy individuals can exceed 300 µg/ml (Lethem et al.,
509 1990; Shak et al., 1990). The revised model was used to investigate this competition.

510 Table 2 lists the frequencies of USS_{10} and USS_{11} in the genomes of various species.
511 Pasteurellaceae species that share *H. influenzae*'s *Hin*-type USS (*e.g. H. parainfluenzae* and
512 *Aggregatibacter actinomycetemcomitans*) typically have about 1000 USS_{10}s per Mb (Redfield
513 2006); most of these are strong USS with scores ≥ 11. Pasteurellaceae with the *Apl*-type USS
514 (*e.g. H. ducreyi* and *Mannheimia haemolytica*) have about tenfold fewer *Hin*-type USS_{10}s, and
515 fewer than 20% of these are strong. Other common respiratory tract bacteria (*e.g. N.*
516 *meningitidis, Streptococcus pneumonia* and *Pseudomonas aeruginosa*) have USS only at the
517 frequencies expected for their base compositions. The human genome is exceptional in having
518 about 5-fold *fewer* USS_{10} per Mb than expected from simulated sequences with the 41% GC
519 content of human DNA (observed frequency 4.6/Mb, simulated frequency 30/Mb, both with
520 mean scores of only 10.3 bits). Since the USS inner-core motif includes a CpG, the

521 underrepresentation of USS_{10} in human DNA relative to simulated sequences is probably a
522 consequence of the 4-5 fold depletion of CpGs in the human genome caused by deamination of
523 methylated cytosines (Babenko et al., 2017; Bird, 1986).

524 To estimate how strongly these DNAs would compete with *H. influenzae* DNA for uptake by *H.*
525 *influenzae* cells, we concatenated the 86-028NP genome with each of two Pasteurellacean
526 genomes and with three 1.9 Mb segments of the human genome, and used the revised uptake
527 model to predict relative uptake. The respiratory pathogen *H. parainfluenzae* was used as the
528 *Hin*-type USS representative. The genital pathogen *H. ducreyi* was used as a representative
529 species with *Apl*-type USSs, since none occur in the human respiratory tract, and segments of
530 human DNA also served to represent non-Pasteurellacean bacteria. To approximate the lengths
531 of DNA fragments in the respiratory tract (Lethem et al., 1990; Shak et al., 1990), the model was
532 run using fixed fragment lengths of 1kb and 10kb, and because human DNA will contain many
533 fragments lacking USSs, the predictions were made with and without the model's *p_uptake*
534 baseline binding probability of 0.005. For each genome in the concatenated sequence, the
535 predicted uptake at every position was then summed to get a total genome uptake value for
536 each fragment length and baseline assumption.

537 Table 3 shows the relative amounts of *H. influenzae* DNA predicted to be taken up under the
538 various competition conditions. As expected from published uptake-competition experiments
539 (Albritton et al., 1984; Redfield et al., 2006), *H. influenzae* and *H. parainfluenzae* DNAs were
540 taken up with equal efficiency in all simulated conditions, and *H. ducreyi* DNA was much less
541 competitive, contributing only 5% of the 1 kb fragments and only 12% of the 10 kb fragments.
542 With human DNA as the 1:1 competitor, more than 99% of the DNA taken up was predicted to

543 be from *H. influenzae*. For all competing DNAs, baseline uptake of fragments containing no USS
544 made only a tiny contribution.
545

546 **DISCUSSION**

547 DNA uptake by competent *H. influenzae* Rd cells was measured at every position in the
548 genomes of two divergent *H. influenzae* strains, using short-fragment and long-fragment DNA
549 preparations. Differences between observed uptake and that predicted by a computational
550 model of USS-dependent uptake revealed the strength of the uptake machinery's bias towards
551 USS and the absence of other sequence biases. These findings increased our understanding of
552 DNA uptake bias and its potential effects on the distribution of recombination.

553 **Implications for DNA uptake**

554 **The USS motif:** The measured discrimination' for USS was very strong; with short DNA
555 fragments, valleys at USS-free segments had ~1000-fold lower uptake ratios than peaks at high-
556 scoring USS. This non-zero baseline is unlikely to be due to residual contamination by recipient
557 DNA, since valley depths were similar for segments with and without Rd homology (Figure 4).
558 One surprising finding of Mell et al.'s (2012) work was the difference between the inner-core
559 uptake motif identified by their degenerate-USS experiments (Figure 1A) and the extended
560 motif of USS sequences in the *H. influenzae* genome (Figure 1B). Our analysis confirmed that
561 uptake absolutely requires a perfect match to the inner core, but found that this was not
562 sufficient to raise uptake above baseline, even if the rest of the core was perfectly matched.

563 **Effect of USS shape:** The predicted shape differences between similarly-scoring USSs that gave
564 strong or weak uptake (Figure 8) suggest a preference for USS that are rigidly bent at AT-tracts
565 and the outer core (Harteis and Schneider, 2014; Rohs et al., 2009). Similar preferences have
566 been described for several DNA binding proteins and have been associated with specific binding
567 by arginine or lysine residues to narrow minor grooves (Rohs et al., 2009; Stella et al., 2010).
568 These features have been integrated successfully in some transcription factor binding models
569 (Li et al., 2017), but using them to improve uptake prediction will require more comprehensive
570 investigation into the effects of DNA shape on uptake. The stiffness also suggests that the initial
571 passage of DNA through the secretin pore may be facilitated by transient strand melting at or
572 beside the USS rather than by bending (Danner et al., 1982).

573 **Other USS effects on DNA uptake:** Comparison of predicted and measured heights of uptake
574 peaks suggested that USSs were ineffective when located very close to fragment ends. This
575 would be consistent with experimental evidence that uptake initiates internally rather than at
576 fragment ends(Barany et al., 1983), but would need to be experimentally investigated. The
577 presence of two USSs within 100 bp did not detectably increase uptake, a finding consistent
578 with Ambur *et al.*'s (Ambur et al., 2007) study of very close uptake sequences in *Neisseri*a.

579 **Lack of USS-independent effects:** The very low valleys between short-fragment uptake peaks
580 allowed us to examine more than 400,000 positions for USS-independent increases in uptake.
581 None were found; the only positions with distinct uptake ratio peaks >0.2 were 9 weak USSs.

