Genome-wide analysis of DNA uptake by 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]

AUTHOR AFFILIATIONS:

1. Department of Zoology, University of British Columbia, Vancouver, BC, 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, USA.
3. Department of Otolaryngology – Head and Neck Surgery, Drexel University College of Medicine, 12 Philadelphia PA, USA.

¶ corresponding author

AUTHOR EMAIL ADDRESSES:

MM: mora@zoology.ubc.ca

JCM: Joshua.Mell@DrexelMed.edu

GDE: ge33@drexel.edu

RLE: rle36@drexel.edu

RJR: redfield@zoology.ubc.ca.

ABSTRACT

BACKGROUND

DNA uptake is the first step in natural transformation of bacteria, leading to DNA internalization and recombination. It is, therefore, a key determinant in genome evolution. Most bacteria take up DNA indiscriminately, but in two families of Gram-negative bacteria the uptake machinery binds preferentially to short sequences called uptake signal sequences (USS). These sequences are highly enriched in their genomes, which causes preferential uptake of self-DNA over foreign DNA.

RESULTS

To fully characterize the effects of this preference, and to identify other sequence factors affecting uptake, we carried out a genome-wide analysis of DNA uptake using both measured uptake and the predictions from a sequence-based uptake model. Maps of DNA uptake were developed by recovering and deep sequencing genomic DNA that had been taken up by competent *Haemophilus influenzae* cells, and comparing sequencing coverage from recovered samples to coverage of the input DNA. Chromosomal DNA that had been sheared into short fragments (50-800bp) produced sharp peaks of uptake centered at USS, separated by valleys with 1000-fold lower uptake. Peaks heights were proportional to the USS scores predicted by the

previously measured contribution to uptake of individual bases in each USS, as well as by predicted differences in DNA shape. Uptake of a long-fragment DNA preparation (1.5-17kb) had much less variation, with 90% of positions having uptake within 2-fold of the mean. Although the presence of a second USS within 100bp had no detectable effect on uptake of short fragments, uptake of long fragments increased with the local density of USS. Simulation of the uptake competition between *H. influenzae* DNA and the abundant human DNA in the respiratory tract DNA showed that the USS-based system allows *H. influenzae* DNA to prevail even when human DNA is present in 100-fold excess.

CONCLUSION

All detectable DNA uptake biases arose from sequences that fit the USS uptake motif, and presence of such sequences increased uptake of short DNA fragments by about 1000-fold. Preferred sequences also had rigidly bent AT-tracts and outer cores. Uptake of longer DNA fragments was much less variable, although detection of uptake biases was limited by strong biases intrinsic to the DNA sequencing process.

Keywords

DNA uptake, uptake bias, natural transformation, competence, uptake signal sequences, deep sequencing.

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 (1). The genetic exchange associated with this latter process contributes to adaptation and is known to have promoted resistance to antibiotics (2) and increased strains' intracellular invasiveness (3) and vaccine resistance (4,5). Thus, understanding how different genomic regions evolve via natural transformation processes could be used to predict the spread of pathogenic traits.

Most competent bacteria that have been tested take up DNA regardless of sequence, but species in two families, the Pasteurellaceae and the Neisseriaceae, exhibit strong sequence biases for short motifs (6). 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 (7,8). The distribution of the preferred sequences around the chromosome is uneven (9), which may cause different genes to experience quite different rates of genetic exchange.

Most steps in the DNA uptake process are highly conserved among naturally transformable species (6). 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 dsDNA at the cell surface. This is followed by the retraction of the pilus, which pulls the DNA across the outer membrane into the periplasm. Uptake is thought to begin internally on DNA fragments, not at an end, because circular DNAs are taken up as efficiently as linear DNAs (10). 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. 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 (11,12). Subsequent translocation of the DNA into the cytoplasm requires a free DNA end; only the 3'-leading strand passes through an inner membrane pore encoded by the *rec2*/*comEC* gene, while the other

strand is degraded and its nucleotides dephosphorylated and imported as nucleosides (13). Circular DNA molecules are not transported from the periplasm into the cytoplasm because they lack free ends (13).

Direct measures of DNA uptake bias: Uptake-competition experiments in the Pasteurellacean *Haemophilus influenzae* and in *Neisseria gonorrhoeae* showed that uptake of genetically marked 'self' DNA was inhibited by unmarked self DNA but not by DNA from unrelated sources (7,8). Subsequent DNA uptake experiments using cloned radiolabeled DNA fragments found that the *H. influenzae* self-preference is caused by the uptake machinery's strong bias for a short sequence motif, the uptake signal sequence (USS) (14). Sequence comparisons and site-directed mutagenesis initially identified an 11bp sequence, with a strong contribution by flanking AT-rich sequences (15,16), and genome sequencing identified 1465 occurrences of a 9bp USS core in *H. influenzae*, and 1892 occurrences of an unrelated 10bp 'DUS' in *N. meningitidis* (9,17). Later analyses using mutagenesis and sequencing of pools of degenerate USS identified the contribution of each position, which are summarized by the sequence logo in Figure 1 (18,19). This study found the central GCGG bases to be crucial for uptake, with smaller and synergistic contributions made by flanking bases and two adjacent AT-rich segments. The Pasteurellacean USS is unrelated to the Neisseriacean DNA uptake sequence (20), and different lineages within each of the families can have variant preferred motifs (21,22).

Evolution of uptake sequences in the genome: Alignment of distinct homologous genomic regions between distantly related Pasteurellaceae species showed that the USS evolve by point mutations (22); *i.e.* they are not insertion elements. Danner et al. (15) proposed that the combination of uptake bias and genomic recombination creates an evolutionary pressure that will cause the preferred uptake sequences to accumulate throughout the genome, with locations limited mainly by interference with gene functions. Consistent with this, both USS and DUS are

112 underrepresented in newly acquired segments, in rRNA genes, and in coding sequences,
113 especially those with strong functional constraints (17,23). Modeling by Maughan et al. (24)
114 confirmed that this molecular drive process could produce uptake sequence distributions like
115 those of real genomes, with no need for direct selection for these sequences or for the
116 chromosomal recombination they promote. Thus, the presence of biased DNA uptake machinery
117 may be sufficient in itself to explain the abundance of uptake sequences. Such sequence biases
118 may have arisen solely by direct selection on the DNA uptake machinery for more efficient DNA
119 binding, or by this in combination with indirect selection for preferential uptake of conspecific
120 DNA.

121 Pasteurellaceae and Neisseriaceae species occur primarily in respiratory tracts and other
122 mucosal environments (25), where transformation can only occur if the released bacterial DNA is
123 able to compete with abundant host DNA for binding to the uptake machinery (26,27). These host
124 DNAs are not expected to be enriched for uptake sequences and, since recombination requires
125 strong sequence similarity between incoming DNA with a genomic segment, any nonhomologous
126 DNA sequences that are taken up will usually be degraded rather than recombining with the
127 bacterial genome (28).

128 The goal of the present study was to measure DNA uptake at every position in the *H. influenzae*
129 genome, and to use this data to characterize the DNA uptake biases caused by the USS and any
130 other sequence factors. We first developed a computational model that predicted the effect of
131 uptake sequences on DNA uptake across the *H. influenzae* genome. This model's predictions were
132 then compared with actual measurements of DNA uptake produced by sequencing genomic DNA
133 fragments that had been recovered after being taken up by competent *H. influenzae* cells.
134 Discrepancies between predicted and observed uptake revealed the strength of the bias, effects
135 of USS sequence differences, and the influence of the distribution of USS locations. These factors

in turn increase 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 simulated, the fragment-size distribution of this DNA, and an uptake function that describes how uptake probability depends on USS presence and strength. The output is the expected relative uptake of every position in the genome.

In developing the model we were guided by basic principles of how sequence-specific DNA-binding proteins interact with DNA (29,30). 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 enables uptake of the DNA fragment across the cell's outer membrane.

The model did not explicitly simulate the first step, non-specific binding, since this is expected to be equally probable for all DNA positions. The specific binding and DNA uptake steps were separately modeled since 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 (18,23). Our strategy was to score genome positions with the uptake-prediction matrix from Mell et al.'s degenerate-sequence uptake experiment (19) (Table S1), and to use overrepresentation of high-scoring sequences as the USS criterion. We scored every position in the genomes of the standard *H. influenzae* reference strain Rd, of the two strains whose uptake we investigated, 86-028NP ('NP') and PittGG ('GG'), and of four randomly generated genome-length sequences with the same base composition (Supp Figure 1). 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 Supp. Figure 1). DNA uptake analyses used a USS cutoff score of 10 bits ('USS_{10}', n=1941) or a less stringent 9.5 bits ('$USS_{9.5}$', n=2248).

The binding step of the computational model evaluated whether the fragment under consideration contains any USS_{10}, and their probability of being encountered by the uptake machinery receptor. In Model version I, fragments with no USS had a baseline binding probability of 0.2; this was reduced in Model versions II and III. In Model versions I and II the encounter probability decreased linearly with fragment length, but increased if more than one USS_{10} was present in proportion to their separation. In Model III binding was instead a function of the number of USS_{10} in the fragment. The probability that this binding led to DNA uptake was a function of the USS score, from a baseline of 0.2 at a score of 9 bits to a maximum of 1 at 12.6 bits.

In Model version 1 this function was linear, but it was replaced with a sigmoidal function in Models II and III.

Once the contributions of every size class of fragment had been calculated for each position (Figure 2A), the model combined all the contributions, taking into account the frequency of each size class in the input DNA. The position-specific uptake predictions were then normalized to a genome-wide mean uptake probability of 1.0.

Model results: Figure 2B and C 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 300bp). 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 25-300bp in length (Figure 2C and D), 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. The grey peak in Figure 2C shows that model versions with a baseline of USS-independent uptake caused valleys to be higher and peaks correspondingly lower. With this original version of the model ('Model I' Table S4), heights of predicted peaks were linearly proportional to USS scores (dashed red line in Figure 2D). In simulated genomes with more than one USS (Figures 2D, 2E, 2F), isolated peaks were only seen when the DNA fragments being taken up were substantially shorter than the spacing of the USSs, and disappeared entirely when the fragments were long enough that almost all contained at least one USS (Figure 2F).

