Original Article

# A cross-species genome-wide analysis of sequences similar to those involved in DNA uptake bias in the Pastuerellaceae and Neisseriaceae families of pathogenic bacteria

M. Bakkali [*]

*Departamento de Genética, Facultad de Ciencias, Universidad de Granada, Granada, Spain*

## ABSTRACT

Acquiring new DNA allows the emergence of drug resistance in bacteria. Some Pasteurellaceae and Neisseriaceae species preferentially take up specific sequence tags. The study of such sequences is therefore relevant. They are over-represented in the genomes of the corresponding species. I found similar sequences to be present only in, but not in all, the genomes of the Pasteurellaceae and Neisseriaceae families. The genomic densities of these sequences are different both between species and between families. Interestingly, the family whose genomes harbor more of such sequences also shows more sequence types. A phylogenetic analysis allowed inferring the possible ancestral Neisseriacean sequence and a nucleotide-by-nucleotide analysis allowed inferring the potential ancestral Pasteurellacean sequence based on its genomic footprint. The method used for this work could be applied to other sequences, including transcription factor binding and repeated DNAs.

## 1. Introduction

DNA acquisition by bacteria has not only significant clinical, epidemiological and economical implications, it is also of interest to different research areas including research on bacterial acquisition of drug resistance and virulence, recombination, genome dynamics and evolution, speciation, *etc.* In fact, when it was discovered by Griffith in 1928 [1], bacterial ability to acquire DNA provided the first proof of the physical nature of the hereditary material. Bacteria can acquire new DNA *via* three mechanisms. Conjugation and transduction result in DNA entry into new bacterial cells thanks to the transfer of plasmids between cells and to the infection by bacteriophages, respectively. The third mechanism, competence for transformation, relays on uptake of extracellular DNA by the bacterial cell itself. Transforming competent bacteria in laboratories is usually artificial and requires the use of chemical (*e.g.*, calcium chloride) or physical treatments (heat choc, electric pulses)—the treated bacteria are thus artificially made competent for transformation. However, many bacterial species are naturally competent for spontaneous transformation after uptake of extracellular DNA (for a review see [2]). Some of these bacteria seem constitutively competent, whereas others are competent only when certain conditions (generally stressful) are met.

Among the naturally competent bacteria for transformation, species from the Pasteurellaceae and Neisseriaceae families are special since they exhibit bias in DNA uptake. They preferentially take up DNA that contains specific short sequences (tags) called Uptake Signal Sequences for the Pasteurellaceae and DNA Uptake Sequences for the Neisseriaceae. Following the logic explained in [3] we will refer to both of these sequences using the unifying name of DNA Uptake Enhancing Sequences (DUESs). The protein that possibly binds the extracellular DNA has been pinpointed relatively recently [4,5] but the mechanistic reason for the bias towards the DUESs is still matter of ongoing research. Still, we know that the preferential uptake of DNA fragments that contain the preferred DUES leads to accumulation of these sequences in the genome which, in turn, results in preferential uptake of those DNAs from conspecific cells (see [3,6]).

The Pasteurellacean and Neisseriacean DUESs so play a crucial role in the horizontal movement of DNAs between competent cells, strains and species of these two important, pathogen-containing, bacterial families. DUESs therefore have a fundamental impact on all the practical and scientific aspects relating to the acquisition of new genetic materials by these bacteria (among which acquisition of resistance to antibiotics and virulence are to highlight). Similarity in DUESs would result in an ease of uptake of the DNA across competent cells, strains and species whereas differences between DUESs would form partial or even full barriers that would hamper cross-species DNA uptake and horizontal transfer of DNAs between species. Determining the nature of the DUESs could therefore be of importance to the prediction and planning of strategies to deal with the exchange of genetic material between species of these bacterial families. Surprisingly, the real nature of these

---

* Corresponding author.
*E-mail address:* mbakkali@ugr.es.

