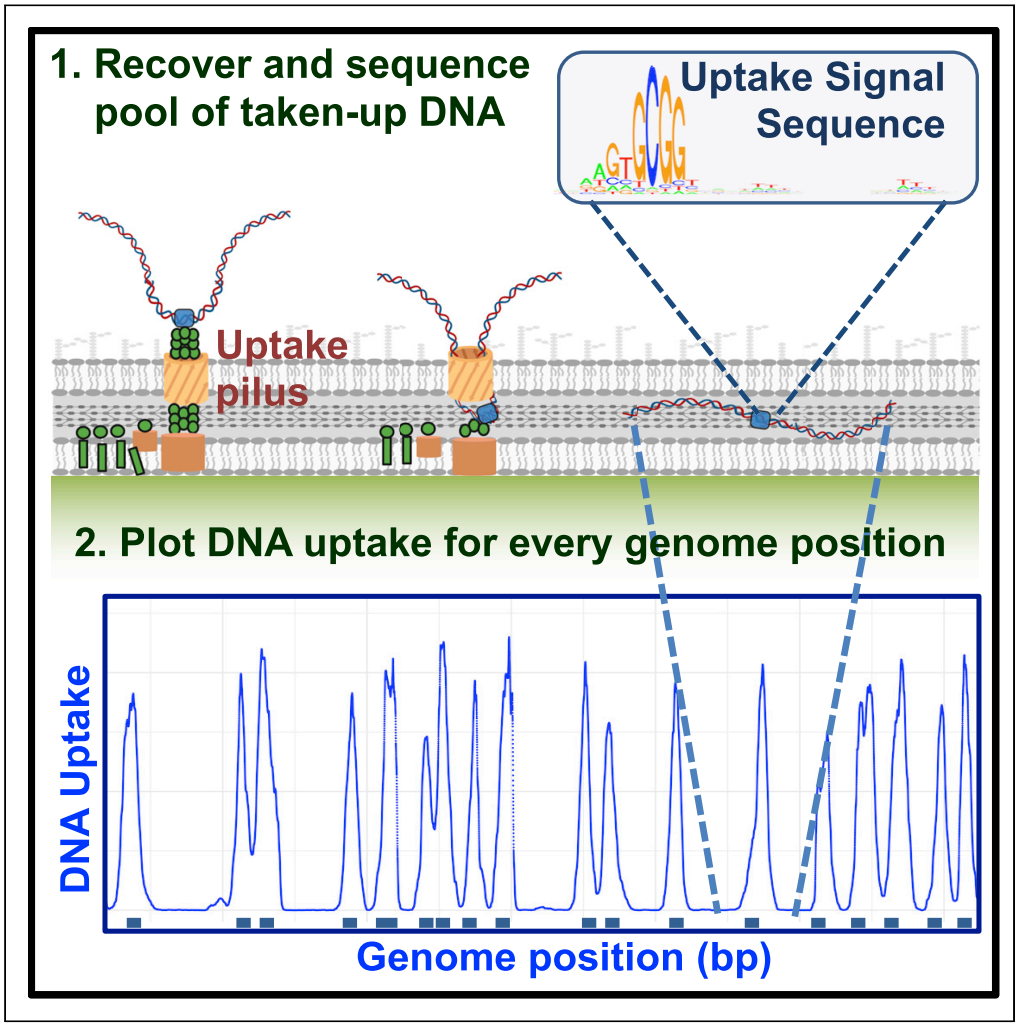
iScience

CellPress
OPEN ACCESS

Article

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



Marcelo Mora, Joshua Chang Mell, Garth D. Ehrlich, Rachel L. Ehrlich, Rosemary J. Redfield

redfield@zoology.ubc.ca

HIGHLIGHTS

For short DNA fragments, an uptake sequence (USS) improves DNA uptake 1000-fold

Most longer *H. influenzae* fragments have USS, giving even uptake across the genome

Preferred USS are stiff, so strand melting may facilitate kinking for uptake

H. influenzae will take up its own DNA 100-fold better than human DNA

Mora et al., iScience *24*, 102007
January 22, 2021
https://doi.org/10.1016/j.isci.2020.102007

iScience

CellPress
OPEN ACCESS

Article

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[1,4,*]

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 ~1,000 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 2-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 effectively limit uptake to DNA from close relatives with the same uptake specificity (Dougherty et al., 1979; Scocca et al., 1974). The distribution of the preferred sequences around the chromosome is uneven (Smith et al., 1995), which may cause different genes to experience quite different rates of genetic exchange.

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

[1]Department of Zoology, University of British Columbia, Vancouver, BC V6T 1Z3, Canada

[2]Department of Microbiology & Immunology, Center for Genomic Sciences, Institute of Molecular Medicine and Infectious Disease, Drexel University College of Medicine, 12 Philadelphia, PA 19102, USA

[3]Department of Otolaryngology – Head and Neck Surgery, Drexel University College of Medicine, 12 Philadelphia, PA 19102, USA

[4]Lead contact

*Correspondence: redfield@zoology.ubc.ca

https://doi.org/10.1016/j.isci.2020.102007

2016). Below we use "DNA uptake" to refer to the combined binding and membrane-transport steps that move DNA from the extracellular environment across the outer membrane into the periplasm. In Gram-positive bacteria similar machinery acts to pull the DNA through the thick cell wall (Chen and Dubnau, 2004). The next step in natural transformation is translocation of the DNA out of the periplasm and into the cytoplasm. Only the 3′-leading strand remains intact, passing through an inner membrane pore encoded by the *rec2*/*comEC* gene, whereas the other strand is degraded in the periplasm and its nucleotides are dephosphorylated and imported as nucleosides (Pifer and Smith, 1985). Circular DNA molecules are efficiently taken up across the outer membrane but remain the periplasm because they lack free ends (Pifer and Smith, 1985). As the single strand enters the cytoplasm it undergoes limited exonucleolytic degradation before being complexed with cellular proteins. If sequence similarity permits, the strand may then recombine with homologous chromosomal sequences; otherwise the strand is degraded to its constituent nucleotides (de Vries et al., 2001). Both linkage and sequencing studies indicate that fragments much longer than the cell are readily taken up (Goodgal, 1982; Mell and Redfield, 2014), as are fragments as short as 200 bp, although a lower limit has not been established (Mell et al., 2012; Maughan and Redfield, 2009). However, recombination of short fragments is limited because they are usually degraded by cytoplasmic nucleases before they can recombine (Pifer and Smith, 1985). Uptake speed has been estimated at 500–1,000 bp/s, with transformation essentially complete by 15 min (Deich and Smith, 1980).

Direct measures of DNA uptake bias

Uptake-competition experiments in the Pasteurellacean *Haemophilus influenzae* and *Neisseria gonorrhoeae* showed that genetically marked "self-derived" DNA competes for uptake with unmarked self-derived DNA but not with DNA from unrelated sources (Dougherty et al., 1979; Scocca et al., 1974). Subsequent DNA uptake experiments using cloned radiolabeled DNA fragments found that these self-preferences are caused by the uptake machineries' strong biases for short sequence motifs, called uptake signal sequences (USS) in *H. influenzae* and DNA uptake sequences (DUS) in *Neisseria* species (Sisco and Smith, 1979) (Davidsen et al., 2004). Sequence comparisons and site-directed mutagenesis initially identified the *H. influenzae* motif as an 11 bp sequence with a strong contribution by flanking AT-rich sequences (Danner et al., 1980, 1982), and later genome sequencing identified 1,465 occurrences of a 9 bp USS core in *H. influenzae* and 1,892 occurrences of an unrelated 10 bp DUS motif in *N. meningitidis* (Smith et al., 1995, 1999). Subsequent motif search analyses by Maughan et al. (Maughan et al., 2010) expanded this number to 2,206 USSs in the genome of the standard *H. influenzae* lab strain Rd, sharing the motif shown in Figure 1A. These genomic analyses were later complemented by direct uptake experiments using mutated and degenerate USS variants (Maughan et al., 2010; Mell et al., 2012). Mell et al. used mutagenesis and sequencing of pools of degenerate USS-containing fragments that had been recovered after uptake to identify the contribution to uptake of each USS position and found that the central GCGG bases are crucial for uptake, with much smaller contributions made by the flanking bases and AT-rich segments. The motif in Figure 1B shows the contribution of each base considered independently. Interaction effects between bases of the AT-tracts and the core were also found to make important contributions to uptake, but only a few of these have been directly measured. Although characterization of the unrelated Neisseriacean DUS has not reached this level of detail (Mathis and Scocca, 1982), the two uptake systems shares many features, apparently convergently evolved, including the presence within each family of lineages with slightly different preferred motifs (Frye et al., 2013; Redfield et al., 2006).

Evolution of uptake sequences in the genome

Alignment of homologous genomic regions from different Pasteurellaceae species showed that USS evolve by point mutations (Redfield et al., 2006); i.e. they are not inserted elements. As a mechanism for this evolution, (Danner et al., 1980) proposed that the combination of uptake bias and genomic recombination creates an evolutionary pressure that causes the preferred uptake sequences to accumulate throughout the genome, with their numbers limited by their eventual interference with gene function. Consistent with this, uptake sequences in both *Haemophilus* and *Neisseria* are underrepresented in newly acquired segments, rRNA genes, and coding sequences, especially those with strong functional constraints (Findlay and Redfield, 2009; Smith et al., 1999). Modeling by Maughan et al. (Maughan et al., 2010) confirmed that this molecular drive process could produce uptake sequence distributions as those of real genomes, with no need for any fitness benefit from either the uptake sequences or the recombination they promote. Thus, the presence of biased DNA uptake machinery may be sufficient in itself to explain the abundance of uptake sequences. Such biases may be solely a consequence of direct selection on the DNA uptake machinery for more effective DNA binding or may have been reinforced by indirect selection for preferential uptake of conspecific DNA. These processes might be especially important in respiratory tracts and other

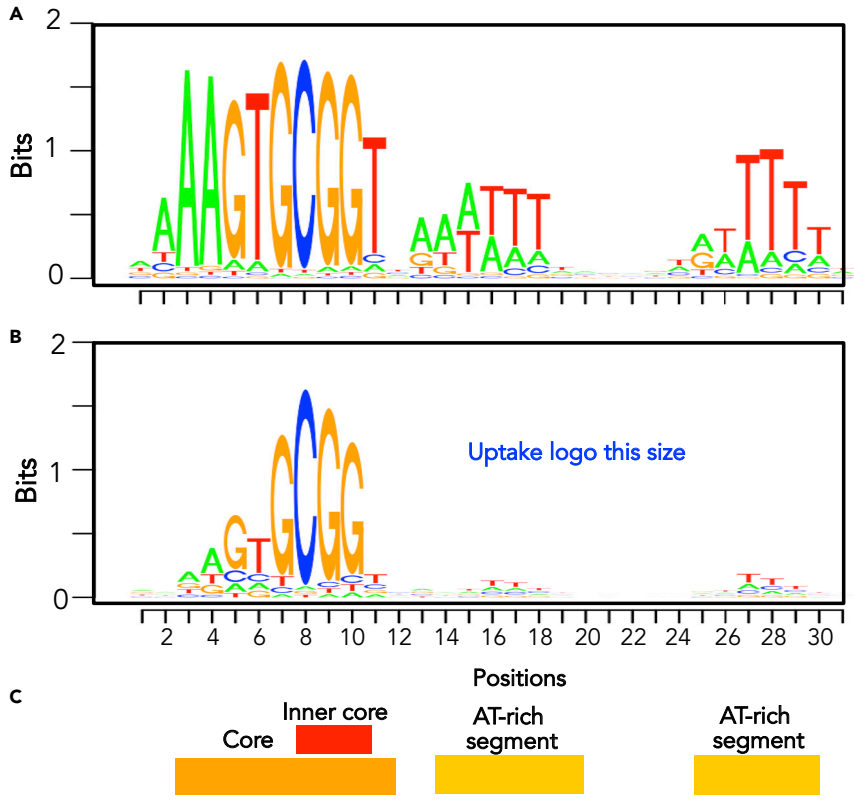


Figure 1. The *H. influenzae* uptake signal sequence

(A) Sequence logo showing the individual contributions to genomic abundance of bases in the USS motif (Maughan et al., 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.

mucosal environments where Pasteurellaceae and Neisseriaceae species mainly occur (Man et al., 2017). These environments contain abundant host DNA, and transformation can only occur if the released bacterial DNA competes successfully for binding to the uptake machinery (Lethem et al., 1990; Shak et al., 1990).

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

RESULTS

A computational model of DNA uptake

As a framework for interpreting DNA uptake data we developed a simulation model of USS-dependent DNA uptake. It takes as input the locations and strengths of USSs in the DNA whose uptake is to be

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

simulated, the fragment-size distribution of this DNA, and binding and uptake functions that specify how uptake probability depends on USS presence and strength. The output is the expected relative uptake of every position in the genome.

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

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

Simulating these steps required first specifying the genomic sequences that should be treated as USS. This was not straightforward because genomes contain many USS variants that differ in how well they promote DNA uptake (Findlay and Redfield, 2009; Maughan and Redfield, 2009). Our strategy was to score the information content (in bits) of every genome position, using the Position-Specific Scoring Matrix (PSSM) from Mell et al.'s degenerate-sequence uptake experiment (Mell et al., 2012) (Table S1) and to use overrepresentation of high-scoring sequences as the USS criterion. We scored every genome position in the three *H. influenzae* strains used for the experiments described below (Table 1) and in four randomly generated sequences with the same length and base composition (Figure S1 shows the score distributions). In the *H. influenzae* genomes, overrepresentation of high-scoring sequences was detectable above a score of 7.0 bits and became dramatic above 10.0 bits, where the numbers of high-scoring positions increased in *H. influenzae* genomes but became vanishingly small in the random-sequence controls (see inset in Figure S1). Because the slight overrepresentation of scores between 7 and 10 bits was hypothesized to be not a direct effect of DNA uptake but an indirect consequence of mutational degeneration of high-scoring USSs, the model initially used a USS cut-off score of 10 bits ("USS_{10}," n = 1941 in strain 86-028NP). This was later reduced to a less stringent 9.5 bits after weak uptake effects had been examined ("$USS_{9.5}$," n = 2,248 in strain 86-028NP).