582 **Implications for genetic exchange**

583 Preferential uptake of USS can create variation in recombination at all levels: across a single
584 genome, between strains of one species, and between both closely related and unrelated

585 species. Pifer & Smith showed that transformation frequency in *H. influenzae* decreased
586 exponentially when fragments were smaller than 3.5 kb. This decrease was attributed to
587 exonuclease degradation of fragments from their 3' ends, since similar numbers of short and
588 long fragments were taken up, only 5'-end label was incorporated into the chromosome, short
589 fragments transformed more efficiently when the selected marker was far from one end. Mell
590 (2014) used genome sequencing of inter-strain recombinants to examine the distribution of
591 recombination tract lengths; despite the presence of 2-3% SNPs, the mean tract length was 6.9
592 kb and the longest was 43 kb. The efficient uptake and recombination of long fragments allows
593 fragments containing non-homologous segments to still recombine well, provided both
594 fragment ends are homologous. Fragments with only one homologous end recombine much
595 less efficiently ('homology-facilitated recombination' (de Vries and Wackernagel, 2002)), and
596 integration of fragments with no homology is very rare.

597 **Across the genome:** Across the genome of a *H. influenzae* strain, the genetic consequences of
598 USS-dependent DNA uptake depend on USS locations and on the lengths of the available DNA
599 fragments. If only short fragments are available, the limitation to positions close to a USS may
600 be obscured by limitation caused by degradation of incoming DNA in the cytoplasm. If most
601 fragments are long, recombination will be both more frequent and more evenly distributed
602 across the genome, because long fragments are more likely to both contain USS and
603 recombine. The result will be that almost all recombination is caused by fragments long enough
604 to usually contain at least one strong USS. This situation is caused by the high abundance and
605 relatively even distribution of genomic USS, itself an expected consequence of the functionally

606 neutral accumulation of USS under a molecular drive caused by USS-biased DNA uptake
607 (Danner et al., 1982; Maughan et al., 2010).

608 **Recombination between *H. influenzae* strains:** On average about 85% of the genomes of *H.*
609 *influenzae* strains are homologous, with sequence divergence low enough to have only a
610 modest effect on recombination frequencies (Mell et al., 2011). On average, genome segments
611 that are absent from other strains have lower density of USS (0.58/kb vs 1.08/kb for sequences
612 present in 86-028NP but absent from Rd. Only some of these will be segments newly acquired
613 from species without USS. If a non-homologous segment introduced into one strain by
614 conjugation or transduction is beneficial, the USSs in adjacent DNA will help it efficiently spread
615 to other strains by transformation.

616 **Recombination between Pasteurellaceae species:** Uptake of DNA from related species can also
617 influence recombination, either directly if the DNA is sufficiently similar to recombine with the
618 *H. influenzae* genome, or indirectly if it successfully competes with *H. influenzae* DNA for
619 uptake or, once inside the cell, for access to nucleases or recombination machinery. For *H.*
620 *influenzae* the most important competition will be with other Pasteurellaceae that share both
621 the respiratory tract niche and the *Hin*-type USS, but similar effects are expected for
622 Pasteurellaceae in other host species.

623 **Competition with DNAs from human cells and other respiratory bacteria:**
624 In the respiratory tract, the most important source of competing DNA is human cells. However,
625 our analysis suggests that *H. influenzae*'s uptake specificity allows its DNA to outcompete
626 human and other foreign DNAs even if these are in 100-fold excess, a combined effect of the
627 low number of USS_{10}s and their poor match to the uptake motif. Note that efficient self-uptake

628 does not necessarily imply a selective advantage, since USS accumulation in *H. influenzae*'s
629 genome may simply be due to the molecular drive process.

630 Uptake of DNA in the respiratory track could also be influenced by the presence of chromatin
631 and nucleoid proteins stably bound to the DNA. Although laboratory experiments typically use
632 highly purified DNA, DNA released by cell death will be coated with these proteins, which can
633 contribute significantly to biofilm stability (Brockman et al., 2018). Because such proteins could
634 interfere with uptake both directly, by blocking binding to the USS, and indirectly, by blocking
635 sliding of non-specifically bound uptake machinery along the DNA, it will be important to re-
636 examine DNA uptake using DNA that retains its bound proteins.

637 **LIMITATIONS OF THE STUDY**

638 Three factors limited measurements of uptake ratios: low sequencing read coverage,
639 contamination of recovered donor DNA with recipient DNA, and segments of sequence identity
640 between donor and recipient. Many reads had to be excluded at the contamination-correction
641 step, because they were in segments that were either identical between donor and recipient or
642 were repeated within the donor genome. Strong sequence-dependent variation in read
643 coverage caused other positions to be excluded from analysis because they had no coverage in
644 the control input sample. Overall, 2.3% of the genome was excluded from analysis, and an
645 additional 1.7% was flagged as unreliable due to low coverage. In addition, the stochastic
646 variation at low coverage positions introduced substantial noise into the calculation of
647 experimental uptake ratios, especially at uptake valleys. On the other hand, the model
648 predictions for long-fragment may be more accurate than indicated by their modest correlation
649 with the noisy measured uptake ratios.

650 **RESOURCE AVAILABILITY**

651 **Lead Contact:**

652 Further information and requests should be directed to and will be fulfilled by the Lead Contact.

653 Rosemary J. Redfield (redfield@zoology.ubc.ca)

654 **Materials Availability:**

655 All bacterial strains are available from Joshua Chang Mell (joshua.mell@drexelmed.edu).

656 **Data and Code Availability:**

657 All fastQ files have been deposited at NCBI under BioProject PRJNA387591. The corresponding

658 BioSamples are listed in Table S2. FastQ files were also deposited at Mendeley data

659 (DOI:10.17632/hcxp9d4zkf.1. Available at https://data.mendeley.com/datasets/hcxp9d4zkf/1).

660 The PacBio-sequenced PittGG genome reference was deposited into Genbank under SRA

661 number SRR10207558. Full calculations, and R scripts are available

662 at: https://github.com/mamora/DNA_uptake.