sequences is not fully studied and discrepancies exist between works. The Neisseria DUES was suggested to be the 10 bp sequence GCCGTCTGAA [7,8] until 2007, when I suggested that the 12 bp sequence ATGCCGTCTGAA seems more likely to be the DUES of *N. gonorrhoeae* 3. Further works [9–11] confirmed that suggestion and, in 2013, other 12 pb DUES-like sequences were identified in other Neisseriaceae species [12]. When it comes to the Pasteurellacean DUES, it was first suggested to be the 11 bp sequence AAGTGCGGTCA [13]. The data later seemed to suggest that only a 9 bp sequence is needed for efficient DNA uptake [14] and, when the *Haemophilus influenza* genome was sequenced—becoming the first free living being genome to be sequenced [15]—the genomic data seemed to confirm that the *H. influenza* DUES is the 9 pb sequence AAGTGCGGT [16–18]. A latter work added the 9 bp sequence ACAAGCGGT as another Pasteurellacean DUES type [19] and, a year later, I suggested that the main Pasteurellacean DUES sequence is probably the 10 bp AAAGTGCGGT [3]. The panorama is therefore still confusing and semi-circularity and inadequacy of the analyses logic and methods might be hiding the real nature of the sequences that naturally competent Neisseriaceae and/or Pasteurellaceae species actually prefer for taking up. To experimentally determine the nature of the preferred sequence would mean tedious, time consuming, to not say little rewarding, combinations of mutagenesis and DNA uptake testings of sequences of different sizes and compositions. It would also face the problem of the currently still uncultivable species and of species whose requirements for becoming competent are not known yet. The availability of genomic sequences and the affordability of adequate computing capacities, mean that a computer-based approach to find out the real DUES should be more accessible than the 'wet-lab' approach—especially if a multi-species analysis is planned. The key issue though should be to have well designed (see innovative) strategies that are based on what we experimentally know about the core sequence and about the function and consequences of the DUES, while minimizing but not abolishing biased guiding of the search target sequences and methods towards the already known DUESs.

Here, I start by screening for the presence of the DUES-like sequences in complete and unfinished genome sequences of the Pasteurellales and Neisseriales species available in Genebank database. This allowed me to determine the 10 or 12 bp potential DUES in each genome based on the logic that it must be the most over-represented DUES-like sequence in the corresponding genome. In a further step, and for each type of potential 10 or 12 bp DUES, I determine what should be the corresponding real DUES based on the logic that it should be the most over-represented among the sequences that either contain or are contained by the potential DUES previously determined for the respective species. I later establish the phylogenetic relationships between the different DUESs and I infer their ancestral sequences.

## 2. Material and methods

Fig. 1 schematizes the main steps of the workflow followed for this work. These were:

### 2.1. Obtaining the genome sequences

The complete and draft genome sequences of the species used in this
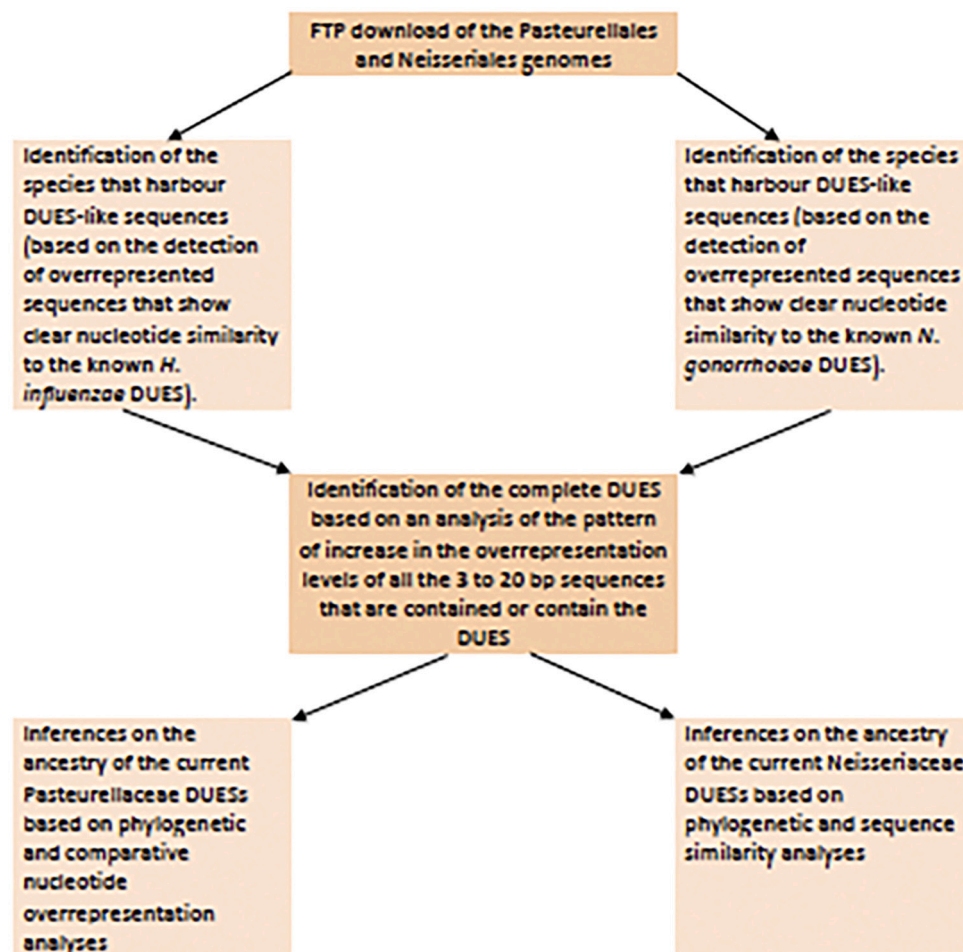