The computational model used these USS scores to predict DNA uptake for every position in the genome, summing the contributions of binding and uptake probabilities from DNA fragments of different sizes (Figure 2A gives an overview). Each fragment under consideration was first checked for the locations of any USS_{10}s, and the maximum fragment length (20 kb) and the mean length of USS-free segments (*mean_gap*)

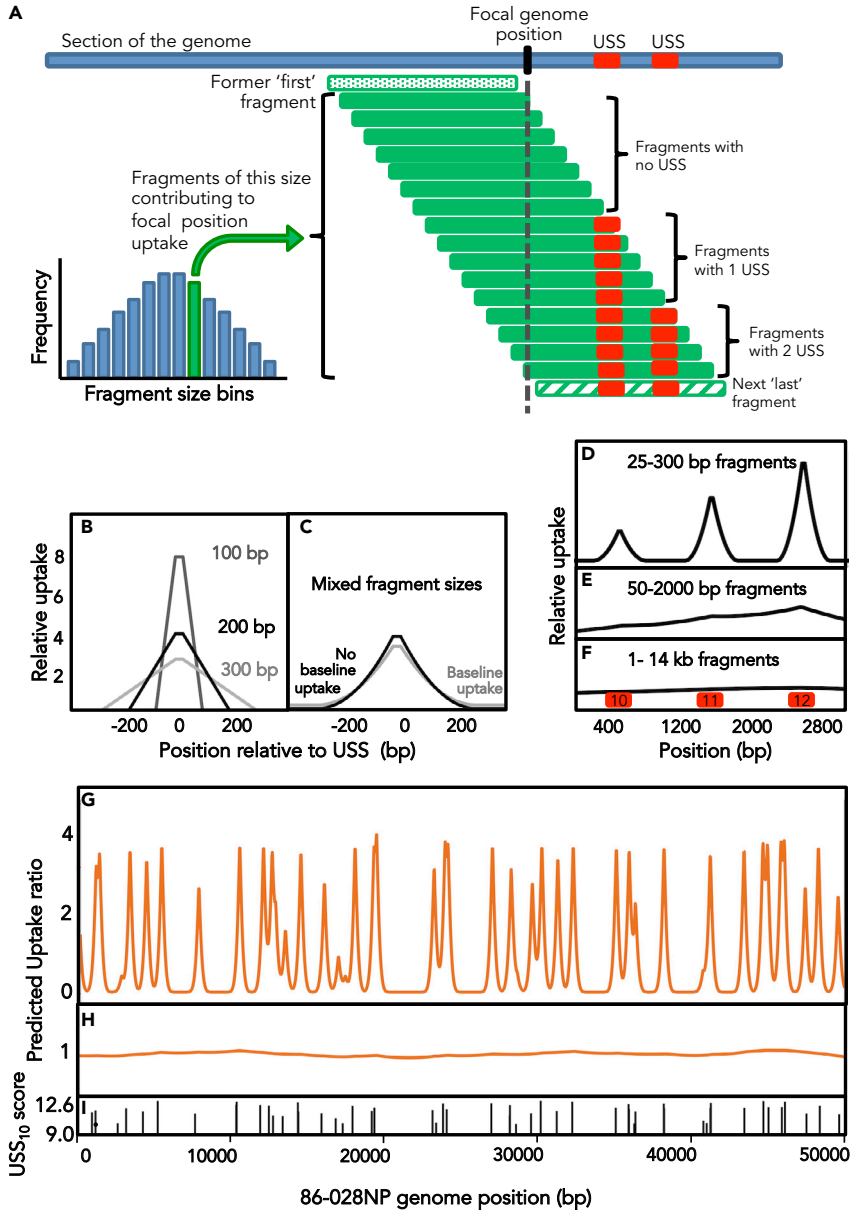


Figure 2. A computational model to predict DNA uptake

(A) Components of the DNA uptake model (see Transparent methods and Results for details).

(B and 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 and 300 bp with and without baseline uptake.

(D–F) Model predictions for uptake of a 3,000 bp region with 3 USSs (red squares, scores in black) using different fragment-length distributions: (D) 50–300 bp fragments, (E) 50–2000 bp fragment, (F) 1–14 kb fragments.

(G and H) Predicted DNA uptake of a 50 kb segment of the 86-028NP genome using (G) short- and (H) long-fragment length distributions:

(I) Locations and scores of USS_{10}s in this 50 kb segment.

See also Table S1 and Figures S1 and S2.

were used to calculate the probability that a DNA receptor protein initially encountering a random location in the fragment would then encounter and bind specifically to a USS_{10} rather than disassociating from the fragment: *p_bind = 1 – mean_gap/20000*). Fragments with no USS were initially assigned a baseline *p_bind* of 0.1.

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

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

Model results

Figures 2B and 2C show examples of model predictions for simple situations. Figure 2B shows the uptake predictions for an 800-bp simulated genome containing a single USS with score 12.0 bits, considering three different input DNA fragment sizes (100, 200, and 300 bp). The peaks at the USS have straight sides, a basal width twice the length of the fragments being taken up, and 31-bp flat tops arising from the model's requirement for a full-length USS. When the DNA fragment sizes were evenly distributed between 25 and 300 bp in length (Figure 2C), the peak had steep sides at its tops and gradually flattened at the base; maximum width at the base equaled twice the maximum fragment length. In simulated mini-genomes with more than one USS (Figures 2D–2F), isolated peaks were only seen when the DNA fragments being taken up were substantially shorter than the spacing of the USSs (Figure 2D), and peaks disappeared entirely when the fragments were long enough that almost all contained at least one USS (Figure 2F).

Figures 2G and 2H show the predicted uptake maps when this model analyzed a 50kb segment of the *H. influenzae* 86-028NP genome, using the short-fragment and long-fragment length distributions from the actual uptake experiments described below (Figures S2A and S2B), and Figure 2I shows the distribution of USSs over this segment. Because the "short" DNA fragments are shorter than the typical separation between USSs, the model predicts that uptake will be restricted to sharp peaks at each USS. In contrast, uptake of long DNA fragments is predicted to be much more uniform, because most of these will contain at least one USS.

Generation of experimental DNA uptake data

To obtain high-resolution measurements of actual DNA uptake, we sequenced *H. influenzae* genomic DNA that had been taken up by and recovered from competent *H. influenzae* cells. Competent cells of the standard laboratory strain Rd were first incubated with genomic DNA preparations from strains 86-028NP and PittGG, whose core genomes are readily differentiated from Rd (and each other) because they differ at ~3% of orthologous positions (Hogg et al., 2007). To allow efficient recovery of the taken-up DNA, the Rd strain in which competence was induced carried a *rec2* mutation that blocks translocation of taken-up DNA fragments, causing the DNA to be trapped intact in the periplasm (16). The 86-028NP and PittGG genomic DNAs were pre-sheared to give short (50–800 bp) and long (1.5–17 kb) DNA preparations (size distributions are shown in Figure S2), and three replicate uptake experiments were done with each DNA preparation. After 20-min incubation with competent cells, the taken-up DNA was recovered from the cell periplasm using the cell-fractionation procedure of Kahn et al. (Barouki and Smith, 1985; Kahn et al., 1983; Mell et al., 2012). Taken-up DNA samples were sequenced along with samples of the input 86-028NP and PittGG DNAs and of the recipient Rd DNA. The input and uptake reads were then aligned to the corresponding 86-028NP and PittGG reference sequences, and coverage at every position was calculated. Table S2 provides detailed information about the four input samples, the twelve uptake samples, and the Rd sample.

Removal of contaminating Rd DNA

Preparations of DNA recovered from the periplasm after uptake always included some contaminating DNA from the recipient Rd chromosome. The divergence between the Rd and donor genomes allowed us to estimate the extent of this contamination by competitively aligning the taken-up reads from each sample to an artificial reference "genome" consisting of both recipient and donor genomes as separate

"chromosomes." Reads that uniquely aligned to only one chromosome could then be unambiguously assigned to either donor (taken-up) or recipient (contamination). The resulting estimates of Rd chromosomal contamination were between 3.2% and 19.3% of reads; sample-specific values are listed in Table S2.

The effects of this contamination were not expected to be uniform across each donor genome, because segments of the 86-028NP and PittGG genomes with high divergence from or with no close homolog in Rd would be free of contamination-derived reads. We used the competitive-alignment described earlier to create contamination-corrected uptake coverages by discarding all reads that could not be uniquely mapped to the donor genome; in addition to removing Rd contamination, this also removed reads from segments that are identical between the donor and recipient strains ("double-mapping reads") and reads that mapped to repeats, such as the six copies of the rRNA genes. For consistency, the same changes were applied to the input samples although they did not experience any contamination. This correction removed an average of 18.6% of reads (range 8.9%–28.3%), left some segments of the 86-028NP and PittGG genomes with no coverage in all samples (2.3% and 2.1% respectively), and reduced coverage adjacent to these segments. Figure S3 shows the locations of the missing data. Contamination details for each sample are provided in Table S2, and the impacts are considered below. The uptake analysis described below showed this correction to be effective.

Uptake ratios

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

Figure 3A shows the short-fragment uptake ratio map for the first 50kb of the 86-028NP genome; the ticks in Figure 3C indicate locations and scores of USS_{10}s. The pattern is strikingly similar to that predicted for the same DNA segment by the model (Figure 2G). Sharp uptake peaks are seen at USS_{10} positions, some separated by flat-bottomed valleys and others overlapping. Figures S4A–S4C show similar analysis for the first 50 kb of strain PittGG's genome, and Figures S5A and S5B show expanded maps for two examples of the 86-028NP peaks. The full-genome maps of these uptake ratios are provided in Figure S4D (86-028NP) and S4G (PittGG); they display the consistency of the peak heights across each genome.

As expected, the long-fragment DNA samples (Figures 3B, S4B, S4E, and S4H) had much less variation in uptake ratio than the short-fragment samples; 90% of positions had uptake ratios within 2-fold of the mean, and there were few high peaks or low valleys. The few extended segments with low or no uptake coincided with large gaps between USS_{10}s. The largest gap was in the 86-028NP segment between 95 and 145kb—the site of a genomic island that is absent from the PittGG and Rd strains, has few USS, and has high similarity to an *H. influenzae* plasmid (Harrison et al., 2005). However, uptake ratios did exhibit substantial short-range variation not predicted by the model, which is considered further below.

Sensitivity is limited by low sequencing coverage

At some genome positions our ability to detect USS-dependent uptake biases and possible USS-independent biases in uptake coverage was limited by low sequencing coverage. Although some of this low coverage arose from the contamination-correction step described earlier, segments of low coverage were also seen in the raw data, likely arising from biases in the library preparation and sequencing steps. Figure S6 compares the raw coverage of short and long input samples for a 50 kb segment of the 86-028NP genome, illustrating the strong variation in sequencing coverage that was both broadly reproducible and sequence dependent. Low coverage had similar effects in all samples, precluding calculation of uptake ratios where input coverage was zero and generating high levels of stochastic variation where coverage was low (segments with no coverage and positions with coverage below 20 reads are indicated in uptake ratio maps by gaps and gray dots, respectively).

We used the low uptake ratios of positions at least 1,000 bp from the nearest $USS_{9.5}$ to assess the effectiveness of the contamination correction described earlier, because these positions are where contamination

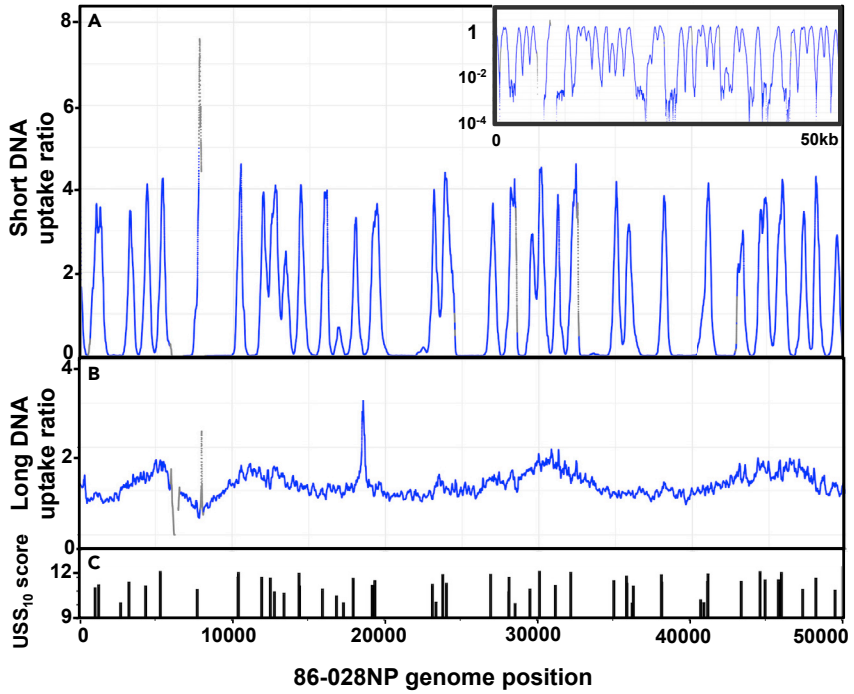


Figure 3. Experimentally determined uptake ratios for a 50 kb segment

The X axis is the same 50 kb segment of the 86-028NP genome as Figures 2G and 2H. Gray points indicate positions with input coverage lower than 20 reads. Gaps indicate unmappable segments.

(A) Uptake ratios of short-fragment DNA. **Inset:** same data with a logarithmic-scale Y axis.

(B) Uptake ratios of long-fragment DNA.

(C) Locations and scores of USS_{10}s.

See also Figures S3 and S4.

would have had its largest effects. This peak-separation distance was chosen because a peak-shape analysis of uptake around 158 USS_{10}s that were separated by at least 1,200 bp from other USS_{10}s and had uptake ratios of at least 3.0 found that mean uptake ratio had fallen to baseline at positions 600 bp from the USS (Figure S7). Figure 4 compares uptake ratios between the valley positions that received the contamination correction described earlier (those with a Rd homolog) and the positions where no correction was needed (those with no Rd homolog). The two distributions had nearly identical medians (0.0022 and 0.0020, respectively), indicating that the correction was sufficient but not excessive. These low values also confirm that the DNase I treatment and washing steps removed at least 99.99% of the donor DNA that had not been taken up.

Uptake ratios show no periodicity across the genome

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

Uptake of short-fragment 86-028NP DNA

To investigate how USSs contribute to the DNA uptake process, our strategy was to first analyze discrepancies between model predictions and observed uptake ratio peaks in the 86-028NP short-fragment dataset, because these would reveal ways in which the simple assumptions underlying the model mischaracterized the actual steps of DNA uptake. Model changes that improved the predictions were considered to better reflect the true constraints on uptake of short DNA fragments. The refined model's predictions

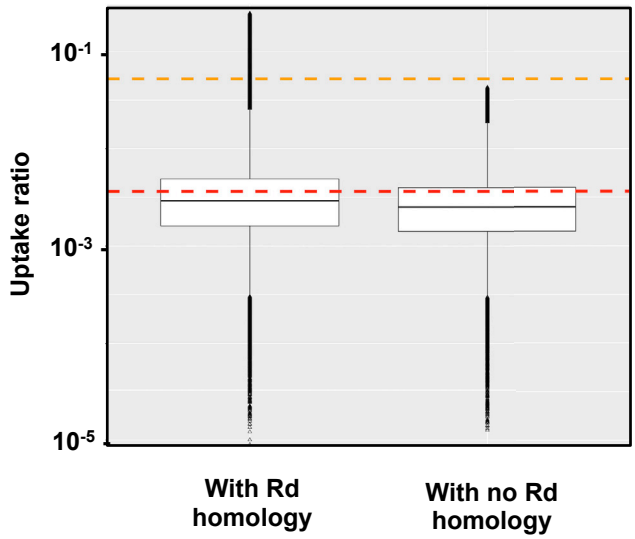


Figure 4. Depths of uptake ratio valleys with and without contamination correction

Uptake ratios of positions in the 86-028NP short-fragment dataset that were at least 1kb from the nearest $USS_{9.5}$ and whose uptake coverages were either corrected for contamination with homologous Rd sequences (left, 272,547 positions) or not corrected for contamination because they had no Rd homology (right, 89,418 positions). Orange and red dashed lines indicate valley uptake predicted by the original and revised models respectively.

See also Figure S6.

were then compared with the real uptake ratios for 86-028NP long-fragment DNA, allowing additional refinement, and finally to the short- and long-fragment uptake ratios for PittGG DNA.

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

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

Second, the score cutoff of 10.0 bits was too high. Analysis of score-subsets within the uptake-valley dataset showed that 9.5 bits was a better cutoff. For the 78 positions with scores between 9.5 and 10.0 bits, the correlation between score and uptake ratio was 0.27. Although a similar correlation (0.27) was seen for the 136 positions with scores between 9.0 and 9.5 bits, it was driven mainly by very small effects, and only four positions had uptake ratios higher than 0.1. For the 461 positions with USS scores between 8.5 and 9.5 bits the correlation was only 0.09.

Third, most of the predicted peaks were too high. Figure 6 complements Figure 4's analysis of uptake valleys with an analysis of similarly isolated uptake peaks. The blue dots show the uptake ratio at every $USS_{9.5}$ position in the genome that is at least 1,000 bp from the nearest USS_{10} (n = 209), and the small orange dots show predicted uptake at the same positions. The lack of scatter in the orange points confirms that the separation distance was sufficient to avoid predicted effects of nearby USS; the least-squares difference between predicted and observed uptake ratios at these 209 USS positions was 0.521.