663 **ACKNOWLEDGEMENTS**

664 Financial support for this work was provided by the National Science and Engineering Research

665 Council of Canada (to RJR) and by the National Institutes of Health (to GDE (5R01DC002148-

666 21)). We thank the Drexel Genomic Core Facility for DNA sequencing, Rachel Simister for her

667 assistance with the bioanalyzer fragment analysis, and Matt Pennell, Sally Otto, Stephen Hallam

668 and David Baltrus for advice on the manuscript. We also thank two anonymous reviewers for

669 their very helpful comments, and the Editor for their patience during the revisions.

670 **AUTHOR CONTRIBUTIONS**

671 MM, JCM and RJR designed the experiments and analysis. MM performed the experiments and
672 most of the analyses; JCM and RJR performed the remaining analyses. Sequencing was done in
673 the laboratory of GDE; data curation by RLE, JCM and MM. RJR and MM wrote the manuscript
674 with input from the other authors.

675 **DECLARATIONS OF INTERESTS**

676 The authors declare no competing interests.

677

678 **REFERENCES**

679 Albritton, W.L., Setlow, J.K., Thomas, M., Sottnek, F., and Steigerwalt, A.G. (1984).
680 Heterospecific transformation in the genus Haemophilus. Mol. Gen. Genet. 193, 358–363.

681 Ambur, O.H., Frye, S.A., and Tønjum, T. (2007). New functional identity for the DNA uptake
682 sequence in transformation and its presence in transcriptional terminators. J. Bacteriol. 189,
683 2077–2085.

684 Babenko, V.N., Chadaeva, I. V, and Orlov, Y.L. (2017). Genomic landscape of CpG rich elements
685 in human. BMC Evol. Biol. 17, 19.

686 Bae, J., Oh, E., and Jeon, B. (2014). Enhanced transmission of antibiotic resistance in
687 Campylobacter jejuni biofilms by natural transformation. Antimicrob. Agents Chemother. 58,
688 7573–7575.

689 Barany, F., Kahn, M.E., and Smith, H.O. (1983). Directional transport and integration of donor
690 DNA in Haemophilus influenzae transformation. Proc Natl Acad Sci U S A 80, 7274–7278.

691 Barouki, R., and Smith, H.O. (1985). Reexamination of phenotypic defects in rec-1 and rec-2
692 mutants of Haemophilus influenzae Rd. J. Bacteriol. 163, 629–634.

693 Bird, A.P. (1986). CpG-rich islands and the function of DNA methylation. Nature 321, 209–213.

694 Brockman, K.L., Azzari, P.N., Taylor Branstool, M., Atack, J.M., Schulz, B.L., Jen, F.E.-C., Jennings,
695 M.P., and Bakaletz, L.O. (2018). Epigenetic regulation alters biofilm architecture and
696 composition in multiple clinical isolates of Nontypeable Haemophilus influenzae. MBio 9,
697 e01682-18.

698 Chen, I., and Dubnau, D. (2004). DNA uptake during bacterial transformation. Nat. Rev.
699 Microbiol. 2, 241–249.

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

702 Danner, D.B., Smith, H.O., and Narang, S.A. (1982). Construction of DNA recognition sites active
703 in Haemophilus transformation. Proc. Natl. Acad. Sci. U. S. A. 79, 2393–2397.

704 Davidsen, T., Rødland, E.A., Lagesen, K., Seeberg, E., Rognes, T., and Tønjum, T. (2004). Biased
705 distribution of DNA uptake sequences towards genome maintenance genes. Nucleic Acids Res.
706 32, 1050–1058.

707 Dougherty, T.J., Asmus, A., and Tomasz, A. (1979). Specificity of DNA uptake in genetic
708 transformation of gonococci. Biochem. Biophys. Res. Commun. 86, 97–104.

709 Findlay, W.A., and Redfield, R.J. (2009). Coevolution of DNA uptake sequences and bacterial
710 proteomes. Genome Biol. Evol. 1, 45–55.

711 Frye, S.A., Nilsen, M., Tønjum, T., and Ambur, O.H. (2013). Dialects of the DNA uptake sequence
712 in Neisseriaceae. PLOS Genet. 9, e1003458.

713 Goodgal, S.H. (1982). DNA uptake in Haemophilus transformation. Annu. Rev. Genet. 16, 169–
714 192.

715 Goodman, S.D., and Scocca, J.J. (1991). Factors influencing the specific interaction of Neisseria
716 gonorrhoeae with transforming DNA. J. Bacteriol. 173, 5921–5923.

717 Halford, S.E., and Marko, J.F. (2004). How do site-specific DNA-binding proteins find their
718 targets? Nucleic Acids Res. 32, 3040–3052.

719 Harrison, A., Dyer, D.W., Gillaspy, A., Ray, W.C., Mungur, R., Carson, M.B., Zhong, H., Gipson, J.,
720 Gipson, M., Johnson, L.S., et al. (2005). Genomic sequence of an otitis media isolate of
721 nontypeable Haemophilus influenzae: Comparative study with H. influenzae serotype d, strain
722 KW20. J. Bacteriol. 187, 4627–4636.

723 Harteis, S., and Schneider, S. (2014). Making the bend: DNA tertiary structure and protein-DNA
724 interactions. Int. J. Mol. Sci. 15, 12335–12363.

725 Hepp, C., and Maier, B. (2016). Kinetics of DNA uptake during transformation provide evidence
726 for a translocation ratchet mechanism. Proc. Natl. Acad. Sci. U. S. A. 113, 12467–12472.

727 Hogg, J.S., Hu, F.Z., Janto, B., Boissy, R., Hayes, J., Keefe, R., Post, J.C., and Ehrlich, G.D. (2007).
728 Characterization and modeling of the Haemophilus influenzae core and supragenomes based
729 on the complete genomic sequences of Rd and 12 clinical nontypeable strains. Genome Biol. 8,
730 R103.

731 Kahn, M.E., Barany, F., and Smith, H.O. (1983). Transformasomes: specialized membranous
732 structures that protect DNA during Haemophilus transformation. Proc. Natl. Acad. Sci. U. S. A.
733 80, 6927–6931.