**Fig. 1.** Schematic representation of the main steps of the workflow leading to the identification and study of the Pasteurellaceae and Neisseriaceae DUESs.

work were downloaded from Genebank's ftp site.

## 2.2. Counts of the DUES-like sequences

Each complete and draft genomic sequence was searched using a 10 or 12 bp sliding window approach and the computer script *Sequence_Extractor.pl* from [3]. The number of observed sequences in each genome was compared to the expected number of the same sequence in the same genome in order to detect the most over-represented DUES-like sequence in the studied genome. The expected number of each given sequence in a genome was calculated the formula used in previous works (*e.g.* Bakkali 2007). That is $2 L(A)^a(T)^t(C)^c(G)^g$, where, for each nucleotide, the capital letter is the genomic proportion of that nucleotide, the super-indexed letter is the number of that nucleotide in the analyzed (potential DUES) sequence and L is the genome's length in bp. The most over-represented DUES-like sequence in each genome was determined based on the value of the Chi-squared in the comparison between the observed and expected numbers of the sequence in question.

## 2.3. Identification of the real DUES

Identification of the real DUES was based on identifying the most over-represented sequence among those that either contain or are contained by the most over-represented 10 or 12 bp sequence in the respective genome. The value of the Chi-squared in the comparison between the observed and expected numbers was used as quantifier of the over-representation. For that, sliding window screening for all the 3 to 20 bases in each of the studied Pasteurellacean and Neisseriacean DUES-containing genomes was carried out. The expected and observed numbers of each 3 to 20 bases sequence that contains or is contained in the corresponding DUES-like sequence were used in order to estimate the over-representation of each potential DUES at each size level.

## 2.4. Phylogenetics

In order to infer the ancestral DUES sequence I mapped the distribution of the different actual DUESs on the 16S-based species phylogeny and complemented the results with an analysis of the alignment of the different DUESs, in the case of the Neisseriaceae, and an analysis of the over-representation of nucleotides, in the case of the Pasteurellaceae.

For the species phylogenies, the sequences of all the 16S ribosomal DNAs of the different strains and species of the Pasteurellaceae and Neisseriaceae families were downloaded from the NCBI database and used for alignment and maximum likelihood phylogenetic tree building using MAFT [20,21]. A similar tree was also built for the DUES-having species. In both cases the 16S ribosomal DNA from *E. coli* was used as outgroup. MEGA [22] was used for tree drawing and editing.

Given their short size and small number, the alignment of the different DUESs was made and edited using BioEdit [23]. The analysis of the nucleotide over-representation for the pasteurellacean DUESs was more laborious. It was carried out in two different ways. In one way *Sequences_Extractor.pl* script from [3] was used to separately extract from *H. influenza* and *A. pleuromoneae* genomes all the 50 bp sequences containing the sequence GCGGT at position 21 (the most conserved nucleotides of the pasteurellacean DUES—see [3]). An over-representation score was calculated for each nucleotide at each position according to the formula EQ1 = *nucs/seqs*AT/2*, for As and Ts and EQ2 = *nucs/seqs*GC/2* for Gs and Cs, where *nucs* is the total number of the particular nucleotide in the position in question, *seqs* is the number of sequences and *AT* and *GC* are the A + T and G + C frequencies in the genome, respectively. The second way was to analyze nucleotide over-representation in the sets of all the 11 bp sequences that carry a mismatched DUES at any of the positions 1 to 10 (the 11th, extra, DUES position was at the 3′ end of the sequence). I hence used the *Sequences_Extractor.pl* script from [3] and separately extracted from *H. influenza* and *A. pleuromoneae* genomes all the 11 bp sequences containing a