In the original-version of the uptake model, the sigmoidal parameters of the *p_uptake* function (baseline and location and slope of inflection point) were set using the frequencies of USS scores in the 86-028NP genome; as expected, the same parameters were obtained when a sigmoidal function was fit to the uptake

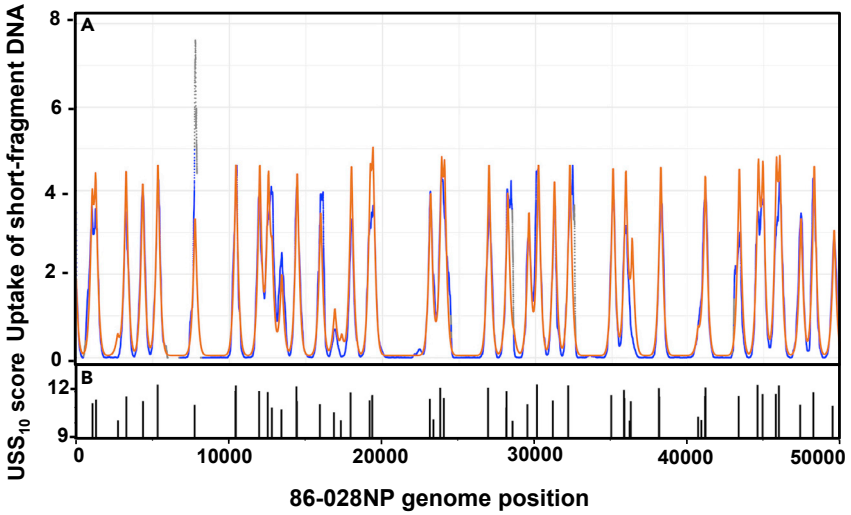


Figure 5. Predicted and observed DNA uptake analysis for short fragments of 86-028NP DNA

(A) Map of uptake ratios and initial model predictions. The blue points show the same uptake ratio map as in Figure 3A. The orange points show the same predicted uptake as in Figure 2G.

(B) Locations and scores of $USS_{9.5}$s.

points predicted by this model (orange line in Figure 6). To improve the predictions we replaced these *p_uptake* parameters with those of a sigmoidal function fitted to the real uptake data (blue line in Figure 6), changed the *p_uptake* baseline to 0.005, lowered the USS score cutoff from 10.0 bits to 9.5 bits, and ran the model again. The purple points in Figure 6 show that although peak-height predictions for low-scoring USS were somewhat improved by the changed *p_uptake* function, those for USS with scores >11.5 bits became slightly worse; the least-squares difference for the 209 positions was modestly reduced from 0.521 to 0.420.

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

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

Weak uptake biases could arise from either low-scoring USS or non-USS sequence factors. Because weak biases would only be detectable in genome segments that lack strong USSs, we searched for them using a far-from-USS_{10} dataset containing only DNA segments whose ends were at least 0.6kb from the closest USS_{10} and that had input coverage of at least 20 reads. This dataset contained 513 segments where weak uptake effects could in principle be detected (22% of the genome); their mean uptake ratio over the 428678 positions was 0.0102. In these segments, the only positions with distinct uptake ratio peaks >0.2 were nine weak USSs with scores between 9.5 and 9.9 bits (an example is shown in panel C of Figure S5). Because this analysis did not find any non-USS positions giving uptake higher than 0.2, it suggests that other sequence factors do not detectably promote uptake in the absence of a USS.

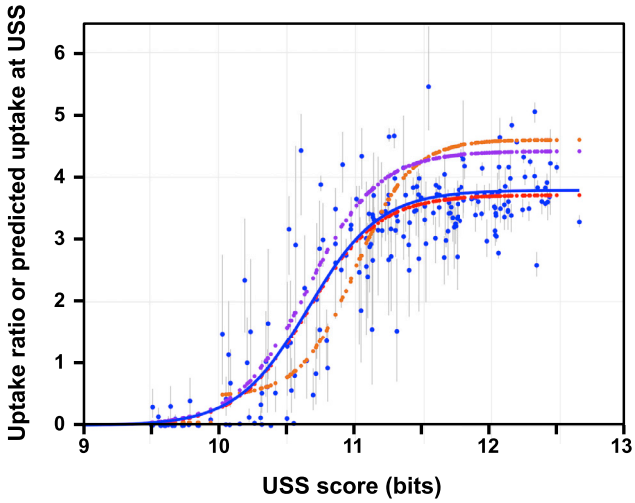


Figure 6. Short-fragment uptake ratios and predicted DNA uptake at isolated 86-028NP USS as a function of USS score

Predicted or measured DNA uptake at 209 $USS_{9.5}$ positions separated by at least 1,000 bp from the nearest USS_{10}s. The blue dots show the measured uptake ratios; gray bars show the ranges of the three replicates at each position. The blue line shows a sigmoidal function fit to these points. The small orange dots and line show uptake predicted by the original model at the same positions, and the small purple and red dots show uptake predicted by the intermediate and revised model versions discussed in the text.

See also Figure S6.

The degenerate-USS analysis of Mell et al. (Mell et al., 2012) found that pairwise interactions between the AT-tract bases and core bases of a USS made substantial contributions to uptake of a 200 bp synthetic fragment. To evaluate the effects of these interactions in predicting uptake of genomic DNA, the USS scores used by our revised model were raised or lowered in proportion to the interaction effects reported in Figure 6 of Mell et al. Figure 7 shows that for the same 209 isolated $USS_{9.5}$s used earlier (Figure 6), this adjustment had little effect on high-scoring USS but further reduced the scores of low-scoring USS. The data points are colored by their uptake ratios, showing that weak USS giving unusually high or low uptake were equally likely to have their scores reduced. However, these scoring changes had no effect on the short-fragment uptake predictions of the revised model (both Pearson correlations 0.93), probably because (1) the scores of strong USS were not significantly changed and (2) the scores of weak USS already gave near-baseline uptake.

DNA shape effects

Most of the interactions in Figure 7 were between bases separated by 10 bp or more, so we complemented that analysis with analysis of DNA shape, which reflects both pairwise and more complex interactions over a 5 bp range. The major shape features that can be predicted from DNA sequence are the minor groove width, the propeller twist between two paired bases, the helix twist between one base pair and the next, and the roll of one base pair relative to the next. In Figure 8, the wide gray line reproduced in each panel shows these feature predictions for the consensus USS. The USS inner core (orange shading) has a relatively wide minor groove and high propeller twist, which would facilitate sequence recognition by proteins (Rohs et al., 2009). To the left of this and in both AT-tracts (yellow shading) the minor groove is narrow with low propeller twist and negative roll, predicting that these segments are both rigid and slightly bent.

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

In all three score subsets the inner-core shape features were very similar for low-uptake and high-uptake subsets (blue and orange lines), probably because this sequence perfectly matches the consensus in 206

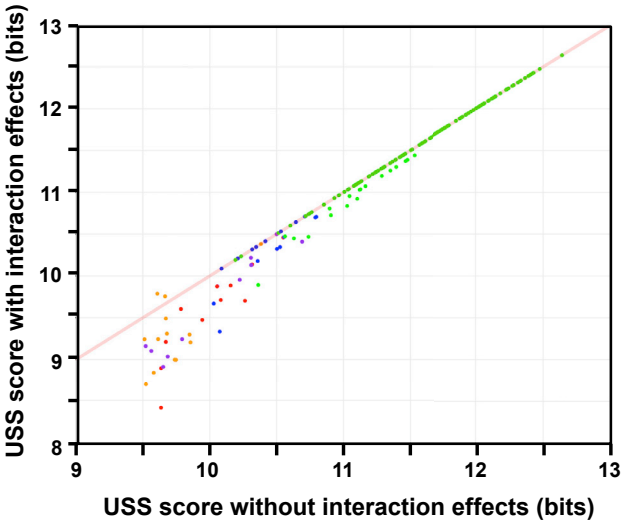


Figure 7. Effects of within-USS interactions on USS scores

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

of the 209 $USS_{9.5}$. (The other three had otherwise-perfect core sequences and good AT tracts but had only baseline uptake ratios.) However, the AT-tract shapes had marked differences, with low-uptake USSs having no distinctive shape features and high-uptake USSs resembling the USS consensus shape. This suggests that the predicted rigidity and slight bend caused by interactions within the consensus AT tracts facilitate DNA uptake.

Uptake of fragments with more than one USS

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

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

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

Overall, these results indicate that the presence of two USS_{10}s within 100 bp does not detectably increase the probability of the receptor binding to a USS, a result consistent with that of Ambur et al. for pairs of closely spaced DUS in *Neisseria meningitidis* (Ambur et al., 2007). Because individual fragments were not tracked, we cannot make any conclusions about uptake of fragments with more widely separated USS.

Uptake of long-fragment 86-028NP DNA

Figure 9A compares the revised model's predictions for long-fragment 86-028NP DNA with the uptake ratios observed over the same 50kb genome segment as in Figure 3B. In contrast to predictions

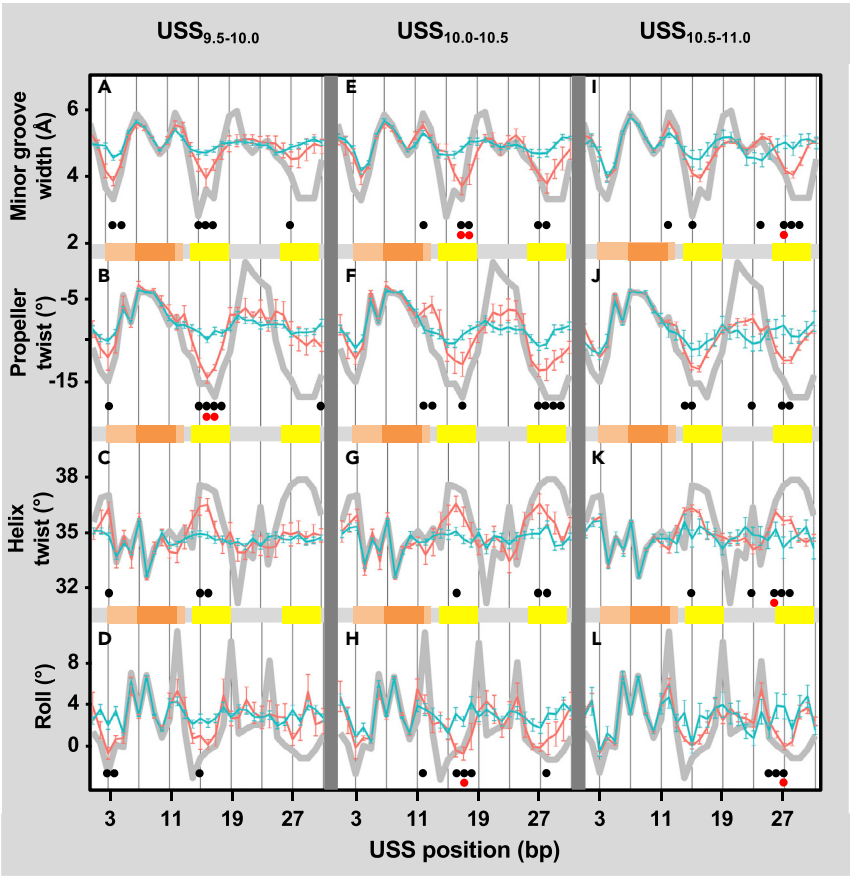


Figure 8. Predicted shape features of USS

The thick gray 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 colored 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.

(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).

(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).

(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. Gray 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.

for short-fragment uptake (correlation of 0.93), it seriously underpredicted the variation in long-fragment uptake ratios (correlation of 0.60). Although 51% of 86-028NP positions had long-fragment uptake ratios lower than 0.8 or higher than 1.2, only 13% had predicted uptake outside these limits.

Because long-fragment uptake ratios lacked the dramatic peaks and valleys seen for short fragments, stochastic noise arising at the regions of low sequencing coverage described earlier was expected to play a larger role. To estimate the magnitude of this effect, we compared the effects of adding different amounts of artificially generated noise to simulated (noise-free) uptake data (Figure S10A shows examples of noise-free and noise-added coverages). Figure S10B shows that, as expected, the correlation between noisy and noise-free data worsened as the arbitrary level of noise increased for both short-fragment (blue) and long-fragment (red) simulations and that increasing noise had a much stronger effect on the long-fragment simulations. For short DNA fragments, simulations with noise levels of 2 and 2.5 gave correlations of 0.94 and

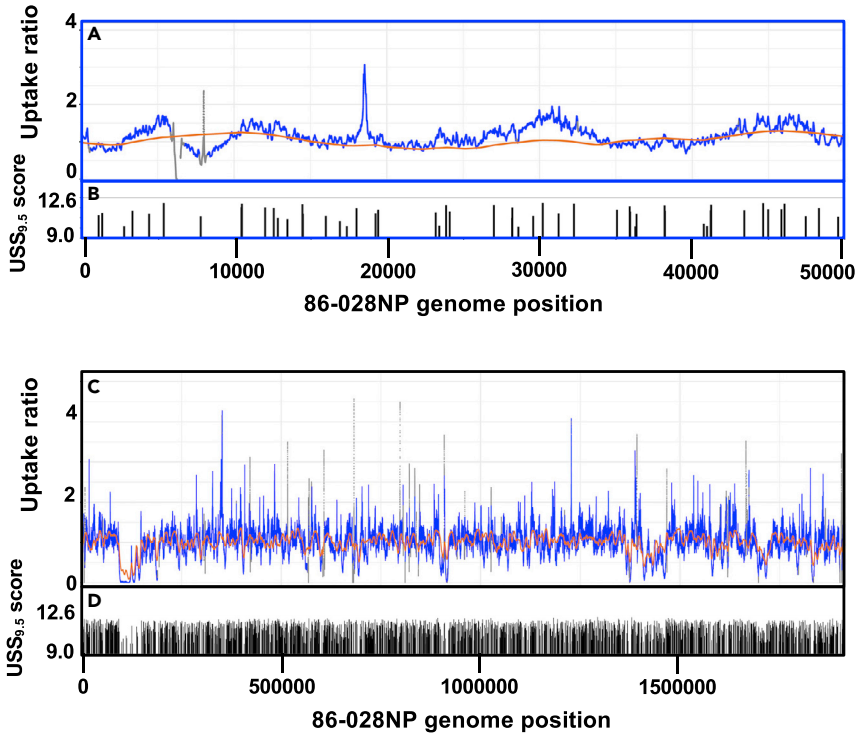


Figure 9. Predicted and observed uptake of long 86-028NP DNA fragments

(A) and (C) Uptake maps for 86-028NP long-fragment DNA. Orange points: USS-dependent uptake predicted by the revised model. Blue points: mean uptake ratios from three replicate experiments (gray indicates input coverage <20 reads, gaps indicate unmappable positions).

(B) and (D) $USS_{9.5}$ positions and scores. (A) and (B) The same 50 kb genome segment shown in previous figures. (C) and (D) Whole genome.

0.92, very close to the 0.93 correlation between the revised model and the real data (dashed blue line in Figure S9B). For long DNA fragments with the same noise levels, more noise was needed, with a 3.0 multiplier giving a correlation of 0.56, slightly lower than the observed correlation of 0.60 (dashed red line in Figure S9B). This confirms that the disparities between measured uptake ratios and USS-based predictions were at least partially due to noise in the sequencing data. However, these correlations are still 21% and 10% higher than the best correlation obtained between the revised model's predictions and real data.