734 Kingsford, C.L., Ayanbule, K., and Salzberg, S.L. (2007). Rapid, accurate, computational discovery
735 of Rho-independent transcription terminators illuminates their relationship to DNA uptake.
736 Genome Biol. 8, 1–12.

737 Kress-Bennett, J.M., Hiller, N.L., Eutsey, R.A., Powell, E., Longwell, J., Hillman, T., Blackwell, T.,
738 Byers, B., Mell, J.C., Post, J.C., et al. (2016). Identification and characterization of msf , a novel
739 virulence factor in Haemophilus influenzae. PLoS One 11, e0149891.

740 Lethem, M., James, S.L., Marriott, C., and Burke, J.F. (1990). The origin of DNA associated with
741 mucus glycoproteins in cystic fibrosis sputum. Eur. Respir. J. 3, 19–23.

742 Li, J., Sagendorf, J.M., Chiu, T.P., Pasi, M., Perez, A., and Rohs, R. (2017). Expanding the
743 repertoire of DNA shape features for genome-scale studies of transcription factor binding.
744 Nucleic Acids Res. 45, 12877–12887.

745 Lorenz, M.G., and Wackernagel, W. (1994). Bacterial gene transfer by natural genetic

746 transformation in the environment. Microbiol. Rev. 58, 563–602.

747 Man, W.H., De Steenhuijsen Piters, W.A.A., and Bogaert, D. (2017). The microbiota of the
748 respiratory tract: Gatekeeper to respiratory health. Nat. Rev. Microbiol. 15, 259–270.

749 Mathis, L.S., and Scocca, J.J. (1982). Recognize different specificity determinants in the DNA
750 uptake step of genetic transformation. J. Gen. Microbiol. 128, 1159–1161.

751 Maughan, H., and Redfield, R.J. (2009). Extensive variation in natural competence in
752 haemophilus influenzae. Evolution. 63, 1852–1866.

753 Maughan, H., Wilson, L.A., and Redfield, R.J. (2010). Bacterial DNA uptake sequences can
754 accumulate by molecular drive alone. Genetics 186, 613–627.

755 Mell, J.C., and Redfield, R.J. (2014). Natural competence and the evolution of DNA uptake
756 specificity. J. Bacteriol. 196, 1471–1483.

757 Mell, J.C., Shumilina, S., Hall, I.M., and Redfield, R.J. (2011). Transformation of natural genetic
758 variation into Haemophilus Influenzae genomes. PLoS Pathog. 7, e1002151.

759 Mell, J.C., Hall, I.M., and Redfield, R.J. (2012). Defining the DNA uptake specificity of naturally
760 competent Haemophilus influenzae cells. Nucleic Acids Res. 40, 8536–8549.

761 Mell, J.C., Viadas, C., Moleres, J., Sinha, S., Fernández-Calvet, A., Porsch, E.A., St. Geme, J.W.,
762 Nislow, C., Redfield, R.J., and Garmendia, J. (2016). Transformed recombinant enrichment
763 profiling rapidly Identifies HMW1 as an intracellular invasion locus in Haemophilus influenza.
764 PLoS Pathog. 12, e1005576.

765 Mrazek, J. (2010). Comparative analysis of sequence periodicity among prokaryotic genomes
766 points to differences in nucleoid structure and a relationship to gene expression. J. Bacteriol.
767 192, 3763–3772.

768 Pifer, M.L., and Smith, H.O. (1985). Processing of donor DNA during Haemophilus influenzae
769 transformation: analysis using a model plasmid system. Proc. Natl. Acad. Sci. U. S. A. 82, 3731–
770 3735.

771 Redfield, R.J., Findlay, W.A., Bossé, J., Kroll, J.S., Cameron, A.D.S., and Nash, J.H.E. (2006).

772 Evolution of competence and DNA uptake specificity in the Pasteurellaceae. BMC Evol. Biol. 6,
773 1–15.

774 Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S., and Honig, B. (2009). The role of DNA
775 shape in protein-DNA recognition. Nature 461, 1248–1253.

776 Rohs, R., Jin, X., West, S.M., Joshi, R., Honig, B., and Mann, R.S. (2010). Origins of specificity in
777 protein-DNA recognition. Annu. Rev. Biochem. 79, 233–269.

778 Salzer, R., Kern, T., Joos, F., and Averhoff, B. (2016). The Thermus thermophilus comEA / comEC
779 operon is associated with DNA binding and regulation of the DNA translocator and type IV pili.
780 Environ. Microbiol. 18, 65–74.

781 Scocca, J.J., Poland, R.L., and Zoon, K.C. (1974). Specificity in deoxyribonucleic acid uptake by
782 transformable Haemophilus influenzae. J. Bacteriol. 118, 369–373.

783 Shak, S., Capon, D.J., Hellmiss, R., Marsters, S.A., and Baker, C.L. (1990). Recombinant human
784 DNase I reduces the viscosity of cystic fibrosis sputum. Proc. Natl. Acad. Sci. U. S. A. 87, 9188–
785 9192.

786 Sisco, K.L., and Smith, H.O. (1979). Sequence-specific DNA uptake in Haemophilus
787 transformation. Proc. Natl. Acad. Sci. U. S. A. 76, 972–976.

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

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

793 Stella, S., Cascio, D., and Johnson, R.C. (2010). The shape of the DNA minor groove directs
794 binding by the DNA-bending protein Fis. Genes Dev. 24, 814–826.

795 Straume, D., Stamsås, G.A., and Håvarstein, L.S. (2015). Natural transformation and genome
796 evolution in Streptococcus pneumoniae. Infect. Genet. Evol. 33, 371–380.

797 de Vries, J., and Wackernagel, W. (2002). Integration of foreign DNA during natural

transformation of Acinetobacter sp. by homology-facilitated illegitimate recombination. Proc. Natl. Acad. Sci. 99, 2094–2099.

de Vries, J., Meier, P., and Wackernagel, W. (2001). The natural transformation of the soil bacteria Pseudomonas stutzeri and Acinetobacter sp. by transgenic plant DNA strictly depends on homologous sequences in the recipient cells. FEMS Microbiol. Lett. 195, 211–215.