mismatched DUES. For this latter analysis I first calculated the expected number of sequences at each mismatch level using Eq. (1) in [3], henceforth EQ3. Then, the number of expected nucleotides at each position was calculated for As and Ts and Gs and Cs both when they were as matches or as mismatches both in AT and GC positions of the DUES. For that, I calculated the number of sequence categories at each mismatch level as EQ4 = *(n)!/(m!(n-m)!)*. I then calculated the expected times of matches per position of the DUES at each mismatch level EQ5 = *((n-1)!/(m!(abs((n-1)-m))!))*. The expected number of A or T matches in each A or T position was then calculated as EQ6 = *EQ3(EQ5/EQ4)(1 + ((AT-GC)/4))*. The expected number of G or C matches in each G or C position was EQ7 = *EQ3(EQ5/EQ4)(1-((AT-GC)/4))*. The expected number of A or T mismatches in each A or T position of the DUES at each mismatch level was calculated as EQ8 = *(EQ3-EQ6)((AT/2)/(1-(AT/2)))*. The expected number of A or T mismatches in each G or C position was EQ9 = *(EQ3-EQ7)((AT/2)/(1-(GC/2)))*. The expected number of G or C mismatches in each A or T position was EQ10 = *(EQ3-EQ6)((GC/2)/(1-(AT/2)))*. Finally, the expected number of G or C mismatches in each G or C position was EQ11 = *(EQ1-EQ7) ((GC/2)/(1-(GC/2)))*.

An estimate of the over-representation of each nucleotide type at each position of the DUES at each mismatch load was based upon the equation EQ10 = *(obs*-exp*)/exp*, where *obs* and *exp* are the observed and expected numbers of the nucleotide in question at the particular position and mismatch level. The overall genomic over-representation of each nucleotide at each DUES position therefore is EQ12 = $\sum_{m=1}^{m=n} EQ10$.

## 3. Results

### 3.1. Distribution of the DUES-like sequences

For this analysis the k-mer sizes 10 and 12 bases were considered for the Pasteurellacean and Neisseriacean DUESs, respectively, as data from [3] seemed to indicate that these are the likely core sequence lengths of the DUESs in species from those two families.

The results (Table 1) show that while no DUES-like sequence was noticeably over-represented in the examined non-Pasteurellacean and non-Neisseriacean genomes, most but not all Pasteurellacean and Neisseriacean species do have DUESs.

Two 10 bp DUES-like variants could be identified among the Pasteurellacean genomes analyzed here. Their frequencies vary between species and their genomic densities ranged between 0.036% of the genome of *Haemophilus pittmaniae* HK 85 ctg1129913985426 and 0.338% of the *Aggregatibacter segnis* ATCC 33393 genome—about a 10 fold difference between the least and the most dense DUES-containing Pasteurellacean genomes (Table 1). There was no ambiguity as to the nature of the most over-represented DUES-like sequence in the Pasteurellacean genomes that contain the DUES type AAAGTGCGGT (henceforth called *H. influenza*-type DUES, in accordance with [19]). In the Pasteurellacean genomes that contain AACAAGCGGT (henceforth called *A. pleuromoneae*-type DUES1, in accordance with [19]), that DUES was consistently closely followed (in terms of frequency) by another sequence, ACAAGCGGTC (which I will henceforth call *A. pleuromoneae*-type DUES2)—the sequence of the DUES of these genomes thus being ambiguous. When compared to the *H. influenzae*-type DUES, the *A. pleuromoneae*-type DUES1 has the same number of differences (four) as does the *A. pleuromoneae*-type DUES2, parsimony hence does not allow discarding any of these sequences as the potential real DUES of the corresponding species.