Figure 2G and 2H show the predicted uptake maps when this model analyzed a 50kb segment of the *H. influenzae* NP genome, using the short-fragment and long-fragment size distributions from

the actual uptake experiments described below (Supp. Figure 2A and B), and Figure 2I shows the distribution of USSs over this segment. Because the short DNA fragments are shorter than the typical separation between USSs, uptake is predicted to be restricted to sharp peaks at each USS. In contrast, uptake of long DNA fragments is predicted to be much more uniform, since 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 NP and GG, whose core genomes differ from Rd and each other at ~3% of orthologous positions (31). To allow efficient recovery of the taken-up DNA, the Rd strain in which competence was induced carried a *rec2* mutation that causes DNA to be trapped intact in the periplasm (16). The NP and GG genomic DNAs were pre-sheared to give short (50-800bp) and long (1.5-17kb) DNA preparations (size distributions are shown in Supp. Figure 2), 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. (19,32,33). Recovered DNA samples were sequenced along with samples of the input NP and GG DNAs and of the recipient Rd DNA. The input and uptake reads were then aligned to the corresponding NP and GG 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.

Effects of contaminating Rd DNA: Preparations of DNA recovered after uptake always included some contaminating DNA from the recipient Rd chromosome. The divergence between the Rd

and donor genomes allowed the extent of this contamination to be estimated by competitively aligning the recovered reads from each sample to a reference that included both recipient and donor genomes as separate chromosomes. Thus, reads that uniquely aligned to only one chromosome could be unambiguously assigned to either donor or recipient. The resulting Rd chromosomal contamination estimates were between 3.2% and 19.3% of reads; specific values for each sample are listed in Table S2.

The effects of this contamination were not expected to be uniform across the donor genome, since segments of the NP and GG genomes with high divergence from or with no close homologs in Rd would be free of contamination-derived reads. We used the competitive-alignment described above to create contamination-corrected uptake coverages, by discarding all reads that preferentially aligned to Rd rather than NP or GG. We also discarded reads that could not be uniquely mapped to the donor genome; this included reads from segments that are identical between the two strains ('double-mapping reads') and reads that mapped to repeats, such as the six copies of the rRNA genes. This removed an average of 18.6% of reads (range 8.9%-28.3%), left some segments of the NP and GG genomes with no coverage in all samples (2.3% and 2.1% respectively) and reduced coverage adjacent to these segments. Contamination details for each sample are provided in Supplementary Table 2, and the impacts are considered below.

Uptake ratios: To control for position-specific differences in sequencing efficiency, read coverage at each position in each uptake sample was divided by read coverage in the corresponding input sample (e.g. each NP-short uptake sample by NP-short input). Normalizing the mean of the three replicates to a genome-wide mean uptake of 1.0 then gave a mean 'uptake ratio' measurement for each genome position for each DNA type. Figure 3 and Supp. Figure 3 show the resulting uptake ratio maps, smoothed using a 31bp sliding window.

Figure 3A shows the short-fragment uptake ratio map for the first 50kb of the NP genome; the ticks in Figure 3C indicate locations and scores of USS_{10}s. The pattern is strikingly similar to that predicted by the model (Figure 2G). Sharp uptake peaks are seen at USS_{10} positions; some peaks are separated by flat-bottomed valleys and others overlap. Supp. Figure 3A shows a similar map for the first 50kb of strain GG's genome. The full-genome maps of these NP and GG uptake ratios in Supp. Figure 3D and 3G display the consistency of the peak heights.

Also as predicted by the model, the long-fragment DNA samples (Figure 3B and Supp. Figs 3B, E and H) had much less variation in uptake than the short-fragment samples; 90% of positions had uptake ratios within two-fold of the mean, and there were few high peaks or low valleys. Extended genome segments with low or no uptake coincided with large gaps between USS_{10}s. The largest gap is in the NP segment between 95 and 145kb —the site of a genomic island with high similarity to an *H. influenzae* plasmid but few USS (34).

Sources of variation: Characterization of USS dependent uptake biases and possible USS-independent biases in uptake coverage was limited by strong variation in sequencing coverage, presumably due to biases in the library preparation and sequencing steps. Supp. Figure 4A compares coverage for the NP short and long input samples, showing that this variation was both reproducible and sequence dependent. These biases are expected to have very similar effects on coverage in all samples, precluding calculation of uptake ratios where input coverage is zero and generating high levels of stochastic variation where coverage is low.

In Supp. Figure 4B and C, the colouring of NP long-fragment uptake ratio points according to input coverage reveals that all of the extreme uptake ratio values occurred in regions of low input coverage. Table S3 extends this analysis to the whole genome, showing that anomalously high uptake was seen mainly at positions with low input coverage, indicating that these values are

likely due to stochastic variation rather than to genuinely high uptake. In contrast, positions with low uptake showed no such bias, indicating that these are mainly due to genuinely low uptake.

Periodicity: Bacterial genomes show periodicity for several features related to DNA curvature and codon usage biases (35), so we examined the distribution of uptake ratios across each genome by Fourier analysis, using the R package TSA. The log-log views in Supp. Figure 5 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 bias analysis: Our strategy to investigate the DNA uptake process was to analyze discrepancies between model predictions and observed uptake ratio peaks in the NP short-fragment dataset, since these revealed ways in which the simple assumptions underlying the model mis-characterized 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. We then compared the refined model's predictions to the real uptake ratios for the NP long-fragment DNA, and finally to the long- and short-fragment uptake ratios for the GG DNA.

Figure 4A compares predicted (orange line) and measured (blue line) uptake of short-fragment DNA for the first 50kb of the NP genome. The model's predictions of peak locations and peak shapes were both extremely accurate, but the predicted baseline uptake in the valleys between peaks was too high, and some predicted peaks were too high or too low.

We next inspected the depths of the valleys between uptake ratio peaks. Although these were quite variable, (see log-scale inset in Figure 3A), the histogram of uptake ratios below 0.1 in Supp. Figure 6 shows that most deep valleys fell to uptake ratios between 0.0005 and 0.005. In Model

297 version I, fragments that lacked USS were arbitrarily assigned binding and uptake probabilities of
298 0.2, resulting in predicted baseline uptake of ~0.08. To improve the model, we lowered the both
299 baseline parameter settings from 0.2 to 0.02, which gave predicted baseline uptake of ~0.002 in
300 regions far from a USS.

301 The peak heights predicted by the initial model were linearly proportional to USS score (Figure
302 2D) reflecting the model's assumed linear dependence of uptake probability on USS score (blue
303 line in Figure 4B). However, analysis of 209 $USS_{9.5}$ that were separated by at least 1000 bp (to
304 minimize effects of overlapping peaks) (black dots in Fig. 4B) showed that experimental uptake
305 ratios instead followed a sigmoidal relationship with score. Very little uptake was seen at
306 isolated USSs with scores between 9.5 and 10 bits, and consistently high peaks were observed at
307 isolated USSs with scores above 11.5 bits. Accordingly, Model version I was further revised to use
308 a sigmoidal function fit to this data (orange line in Figure 4B); the new predictions (Model
309 version II) better matched the observed valley depths and peak heights (Figure 4C; Pearson
310 correlation rose from 0.691 to 0.755).

311 **Symmetry and shape of uptake peaks:** The DNA uptake motif is not palindromic, so
312 asymmetric interactions of DNA with the uptake machinery could polarize DNA uptake by
313 causing one side of the motif to be pulled into the cell more efficiently than the other. The motif's
314 strongly weighted positions are also all on one side, not at its center, which might cause peak
315 centers to be shifted relative to USS centers. Supp. Figure 7 shows that, when all isolated USS_{10}s
316 with high uptake (≥ 3) were analyzed in the same orientation, the mean peak was both centered
317 on the USS and symmetric about it (no significant difference between mean ratios left and right of
318 the USS center at position 16 ($p=0.9$)).

Pairwise base interaction: Mell et al. (19) found evidence for substantial contributions to uptake by long-distance pairwise interactions between AT-tract bases and core bases. To incorporate the effects of these interactions in the model, USS scores were adjusted using the interaction information in Figure 6 of Mell et al. (19). This change had little effect on high-scoring USS, but further reduced the scores of low-scoring USS (Supp. Figure 8). The uptake ratios predicted by the modified scores were no more accurate than those for the original scores (correlation between observed uptake and predicted uptake without interactions: 0.937; with interactions: 0.936), likely because most of the affected scores were already very low (70% were < 10.5).

DNA shape effects: Although the analysis of Mell et al. (19) found no evidence of pairwise interactions between close positions, we used analysis of DNA shape to detect both pairwise and more complex interactions over a 5bp distance. Shape features that can be predicted from DNA sequence includes the minor groove width, the propeller twist between bases in a base pair, the helix twist between one base pair and the next, and roll, the rotation of one base pair relative to the next. The thick grey line in each panel of Figure 5 shows these features 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 (36). 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.

The coloured lines in Figure 5 compare the shape features of subsets of isolated USS with similar scores but different uptake ratios. Panels A-D compare the shape features of low-scoring USS ($USS_{10\text{-}10.5}$) whose uptake ratios were low (<0.6, blue lines) or high (>2.0, orange lines). Similarly, panels E-H show the same comparison for USSs with better scores ($USS_{10.5\text{-}11}$). USSs with scores higher than 11 were not analyzed since they did not exhibit enough uptake variation to reveal

correlations between uptake and DNA shape. Although very similar inner-core shape features were seen for low-uptake and high-uptake subsets, 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 of the AT tracts facilitate DNA uptake.