FIGURES:

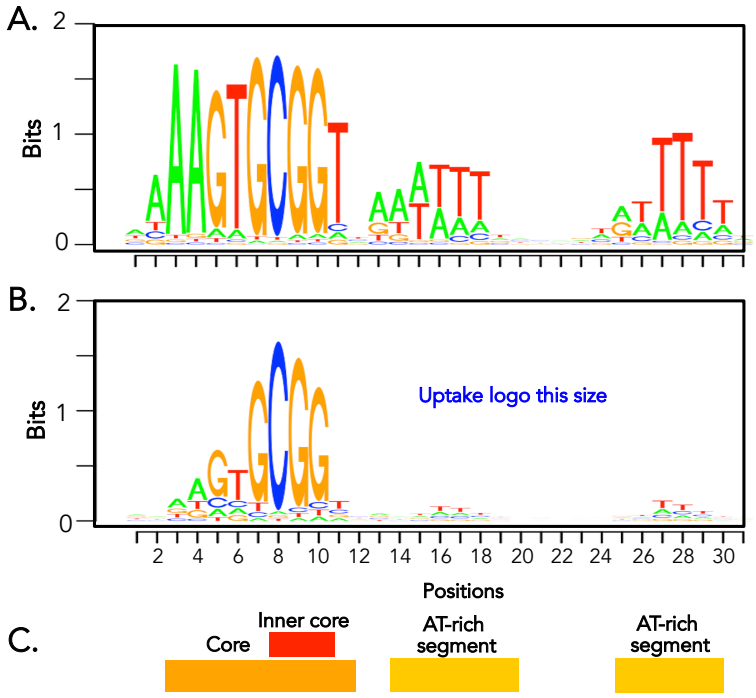


Figure 1. The *H. influenzae* uptake signal sequence.

Legend: A. Sequence logo showing the individual contributions to genomic abundance of bases in the USS motif (Maughan 2010). **B.** Sequence logo showing the individual contributions to uptake of bases in the USS motif, as measured by Mell *et al*. **C.** Conserved USS segments. See also Figure S1.

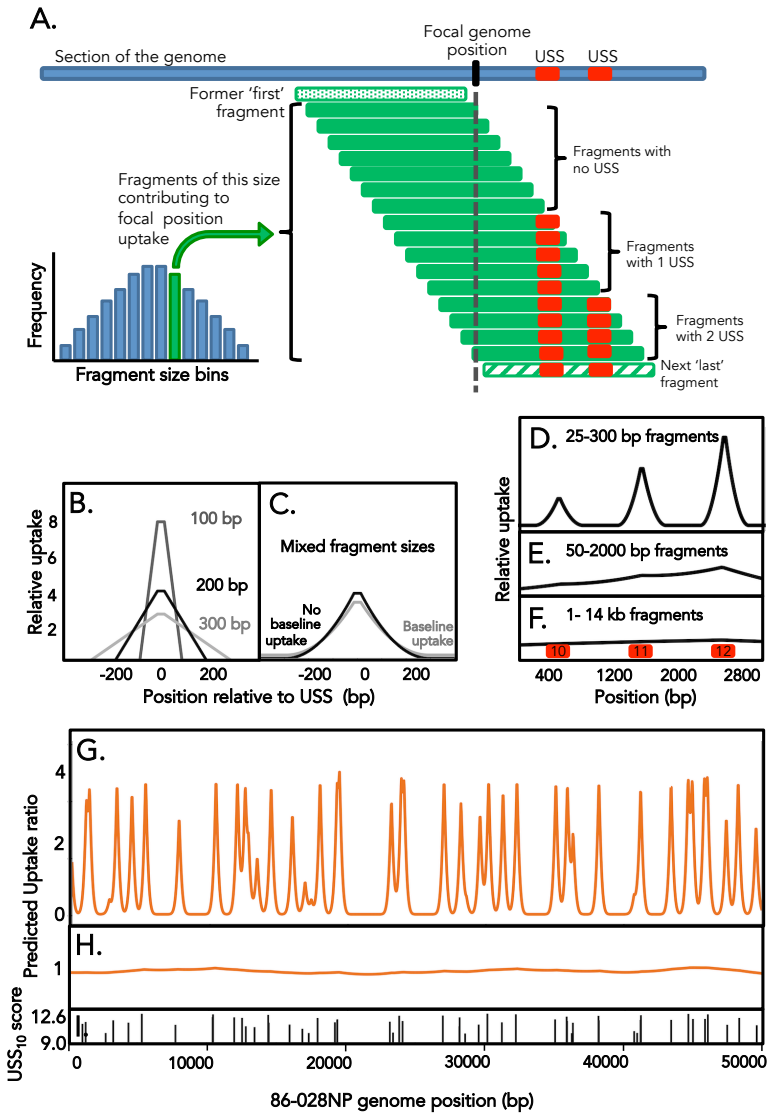


Figure 2. A computational model to predict DNA uptake.

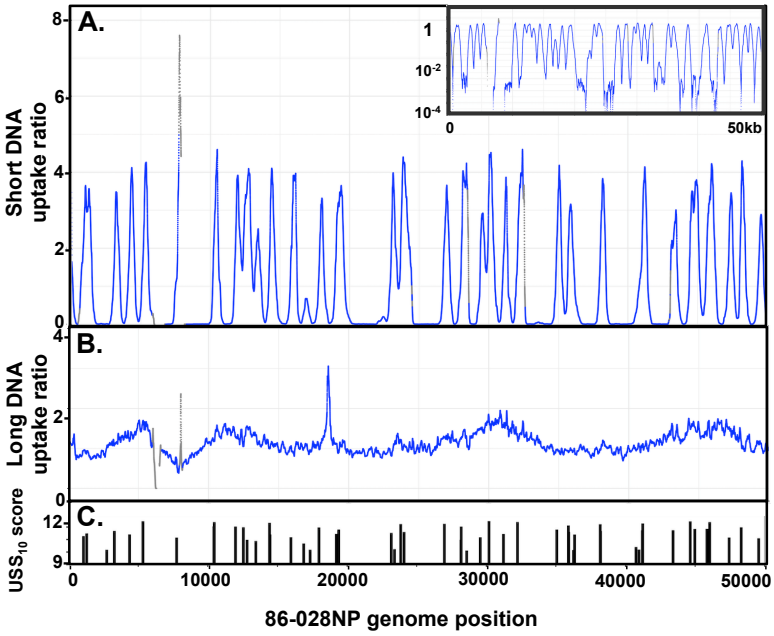
Legend: A. Components of the DNA uptake model (see Transparent Methods and Results for details). **B. & C.** Model predictions for uptake centered at a 12-bit USS for: **B**. 100, 200, and 300 bp fragments, **C**. a mixed distribution of fragments between 25-300 bp with and without baseline uptake. **D. E. & F.** Model predictions for uptake of a 3000 bp region with 3 USSs (red squares, scores in black) using different fragment-length distributions: **D**. 50-300 bp fragments,