For the Neisseriaceae species however, and in wide accordance with the data in Frye et al. (2013), nine different 12-bp DUES-like sequences were identified. However, there are two potential discrepancies between the present work and Frye et al.'s (2013) data: (i) one is the apparent presence of a DUES-like sequence in *Lutiella nitroferrum*—reported in the current work but not in Frye et al. 2013—and (ii) the second discrepancy is the fact that AGGCAGCCTGAA, reported as AG-kingDUS in Frye et al. 2013, is more frequent and over-represented in the genome of

**Table 1**
Sequence and over-representation of the core DUES-like sequences in the finished (F) and unfinished (U) genomes of the Pasteurellales species. No DUES-like sequence was detected in Pasteurellales species other than Pasteurellaceae. Obs.: Observed number of DUESs in the genomic sequence, Exp.: Expected number of DUESs in the genomic sequence, Percent: Percentage of the genomic sequence covered by DUESs.

| Species | Genome size in bp | | Sequence | Obs. | Exp. | Percent |
|---|---|---|---|---|---|---|
| *Avibacterium paragallinarum* 72 | 2,455,947 | U | AAAGTGCGGT | 590 | 4 | 0.240 |
| *Haemophilus influenzae* Rd KW20 | 1,830,138 | F | AAAGTGCGGT | 565 | 3 | 0.309 |
| *Pasteurella multocida* subsp. *multocida* str. Pm70 | 2,257,487 | F | AAAGTGCGGT | 345 | 4 | 0.153 |
| *Mannheimia succiniciproducens* MBEL55E | 2,314,078 | F | AAAGTGCGGT | 637 | 4 | 0.275 |
| *Actinobacillus succinogenes* 130Z | 2,319,663 | F | AAAGTGCGGT | 658 | 4 | 0.284 |
| *Haemophilus somnus* 2336 | 2,263,857 | F | AAAGTGCGGT | 436 | 3 | 0.193 |
| *Aggregatibacter aphrophilus* NJ8700 | 2,313,035 | F | AAAGTGCGGT | 740 | 4 | 0.320 |
| *Haemophilus parainfluenzae* T3T1 | 2,086,875 | F | AAAGTGCGGT | 541 | 3 | 0.259 |
| *Aggregatibacter actinomycetemcomitans* ANH9381 | 2,136,808 | F | AAAGTGCGGT | 697 | 4 | 0.326 |
| *Pasteurella dagmatis* ATCC 43325 | 2,246,530 | U | AAAGTGCGGT | 496 | 3 | 0.221 |
| *Aggregatibacter segnis* ATCC 33393 | 1,992,257 | U | AAAGTGCGGT | 674 | 3 | 0.338 |
| *Haemophilus aegyptius* ATCC 11116 | 1,915,203 | U | AAAGTGCGGT | 577 | 3 | 0.301 |
| *Haemophilus haemolyticus* M21621 M21621_028 | 2,089,344 | U | AAAGTGCGGT | 581 | 3 | 0.278 |
| *Haemophilus pittmaniae* HK 85 ctg1129913985426 | 2,182,748 | U | AAAGTGCGGT | 78 | 4 | 0.036 |
| *Haemophilus* sp. oral taxon 851 str. F0397 | 1,838,769 | U | AAAGTGCGGT | 537 | 3 | 0.292 |
| *Aggregatibacter segnis* ATCC 33393 | 2,012,140 | U | AAAGTGCGGT | 496 | 3 | 0.247 |
| *Haemophilus aegyptius* ATCC 11116 | 1,963,730 | U | AAAGTGCGGT | 577 | 3 | 0.294 |
| *Haemophilus* sp. oral taxon 851 str. F0397 | 1,837,756 | U | AAAGTGCGGT | 537 | 3 | 0.292 |
| *Gallibacterium anatis* UMN179 | 2,687,335 | F | AAAGTGCGGT | 47 | 4 | 0.018 |
| | | | CAGCACAGCA | 65 | 3 | 0.025 |
| *Haemophilus parasuis* 29,755 | 2,224,490 | U | AACAAGCGGT | 88 | 3 | 0.040 |
| | | | ACAAGCGGTC | 100 | 2 | 0.045 |
| *Haemophilus parasuis* SH0165 | 2,269,156 | F | AACAAGCGGT | 113 | 4 | 0.050 |
| | | | ACAAGCGGTC | 101 | 2 | 0.045 |
| *Actinobacillus pleuropneumoniae* serovar 5b str. L20 | 2,274,482 | F | AACAAGCGGT | 187 | 4 | 0.082 |
| | | | ACAAGCGGTC | 218 | 3 | 0.096 |
| *Haemophilus ducreyi* 35000HP | 1,698,955 | F | AACAAGCGGT | 41 | 2 | 0.024 |
| | | | ACAAGCGGTC | 40 | 1 | 0.024 |
| *Actinobacillus minor* NM305 | 2,425,565 | U | AACAAGCGGT | 128 | 4 | 0.053 |
| | | | ACAAGCGGTC | 118 | 2 | 0.049 |
| *Haemophilus* [*parainfluenzae*] CCUG 13788 | 2,048,368 | U | AACAAGCGGT | 155 | 3 | 0.076 |
| | | | ACAAGCGGTC | 197 | 2 | 0.096 |
| *Actinobacillus ureae* ATCC 25976 | 2,295,637 | U | AACAAGCGGT | 219 | 4 | 0.095 |
| | | | ACAAGCGGTC | 252 | 2 | 0.110 |
| *Mannheimia haemolytica* serotype A2 str. BOVINE | 2,479,009 | U | AACAAGCGGT | 165 | 4 | 0.067 |
| | | | ACAAGCGGTC | 248 | 3 | 0.100 |