Detecting weak uptake biases: Any weak uptake biases that exist will only be detectable in genome segments that lack a strong USS, so we searched for biases arising from either low-scoring USS or other factors using a far-from-USS_{10} dataset containing only DNA segments whose ends were at least 0.6kb from the closest USS_{10}. This dataset contained 575 segments where weak uptake effects could in principle be detected (29% of the genome); their mean uptake ratio was 0.0097. Of these segments, 62 were set aside because they had low input coverage (<20 reads). Only 16 of the remaining 513 segments contained positions with uptake ratios >0.2, indicating that sequences conferring weak biases are quite rare. Ten of these segments contained distinct peaks (heights between 0.2 and 1.0) that coincided with weak USSs scoring between 9.47 and 10 bits; the other six lacked distinct peaks but contained shoulders at the extended bases of strong USS_{10} peaks. However, this far-from-USS_{10} dataset also contained 68 other similarly weak USS that were not associated with uptake peaks (scores 9.50-9.99 bits, mean uptake of 0.033). Panels I-L of Figure 5 show that shapes of the 10 USS with uptake > 0.2 (orange lines) were more similar to that of the consensus USS than the shapes of the 68 USS with no peaks (blue lines). The boxplot in Supp. Figure 9 summarizes uptake ratios at these 78 weak USSs, showing that median uptake ratios were very low for all sub-classes of weak USSs. Since this analysis did not find any non-USS positions giving uptake higher than 0.2, it also shows that other sequence factors do not detectably promote uptake in the absence of a USS.

Uptake of fragments with multiple USSs: Fragments containing two or more uptake sequences might be expected to have higher uptake, since they have more targets to which the uptake machinery receptor could bind, but only one of the two previous studies in *Neisseria* found this effect (37,38). Many genomic USSs are sufficiently close that they will co-occur even on short DNA fragments; 23% of NP USS_{10}s are within 100bp of another USS_{10}, and 17% are within 30bp (Supp. Figure 10A**)**. The initial runs of the uptake model assumed that multiple USS on the same fragment decreased the search distance for the specific-binding step but did not affect the uptake step, which used the mean USS score, not the best. This predicted single peaks at pairs of USS_{10} within 100bp of each other, and two distinguishable peaks or a peak with a distinct shoulder at USS with wider separations. Except for very close USSs, the single peaks were about 15% higher than for isolated USSs with the same scores.

Visual examination of uptake ratios at the 230 pairs of NP USS_{10}s within 100bp found single peaks; Supp. Figure 10B) shows that these USS pairs (coloured points) do not have noticeably higher uptake ratios than isolated USSs (grey points, from Figure 4). However, a mean difference in peak heights of less than 12% could not be confidently detected because of the low numbers of USS pairs, especially those with scores lower than 11.0 bits. A special class of USS pairs consists of overlapping oppositely oriented pairs that are located at the ends of genes and act as transcriptional terminators (9,17,39). Supp. Figure 10A shows that the NP genome has 109 USS_{10} pairs whose centers are within 14bp: 69 0-3 bp apart (-/+ orientation) and 40 10-14bp apart (+/- orientation). Supp. Figure 10C shows that uptake ratios at these did not differ from those at isolated USS_{10}s ($P = 0.12$ for the 103 pairs whose mean scores were ≥11.0 bits). These results suggest that the presence of two USS_{10}s in a 100bp segment does not detectably increase the probability of the receptor finding a USS, a result consistent with that of Ambur et al. (37).

Uptake of long-fragment NP DNA:

Next step was to investigate the uptake of longer DNA fragments, using the improved model (Model II, Table S4) and the NP long-fragment dataset. Supp. Figure 11A compares this model's predictions with the observed uptake over the same 50kb genome segment as in Figure 3B. In contrast to the model's accurate prediction of short-fragment uptake ratios (correlation of 0.94), it seriously underpredicted the variation in long-fragment uptake ratios (correlation of 0.61), predicting uptake <0.8 or >1.2 for only 13% of NP positions when experimental uptake ratios were outside these limits at 51% of positions. The likeliest explanations are that: 1) the fragment-length distribution the model used overestimated the actual proportions of long fragments available for uptake, or 2) USS density has an effect on uptake of long fragments that was not detectable in the short-fragment dataset.

To investigate the first explanation, uptake predictions were generated using a shorter fragment-length distribution, that of NP-long DNA recovered after uptake. Although this DNA's substantial depletion of long fragments (Supp. Figure 12) could be due to non-uptake effects (post-uptake steps in the periplasm or biases during DNA purification), it could also be due to preferential binding or uptake of short fragments. This length distribution thus provided a lower-bound estimate on the real sizes of fragments that were taken up. However, when it replaced the input DNA distribution as a parameter in Model II, the model's correlation with observed uptake ratios was only slightly improved (0.63 vs 0.61) (Supp. Figure 11B), suggesting that fragment length differences were not a major factor.

To test the second explanation, we revisited the effect of multiple USS on uptake, this time examining the relationship between DNA uptake and the number of USS_{10}s in a 5kb neighbourhood. Supp. Figure 13A shows that positions with higher local USS densities had higher uptake, and that Model version II only partially accounted for this effect (compare black and red lines). Since mean USS scores did not increase with numbers of USS in the window (Supp. Figure

13B), long fragments with more USSs must instead have a higher uptake probability than predicted by the model. The model was consequently revised again; rather than using fragment length and USS separation to calculate binding of each fragment, Model version III used the observed relationship between USS density and uptake ratio (black line in Supp. Figure 13A) to specify fragment-binding probabilities as a function of the number of USS in the fragment. However, this only slightly improved the USS density analysis and the overall correlation between predicted and observed uptake (Supp. Figures 11C and 13C; correlation 0.65 vs. 0.61), and caused a corresponding decrease in the short-fragment correlation (0.90 vs. 0.94). Predictions were slightly better when the recovered fragment-length distribution was used with Model III, (Supp. Figure 13A, green line; overall correlation =0.67), suggesting that both explanations contribute to the uptake variation.

Because the correlation between USS-based prediction and observation was only modestly improved by these changes, we also investigated the extent to which the correlation was limited by stochastic noise arising at the regions of low sequencing coverage described earlier. 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. Supp. Figure 14 shows that although 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, the effect was much worse for the long-fragment simulations. Simulations with noise levels of 2 and 2.5 gave short-fragment correlations very close to the 0.90 between the model and the real data. For the same noise levels the long-fragment correlations were 0.86 and 0.75 respectively, confirming that much of the disparity between measured uptake ratios and USS-based predictions was due to noise in the data. However, these correlations are still 21% and 10% higher than the best

correlation obtained between Model III predictions and real data, suggesting that one or more factors remain unidentified.

How well does the model predict uptake of PittGG DNA?

Since the final version of the model (Model III) had been refined using uptake data for DNA of strain 86-028NP, we further evaluated it using the measured uptake data for DNA of strain PittGG, which differs from NP by SNPs and indels affecting about 11% of its genome. Supp. Fig. 15 compares the model's uptake predictions with the observed GG uptake ratios. For short-fragment data the correlation between predicted and observed uptake of GG DNA was 0.90, the same as that for NP. However, for long-fragment data, the GG correlation was substantially worse (0.50 compared to 0.65 for NP).

Some of this discrepancy is due to noise arising from low sequencing coverage. For GG DNA the mean uptake ratio was substantially greater, and the variation more extreme, at low coverage positions (Supp. Figure 16A); this was not seen for NP DNA. However, the cause of this is not clear, since NP and GG had similar frequencies of low-coverage positions for both short-fragment and long-fragment input samples (Table S3).

Predicted competition with human DNA:

H. influenzae's natural environment is the human respiratory tract, where *H. influenzae* DNA must compete for uptake with host-derived DNA whose mucus concentration can exceed 300 μg/ml in healthy individuals (26,27). We used the final model (version III) to investigate this competition.

We started by scoring the human genome for USS. This identified 14924 USS_{10}s (density 4.6/Mb), with a mean score of 10.29 bits. For comparison, the NP genome has 1022 USS_{10}s/Mb with a mean score of 11.45 bits, and simulated sequences with the 41% GC content of human DNA had 56 USS_{10}s/Mb with a mean score of 10.31 bits. Since the USS motif includes a CpG, the

underrepresentation of USS_{10} in human DNA is probably a consequence of the 4-5 fold depletion of CpGs in the human genome due to deamination of methylated cytosines (40,41).

To determine how much human DNA would be needed to outcompete *Haemophilus* DNA for uptake, we first determined the proportion of expected uptake from USS_{10} in human vs *Haemophilus* genomes. We did this by using Model III to predict uptake of the *H. influenzae* NP genome and of 4 randomly selected 1.9 Mb segments of human DNA. Because human DNA will contain many fragments lacking USSs the predictions were made using baseline binding and uptake probabilities of 0.0 and 0.02. To approximate the lengths of DNA fragments in the respiratory tract (26,27), the model was run using fixed fragment lengths of 1kb and 10kb. The predicted uptake at each position (without normalization) was then summed across all positions to get a total uptake value for each fragment length and baseline assumption. Table S5 shows these uptake values, in arbitrary units.

In all cases, if *H. influenzae* cells were exposed to an equal mixture of *H. influenzae* and human DNAs, more than 99% of the DNA taken up was predicted to be from *H. influenzae*. Substantial amounts of *H. influenzae* DNA (14-35%) would be taken up even if the human DNA were in 1000-fold excess. For 10 kb fragments, baseline uptake of fragments lacking USS made only a small contribution, but for 1 kb fragments it increased total uptake of human DNA by 86%, and reduced *H. influenzae* DNA's advantage by 36% when human DNA is in 1000-fold excess. The uptake advantage of *H. influenzae* DNA is due to both the much higher frequency of USS_{10}s in its genome and to its USSs' much stronger matches to the USS consensus.

Discussion

We measured DNA uptake by competent *H. influenzae* cells at every position in the genome, using short-fragment and long-fragment DNA preps from two divergent strains. Differences between

predicted and observed uptake revealed the strength of the uptake machinery's bias towards USS, the absence of other sequence biases, and a role for DNA shape. These findings increased our understanding of DNA uptake bias and the role it plays in recombination.

Implications for the molecular mechanism of DNA uptake.