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

Prediction of uptake of PittGG DNA

Because the uptake model was refined using uptake data for DNA of strain 86-028NP, the revised model's predictions were assessed using uptake data for DNA of a different strain, PittGG, which differs from 86-028NP DNA by SNPs and indels affecting about 11% of its genome. Figure S11 compares the uptake predictions with the observed PittGG uptake ratios. For short-fragment data the peak heights and valley

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 actinomycetemcomitans 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[a]	3 x 1.9 Mb	6.6	0.1
Random-sequence DNA[b], *41% G + C*	3 x 1.9 Mb	29.5	0.3

[a]Means for three 1.9 Mb segments of human chromosomes 1, 3, and 12.
[b]Means for three 1.9 Mb "genomes" with the same base composition as human DNA.

depths were similar to those for 86-028NP, as was the Pearson correlation between predicted and observed uptake (0.92). Although the model predicted similar long-fragment variation for 86-028NP and PittGG, the PittGG experimental data showed more extreme variation, and the correlation was only 0.41, substantially worse than the 0.60 obtained for 86-028NP. As with the 86-028NP data, much of this discrepancy may be due to noise in regions of low sequencing coverage. However, this may not fully explain PittGG's lower correlation, because input samples of the two strains had similar frequencies of low-coverage positions (Table 2.2).

Predicted competition with other DNAs in the human respiratory tract

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

Table 2 lists the frequencies of USS_{10} and USS_{11} in the genomes of various species. Pasteurellaceae species that share *H. influenzae*'s *Hin*-type USS (e.g. *H. parainfluenzae* and *Aggregatibacter actinomycetemcomitans*) typically have about 1000 USS_{10}s per Mb (Redfield et al., 2006); most of these are strong USS with scores ≥11. Pasteurellaceae with the *Apl*-type USS (e.g. *H. ducreyi* and *Mannheimia haemolytica*) have about 10-fold fewer *Hin*-type USS_{10}s, and fewer than 20% of these are strong. Other common respiratory tract bacteria (e.g. *N. meningitidis, Streptococcus pneumonia* and *Pseudomonas aeruginosa*) have USS only at the frequencies expected for their base compositions. The human genome is exceptional in having about 5-fold *fewer* USS_{10} per Mb than expected from simulated sequences with the 41% GC content of human DNA (observed frequency 4.6/Mb, simulated frequency 30/Mb, both with mean scores of only 10.3 bits). Because the USS inner-core motif includes a CpG, the underrepresentation of USS_{10} in human DNA relative to simulated sequences is probably a consequence of the 4- to 5-fold depletion of CpGs in the human genome caused by deamination of methylated cytosines (Babenko et al., 2017; Bird, 1986).

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

Table 3. Predicted relative uptake of *H. influenzae* DNA in simulated competition with DNAs of other species

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

[a]*H. influenzae* DNA as a fraction of the total DNA predicted to be taken up.
[b]Competitions were genome:genome.
[c]Means for the three genome 1.9 Mb segments described in Transparent methods.

uptake at every position was then summed to get a total genome uptake value for each fragment length and baseline assumption.

Table 3 shows the relative amounts of *H. influenzae* DNA predicted to be taken up under the various competition conditions. As expected from published uptake-competition experiments (Albritton et al., 1984; Redfield et al., 2006), *H. influenzae* and *H. parainfluenzae* DNAs were taken up with equal efficiency in all simulated conditions, and *H. ducreyi* DNA was much less competitive, contributing only 5% of the 1 kb fragments and only 12% of the 10 kb fragments. With human DNA as the 1:1 competitor, more than 99% of the DNA taken up was predicted to be from *H. influenzae*. For all competing DNAs, baseline uptake of fragments containing no USS made only a tiny contribution.

DISCUSSION

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

Implications for DNA uptake

The USS motif

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

Effect of USS shape

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

Other USS effects on DNA uptake

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

Lack of USS-independent effects

The very low valleys between short-fragment uptake peaks allowed us to examine more than 400,000 positions for USS-independent increases in uptake. None were found; the only positions with distinct uptake ratio peaks >0.2 were nine weak USSs.

Implications for genetic exchange

Preferential uptake of USS can create variation in recombination at all levels: across a single genome, between strains of one species, and between both closely related and unrelated species. Pifer and Smith showed that transformation frequency in *H. influenzae* decreased exponentially when fragments were smaller than 3.5 kb. This decrease was attributed to exonuclease degradation of fragments from their 3′ ends, because similar numbers of short and long fragments were taken up, only 5′-end label was incorporated into the chromosome, short fragments transformed more efficiently when the selected marker was far from one end. Mell and Redfield (2014) used genome sequencing of inter-strain recombinants to examine the distribution of recombination tract lengths; despite the presence of 2%–3% SNPs, the mean tract length was 6.9 kb and the longest was 43 kb. The efficient uptake and recombination of long fragments allows fragments containing non-homologous segments to still recombine well, provided both fragment ends are homologous. Fragments with only one homologous end recombine much less efficiently ("homology-facilitated recombination" (de Vries and Wackernagel, 2002)), and integration of fragments with no homology is very rare.

Across the genome

Across the genome of an *H. influenzae* strain, the genetic consequences of USS-dependent DNA uptake depend on USS locations and the lengths of the available DNA fragments. If only short fragments are available, the limitation to positions close to a USS may be obscured by limitation caused by degradation of incoming DNA in the cytoplasm. If most fragments are long, recombination will be both more frequent and more evenly distributed across the genome, because long fragments are more likely to both contain USS and recombine. The result will be that almost all recombination is caused by fragments long enough to usually contain at least one strong USS. This situation is caused by the high abundance and relatively even distribution of genomic USS, itself an expected consequence of the functionally neutral accumulation of USS under a molecular drive caused by USS-biased DNA uptake (Danner et al., 1982; Maughan et al., 2010).

Recombination between H. influenzae strains

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

Recombination between Pasteurellaceae species

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

Competition with DNAs from human cells and other respiratory bacteria

In the respiratory tract, the most important source of competing DNA is human cells. However, our analysis suggests that *H. influenzae*'s uptake specificity allows its DNA to outcompete human and other foreign

DNAs even if these are in 100-fold excess, a combined effect of the low number of USS_{10}s and their poor match to the uptake motif. Note that efficient self-uptake does not necessarily imply a selective advantage, because USS accumulation in *H. influenzae*'s genome may simply be due to the molecular drive process.

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

Limitations of the study

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

Resource availability

Lead contact

Further information and requests should be directed to and will be fulfilled by the Lead Contact. Rosemary J. Redfield (redfield@zoology.ubc.ca).

Materials availability

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

Data and code availability

All fastQ files have been deposited under NCBI BioProject:PRJNA387591. The corresponding BioSamples are listed in Table S2. FastQ files were also deposited at Mendeley data (https://doi.org/10.17632/hcxp9d4zkf.1. Available at https://data.mendeley.com/datasets/hcxp9d4zkf/1). The PacBio-sequenced PittGG genome reference was deposited into GenBank under SRA number SRR10207558. Full calculations and R scripts are available at: https://github.com/mamora/DNA_uptake.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2020.102007.

ACKNOWLEDGMENTS

Financial support for this work was provided by the National Science and Engineering Research Council of Canada (to RJR) and by the National Institutes of Health (to GDE (5R01DC002148-21)). We thank the Drexel Genomic Core Facility for DNA sequencing; Rachel Simister for her assistance with the bioanalyzer fragment analysis; and Matt Pennell, Sally Otto, Stephen Hallam, and David Baltrus for advice on the manuscript. We also thank two anonymous reviewers for their very helpful comments and the Editor for their patience during the revisions.

AUTHOR CONTRIBUTIONS

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

DECLARATIONS OF INTERESTS

The authors declare no competing interests.

Received: January 15, 2020
Revised: November 30, 2020
Accepted: December 23, 2020
Published: January 22, 2021

REFERENCES

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

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

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

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

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

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

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

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

Cehovin, A., Simpson, P.J., McDowell, M.A., Brown, D.R., Noschese, R., Pallett, M., Brady, J., Baldwin, G.S., Lea, S.M., Matthews, S.J., et al. (2013). Specific DNA recognition mediated by a type IVpilin. Proc. Natl. Acad. Sci. U S A *110* (8), In this issue, 3065–3070.

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

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

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

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

Deich, R.A., and Smith, H.O. (1980). Mechanism of homospecific DNA uptake in *Haemophilus influenzae* transformation. Molecular and General Genetics *177*, 369–374.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Lorenz, M.G., and Wackernagel, W. (1994). Bacterial gene transfer by natural genetic transformation in the environment. Microbiol. Rev. *58*, 563–602.

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

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

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

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

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

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

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

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

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

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

Redfield, R.J., Findlay, W.A., Bossé, J., Kroll, J.S., Cameron, A.D.S., and Nash, J.H.E. (2006). Evolution of competence and DNA uptake specificity in the Pasteurellaceae. BMC Evol. Biol. *6*, 1–15.

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

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

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

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

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

Sinha, S., Mell, J.C., and Redfield, R.J. (2012). Seventeen Sxy-Dependent Cyclic AMP Receptor Protein Site-Regulated Genes Are Needed for Natural Transformation in Haemophilus influenzae. J. Bacteriol. *194*, 5245–5254.

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

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

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

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

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

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.

iScience, Volume *24*

Supplemental Information

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

Marcelo Mora, Joshua Chang Mell, Garth D. Ehrlich, Rachel L. Ehrlich, and Rosemary J. Redfield

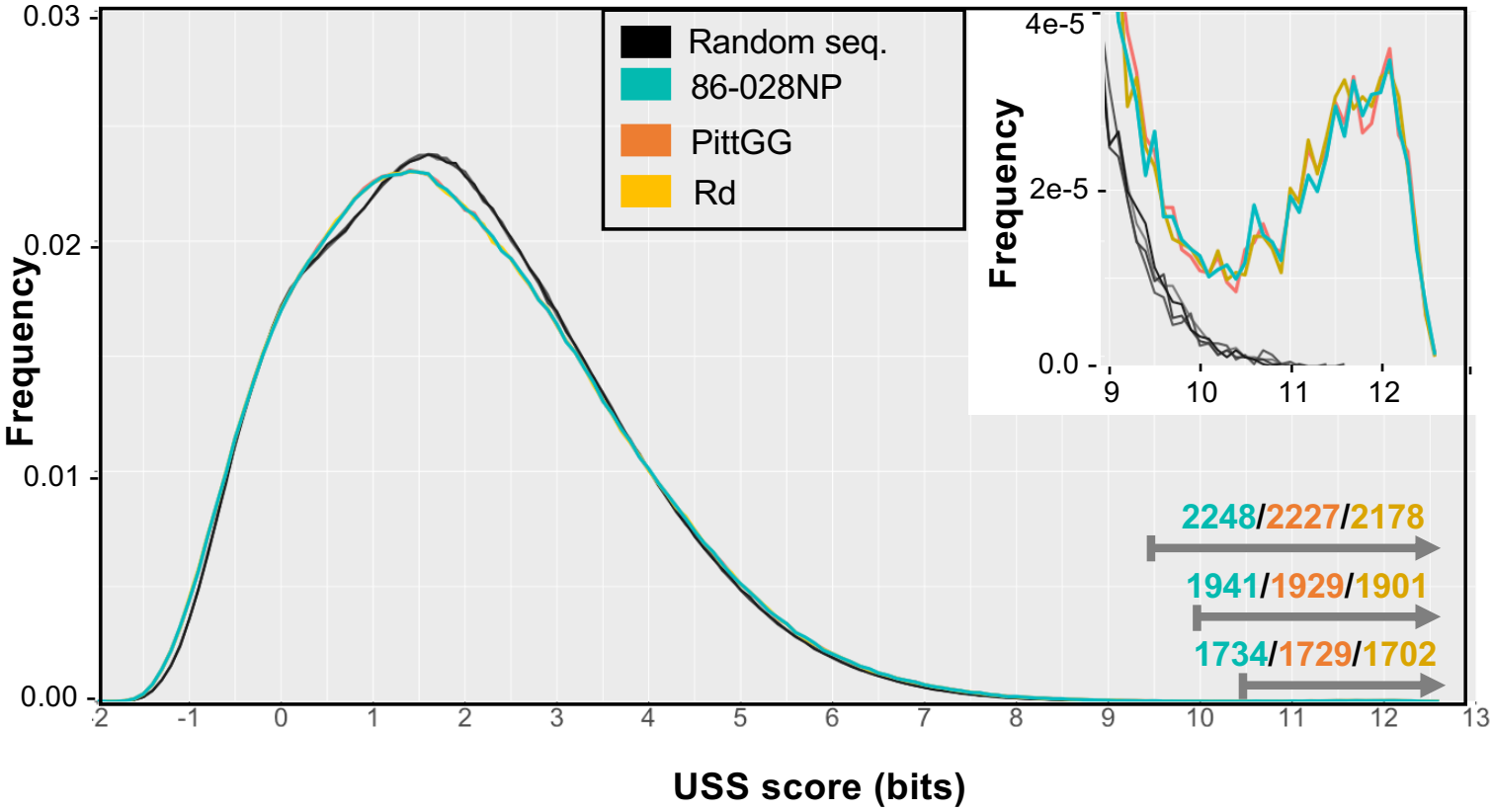


Figure S1. Frequency distribution of USS scores for all positions in *H. influenzae* and random-sequence genomes. Related to Figure 1.

Legend: 86-028 NP (blue), PittGG (orange), Rd (gold), and four random-sequence 1.9 Mb genomes with the same base composition (38%G+C, black and grey). Scores were calculated with the uptake scoring matrix in Table S1. The numbers in the lower right are the numbers of positions meeting cutoff scores of 9.5, 10.0 and 10.5 bits. **Inset:** Expanded view for positions with scores higher than 9 bits.

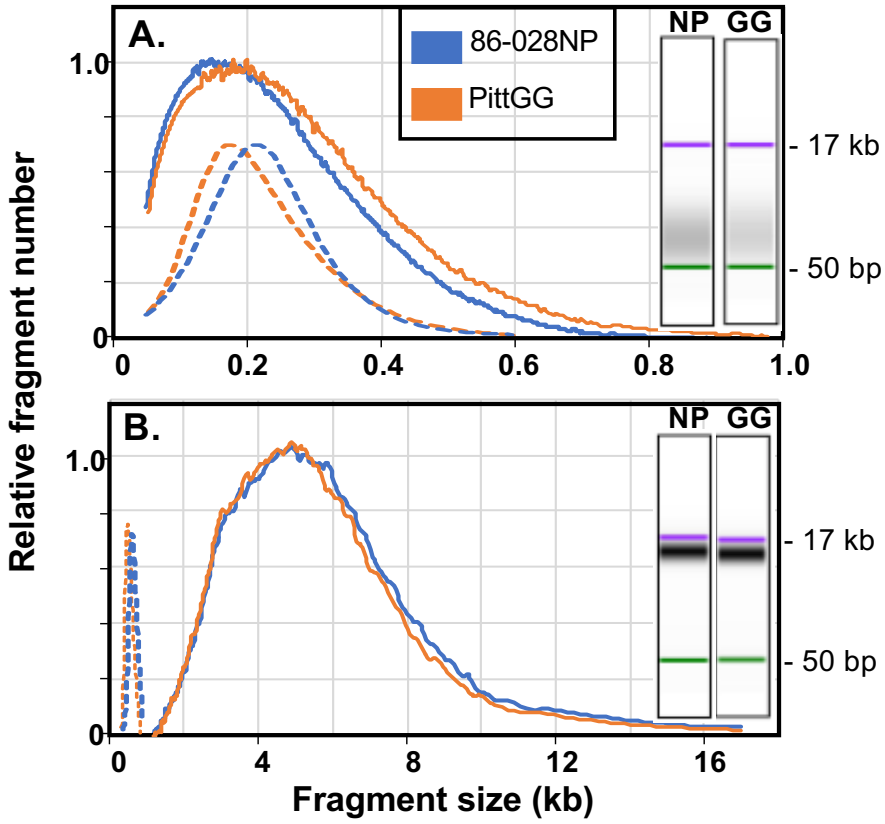


Figure S2: Distributions of fragment lengths in input DNA preparations. Related to Figures 3, 5 and 9.