**Table 2**
Sequence and over-representation of the core DUES-like sequences in the finished (F) and unfinished (U) genomes of the Neisseriales species. No DUES-like sequence was detected in Neisseriales species other than Neisseriaceae. Obs.: Observed number of DUESs in the genomic sequence, Exp.: Expected number of DUESs in the genomic sequence, Percent: Percentage of the genomic sequence covered by DUESs.

| Species | Genome size in bp | | Sequence | Obs. | Exp. | Percent |
|---|---|---|---|---|---|---|
| *Neisseria meningitidis* FAM18 | 2,194,961 | F | ATGCCGTCTGAA | 721 | 0 | 0.394 |
| *Neisseria gonorrhoeae* FA 1090 | 2,153,922 | F | ATGCCGTCTGAA | 791 | 0 | 0.441 |
| *Neisseria cinerea* ATCC 14685 | 1,872,930 | U | ATGCCGTCTGAA | 726 | 0 | 0.465 |
| *Neisseria polysaccharea* ATCC 43768 | 2,031,232 | U | ATGCCGTCTGAA | 874 | 0 | 0.516 |
| *Neisseria lactamica* 020–06 | 2,220,606 | F | ATGCCGTCTGAA | 838 | 0 | 0.453 |
| *Neisseria mucosa* ATCC 25996 | 2,578,935 | U | AGGTCGTCTGAA | 1375 | 0 | 0.640 |
| *Neisseria macacae* ATCC 33926 | 2,685,199 | U | AGGTCGTCTGAA | 1433 | 0 | 0.640 |
| *Neisseria sicca* ATCC 29256 | 2,830,951 | U | AGGTCGTCTGAA | 1450 | 0 | 0.615 |
| *Neisseria* sp. GT4A_CT1 | 2,695,613 | U | AGGTCGTCTGAA | 1368 | 0 | 0.609 |
| *Neisseria weaveri* LMG 5135 NW5135.2_46 | 2,183,521 | U | AGGCCGTCTGAA | 1227 | 0 | 0.674 |
| *Neisseria flavescens* SK114 | 2,204,257 | U | AGGCCGTCTGAA | 1161 | 0 | 0.632 |
| *Neisseria elongata* subsp. *glycolytica* ATCC 29315 | 2,276,246 | U | AGGCCGTCTGAA | 1392 | 0 | 0.734 |
| *Neisseria* sp. oral taxon 014 str. F0314 | 2,502,796 | U | AGGCCGTCTGAA | 1370 | 0 | 0.657 |
| *Neisseria bacilliformis* ATCC BAA-1200 | 2,434,956 | U | AGGCCGTCTGAA | 1969 | 0 | 0.970 |
| *Neisseria subflava* NJ9703 | 2,293,169 | U | AGGCCGTCTGAA | 1185 | 0 | 0.620 |
| *Neisseria shayeganii* 871 | 2,685,199 | U | AGGCTACCTGAA | 864 | 0 | 0.386 |
| *Eikenella corrodens* ATCC 23834 | 2,135,544 | U | AGGCTACCTGAA | 1419 | 0 | 0.797 |
| *Neisseria wadsworthii* 9715 | 2,406,728 | U | TGCCTGTCTGAA | 1056 | 0 | 0.527 |
| *Lutiella nitroferrum* | 4,228,906 | U | AGGTTGGCCGAA[a] | 223 | 1 | 0.063 |
| *Kingella oralis* ATCC 51147 | 2,406,795 | U | AGGCAGCCTGAA | 2126 | 0 | 1.060 |
| *Simonsiella muelleri* ATCC 29453 | 2,390,631 | U | AGGCAGCCTGAA | 924 | 0 | 0.464 |
| *Kingella denitrificans* ATCC 33394 | 2,186,763 | U | AAGCAGCCTGCA | 1356 | 0 | 0.744 |
| *Kingella kingae* ATCC 23330 | 1,916,963 | | AAGCAGCCTGCA | 1152 | 0 | 0.721 |
| *Pseudogulbenkiania* sp. NH8B | 4,332,995 | F | AGGTTGGCCGAA[a] | 250 | 1 | 0.069 |
| *Chromobacterium violaceum* ATCC 12472 | 4,751,080 | F | CAGCGCCAGCAG | 102 | 2 | 0.026 |
| *Laribacter hongkongensis* HLHK9 | 3,169,329 | F | TTTTATTAAACA | 7 | 0 | 0.003 |