The uptake specificity for USS is very strong. With short fragments, valleys at USS-free segments had ~1000-fold lower uptake ratios than peaks at USS. Although the non-zero uptake ratios in USS-free regions could mean that fragments lacking USS are occasionally taken up, they are also consistent with no uptake at all of fragments lacking a USS, since this low coverage could have arisen artefactually, either from low-level contamination of the recovered DNA with 0.2%-0.8% donor DNA that had not been taken up, or from under-correction of the contamination of recovered-DNA samples by recipient DNA.

The correlation of the model predictions with measured uptake ratios was excellent for short fragments but modest for long fragments. However, the model's predictions may be more accurate than indicated by the correlation coefficients, since stochastic variation at low coverage positions introduced substantial noise into the calculation of experimental uptake ratios. Similar errors associated with changes in coverage have been detected in ChIP-seq and RNAseq studies (42–44).

Previous analyses of the effects of multiple uptake-sequences on the amount of DNA taken up gave seemingly contradictory results, which ours help resolve. Consistent with Ambur et al.'s (37) study of very close uptake-sequences in *Neisseri*a, we did not detect any increased uptake when a second USS_{10} was within 100 bp of the first. Consistent with Goodman and Scocca's (38) results, also in Neisseria, we found that, for larger DNA fragments, a higher local density of USS gave higher uptake. Since DNA-binding proteins can search for their sequence-target by 1-

dimensional sliding (30), this discrepancy might arise from the effect of fragment length and USS number on the chance that the uptake receptor will detach from the fragment without having found a USS. Density of USSs might thus have a greater impact in long fragments where the probability of detaching will be greater.

The predicted shape differences between USSs with strong or weak uptake suggest strong uptake bias for USS that are rigidly bent at AT-tracts and outer core (36,45). 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 (36,46). These features have been integrated successfully in some transcription factor binding models (47), but using them to improve uptake prediction will require more comprehensive investigation into the effects of DNA shape on uptake.

Implications for recombination.

Davidsen et al. (48) found higher densities of USS in genes for DNA replication, repair and recombination, and suggested that this distribution resulted from selection for preferential recombination of genes involved in maintenance of the genome. However, any effects of DNA uptake biases on the distribution of recombination across the *H. influenzae* genome are likely to be weak. Although uptake of short DNA fragments (<800 bp) depends dramatically on USSs, this will have little genetic consequence since such short fragments typically are degraded before they can recombine (13). On the other hand, 96% of fragments long enough to participate efficiently in recombination (~3.5kb) contain at least one USS_{10} (13). The major exceptions are the few genomic regions lacking USS. In NP these include the aforementioned 50 kb genomic island and eleven 5-9 kb segments. The GG genome has no large segments without USS but has twelve 5-9 kb segments and a 12 kb segment containing several integrases. Uptake ratios of 90%

of the genome were within two-fold of the mean, which is consistent with previous analysis showing that USS distributions are not strongly correlated with gene functions (23).

Uptake of DNA from other species can also influence recombination, either directly if the DNA is sufficiently similar to *H. influenzae* DNA or indirectly if it competes for uptake of *H. influenzae* DNA. Since the *H. influenzae* USS is shared with other Pasteurellacean species, both factors will be important when *H. influenzae* shares the respiratory tract with coinfecting Pasteurellaceae. Species that share the *Hin*-USS type of USS are expected to compete efficiently for uptake, with recombination limited by sequence similarity, but uptake of DNA from species with the variant *Apl*-USS type is known to be inefficient (22). Although this variant has the same inner core GCGG, the first AT-tract and two outer core bases as the *Hin*-USS, these matches would only give an average score of ~9.1 bits, too low for effective uptake by *H. influenzae*.

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 DNA, even if this is in 100-fold excess. This does not necessarily imply a selective advantage for self-uptake, since USS accumulation in *H. influenzae*'s genome may simply be due to the molecular drive process.

Uptake of DNA in the respiratory track 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, cell death will release high concentrations of these proteins, which can contribute significantly to biofilm stability (49). 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 reexamine DNA uptake using DNA that retains its bound proteins.

552

553 Methods

554 **Bacterial strains, culturing, and competent cell preparations:** Growth and culturing of
555 *Haemophilus influenzae* strains that were used as donor (RR3133 and RR1361) and recipients
556 (RR3117 and RR3125) in the DNA uptake experiments followed standard methods (50).
557 Recipient strains RR3117 and RR3125 are both *rec2* derivatives of strain Rd KW20, with and
558 without a spectinomycin resistance gene respectively (Table S2). Donor strain RR3133 is an 86-
559 028NP derivative with a nalidixic acid resistance gene; and RR1361 is an unmodified PittGG
560 isolate. Strains were grown at 37 °C on brain-heart infusion broth supplemented with NAD
561 (2μg/ml) and hemin (10 μg/ml) (sBHI) with or without 1% agar to isolate single colonies on
562 plates or grow liquid cultures. To prepare naturally competent cells, cultures were maintained in
563 exponential growth for at least 2 hr, and at OD_{600} = 0.2, cells were collected by filtration from 10
564 ml of culture, transferred into starvation medium M-IV, and incubated at 37 °C for 100 minutes
565 before DNA uptake experiments (51).

566 **Input DNA preparations.** Donor DNA was purified using standard phenol:chloroform
567 extractions (52) from 10 ml overnight cultures of clinical strains 86-028NP and PittGG carrying
568 selectable markers (Table S2). High molecular weight DNA was then sheared into separate 'long
569 fragment' (1.5-9kb) and 'short fragment' (50-500bp) preparations using Covaris G-tubes and
570 sonication respectively. The fragment size distributions were measured using a Bioanalyzer with
571 the DNA 12000 kit (Agilent), dividing the relative fluorescence of each time point by its fragment
572 length estimated from the size standards. Fragment lengths were then grouped in classes of 10bp
573 for short fragments and of 200bp for large fragments. Using the large fragment distribution, in
574 our predictive model, grouped in 200bp bins took very long (an average of 184 seconds for

100bp on a macOS Mojave v.10.14.5 with 8Gb of memory and a 2.2 GHz processor). For this reason, we grouped fragment sizes in 1000bp bins, which reduced running times 12-fold. Results with 1kb binds were nearly identical then when using 200bp bins.

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 10mM ph 7.5, EDTA 10mM, CsCl 1.0 M). Periplasmic DNA was extracted using the organic phenol:acetone extraction method as described by (19,32,33) followed by an ethanol precipitation. DNA was resuspended by using 20 µl of T10E10 buffer (Tris-HCl 10mM ph 7.5, EDTA 10mM). 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 both a Qubit dsDNA HS Assay Kit (absolute DNA concentration) and by transformation into Rd (concentration of NalR donor DNA).

DNA sequencing and data processing. Sequencing libraries of the input and recovered DNA samples were prepared using the Illumina Nextera XT DNA library prep kit according to manufacturer recommendations. An Illumina NextSeq500 was used to collect 1-10 million paired-end reads of 2x150nt for each library (for >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 v2.0, followed by Circlator (53), and then Quiver to polish the circular junction). Sequence references for this new PittGG reference, as well as the genome references for 86-028NP (NC_007146.2) and Rd KW20 (NC_000907.1) were then corrected from input and control reads based on Illumina sequencing using Pilon v1.22 (59). This was particularly important for the Rd KW20 recipient reference, since the original (60) sequence dates from 1995 and contains several hundred ambiguous bases and errors (10). This also accommodated differences between the sequence references and the donor strains, which carried antibiotic resistance markers (Table S2).

Chromosomal contamination measurements: To identify and remove contaminating genomic recipient reads in the recovered-DNA datasets, reads were aligned (via bwa mem v0.7.15, samblaster v0.1.24, and sambamba v0.5.0) competitively to a concatenated reference sequence consisting of the recipient Rd genome and the donor genome (NP or GG). Because the donor and recipient genomes are distinguished by a high density of SNVs, as well as structural variation and large indels (31,34,54), most contaminating Rd reads in uptake samples aligned to the Rd reference while the desired periplasmic donor reads aligned to the donor reference. Reads that mapped equally well to both genomes or to repetitive sequences within a genome were flagged as low quality. The levels of uniquely aligned reads with quality > 0 that mapped to donor and recipient chromosomes were used to calculate the percentage of contamination with recipient Rd DNA (Table S2), and only the former were used for calculation of uptake ratios. 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.

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. (19); 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 is written in R v.3.5.1. Given a list of USS positions and scores in a DNA genome of specified length, it uses a specified distribution of DNA fragment sizes (over 10bp bins) to calculate the relative uptake of every position in the genome. The genome is assumed to be circular. At each DNA position in turn, for each 10bp bin of DNA fragment sizes, the model sums the predicted uptake contributions for every fragment of that size that overlaps the position. For efficiency, the full calculation is only done for the first position. At each subsequent position, the model calculates 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 2A).

Each fragment's contribution depends on the number of USS it contains, and on the scores and separation of these USS. Fragments with no or incomplete USSs have baseline probabilities of being bound (p_bind) and taken up (p_uptake); initial values for both = 0.02. For fragments with one or more complete USS, p_bind depends on the fragment length (L) and on the separation of the USSs if more than one is present, and p_uptake depends on the USS score(s). Initially p_bind

for a fragment with one USS = 1 - L/17000, assuming a maximum fragment length of 17kb. For a fragment with 2 or more USS, the effective value of L was initially decreased by the separation between the USS, so p_bind = 1 - (L - separ)/17000. Initially p_uptake for a fragment with one USS = (score - 10)/maxScore - 10), so p_uptake increased linearly from 0 for score = 10 to 1.0 for score = 12.6. For a fragment with two or more USS, the mean of the USS scores was initially used.

After the experimental uptake ratios had been analyzed, both p_bind and p_uptake were modified to use sigmoidal functions. The revised p_bind = 1/(1 + exp(7000 - L/-1500), where 7000bp is the DNA length at the inflection point of the function and -1500 specifies the slope at this point. The revised p_uptake = 1/(1 + exp(3.48(score-10.6), where 10.6 is the USS score at the inflection point of the function and 3.84 is a value determining the slope at this point, estimated with the R package Sicegar v. 0.2.2 (55), using USS scores and corresponding uptake ratios for a set of 209 $USS_{9.5}$ isolated by at least 1000bp. A summary of each model parameters and equations is included in Table S4.