820 **E.** 50-2000 bp fragment, **F.** 1-14 kb fragments. **G. & H.** Predicted DNA uptake of a 50 kb

821 segment of the 86-028NP genome using **G.** short and **H.** long fragment-length distributions: **I.**

822 Locations and scores of USS_{10}s in this 50 kb segment. See also Table S1 and Figures S1 and S2.

823



824

825 **Figure 3.** Experimentally determined uptake ratios for a 50 kb segment.

826 **Legend:** The X-axis is the same 50 kb segment of the 86-028NP genome as Figure 2G and H.

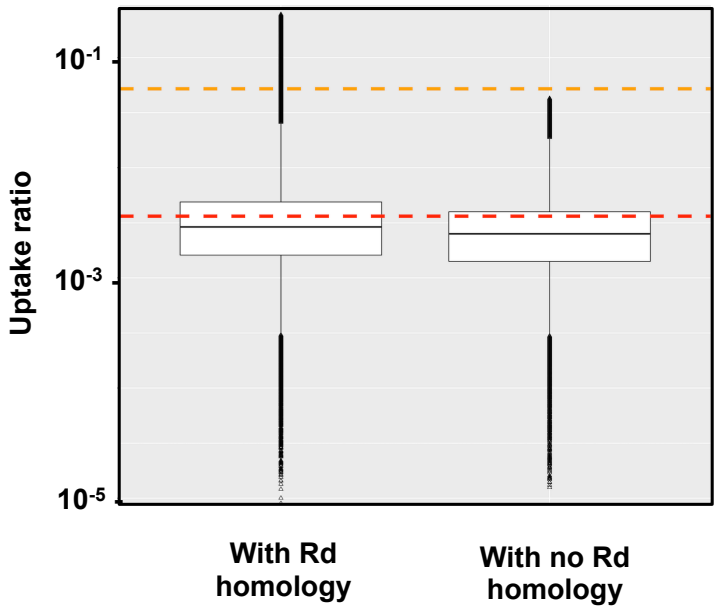
827 Grey points indicate positions with input coverage lower than 20 reads. Gaps indicate

828 unmappable segments. **A.** Uptake ratios of short-fragment DNA. **Inset:** Same data with a

829 logarithmic-scale Y-axis. **B.** Uptake ratios of long-fragment DNA. **C.** Locations and scores of

830 USS_{10}s. See also Figures S3 and S4.

831



832

833 **Figure 4.** Depths of uptake ratio valleys with and without contamination correction.

834 **Legend:** Uptake ratios of positions in the 86-028NP short-fragment dataset that were at least

835 1kb from the nearest $USS_{9.5}$ and whose uptake coverages were either corrected for

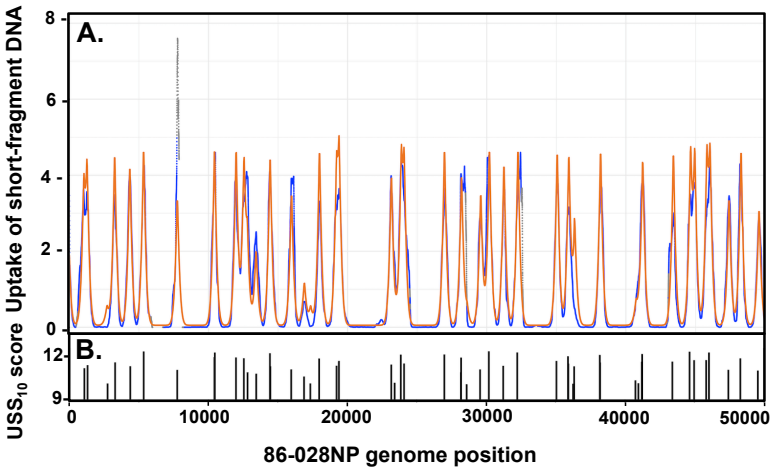
836 contamination with homologous Rd sequences (left, 272,547 positions) or not corrected for

837 contamination because they had no Rd homology (right, 89,418 positions). Orange and red

838 dashed lines indicate valley uptake predicted by the original and revised models respectively.

839 See also Figure S6.

840



841

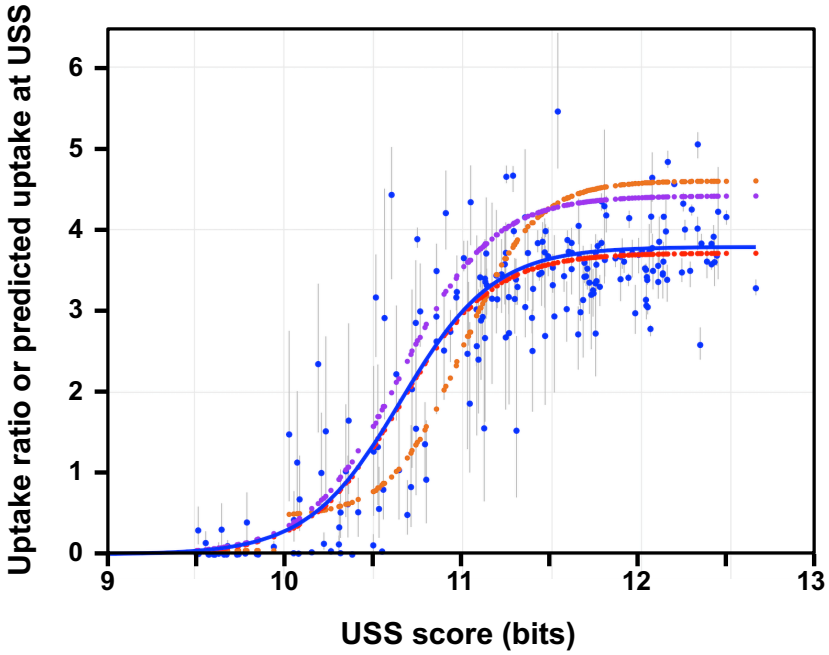
842 **Figure 5.** Predicted and observed DNA uptake analysis for short fragments of 86-028NP DNA.