Legend: Relative abundances of DNA fragment lengths were estimated from Bioanalyzer data for input DNA samples from strains 86-028NP (blue) and PittGG (orange). **A**. Short-fragment preparations. **B**. Long-fragment preparations. Solid lines: length distributions of fragments in input DNA preparations, normalized to most frequent length. Dashed lines: Length distributions of sequenced fragments, with arbitrary scaling. Insets: Bioanalyzer pseudo-gel images of sheared DNAs (NP: 86-028NP, GG: PittGG). Bioanalyzer molecular weight markers are shown in purple and green.

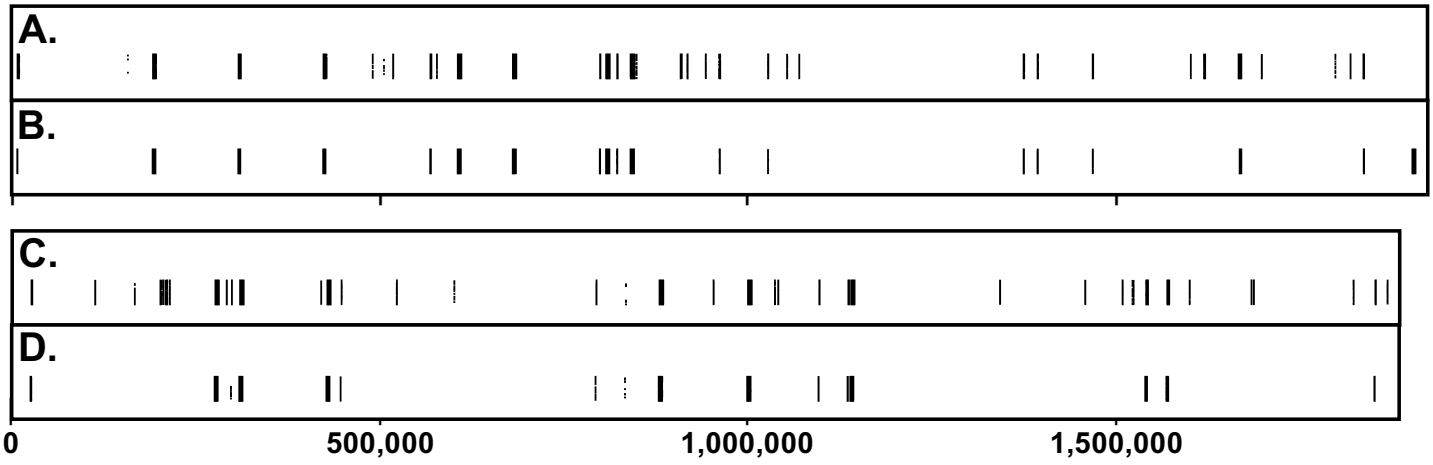
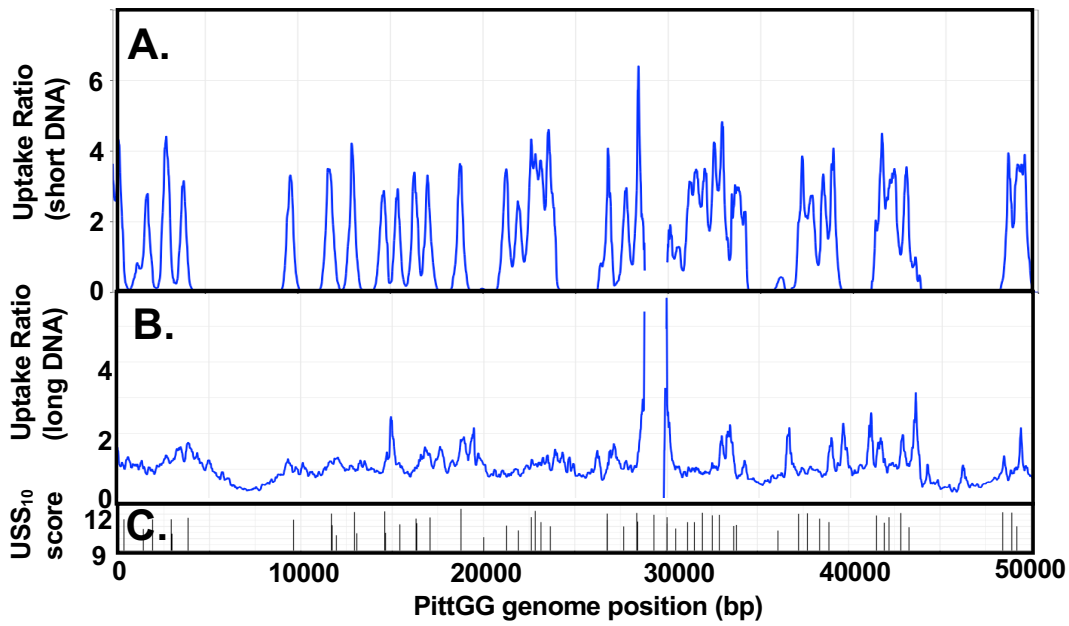
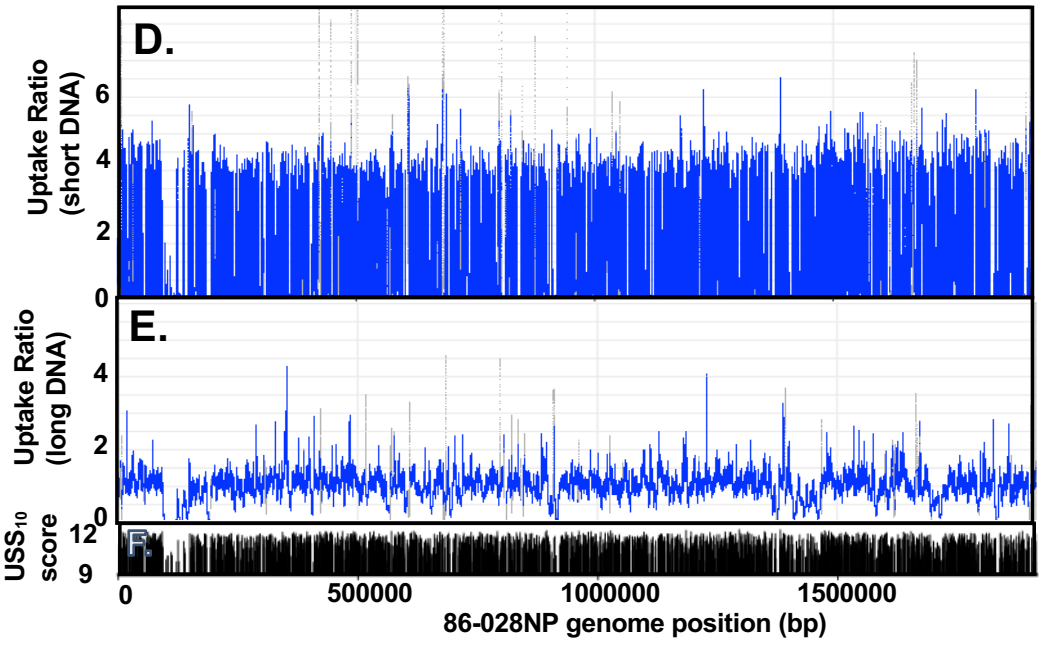


Figure S3 Locations of positions with missing uptake ratio data. Related to Figures 3, 5 and 9.

Legend: Each point represents a genome position for which an uptake ratio could not be calculated. The points are vertically jittered, so segments with no coverage appear as black rectangles. **A.** 86-028NP, short-fragment data. **B.** 86-028NP, long-fragment data. **C.** PittGG, short-fragment data. **D.** PittGG, long-fragment data.


A.
Uptake Ratio (short DNA)
6
4
2
0
B.
Uptake Ratio (long DNA)
4
2
0
C.
USS10 score
12
9
0
10000
20000
30000
40000
50000
PittGG genome position (bp)


D.
Uptake Ratio (short DNA)
6
4
2
0
E.
Uptake Ratio (long DNA)
4
2
0
F.
USS10 score
12
9
0
500000
1000000
1500000
86-028NP genome position (bp)

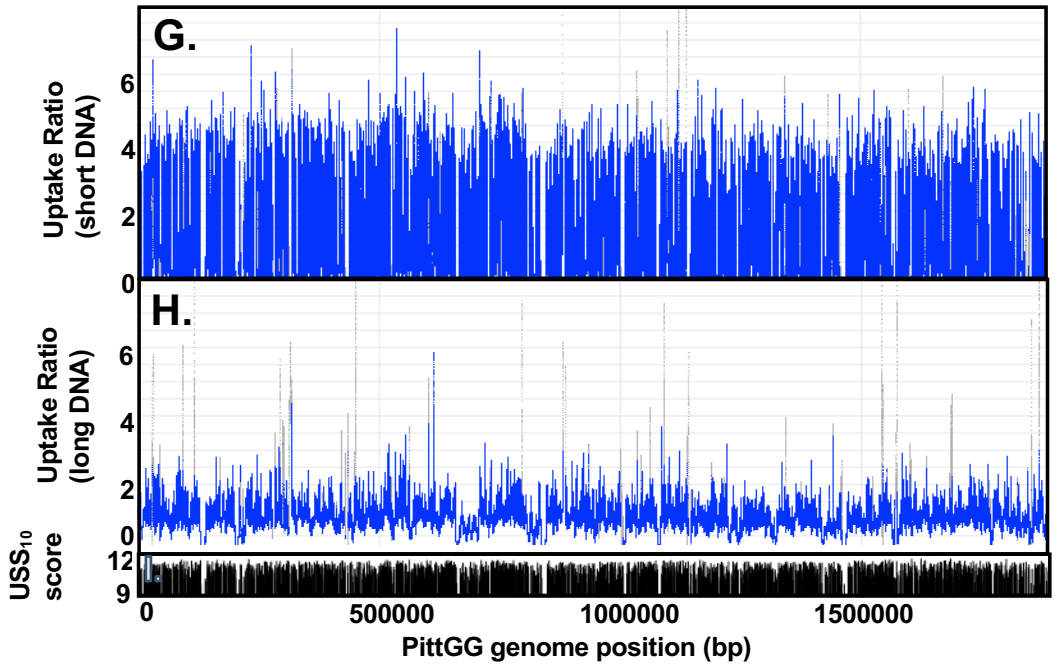

G.
Uptake Ratio (short DNA)
6
4
2
0
H.
Uptake Ratio (long DNA)
6
4
2
0
I.
USS10 score
12
9
0
500000
1000000
1500000
PittGG genome position (bp)

Figure S4. Experimentally determined uptake ratio maps and USS_{10} maps. Related to Figure 3.

Legend: Grey points indicate positions with input coverage lower than 20 reads. Gaps indicate unmappable positions. **A-C:** Maps of a 50 kb segment of the PittGG genome. **A.** Uptake ratios of short-fragment PittGG DNA. **B.** Uptake ratios of long-fragment PittGG DNA. **C.** PittGG USS_{10} positions and scores. **D-I**: Whole-genome maps. **D.** Uptake ratios of short-fragment 86-028NP DNA. **E.** Uptake ratios of long-fragment 86-028NP DNA. **F.** 86-028NP USS_{10} positions and scores. **G.** Uptake ratios of short-fragment PittGG DNA. **H.** Uptake ratios of long-fragment PittGG DNA. **I.** PittGG USS_{10} positions and scores.

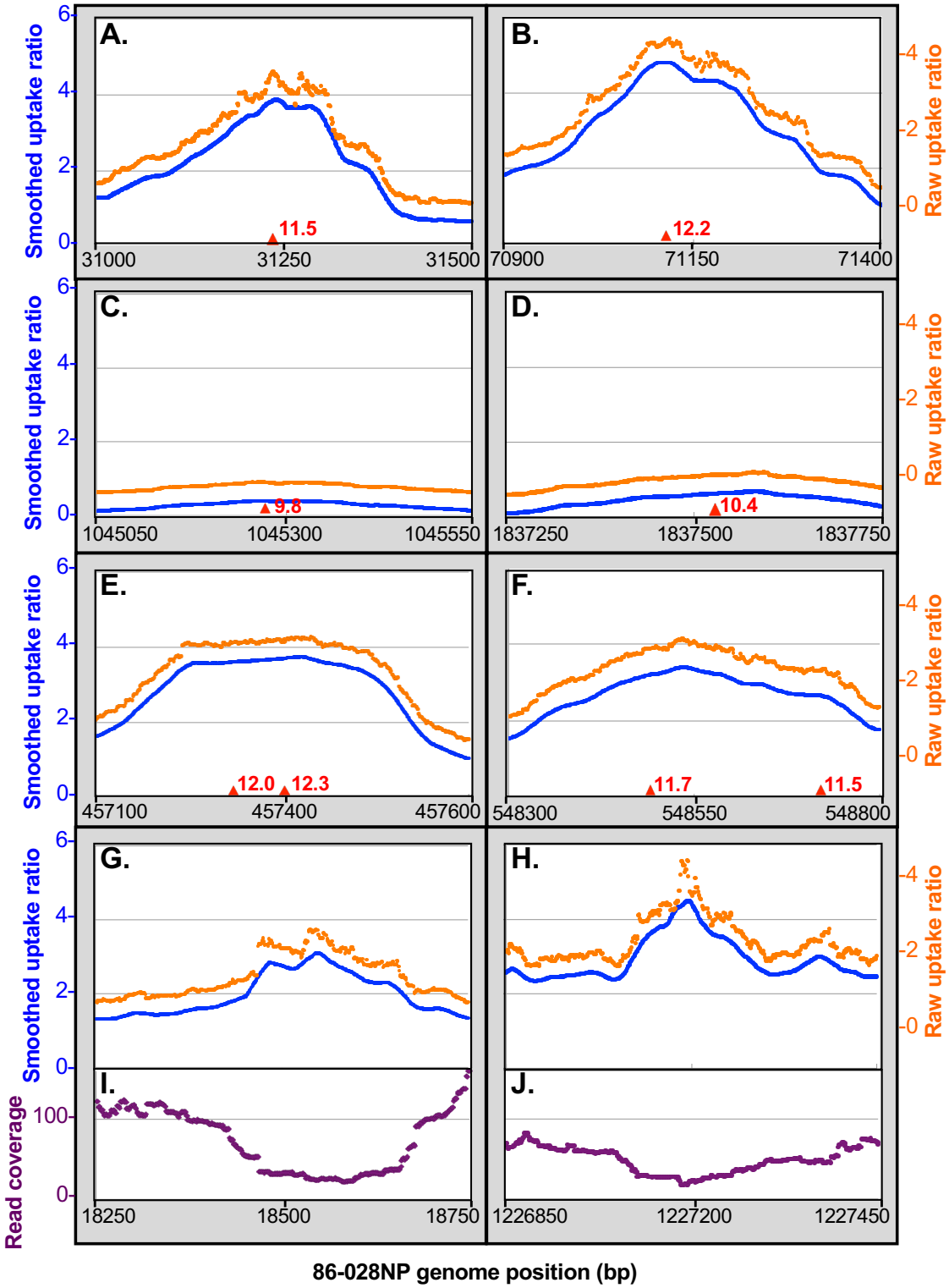

A.
B.
C.
D.
E.
F.
G.
H.
I.
J.
Smoothed uptake ratio
Raw uptake ratio
Read coverage
11.5
12.2
9.8
10.4
12.0
12.3
11.7
11.5
31000
31250
31500
70900
71150
71400
1045050
1045300
1045550
1837250
1837500
1837750
457100
457400
457600
548300
548550
548800
18250
18500
18750
1226850
1227200
1227450
86-028NP genome position (bp)

Figure S5. Shapes of typical 86-028NP uptake peaks. Related to Figures 3 and 9.