Once the model has calculated the contributions of a specific fragment size to uptake of every genome position, it moves on to the next size class. Once the contributions of every size class have been calculated, the model combines all the contributions for each position, taking into account the frequency of each size class in the input DNA. These position-specific uptake predictions are then normalized to a mean uptake value of 1.

Calculation of experimental uptake ratios from sequence coverage. Uptake maps for each donor DNA were created by dividing the mean of the three normalized recovered-DNA coverages for each position by the corresponding normalized input-DNA coverage. Finally, uptake ratios were normalized to a mean uptake of 1 over the entire genome and smoothed by calculating the mean uptake over a 31bp central-oriented sliding window using function rollapply from R

package zoo v. 1.8-5. Because the replicates were extremely reproducible by Pearson correlation, most plotting and analyses used the mean values.

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, 600bp), excluding positions with missing data and USS_{10} that were 400bp or less from positions with low input coverage (≤ 20 reads). Separation of ≥1000bp was found to give the best compromise between good peak separation and the number of USS_{10}s or $USS_{9.5}$s meeting the separation criterion (n=237 and n = 209 respectively). To assess USS peak centrality and symmetry, we used the sequences of 158 isolated USS_{10}s, that were at least 1000bp from the nearest $USS_{9.5}$ and had uptake ratios ≥3. These sequences were aligned at position 16 of their USS after reverse-complementing those with reverse-orientation USSs, and the mean and standard deviation of uptake ratios at each position was calculated out to 100bp on either side of the USS. Differences between the left and right sides were assessed with a Student's t-test ($P > 0.05$).

Contributions of weak USSs: To look for weak uptake effects in the valleys between USS-associated uptake-ratio peaks, a 'far from USS_{10}' subset of NP positions was created, consisting of positions that were at least 0.6kb from the closest USS_{10}. This gave 575 segments summing to 29% of the genome. Each of these regions was searched for positions with uptake ratios > 0.2. Uptake maps containing positions with uptake ratios > 0.2 were plotted, including flanking positions out to 2kb, to identify effects of USS with scores < 10. Uptake ratios at all the $USS_{9.5\text{-}10}$s in the 'far from USS_{10}' dataset (n=99) were then examined. Mean uptake of each $USS_{9.5\text{-}10}$was

calculated and a boxplot was built grouping $USS_{9.5\text{-}10}$s by score. Significance of differences between the score groups was evaluated using a t-test. A two-proportions power analysis was used to measure the effect that could be detected with the current number of $USS_{9.5\text{-}9.6}$ and $USS_{9.6\text{-}9.7}$, using R package pwr v. 1.2-2.

Incorporating within-USS interaction effects into uptake predictions: Figure 6 of Mell et al. (19) shows the strength and direction of pairwise interaction effects between USS positions. 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 NP $USS_{9.5}$ 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: Noise-free uptake data for short and long fragments was simulated by raw input coverage data for NP-short (sample UP07) and NP-long (sample UP03) that had been smoothed using a LOESS regression and normalized to a mean coverage of 1.0. Amplitude of three types of noise ('white', 'pink', and 'red') were generated for every genome position using the 'tuneR' R-package (56).

The determine the level of noise to be added, for each genomic position, we first calculated the difference in normalized coverage (depth per millon reads) of each replicate from the mean of the three replicates, grouping the normalized coverage-differences by the normalized mean coverage at the 3 replicates that was used in the subtraction before. Next, we calculated the maximum coverage-differences according to each mean normalized coverage value. This number was multiplied by the simulated amplitude of red noise and by 1, 1.5, 2, 2.5 or 3 to estimate the level of noise to be added to each position according to its coverage. The most appropriate type

713 of noise was identified by examining the autocorrelations of simulated coverages after noise was
714 added. Adding red noise to each position at levels proportional to the observed coverage-
715 dependent variation gave an autocorrelation of 0.999, identical to that of the real data.

716 **Data availability:** All short read data have been deposited at NCBI under BioProject
717 PRJNA387591 and BioSamples are listed in Table S2. The PacBio-sequenced PittGG genome
718 reference was deposited into Genbank under SRA number SRR10207558. Full calculations,
719 processed datasets, and Rscripts available at: https://github.com/mamora/DNA_uptake.

720 ## Ethics approval and consent to participate

721 Not applicable

722 ## Consent for publication

723 Not applicable

724 ## Competing interests

725 The authors declare that they have no competing interests

726 ## Funding

727 The study was funded by a NSERC Discovery Grant and by an NIH grant to GDE (5R01DC002148-
728 21) and support from the Drexel University Center for Genomic Sciences.
729

730 ## Authors' contributions

731 RR and JCM conceived the study. MM and RR wrote the manuscript and performed the bioinformatic
732 analysis. MM did the DNA uptake experiments. JCM did the library preparation, sequencing and
733 sequence alignments. GE and RE did the genome assembly of the PittGG genome.

734 ## Acknowledgements

The authors would like to thank Dr. Rachel Simister for her assistance with the bioanalyzer fragment analysis and Dr. Matthew Pennell for the statistical advice.

Bibliography:

1. Lorenz MG, Wackernagel W. Bacterial gene transfer by natural genetic transformation in the environment. Microbiol Rev. 1994;58(3):563–602.

2. Bae J, Oh E, Jeon B. Enhanced transmission of antibiotic resistance in Campylobacter jejuni biofilms by natural transformation. Antimicrob Agents Chemother. 2014;58(12):7573–5.

3. Mell JC, Viadas C, Moleres J, Sinha S, Fernández-Calvet A, Porsch EA, et al. Transformed recombinant enrichment profiling rapidly Identifies HMW1 as an intracellular invasion locus in Haemophilus influenza. PLoS Pathog. 2016;12(4):e1005576.

4. Straume D, Stamsås GA, Håvarstein LS. Natural transformation and genome evolution in Streptococcus pneumoniae. Infect Genet Evol. 2015;33(1432):371–80.

5. Kress-Bennett JM, Hiller NL, Eutsey RA, Powell E, Longwell J, Hillman T, et al. Identification and characterization of msf , a novel virulence factor in Haemophilus influenzae. PLoS One. 2016;11(3):e0149891.

6. Chen I, Dubnau D. DNA uptake during bacterial transformation. Nat Rev Microbiol. 2004;2(3):241–9.

7. Scocca JJ, Poland RL, Zoon KC. Specificity in deoxyribonucleic acid uptake by transformable Haemophilus influenzae. J Bacteriol. 1974;118(2):369–73.

8. Dougherty TJ, Asmus A, Tomasz A. Specificity of DNA uptake in genetic transformation of gonococci. Biochem Biophys Res Commun. 1979;86(1):97–104.

9. Smith HO, Tomb JF, Dougherty BA, Fleischmann RD, Venter JC. Frequency and distribution

of DNA uptake signal sequences in the Haemophilus influenzae Rd genome. Science (80-). 1995;269(5223):538–40.

10. Barany F, Kahn ME, Smith HO. Directional transport and integration of donor DNA in Haemophilus influenzae transformation. Proc Natl Acad Sci U S A. 1983;80(23):7274–8.

11. Salzer R, Kern T, Joos F, Averhoff B. The Thermus thermophilus comEA / comEC operon is associated with DNA binding and regulation of the DNA translocator and type IV pili. Environ Microbiol. 2016;18(1):65–74.

12. Hepp C, Maier B. Kinetics of DNA uptake during transformation provide evidence for a translocation ratchet mechanism. Proc Natl Acad Sci U S A. 2016;113(44):12467–72.

13. Pifer ML, Smith HO. Processing of donor DNA during Haemophilus influenzae transformation: analysis using a model plasmid system. Proc Natl Acad Sci U S A. 1985;82(11):3731–5.

14. Sisco KL, Smith HO. Sequence-specific DNA uptake in Haemophilus transformation. Proc Natl Acad Sci U S A. 1979;76(2):972–6.

15. Danner DB, Deich R a, Sisco KL, Smith HO. An eleven-base-pair sequence determines the specificity of DNA uptake in Haemophilus transformation. Gene. 1980;11:311–8.

16. Danner DB, Smith HO, Narang SA. Construction of DNA recognition sites active in Haemophilus transformation. Proc Natl Acad Sci U S A. 1982;79:2393–7.

17. Smith HO, Gwinn ML, Salzberg SL. DNA uptake signal sequences in naturally transformable bacteria. Res Microbiol. 1999 Nov;150(9–10):603–16.

18. Maughan H, Redfield RJ. Extensive variation in natural competence in haemophilus influenzae. Evolution (N Y). 2009;63(7):1852–66.

779 19. Mell JC, Hall IM, Redfield RJ. Defining the DNA uptake specificity of naturally competent
780 Haemophilus influenzae cells. Nucleic Acids Res. 2012;40(17):8536–49.

781 20. Mathis LS, Scocca JJ. Recognize different specificity determinants in the DNA uptake step of
782 genetic transformation. J Gen Microbiol. 1982;128:1159–61.

783 21. Frye SA, Nilsen M, Tønjum T, Ambur OH. Dialects of the DNA uptake sequence in
784 Neisseriaceae. PLOS Genet. 2013;9(4):e1003458.

785 22. Redfield RJ, Findlay WA, Bossé J, Kroll JS, Cameron ADS, Nash JHE. Evolution of competence
786 and DNA uptake specificity in the Pasteurellaceae. BMC Evol Biol. 2006;6:1–15.

787 23. Findlay WA, Redfield RJ. Coevolution of DNA uptake sequences and bacterial proteomes.
788 Genome Biol Evol. 2009;1:45–55.

789 24. Maughan H, Wilson LA, Redfield RJ. Bacterial DNA uptake sequences can accumulate by
790 molecular drive alone. Genetics. 2010;186(2):613–27.