843 **Legend: A.** Map of uptake ratios and initial model predictions. The blue points show the same

844 uptake ratio map as in Figure 3A. The orange points show the same predicted uptake as in

845 Figure 2G. **B.** Locations and scores of $USS_{9.5}$s.

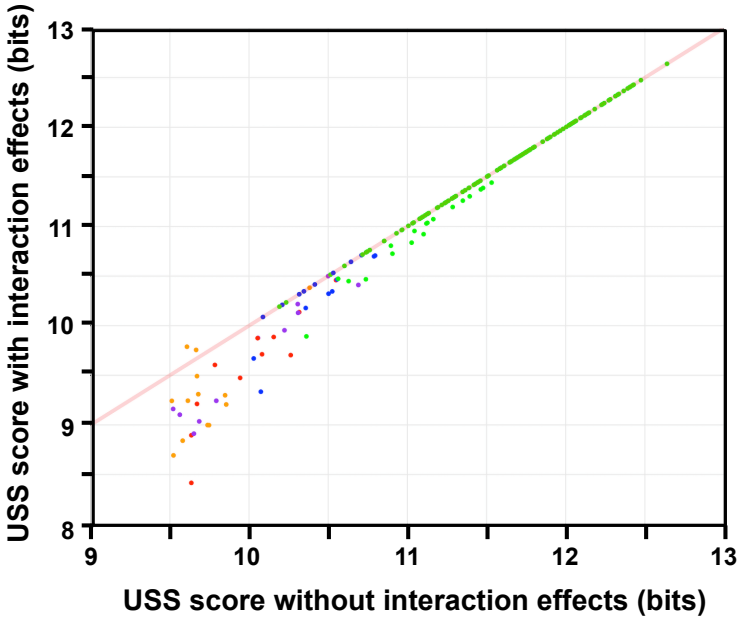
846



847

848 **Figure 6.** Short-fragment uptake ratios and predicted DNA uptake at isolated 86-028NP USS as a
849 function of USS score.

850 **Legend:** Predicted or measured DNA uptake at 209 $USS_{9.5}$ positions separated by at least 1000
851 bp from the nearest USS_{10}s. The blue dots show the measured uptake ratios; grey bars show
852 the ranges of the three replicates at each position. The blue line shows a sigmoidal function fit
853 to these points. The small orange dots and line show uptake predicted by the original model at
854 the same positions, and the small purple and red dots show uptake predicted by the
855 intermediate and revised model versions discussed in the text. See also Figure S6.



856

857 **Figure 7.** Effects of within-USS interactions on USS scores.

858 **Legend:** USS scores calculated with and without interactions effects for 209 isolated 86-028Np
859 $USS_{9.5}$s (see Methods for details). Red line shows expected scores if the interactions had no
860 effect. Point color indicates the uptake ratios at the USS: yellow, <0.01; red, 0.01-0.1; purple,
861 0.1-0.5; blue 0.5-1.5; green, >1.5.

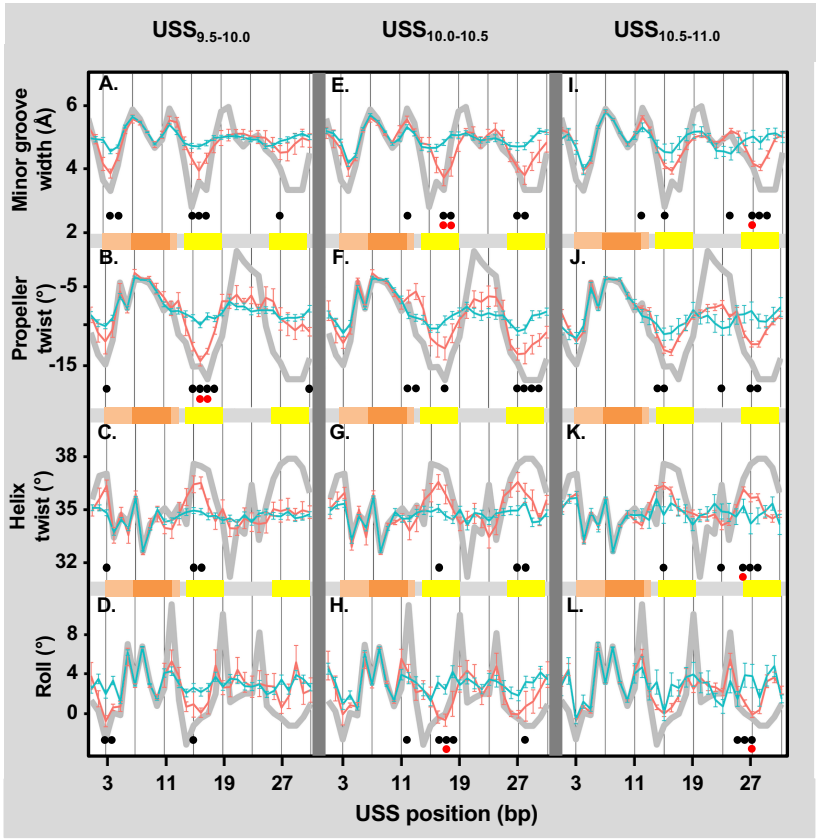
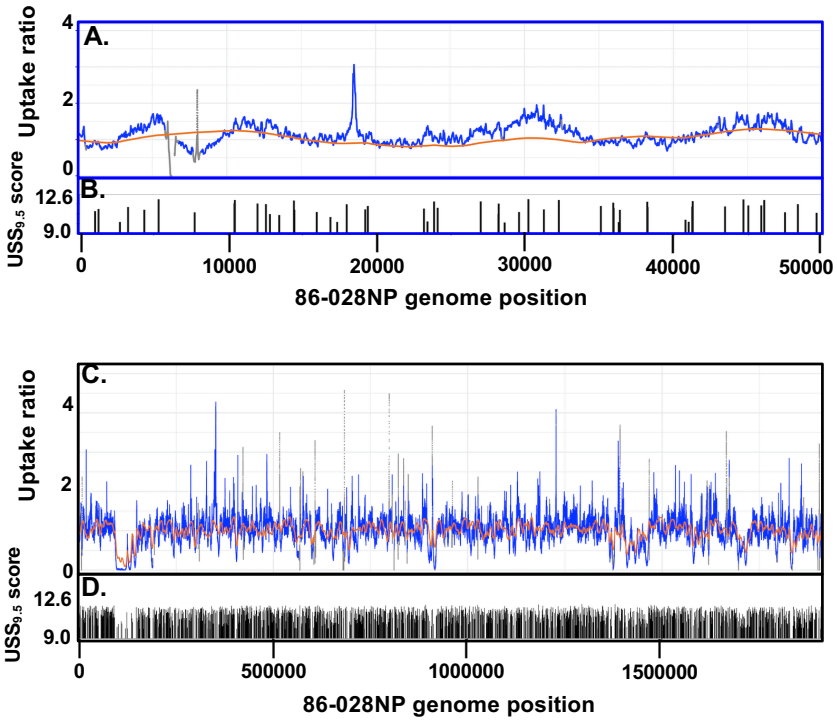