[a] Both are the same species and the sequence does not resemble Neisseriaceae DUESs.

*Simonsiella muelleri* ATCC 29453 (924 sequences) than AGGCTGCCT-GAA, reported by Frye et al. (2013) as the DUES-like sequence in that species (AG-simDUS), and which I found in just 857 copies in the same genome. As in the case of the DUES-containing Pasteurellacean genomes, the densities of the DUES-like sequences were very different between Neisseriacean species and ranged between 0.063% of the *Lutiella nitroferrum* genome and 1.06% of the *Kingella oralis* ATCC 51147 genome—about a 20 fold difference between the least and most dense DUES-containing Neisseriacean genomes (Table 2).

From Fig. 2 one can see how, in both bacterial families, the overall results were similar between the finished and unfinished genomes used for the current work. Nonetheless, noticeable differences, in terms of genomic DUES densities, can be observed both between the two bacterial families and between the DUES types in each family. The genomes of the DUES-containing Neisseriaceae species seem to have more DUES density than the genomes of the DUES-containing Pasteurellaceae species, so that DUES-containing Neisseriaceae species, as a whole, have over 60 folds more over-representation of the DUES-like sequences than DUES-containing Pasteurellaceae species. The latter show higher densities for the *H. influenza*-type of DUES than for the *A. pleuromoneae*-types. In the case of the DUES-containing Neisseriaceae, the *N. gonorrhoeae*-type of DUES seems slightly less frequent than the other variants, being *Kingella oralis* ATCC 51147 the Neisseriacean bacterium whose genome has the most DUESs. The Pasteurellaceae species *Haemophilus pittmaniae* HK 85 ctg1129913985426 is also notorious for its very low DUES density.

### 3.2. Identification of the real DUES

Because of the preferential uptake of DNA fragments that contain the DUES (see the introduction section and the references therein), it was shown that natural competence for transformation would inevitably enrich the bacterial genome with the preferred DUES sequence [3,6].

The present analysis is based on the logical assumption that the real DUES would be the most over-represented sequence among the sequences contained in or containing the DUES-like sequence in a genome. I thus screened all the Pasteurellacean and Neisseriacean DUES-containing genomes for all the 3 to 20 bases sequences that contain or are contained in the corresponding DUES-like sequence. This allowed estimating the over-representation degree of each sequence at each of those sequence (window) size levels.

For the DUES-containing Pasteurellaceae species, the over-representation numbers kept increasing with the increase of the size of the sliding window until the 10 bases window size. The result was the same both for the *H. influenza*- and for the *A. pleuromonieae*-type DUES variants (Fig. 3a and Supplemental Table 1).