791 25. Man WH, De Steenhuijsen Piters WAA, Bogaert D. The microbiota of the respiratory tract:
792 Gatekeeper to respiratory health. Nat Rev Microbiol. 2017;15(5):259–70.

793 26. Lethem M, James SL, Marriott C, Burke JF. The origin of DNA associated with mucus
794 glycoproteins in cystic fibrosis sputum. Eur Respir J. 1990;3:19–23.

795 27. Shak S, Capon DJ, Hellmiss R, Marsters SA, Baker CL. Recombinant human DNase I reduces
796 the viscosity of cystic fibrosis sputum. Proc Natl Acad Sci U S A. 1990;87:9188–92.

797 28. de Vries J, Meier P, Wackernagel W. The natural transformation of the soil bacteria
798 Pseudomonas stutzeri and Acinetobacter sp. by transgenic plant DNA strictly depends on
799 homologous sequences in the recipient cells. FEMS Microbiol Lett. 2001;195(2):211–5.

800 29. Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS. Origins of specificity in protein-DNA

recognition. Annu Rev Biochem. 2010;79:233–69.

30. Halford SE, Marko JF. How do site-specific DNA-binding proteins find their targets? Nucleic Acids Res. 2004;32(10):3040–52.

31. Hogg JS, Hu FZ, Janto B, Boissy R, Hayes J, Keefe R, et al. 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. 2007;8:R103.

32. Kahn ME, Barany F, Smith HO. Transformasomes: specialized membranous structures that protect DNA during Haemophilus transformation. Proc Natl Acad Sci U S A. 1983;80(22):6927–31.

33. Barouki R, Smith HO. Reexamination of phenotypic defects in rec-1 and rec-2 mutants of Haemophilus influenzae Rd. J Bacteriol. 1985;163(2):629–34.

34. Harrison A, Dyer DW, Gillaspy A, Ray WC, Mungur R, Carson MB, et al. Genomic sequence of an otitis media isolate of nontypeable Haemophilus influenzae: Comparative study with H. influenzae serotype d, strain KW20. J Bacteriol. 2005;187(13):4627–36.

35. Mrazek J. Comparative analysis of sequence periodicity among prokaryotic genomes points to differences in nucleoid structure and a relationship to gene expression. J Bacteriol. 2010;192(14):3763–72.

36. Rohs R, West SM, Sosinsky A, Liu P, Mann RS, Honig B. The role of DNA shape in protein-DNA recognition. Nature. 2009;461(7268):1248–53.

37. Ambur OH, Frye SA, Tønjum T. New functional identity for the DNA uptake sequence in transformation and its presence in transcriptional terminators. J Bacteriol. 2007;189(5):2077–85.

823 38. Goodman SD, Scocca JJ. Factors influencing the specific interaction of Neisseria
824 gonorrhoeae with transforming DNA. J Bacteriol. 1991;173(18):5921–3.

825 39. Kingsford CL, Ayanbule K, Salzberg SL. Rapid, accurate, computational discovery of Rho-
826 independent transcription terminators illuminates their relationship to DNA uptake.
827 Genome Biol. 2007;8(2):1–12.

828 40. Babenko VN, Chadaeva I V, Orlov YL. Genomic landscape of CpG rich elements in human.
829 BMC Evol Biol. 2017;17:19.

830 41. Bird AP. CpG-rich islands and the function of DNA methylation. Nature. 1986;321:209–13.

831 42. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput
832 sequencing. Nucleic Acids Res. 2012;40:e72.

833 43. Teng M, Irizarry RA. Accounting for GC-content bias reduces systematic errors and batch
834 effects in ChIP-seq data. Genome Res. 2017;27:1930–8.

835 44. Love MI, Hogenesch JB, Irizarry RA. Modeling of RNA-seq fragment sequencing bias
836 reduces systematic errors in transcript abundance estimation. Nat Biotechnol.
837 2017;34:1287–91.

838 45. Harteis S, Schneider S. Making the bend: DNA tertiary structure and protein-DNA
839 interactions. Int J Mol Sci. 2014;15(7):12335–63.

840 46. Stella S, Cascio D, Johnson RC. The shape of the DNA minor groove directs binding by the
841 DNA-bending protein Fis. Genes Dev. 2010;24(8):814–26.

842 47. Li J, Sagendorf JM, Chiu TP, Pasi M, Perez A, Rohs R. Expanding the repertoire of DNA shape
843 features for genome-scale studies of transcription factor binding. Nucleic Acids Res.
844 2017;45(22):12877–87.

845 48. Davidsen T, Rødland EA, Lagesen K, Seeberg E, Rognes T, Tønjum T. Biased distribution of
846 DNA uptake sequences towards genome maintenance genes. Nucleic Acids Res.
847 2004;32(3):1050–8.

848 49. Brockman KL, Azzari PN, Taylor Branstool M, Atack JM, Schulz BL, Jen FE-C, et al.
849 Epigenetic regulation alters biofilm architecture and composition in multiple clinical
850 isolates of Nontypeable Haemophilus influenzae. MBio. 2018;9(5):e01682-18.

851 50. Poje G, Redfield RJ. General methods for culturing Haemophilus influenzae . In: Herbert M,
852 editor. Methods in Molecular Medicine, Haemophilus influenzae Protocols. Totowa, NJ:
853 Humana Press Inc.; 2003. p. 2–5.

854 51. Poje G, Redfield RJ. Transformation of Haemophilus influenzae. In: Herbert M, editor.
855 Methods in Molecular Medicine, Haemophilus influenzae Protocols. Totowa, NJ: Humana
856 Press Inc.; 2003. p. 57–70.

857 52. Sambrook J. Molecular cloning : a laboratory manual. Third edition. N.Y.: Cold Spring
858 Harbor Laboratory Press; 2001.

859 53. Hunt M, Silva N De, Otto TD, Parkhill J, Keane JA, Harris SR. Circlator : automated
860 circularization of genome assemblies using long sequencing reads. Genome Biol.
861 2015;16:294.

862 54. Mell JC, Shumilina S, Hall IM, Redfield RJ. Transformation of natural genetic variation into
863 Haemophilus Influenzae genomes. PLoS Pathog. 2011;7(7):e1002151.

864 55. Caglar MU, Teufel AI, Wilke CO. Sicegar: R package for sigmoidal and double-sigmoidal
865 curve fitting. PeerJ. 2018;6:e4251.

866 56. Ligges U, Krey S, Mersmann O, Schnackenberg S. TuneR: Analysis of music and speech.

867 2018. Available from: https://cran.r-project.org/package=tuneR

868

869

870 Figure legends

871 **Figure 1. A.** USS sequence logo based on the DNA-uptake position weight matrix from Mell et al.
872 (19) uptake bias sequence logo. **B.** Conserved USS segments

873 **Figure 2. A.** Components of the DNA uptake model (see Methods for details). **B. & C.** Model I
874 predictions for uptake centered at a 12 bit USS for: **B**. 100, 200, and 300 bp fragments, **C**. a mixed
875 distribution of fragments between 25-300 bp with and without baseline uptake. **D. E. & F.** Model
876 predictions for uptake of a 3000 bp region with 3 USSs (black squares, scores in red) using
877 different fragment-length distributions: **D**. 50-300 bp fragments, **E.** 50-2000 bp fragment, **F.** 1-14
878 kb fragments. **G. & H.** Predicted DNA uptake of a 50 kb segment using different fragment-length
879 distributions: **G**. NP-short fragment length distribution, **H.** NP-long fragment length distribution.
880 **I.** Locations and scores of USS_{10}s in this 50 kb segment.

881 **Figure 3.** Local uptake ratios (smoothed over 31 bp) for the same 50 kb segment of the NP
882 genome as Fig. 2G & H. Grey points indicate positions with input coverage lower than 20 reads.
883 Gaps indicate unmappable positions. **A.** Uptake ratios of short-fragment DNA. **Inset:** Same data
884 with a logarithmic-scale Y-axis. **B.** Uptake ratios of long-fragment DNA. **C.** Locations and scores
885 of USS_{10}s.

886 **Figure 4:** Predicted and observed DNA uptake analysis for different model versions. **A.** and **C.**
887 Blue lines show the same uptake ratio maps as in Fig. 3A. **A.** Orange line shows the same
888 predicted uptake as in Fig. 2G, using model I settings (baseline binding and uptake p=0.2, linear

uptake function). **C.** Orange line shows predicted uptake using model II settings (baseline binding and uptake p=0.02, sigmoidal uptake function). **B**. Relationship between USS score and uptake ratio peak height in NP-short dataset for isolated $USS_{9.5}$s (black points, N= 209 USSs separated by at least 1000 bp), and uptake functions used to predict uptake. **Blue line**, linear uptake function used in the model I; **orange line**, sigmoidal uptake function used in the model II.

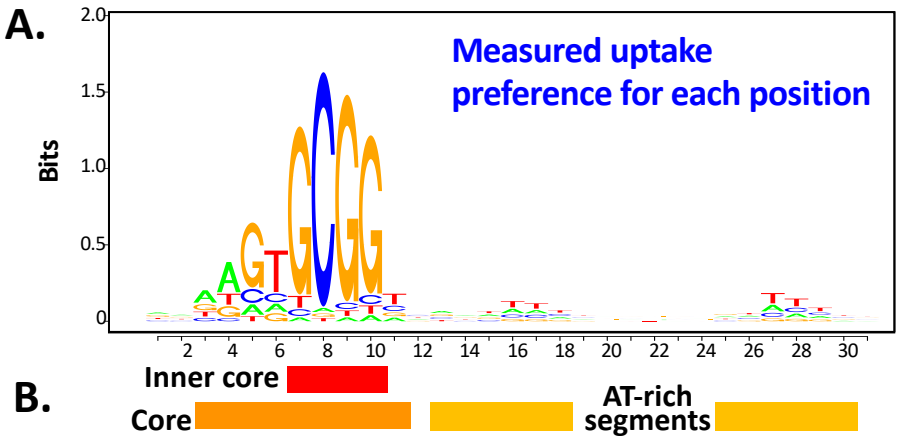
Figure 5. Predicted shape features of USS with strong and weak peaks. Thick grey lines: shape analysis of consensus USS sequence. Blue and orange lines: shape analysis of genomic USS separated by at least 500 bp, grouped by uptake ratio. **A-D:** $USS_{10.0\text{-}10.5}$. Blue: USS with weak peaks (uptake ratios <0.6, n=47, mean score=10.22). Orange: USS with strong peaks (uptake ratios >2.0, n=10, mean score=10.26). **E-H:** $USS_{10.5\text{-}11.0}$. Blue: USS with weak peaks (uptake ratios <0.6, n=14, mean score=10.64). Orange: USS with strong peaks (uptake ratios >2.0, n=59, mean score=10.79). **I-L** DNA shape of the $USS_{9.5\text{-}10}$ with uptake higher (red, n = 10) and lower (blue n = 68) than 0.2. **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. Coloured bars indicate components of the USS (see Figure 1): light orange, outer core; dark orange, inner core; green, AT tracts.