Figure 8. Predicted shape features of USS.

Legend: The thick grey line reproduced in each box shows shape analysis of the consensus USS sequence. Blue and orange lines in each box show shape analysis of genomic USS separated by at least 500 bp, grouped by score and coloured by uptake ratio. The orange and yellow bars below each box indicate components of the USS (see Figure 1): light orange: outer core; dark orange: inner core; yellow: AT tracts. **Panels A-D** $USS_{9.5\text{-}10}$: Blue: uptake ratios <0.2 (n=68, mean USS score =9.7). Orange: uptake ratios >0.2 (n = 10, mean USS score =9.8). **Panels E-H:** $USS_{10.0\text{-}10.5}$: Blue: uptake ratios <0.6 (n=47, mean USS score=10.22). Orange: uptake ratios >2.0 (n=10, mean USS score=10.26). **Panels I-L:** $USS_{10.5\text{-}11.0}$: Blue: uptake ratios <0.6 (n=14, mean USS score=10.64). Orange: uptake ratios >2.0 (n=59, mean USS score=10.79). **A, E and I.** Minor groove width, in Å. **B, F and J**. Propeller twist, in degrees. **C, G and K**. Helix twist, in degrees. **D, G and L**. Base pair roll, in degrees. See also Figure S6. Grey bars for each point show the standard error. Dots indicate significant differences by Kolmogorov-Smirnov between high-uptake and low-uptake positions with (red dots) and without (black dots) Bonferroni correction.

877



878

879 **Figure 9.** Predicted and observed uptake of long 86-028NP DNA fragments.

880 **Legend: A. and C.** Uptake maps for 86-028NP long-fragment DNA. Orange points: USS-

881 dependent uptake predicted by the revised model. Blue points: mean uptake ratios from 3

882 replicate experiments (grey indicates input coverage <20 reads, gaps indicate unmappable

883 positions). **B.** and **D.** $USS_{9.5}$ positions and scores. **A. and B.** The same 50 kb genome segment

884 shown in previous figures. **C. and D.** Whole genome.

885
886

887
888 **TABLES:**

889

890 **Table 1. Bacterial strains used in this study**

Strain number	Strain name	Phenotype	Source
RR3117	*rec2::spec*	Spectinomycin resistant Rd derivative. No translocation of taken-up fragments to the cytoplasm	Sinha *et al.* 2012
RR3125	*Rd Δrec2*	Unmarked Rd derivative. No translocation of taken-up fragments to the cytoplasm	Sinha *et al.* 2012
RR3133	86-028NP NalR	Otitis media clinical isolate. Nalidixic acid resistant	Mell *et al.* 2011
RR1361	*PittGG*	Nontypeable clinical isolate	G. Ehrlich
RR722	*Rd*	Rd KW20, rough (unencapsulated) derivative of type d	H. O. Smith

891

892

893

894 **Table 2. Frequencies of *Hin*-type USS in genomes of other species**

Species	Genome size	USS_{10}/Mb	USS_{11}/Mb
H. influenzae 86-028NP	1.914	1014	754
H. parainfluenzae T3T1	2.087	867	683
Aggregatibacter actino-mycetemcomitans VT1169	2.129	970	737
H. ducreyi 35000HP	1.699	102	16
Mannheimia haemolytica M42584	2.732	96	18
N. meningitidis	2.272	52	4
Streptococcus pneumoniae	2.039	16	1
Pseudomonas aeruginosa	6.264	8	0
Homo sapiens[1]	3 x 1.9 Mb	6.6	0.1
Random-sequence DNA[2], *41% G+C*	3 x 1.9 Mb	29.5	0.3

895 1. Means for three 1.9 Mb segments of human chromosomes 1, 3 and 12.

896 2. Means for three 1.9 Mb ‘genomes’ with the same base composition as human

897 DNA.

898 **Table 3. Predicted relative uptake[1] of *H. influenzae* DNA in simulated competition**

899 **with DNAs of other species[2].**

Assumptions:	**1 kb fragments, *p_uptake*=0**	**1 kb fragments, *p_uptake* =0.005**	**10 kb fragments, *p_uptake*=0**	**10 kb fragments, *p_uptak e*=0.005**
86-028NP DNA in competition with:				
***H. parainfluenzae DNA* (*Hin*-type USS)**	0.500	0.500	0.475	0.475
***H. ducreyi DNA* (*Apl*-type USS)**	0.952	0.951	0.884	0.883
***Homo sapiens DNA*[3] (no USS enrichment)**	0.998	0.997	0.991	0.990

900 [1.] *H. influenzae* DNA as a fraction of the total DNA predicted to be taken up.

901 [2.] Competitions were genome:genome

902 [3.] Means for the three genome 1.9 Mb segments described in Transparent Methods.