Legend. Blue dots: uptake ratios after smoothing with a 31 bp window. **A.-H.**: Orange dots: uptake ratios without smoothing (note that Y axis is offset by 0.5 units). Red triangles and numbers: locations and scores of USS. **A.-F**. 86-028NP short-fragment DNA: **A.** and **B.**: peaks at strong USS. **C.** and **D.**: peaks at weak USS. **E.** and **F.**: peaks at pairs of USS separated by **A.** 69 bp and **B.** 230?? bp. **G.** and **H.** Uptake ratio spikes not at USS in 86-028NP long-fragment DNA. **I.** and **J.** Purple dots: sequencing coverage of input 86-028NP long-fragment DNA.

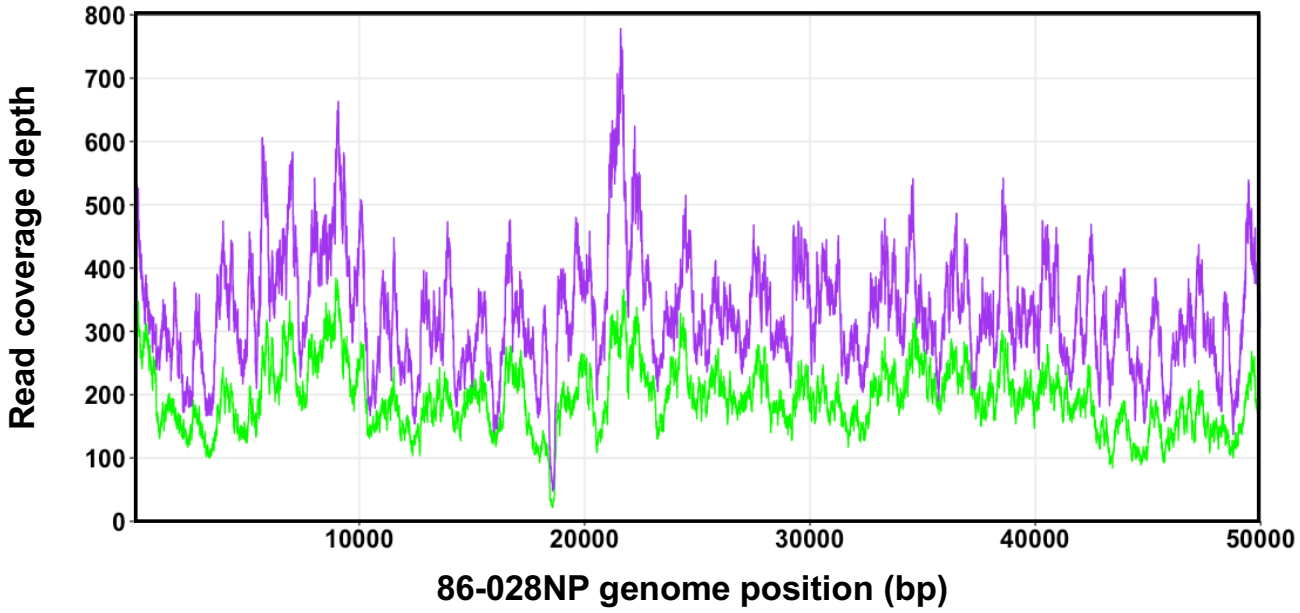


Figure S6. Variation in read coverage. Related to Figures 3 and 9.

Legend: Read coverage of the 86-028NP long-fragment (green) and short-fragment (purple) input samples over a 50 kb genome segment.

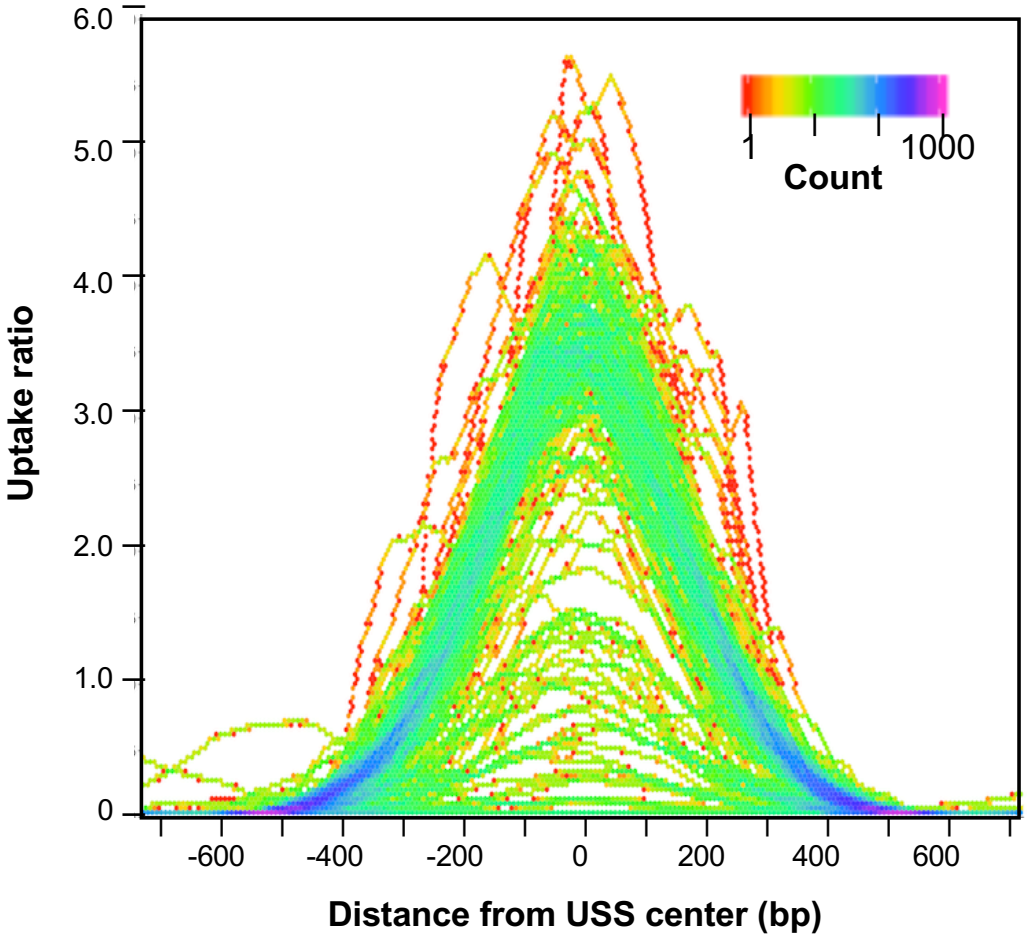


Figure S7. Shape analysis of isolated USS_{10} peaks. Related to Figures 4 and 6.

Legend: Short-fragment uptake ratio data for positions around 158 86-028NP USS_{10}s that were separated by at least 1200 bp from other USS_{10}s and had uptake ratios of at least 3.0.

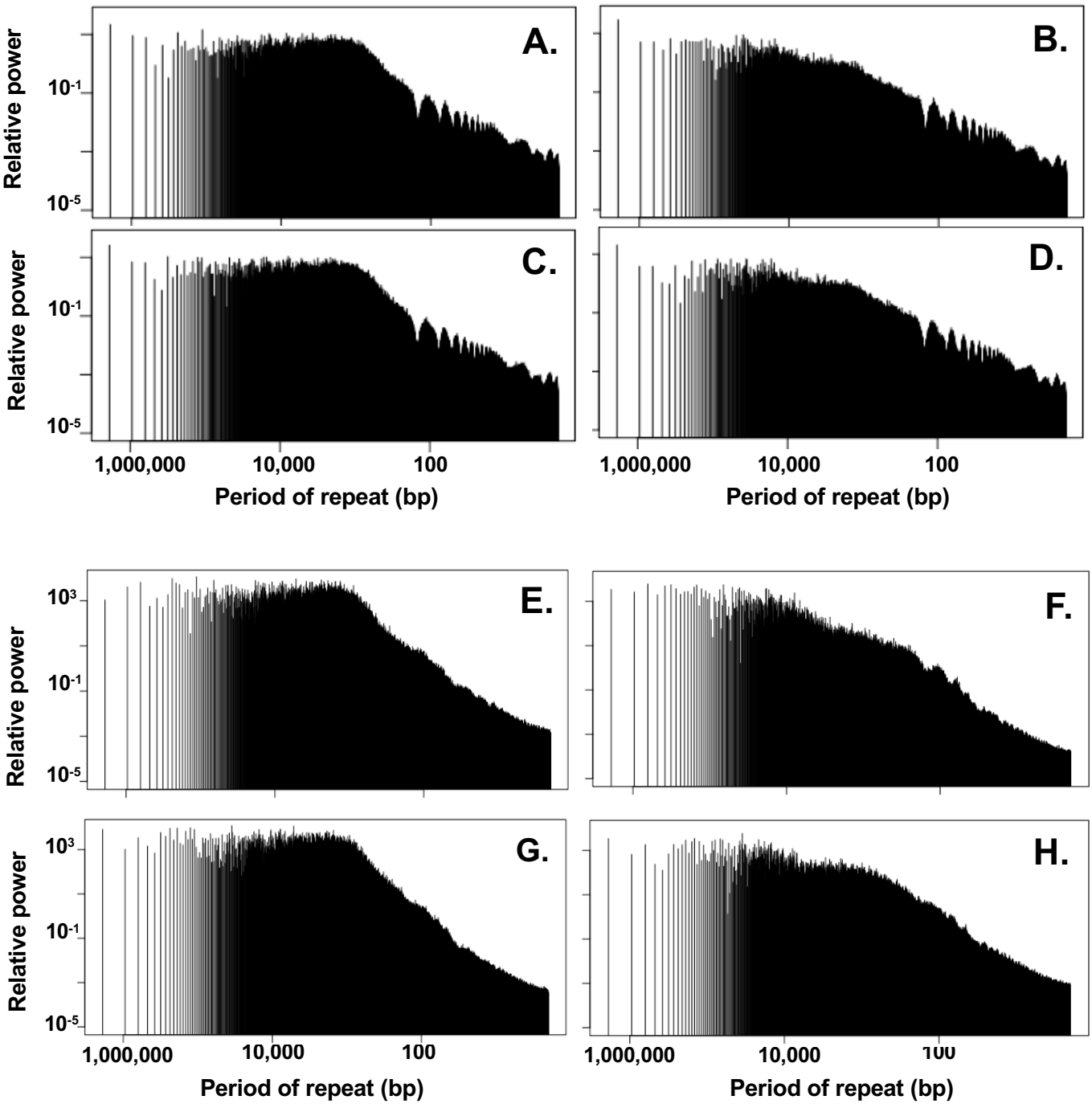


Figure S8. Tests of periodicity. Related to Figures 3 and 9.

Legend: Fourier-transform analyses were performed using R-package RCA. The X-axes are $\log_{10}$ of repeat period in bp; Y-axes are $\log_{10}$ of the relative periodicity at each repeat period. **A-D.** Tests using coverage in input samples. **E-H.** Tests using uptake ratios. Samples: **A** & **E**: 86-028NP short fragments; **B** & **F**: 86-028NP long-fragments; **C** & **G**:Pitt GG short fragments; **D** & **H**: PittGG long fragments.

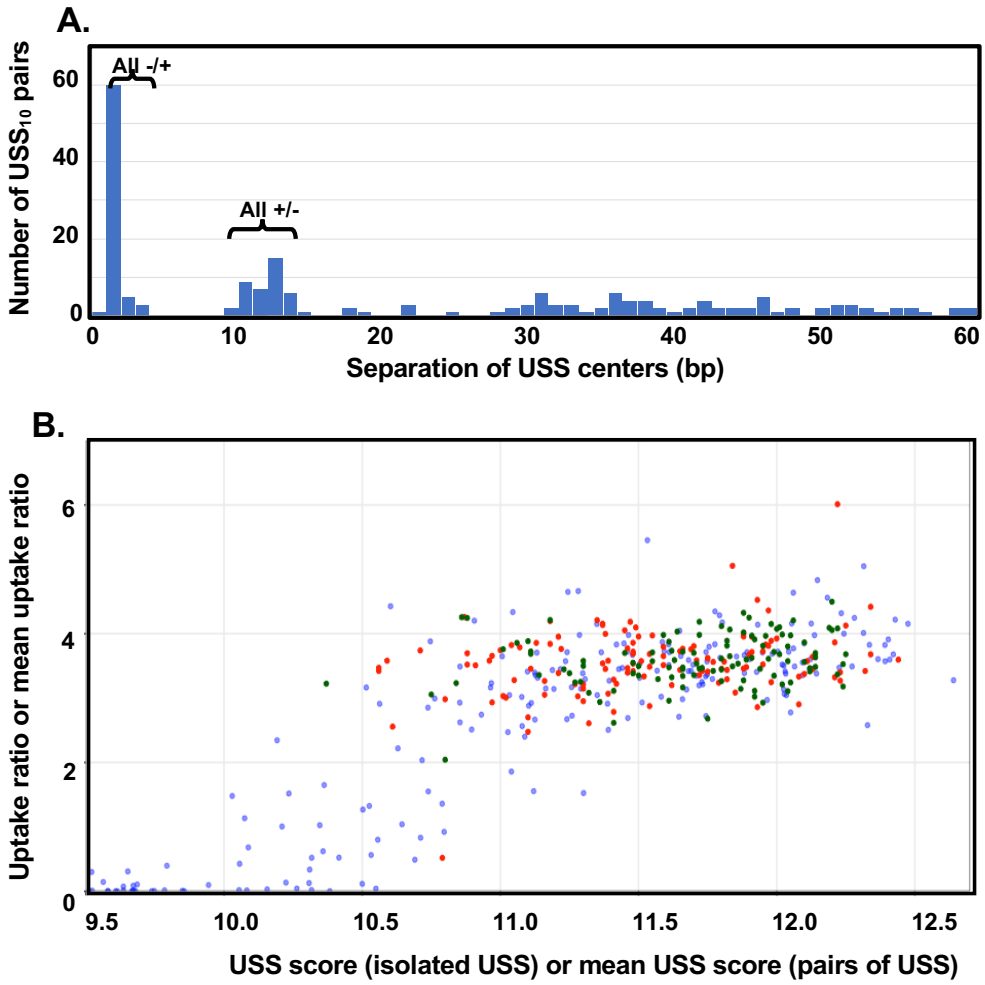


Figure S9. Analysis of DNA uptake effects of USS_{10} pairs in the 86-028NP genome. Related to Figure 3 and 6.

Legend: A. Frequencies of spacings between close USS_{10} pairs. **B.** Uptake ratios for isolated USS_{10} (blue points, data from Figure 6) and for centers of pairs of USS_{10} whose centers are 14-100 bp apart (red points) or 0-13 bp apart (dark green points).