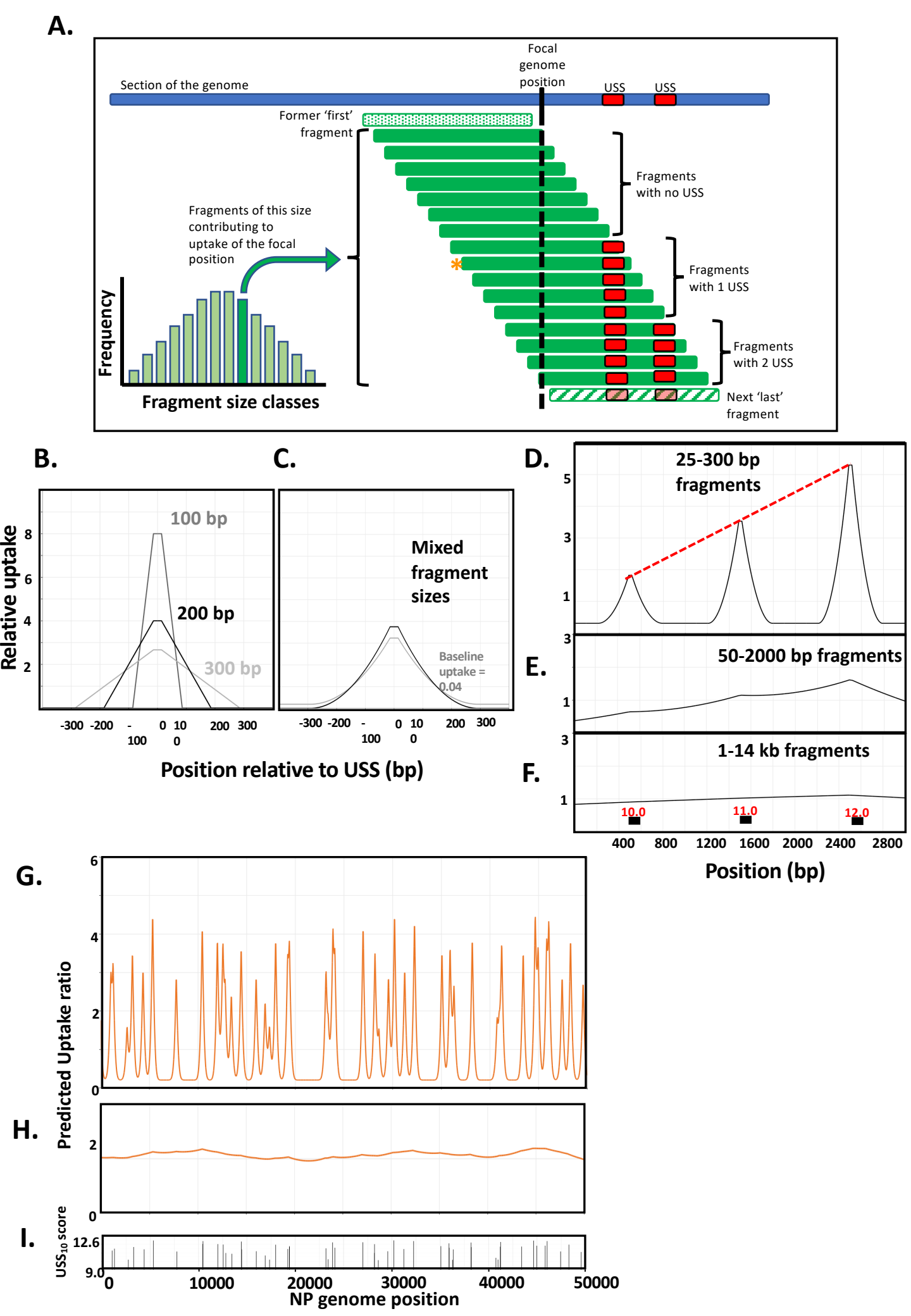
Supplementary information

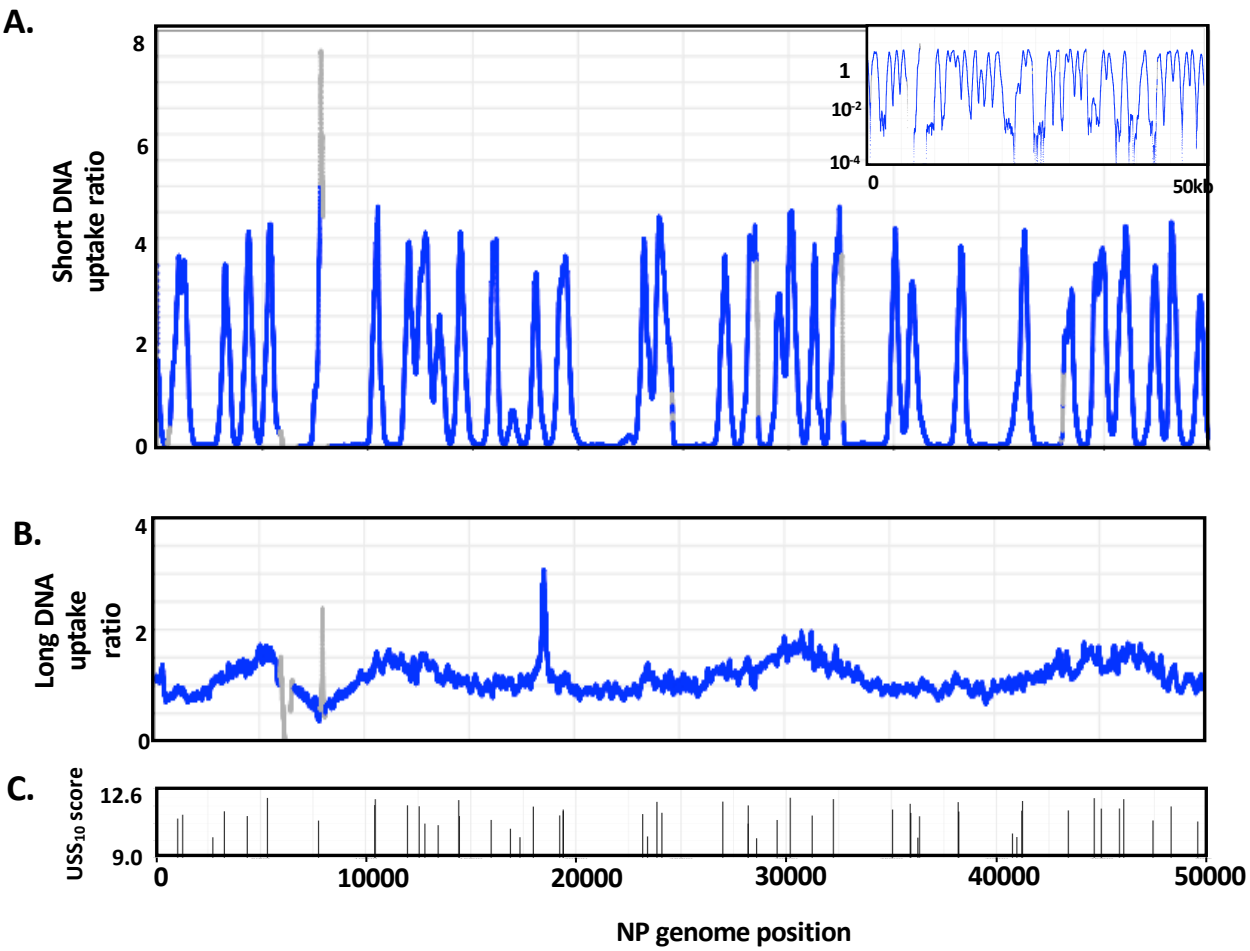
Additional file 1: Supp. Figure 1. Frequency distribution of USS scores for all positions in the NP, GG, and Rd genomes and for four random-sequence genomes with the same base composition. **Supp. Figure 2:** Distributions of DNA fragment lengths. **Supp. Figure 3.** Local uptake ratio maps. **Supp. Figure 4.** Sources of variation in read coverage. Read coverage maps for NP long-fragment samples over a 50 kb segment of the genome. **Supp. Figure 5.** Tests of periodicity by Fourier-transform analyses performed with R-package RCA. **Supp. Figure 6.** Frequencies of uptake ratios below 0.1. **Supp. Figure 7.** Symmetry and centrality of uptake peaks. **Supp. Figure 8.** Effects of interactions between USS positions on USS scores. **Supp. Figure 9.** Uptake ratios at weak USSs. **Supp. Figure 10.** Analysis of DNA uptake effects of USS_{10} pairs. **Supp. Figure 11.** Predicted and observed uptake of long NP DNA fragments. **Supp. Figure 12:** Distributions of long fragment input (blue) and recovered (purple) DNA fragment lengths. **Supp. Figure 13.** Uptake of long NP DNA fragments as a function of local USS_{10} density. **Supp.**