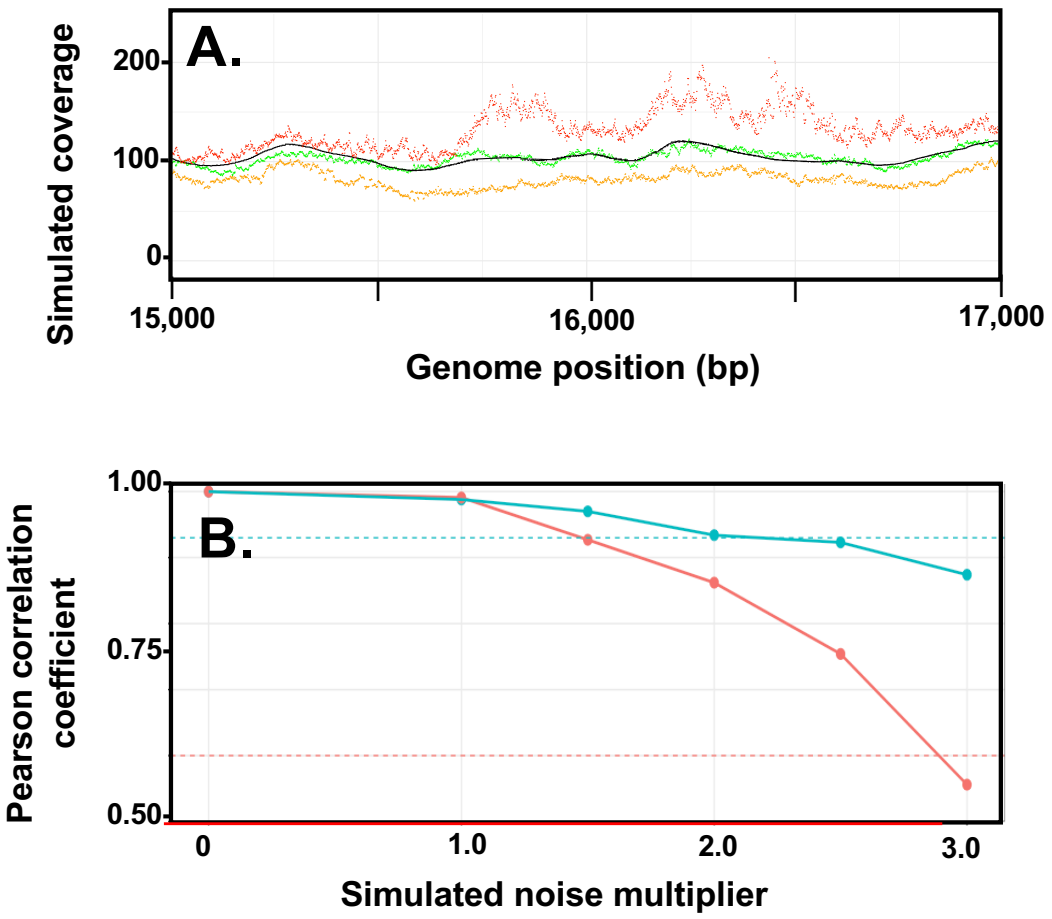


Figure S10. Simulated noise analysis. Related to Figures 5 and 9.

Legend: A. Effects of added red noise on simulated noise-free coverage. Black points: no added noise; Green, yellow and red points: simulated noise added with multipliers of 1.0, 2.0 and 3.0 respectively. **B**. Correlation coefficient from simulated uptake ratios with and without different levels of noise. X-axis represents the multiplicative factor applied to the coverage-dependent amount of noise added. Simulated results for 86-028NP-short and 86-028NP-long DNA fragments are shown in blue and red, respectively. Dashed lines at 0.93 (blue) and 0.60 (red) indicate the real-data correlation of predicted uptake with observed uptake ratios for 86-028NP-short (top) and long (bottom) fragment size distributions, respectively.

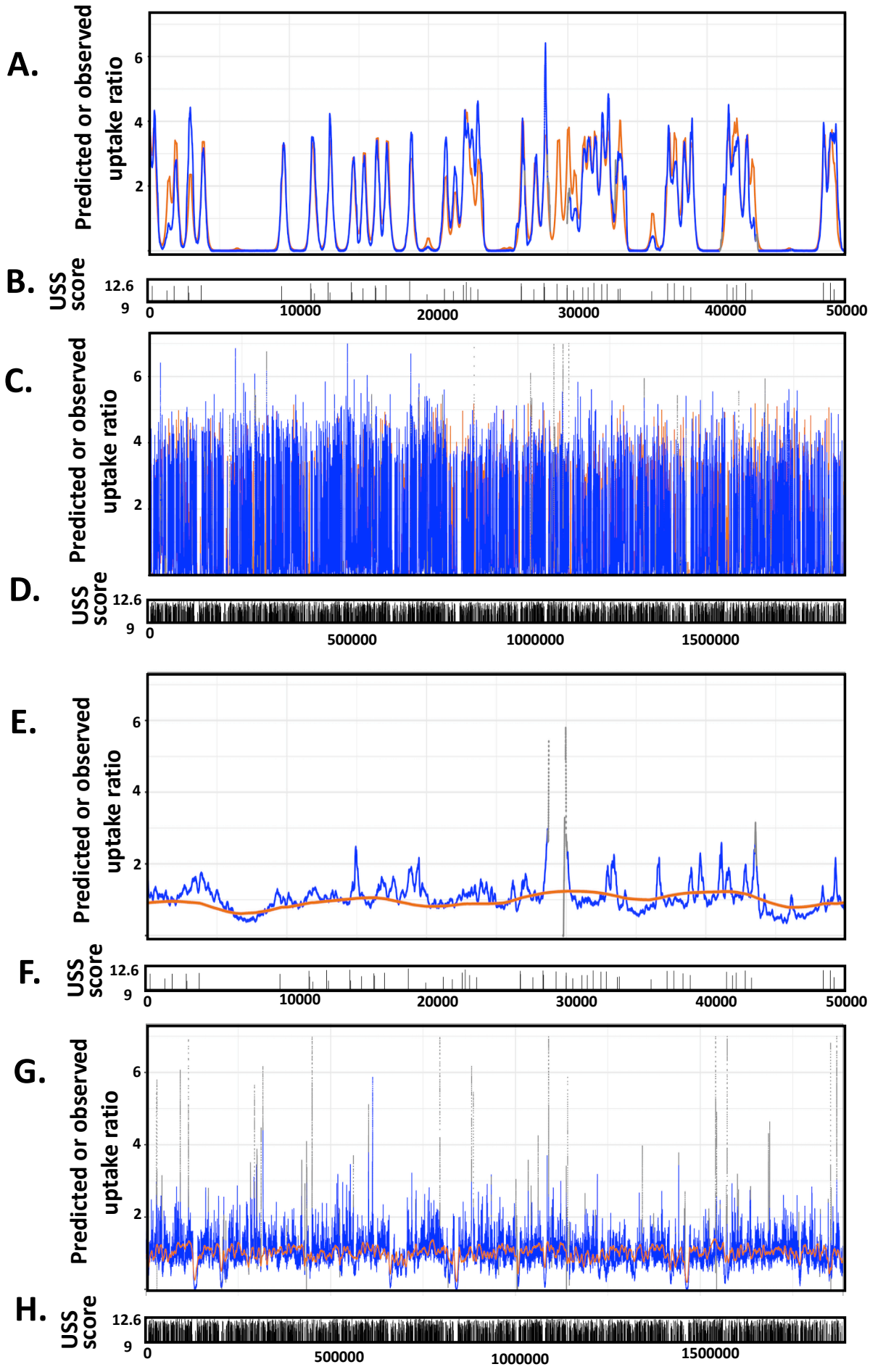

A.
Predicted or observed uptake ratio
6
4
2
B.
USS score
12.6
9
0
10000
20000
30000
40000
50000
C.
Predicted or observed uptake ratio
6
4
2
D.
USS score
12.6
9
0
500000
1000000
1500000
E.
Predicted or observed uptake ratio
6
4
2
F.
USS score
12.6
9
0
10000
20000
30000
40000
50000
G.
Predicted or observed uptake ratio
6
4
2
H.
USS score
12.6
9
0
500000
1000000
1500000

Figure S11. Predicted and observed uptake ratios for PittGG short and long DNA fragments. Related to Figure 3.

Legend: A and **E**: Uptake maps for the first 50kb of the PittGG genome; **C.** and **G.**: Uptake maps for the full PittGG genome. **B.**, **D.**, **F.** and **I.**: Vertical tick marks indicate locations and scores of USS_{10}s. Orange lines: USS-dependent uptake predicted by the revised model. Blue lines: Mean uptake ratios from 3 replicate experiments (grey points indicate positions with <20 reads input coverage, gaps indicate unmapable possitions). Note that some grey dots are beyond the tops of some panels.

Table S1: USS uptake scoring matrix: Related to Figure 1.

	1	2	3	4	5	6	7	8	9	10
A	0.07	0.09	0.42	0.63	-0.13	-0.13	-0.07	-0.06	-0.07	-0.10
C	-0.09	-0.04	-0.13	-0.12	-0.11	-0.12	-0.10	1.86	-0.09	-0.12
G	0.06	-0.02	-0.03	-0.07	1.01	-0.13	1.61	-0.05	1.76	1.57
T	-0.01	-0.03	-0.08	-0.06	-0.12	0.83	-0.12	-0.05	-0.08	-0.11

	11	12	13	14	15	16	17	18	19	20
A	-0.11	-0.01	0.15	0.07	0.04	0.04	-0.02	-0.05	-0.02	0.00
C	-0.03	0.05	-0.07	-0.05	-0.08	-0.08	-0.04	0.01	-0.02	-0.01
G	-0.10	-0.06	0.03	0.00	-0.06	-0.11	-0.11	-0.08	-0.02	0.00
T	0.40	0.03	-0.07	-0.02	0.13	0.27	0.26	0.16	0.07	0.00

	21	22	23	24	25	26	27	28	29	30	31
A	0.00	0.00	-0.01	0.00	0.05	-0.01	0.04	-0.06	-0.06	-0.03	-0.03
C	0.01	0.00	0.00	0.01	-0.06	-0.04	-0.10	-0.01	-0.03	-0.01	-0.01
G	0.01	0.01	0.01	0.00	0.06	-0.06	-0.13	-0.12	-0.08	-0.03	0.00
T	-0.02	-0.01	0.00	-0.01	-0.04	0.13	0.35	0.32	0.23	0.08	0.04

Table S2. Sample metadata. Related to Table 1 and Figures 3 and 9.

Sample name[1]	Sample type	Biosample ID	Bioproject accession	Donor strain	Genome size	Fragment size range (bp)	Recipient strain	% of DNA recovered
UP1 (NP lg)	Taken-up DNA	SAMN07187224	PRJNA387591	86-028NP NalR	1,914,387	1500-17000	RR3117 (*rec2*::*spec*)	2.06
UP2 (NP lg)	Taken-up DNA	SAMN07187225	PRJNA387591	86-028NP NalR	1,914,387	1500-17000	RR3125 (*rec2*-)	1.38
UP3 (NP lg)	Taken-up DNA	SAMN07187226	PRJNA387591	86-028NP NalR	1,914,387	1500-17000	RR3125 (*rec2*-)	1.38
UP4 (GG lg)	Taken-up DNA	SAMN07187227	PRJNA387591	PittGG)	1,887,046	1500-17000	RR3117 (*rec2*::*spec*)	1.60
UP5 (GG lg)	Taken-up DNA	SAMN07187228	PRJNA387591	PittGG	1,887,046	1500-17000	RR3125 (*rec2*-)	1.24
UP6 (GG lg)	Taken-up DNA	SAMN07187229	PRJNA387591	PittGG (RR1361)	1,887,046	1500-17000	RR3125 (*rec2*-)	2.48
UP7 (NP sh)	Taken-up DNA	SAMN07187230	PRJNA387591	86-028NP NalR	1,914,387	50-800	RR3117 (*rec2*::*spec*)	0.64
UP8 (NP sh)	Taken-up DNA	SAMN07187231	PRJNA387591	86-028NP NalR	1,914,387	50-800	RR3125 (*rec2*-)	0.32
UP9 (NP sh)	Taken-up DNA	SAMN07187232	PRJNA387591	86-028NP NalR	1,914,387	50-800	RR3125 (*rec2*-)	0.41
UP10 (GG sh)	Taken-up DNA	SAMN07187233	PRJNA387591	PittGG	1,887,046	50-800	RR3117 (*rec2*::*spec*)	0.77
UP11 (GG sh)	Taken-up DNA	SAMN07187234	PRJNA387591	PittGG	1,887,046	50-800	RR3125 (*rec2*-)	0.91
UP12 (GG sh)	Taken-up DNA	SAMN07187235	PRJNA387591	PittGG	1,887,046	50-800	RR3125 (*rec2*-)	1.03
UP13 (NP lg)	Input DNA	SAMN07187236	PRJNA387591	86-028NP NalR	1,914,387	1500-17000	N/A	N/A
UP14 (GG lg)	Input DNA	SAMN07187237	PRJNA387591	PittGG	1,887,046	1500-17000	N/A	N/A
UP15 (NP sh)	Input DNA	SAMN07187238	PRJNA387591	86-028NP NalR	1,914,387	50-800	N/A	N/A
UP16 (GG sh)	Input DNA	SAMN07187239	PRJNA387591	PittGG	1,887,046	50-800	N/A	N/A
Rd	recipient DNA	SAMN12049038	PRJNA387591	Rd (recipient)	1,831,585	Not sheared	Rd KW20	N/A

Sample name[1]	Mean size of library fragments	Un-mapped reads	Mapped reads	Mean MAPQ score	Reads mapping only to recipient (MAPQ>0)	Reads mapping only to donor (MAPQ>0)	% contam-ination with Rd DNA	% of reads removed	Mean read coverage
UP1 (NP lg)	342.4	22,982	2,6468,71	47.2	122,712	2,253,377	5.16	14.9	174
UP2 (NP lg)	334.0	29,276	2,386,785	47.1	151,131	1,987,921	7.07	16.7	154
UP3 (NP lg)	341.3	29,111	2,962,387	47.1	226,132	2,428,674	8.52	18.0	188
UP4 (GG lg)	348.8	11,134	2,183,783	46.3	145,426	1,792,255	7.51	17.9	140
UP5 (GG lg)	348.2	10,423	2,149,185	46.4	366,986	1,539,930	19.25	28.3	121
UP6 (GG lg)	309.5	7,732	1,227,888	46.0	97,875	978,918	9.09	20.3	77
UP7 (NP sh)	227.8	203,460	5,011,237	45.8	363,683	4,006,647	8.32	20.0	303
UP8 (NP sh)	219.5	211,676	3,592,281	44.7	522,918	2,591,844	16.79	27.8	197
UP9 (NP sh)	247.2	201,601	7,0983,09	46.5	801,198	5,452,848	12.81	23.2	417
UP10 (GG sh)	240.9	171,799	4,746,454	45.4	323,716	3,807,524	7.84	19.8	293
UP11 (GG sh)	241.0	124,946	5,563,501	46.1	157,046	4,687,429	3.24	15.7	361
UP12 (GG sh)	233.8	234,499	5,298,060	44.9	205,997	4,389,741	4.48	17.1	337
UP13 (NP lg)	344.5	11,787	2,715,317	46.8	1,333	24,04,540	0.06	11.4	185
UP14 (GG lg)	388.4	7,774	5,836,702	46.4	1,345	5,314,003	0.03	9.0	192
UP15 (NP sh)	222.4	13,750	4740911	45.3	1154	3,961,509	0.03	16.4	296
UP16 (GG sh)	234.0	9531	6627713	45.3	1668	4,938,510	0.03	25.5	395
Rd	241.0	4564	4083313	N/A	N/A	N/A	N/A	N/A	298

1. **NP lg:** Long-fragment 86-028NP DNA, **GG lg:** Long-fragment PittGG DNA, **NP-sh:** Short-fragment 86-028NP DNA, **GG sh:** Short-fragment PittGG DNA.

Table S3. Analysis of low-coverage positions. Related to Figures 3 and 9.

86-028NP-long:							
Input coverage:	**≤10**	**10-20**	**20-50**	**50-100**	**>100**	**sum**	**% in the genome**
% of positions with this coverage	2.36	0.45	1.78	6.14	89.27	100	
% of positions with uptake ratios >2.0	7.53	4.73	20.75	37.81	29.18	100	0.82%
% of positions with uptake ratios >3.0	30.22	12.07	38.38	19.33	0	100	0.06%
% of positions with uptake ratios <0.5	1.78	0.85	2.7	6.6	88.07	100	8.90%
% of positions with uptake ratios <0.25	2.19	0.69	2.08	4.82	90.22	100	4.00%
86-028NP-short							
Input coverage	**≤10**	**10-20**	**20-50**	**50-100**	**>100**	**sum**	
% of positions with this coverage	3.37	0.63	1.68	2.88	91.44	100	
% of positions with uptake ratios >4.0	9.13	3.93	7.11	9.2	70.63	100	1.11%
% of positions with uptake ratios >5.0	45.46	16.48	7.87	9.11	21.08	100	0.13%
% of positions with uptake ratios >0.1	1.19	0.75	1.91	3.23	92.92	100	44.40%
% of positions with uptake ratios >0.01	1.45	0.76	1.93	3.22	92.63	100	28.15%
PittGG-long							
Input coverage	**≤10**	**10-20**	**20-50**	**50-100**	**>100**	**sum**	
% of positions with this coverage	2.33	0.43	1.36	2.85	93.03	100	
% of positions with uptake ratios >2.0	16.85	10.76	25.65	27.8	18.94	100	1.57%
% of positions with uptake ratios >3.0	61.87	17.07	15.62	5.44	0	100	0.26%
% of positions with uptake ratios <0.5	0.58	0.32	0.97	2.87	95.26	100	6.90%
% of positions with uptake ratios <0.25	0.86	0.46	1.65	4.41	92.62	100	2.57%
PttGG-short							
Input coverage	**≤10**	**10-20**	**20-50**	**50-100**	**>100**	**sum**	
% of positions with this coverage	3.11	0.73	1.98	2.9	91.28	100	
% of positions with uptake ratios >4.0	1.73	1.62	6.77	11.33	78.55	100	2.50%
% of positions with uptake ratios >5.0	8.8	7.41	20.2	28.53	35.06	100	0.24%
% of positions with uptake ratios >0.1	1.25	0.77	2.25	3.02	92.72	100	39.38%
% of positions with uptake ratios >0.01	1.54	0.76	2.06	3.02	92.62	100	26.50%

Transparent Methods

Identifying USSs in the genomes. Genomic USSs were identified by scoring each genome position with the position-specific scoring matrix (PSSM) of Mell et al. (2012); this is based on uptake of synthetic fragments containing degenerate USS sequences. Positions scoring ≥ 10.0 or ≥ 9.5 (maximum score is 12.6) were included in the standard (USS_{10} and $USS_{9.5}$) lists of USS locations. Since USS are asymmetric, USS positions in both orientations were specified by the location of their central base 16. Sequence logos of USSs were generated using R package seqLogo v. 3.8.

Predicting DNA uptake from DNA sequence. The predictive model was written in R v.3.5.1. Given a list of USS positions and scores in a DNA genome of specified length, it used a specified distribution of DNA fragment lengths or length bins (e.g. 1-100 bp, 101-200 bp, etc.) to calculate the relative uptake of every position in a circular genome. At each DNA position in turn, for each fragment length or bin, the model summed the predicted uptake contributions for every fragment of that length that overlapped the position. For efficiency, the full calculation was only done for the first position. At each subsequent position, the model calculated the new sum from the previous position's sum by subtracting the contribution of the formerly leftmost fragment and adding the contribution of the new rightmost fragment (Figure 2.3A).

Each fragment's predicted contribution to uptake depended on the number of USS it contained, and on the scores and relative locations of these USS. Fragments with no or incomplete USSs were assigned baseline values for the probabilities of being bound (*p_bind*) and taken up

(*p_uptake*); initial values for both were arbitrarily set to 0.1. For fragments with one or more complete USS_{10}, *p_bind* was calculated as *1 – mean_gap/20000,* where *mean_gap* was the mean length of USS-free segments in the fragment and *20000* the maximum fragment length in bp. The uptake function *p_uptake* was initially specified as *p_uptake = 0.1 + (1 - 0.1)/(1 + exp(-5 * (score – 11)))*.

Once the model had calculated the contributions of a specific fragment length or length bin to uptake of every genome position, it moved on to the next length or bin. Once the contributions of every length or bin had been calculated, the model combined all the contributions for each position, taking into account the frequency of each length or bin in the input DNA. These position-specific uptake predictions were then normalized to a mean genome-wide uptake value of 1.0.

In response to ongoing analysis of the 86-028NP DNA uptake data, the initial model underwent modifications to improve its predictions for a 'far from USS_{10}' subset of positions that were at least 0.5kb from a $USS_{9.5.}$ (n = 361965 positions), combined with 209 $USS_{9.5}$ peak positions separated from the nearest USS_{10}s by at least 1000 bp. This reduced the baseline *p_uptake* of USS-free fragments from 0.1 to 0.005 and the USS cutoff score from 10 to 9.5, excluded from consideration $USS_{9.5}$ that were within 50 bp of fragment ends, and identified better slope and inflection point values for the sigmoidal uptake function using the R function "nls" from the stats-package. These changes replaced the previous uptake function with *p_uptake = 0.005 + (1 - 0.005)/(1 + exp(-3.8 * (score – 10.6)))*.

A subsequent change adjusted uptake predictions according to the GC content around each position. First the observed effect of GC content on uptake was approximated by a linear

function describing how the 86-028NP long fragment uptake ratios depended on their local GC contents calculated with a 2 kb window (see inset in Fig. GC). Each genomic position was assigned the GC content of a 1001 bp window centered on it. The predicted uptake at each position was then modified using the local GC content and the function.

Bacterial strains, culturing, and competent cell preparations: The KW20 recipient strains were *rec2* derivatives of the standard *H. influenzae* lab strain Rd KW20, with (RR3117) and without (RR3125) a spectinomycin resistance allele (Mell et al., 2012; Sinha et al., 2012). The 86-028NP donor strain (RR3133) was a derivative with a nalidixic acid resistance allele (Mell *et al.* 2011); the PittGG isolate was unmodified. Standard growth and culturing methods were used (Poje and Redfield, 2003); liquid cultures were grown with shaking at 37 °C in brain-heart infusion broth supplemented with NAD (2μg/ml) and hemin (10 μg/ml) (sBHI), with 1.2% agar added for plate cultures. To prepare naturally competent cells, cultures were first maintained in exponential growth at OD_{600} below 0.2 for at least 2 hr, and at OD_{600} = 0.2 cells were collected by filtration from 10 ml of culture, transferred into 10 ml of starvation medium M-IV, and incubated at 37 °C for 100 minutes before DNA uptake experiments (Poje and Redfield, 2003).

Input DNA preparations. High molecular weight donor DNA was purified using standard phenol:chloroform extractions (Sambrook, 2001) from 10 ml overnight cultures of the 86-028NP derivative, and PittGG carrying selectable markers (Table S2). This DNA was then sheared into separate 'long fragment' (1.5-9 kb) and 'short fragment' (50-500 bp) preparations using Covaris G-tubes and sonication respectively. The fragment length distributions were measured using a Bioanalyzer with a DNA 12000 kit (Agilent), dividing the relative fluorescence at each time point by its fragment length estimated from the size standards.

DNA uptake and recovery. 10 ml of competent *rec-2* mutant Rd cells in MIV were incubated with 10 µg of sheared donor DNA for 20 min at 37 °C. To degrade remaining free DNA, the culture was incubated with 1 ug/ml of DNase I for 5 minutes. Cells were washed twice by pelleting and resuspension in cold MIV, and the final pellet was rinsed twice with cold MIV before resuspension in 0.5 ml of extraction buffer (Tris-HCL 10 mM pH 7.5, EDTA 10 mM, CsCl 1.0 M). Periplasmic DNA was extracted using the organic phenol:acetone extraction method as described by Mell *et al.* 2012 (Barouki and Smith, 1985; Kahn et al., 1983; Mell et al., 2012) followed by an ethanol precipitation. DNA was resuspended in 20 µl of $T_{10}E_{10}$ buffer (Tris-HCl 10 mM pH 7.5, EDTA 10 mM). The DNA was then incubated at 37 °C with 400 ng of RNase A for 1 hour, followed by 30 min incubation with 30 ng of proteinase K to remove RNase A. Recovered DNA was then separated from longer fragments of contaminating genomic DNA by electrophoresis in a 0.8% agarose gel and recovered from the gel slice with a Zymo gel DNA recovery kit. Recovered periplasmic DNA was quantified using a Qubit dsDNA HS Assay Kit (absolute DNA concentration).

DNA sequencing and data processing. Sequencing libraries of the input and taken-up DNA samples were prepared using Illumina Nextera XT DNA library prep kits according to manufacturer recommendations. An Illumina NextSeq500 was used to generate 1-10 million paired-end reads of 2x150 nt for each library (giving >100-fold genomic coverage). Summary statistics for each sample are provided in Table S2.

Reference sequences*:* The original PittGG reference (NC_009567.1) generated by pyrosequencing had many indel errors, so a new reference was constructed by Pacific Biosciences RSII of our laboratory version of this strain (RR1361) (assembly by HGAP2 v2.3,

followed by Circlator (Hunt et al., 2015), and then Quiver to polish the circular junction). For our analysis the NCBI sequence references for PittGG, 86-028NP, (NC_007146.2) and Rd KW20 (NC_000907.1) were then further corrected using Pilon v1.22 and the new Illumina reads of input or control samples. This was particularly important for the Rd KW20 recipient reference, since the original sequence dates from 1995 (Fleischmann et al., 1995) and contains several hundred ambiguous bases and errors. This correction step also accommodated the presence of the nalidixic acid resistance marker in 86-028NP.

Competition essays simulations of *H. influenzae* with human DNA, used 3 random segments of the same size as *H. influenzae* 86-028NP genome (1914386 bp) the Chromosome 1 (CM000663.2, positions 33610150 – 35524535), Chromosome 3 (CM000665.2, positions 1348752 – 3263137), and Chromosome 12 (CM000674.2, positions 18170588 – 20084973). Respiratory bacterial genomes used to score USS_{10} and USS_{11} in table 2 and competitions essays were *Streptococcus pneumoniae* R6 (NC_003098.1), *Neisseria meningitidis* MC58 (NC_003112.2), *Pseudomonas aeruginosa* PAO1 (NC_002516.2), *Aggregatibacter actinomycetemcomitans* VT1169 (NZ_CP012958.1), *Haemophilus parainfluenzae* T3T1 (NC_015964.1), *Haemophilus ducreyi* 35000HP (NC_002940.2), *Mannheimia haemolytica* M42548 (NC_021082.1).

Chromosomal contamination measurements and corrections: Reads from the recipient genomic DNA that contaminated taken-up DNA samples were identified by a 'competitive alignment' step that aligned all the sample's reads (using bwa mem v0.7.15, samblaster v0.1.24, and sambamba v0.5.0) to a concatenated double-reference sequence consisting of the recipient Rd genome and the donor genome (86-028NP or PittGG). Because the donor and recipient

genomes are distinguished by a high density of SNVs, as well as structural variation and large indels (Harrison et al., 2005; Hogg et al., 2007; Mell et al., 2011), most contaminating Rd reads in uptake samples aligned only to the Rd reference while most of the desired donor-derived reads aligned only to the donor reference. Reads that mapped equally well to both genomes or to repetitive sequences within a genome were identified by their MAPQ scores of 0 and were removed from the analysis. The numbers of reads mapping uniquely to either donor or recipient genome were then used to calculate the contamination level of each taken-up DNA sample, as the ratio of recipient-mapping reads to total uniquely mapping reads (Table S2). Subsequent depth of coverage values and summary statistics were extracted for all positions or specific intervals using bedtools coverage v2.16.2 or sambamba flagstat (Table S2). All subsequent analyses and plotting used the R statistical programming language, including standard add-on packages dplyr, tidyr, plyr, ggplot2, data.table. Other packages used are specified below. Code is available at https://github.com/mamora/DNA_uptake.

Calculation of experimental uptake ratios from sequence coverage. After contaminating reads had been removed from each sample, uptake maps for each donor DNA were created by dividing the mean of the three normalized taken-up-DNA coverages for each position by the corresponding normalized input-DNA coverage. Finally, uptake ratios were normalized to a genome-wide mean uptake of 1.0 and smoothed by calculating the mean uptake over a 31 bp central-oriented sliding window using function rollapply from R package zoo v. 1.8-5. The effects of this smoothing are shown for the peak examples in Figure S5.

Periodicity analysis: To detect possible periodic patterns in coverage depth and in uptake ratios for the four datasets, periodograms were created using the R package TSA v. 1.2.

Analysis of uptake ratio data: To obtain a set of well-isolated USS_{10}s for analysis of peak shapes, we identified the closest peak separation at which USS effects did not overlap by examining sets of USS_{10} that were separated by different distances (1200, 1000, 800, 600 bp), excluding positions with missing data and USS_{10} that were 400 bp or less from positions with low input coverage (≤ 20 reads). Separation of ≥1000 bp was found to give the best compromise between good peak separation and the number of USS meeting the separation criterion (237 USS_{10}s and 209 $USS_{9.5}$s).

The search for non-USS sequences causing weak uptake effects used a subset of positions that were at least 0.6kb from the closest USS_{10}. This gave 575 'far from USS' segments summing to 29% of the genome. Uptake maps of segments containing positions with uptake ratios > 0.2 were examined visually to distinguish between (i) shoulders of adjacent USS_{10} peaks, (ii) increased uptake at USS with scores between 9.5 and 10, and (iii) increased uptake at non-USS sequences.

Incorporating within-USS interaction effects into uptake predictions: Figure 6 of Mell et al. (Mell et al., 2012) shows the strength and direction of pairwise interaction effects between positions on the same USS , inferred from uptake analysis of synthetic degenerate USS. From this figure we extracted the mid-range value of the interaction effect at each interacting pair of USS positions (only some pairs of positions showed such effects). For each 86-028NP $USS_{9.5}$ whose sequence differed from the USS consensus at both positions of such a pair, the USS score was modified by adding or subtracting the corresponding interaction value. The modified scores were then used by the model to predict DNA uptake, as described above.

Simulated noise analysis: Simulated noise-free uptake data for short and long fragments was first generated by smoothing raw uptake coverage data for 86-028NP-short (sample UP7) and 86-028NP-long (sample UP3) using a LOESS regression, and normalizing the results to a mean coverage of 1.0. Simulated relative-noise amplitudes for every genome position were generated using the 'tuneR' R-package (Ligges et al., 2018). Before being added to the noise-free data, the noise amplitude at each position was adjusted in proportion to the noise-free simulated coverage at that position, with the maximum noise range for each coverage level set by a multiplier (1.0, 1.5, 2.0, 2.5 or 3.0) and by the range of all experimental replicates for positions with that mean coverage. Red noise was used because, when added to the simulated noise-free coverage it gave an autocorrelation of 0.999, identical to that of the experimental data. Other noise types were evaluated but not used, since their autocorrelations were lower (0.975 for pink noise and 0.836 for white noise).

SUPPLEMENTAL REFERENCES:

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M., et al. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science. 269, 496–512.

Hunt, M., Silva, N. De, Otto, T.D., Parkhill, J., Keane, J.A., and Harris, S.R. (2015). Circlator : automated circularization of genome assemblies using long sequencing reads. Genome Biol. 16, 294.

Ligges, U., Krey, S., Mersmann, O., and Schnackenberg, S. (2018). TuneR: Analysis of music and speech. Available at: https://cran.r-project.org/package=tuneR.

Poje, G., and Redfield, R.J. (2003). Transformation of Haemophilus influenzae. In Methods in Molecular Medicine, Haemophilus Influenzae Protocols, M. Herbert, ed. (Totowa, NJ: Humana Press Inc.), pp. 57–70.

Sambrook, J. (2001). Molecular cloning : a laboratory manual (N.Y.: Cold Spring Harbor Laboratory Press).

Sinha, S., Mell, J.C., and Redfield, R.J. (2012). Seventeen sxy-dependent cyclic amp receptor protein site-regulated genes are needed for natural transformation in Haemophilus influenzae. J. Bacteriol. 194, 5245–5254.