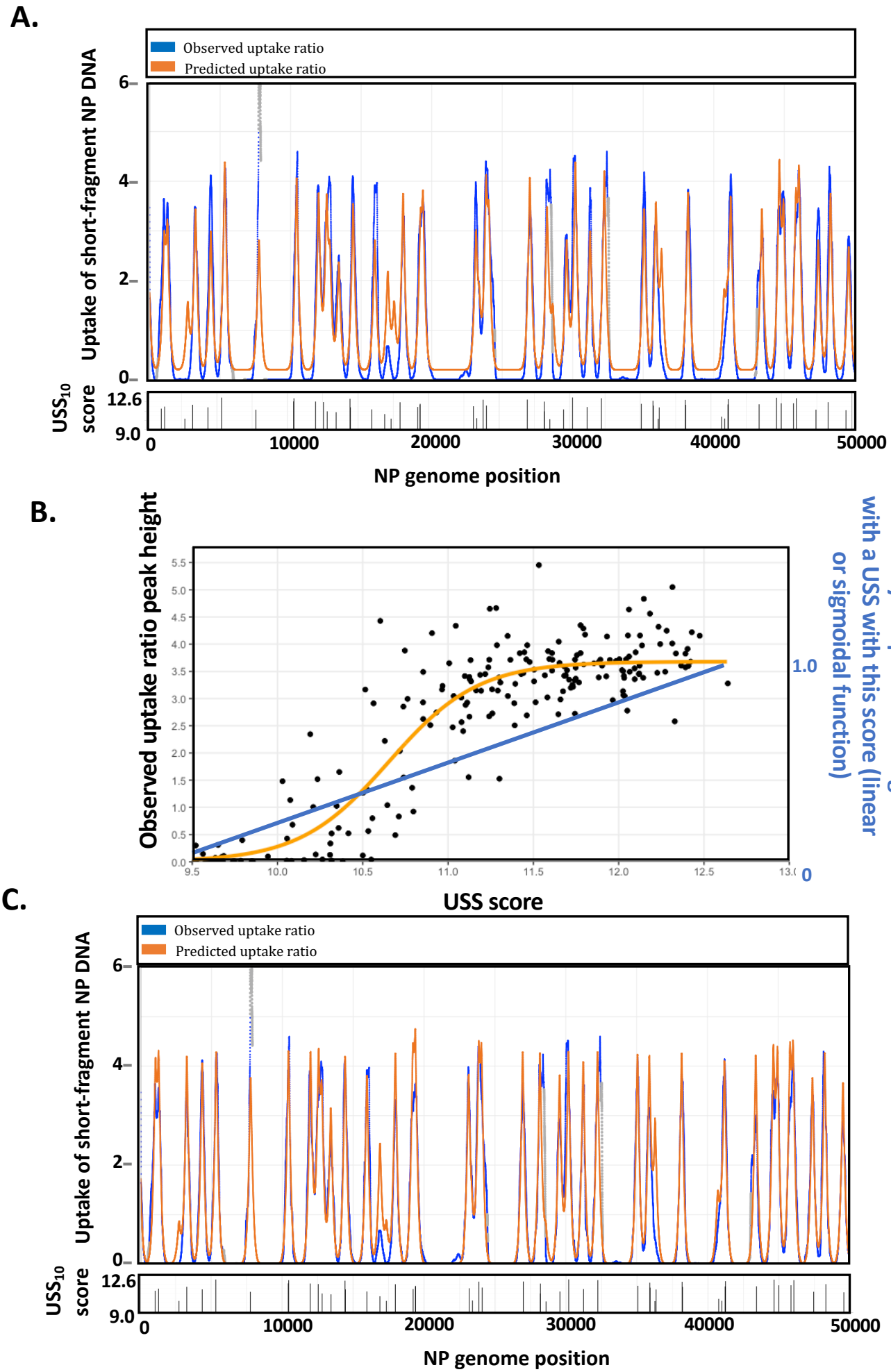
Figure 14. Correlation coefficient between simulated signal with and without different levels of noise. **Supp. Figure 15.** Predicted and observed uptake of GG short and long fragments. **Supp. Figure 16.** Uptake ratio of NP and GG positions with input coverage higher or lower than 20 reads

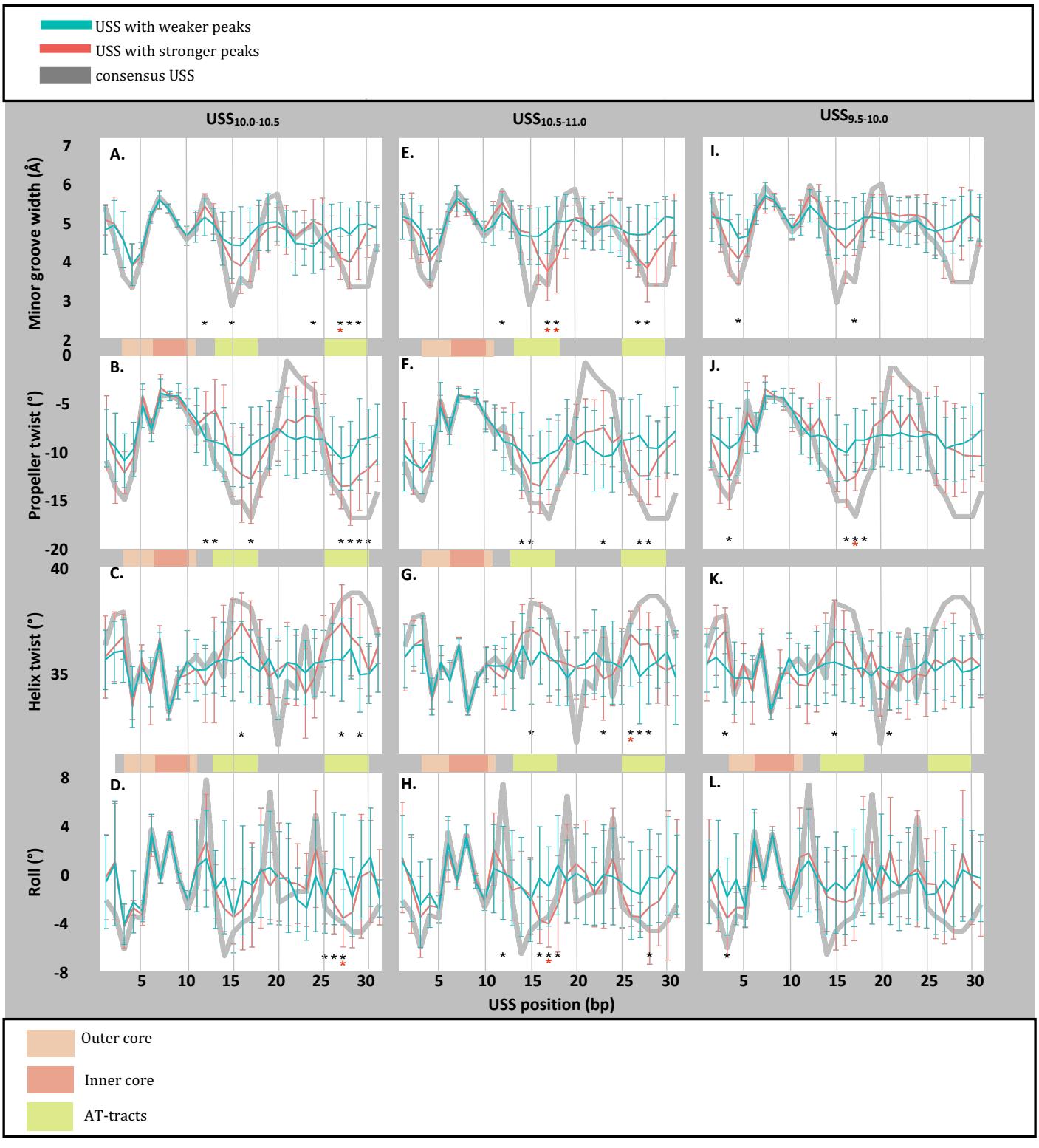
Additional file 2: TableS1 uptake-prediction position-specific scoring matrix from Mell *et al.*'s degenerate-sequence uptake experiment. **Table S2** Detailed sequencing information about all samples. **Table S3** Proportion of positions in NP and GG with high and low uptake that are biased towards low input coverages. **Table S4** Summary of the three models and their parameters. **Table S5** Relative simulated uptake of human and NP 1kb and 10kb DNA fragments.




A.
Focal genome position
Section of the genome
USS
USS
Former 'first' fragment
Fragments with no USS
Fragments of this size contributing to uptake of the focal position
Fragments with 1 USS
Fragments with 2 USS
Frequency
Fragment size classes
Next 'last' fragment
B.
C.
D.
Relative uptake
100 bp
200 bp
300 bp
Mixed fragment sizes
Baseline uptake = 0.04
25-300 bp fragments
E.
50-2000 bp fragments
F.
1-14 kb fragments
10.0
11.0
12.0
Position relative to USS (bp)
Position (bp)
G.
Predicted Uptake ratio
H.
I.
USS10 score
NP genome position


A.
Short DNA uptake ratio
8
6
4
2
0
1
10-2
10-4
0
50kb
B.
Long DNA uptake ratio
4
2
0
C.
USS10 score
12.6
9.0
0
10000
20000
30000
40000
50000
NP genome position


A.
Observed uptake ratio
Predicted uptake ratio
Uptake of short-fragment NP DNA
6
4
2
0
USS10 score
12.6
9.0
0
10000
20000
30000
40000
50000
NP genome position
B.
Observed uptake ratio peak height
Probability of uptake of a fragment with a USS with this score (linear or sigmoidal function)
1.0
0
USS score
C.
Observed uptake ratio
Predicted uptake ratio
Uptake of short-fragment NP DNA
USS10 score
NP genome position


USS with weaker peaks
USS with stronger peaks
consensus USS
USS10.0-10.5
USS10.5-11.0
USS9.5-10.0
A.
E.
I.
B.
F.
J.
C.
G.
K.
D.
H.
L.
Minor groove width (Å)
Propeller twist (°)
Helix twist (°)
Roll (°)
USS position (bp)
Outer core
Inner core
AT-tracts