In the case of the Neisseriaceae species, however, the picture was more complex; although the patterns were similar to those observed for the analyzed Pasteurellaceae genomes. Here also, the over-representation values kept increasing with the increase in the size of the sliding window up to a limit. However, in the Neisseriaceae case, that limit was different between the different genomes (*i.e.*, DUES-like sequences). It ranged between 14 and 19 bases (Fig. 3b, Table 3 and Supplemental Table 2). It is notorious that the seemingly different sizes of the Neisseriacean DUESs affect even the same core DUES-types so that, for instance, the *N. gonorrhoeae* and *N. lactamica* DUES seems to be the sequence AAAAATGCCGTCTGAAAC whereas the *N. meningitides* DUES seems to be the sequence AAAAATGCCGTCTGAAA, although both species show the same 12 pb core sequence identified in [12] and confirmed here in Table 2.

It is relevant to note that the same analysis when applied to non DUESs in DUES-lacking genomes (negative controls) did not show the size-correlated increase in sequence over-representation seen for the sequences containing or contained in the DUESs of the Pasteurellacean and Neisseriacean genomes (Fig. 3c).
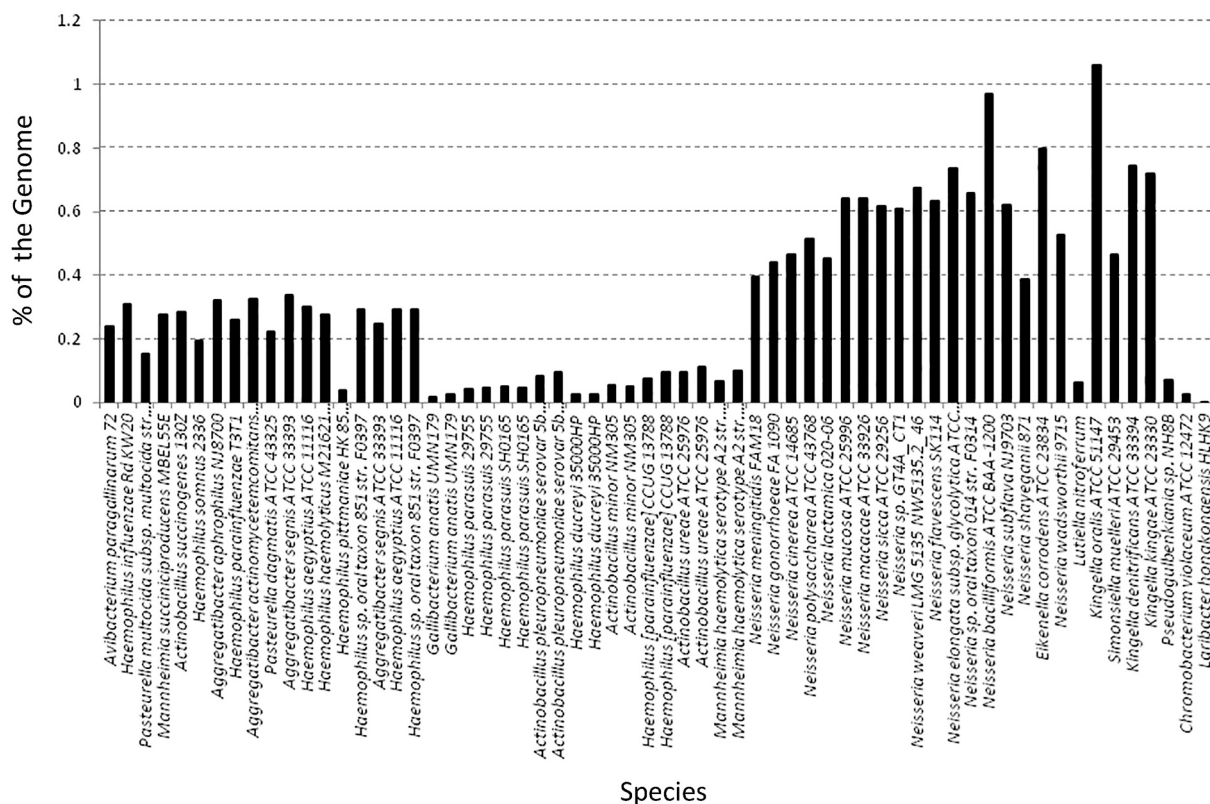


**Fig. 2.** Density of the DUES-like sequences in the analyzed genomes of the Pasteurellaceae and Neiseriaceae spacies. X-axis: Species, Y-axis: Percentage of the genome covered by the DUESs nucleotides.
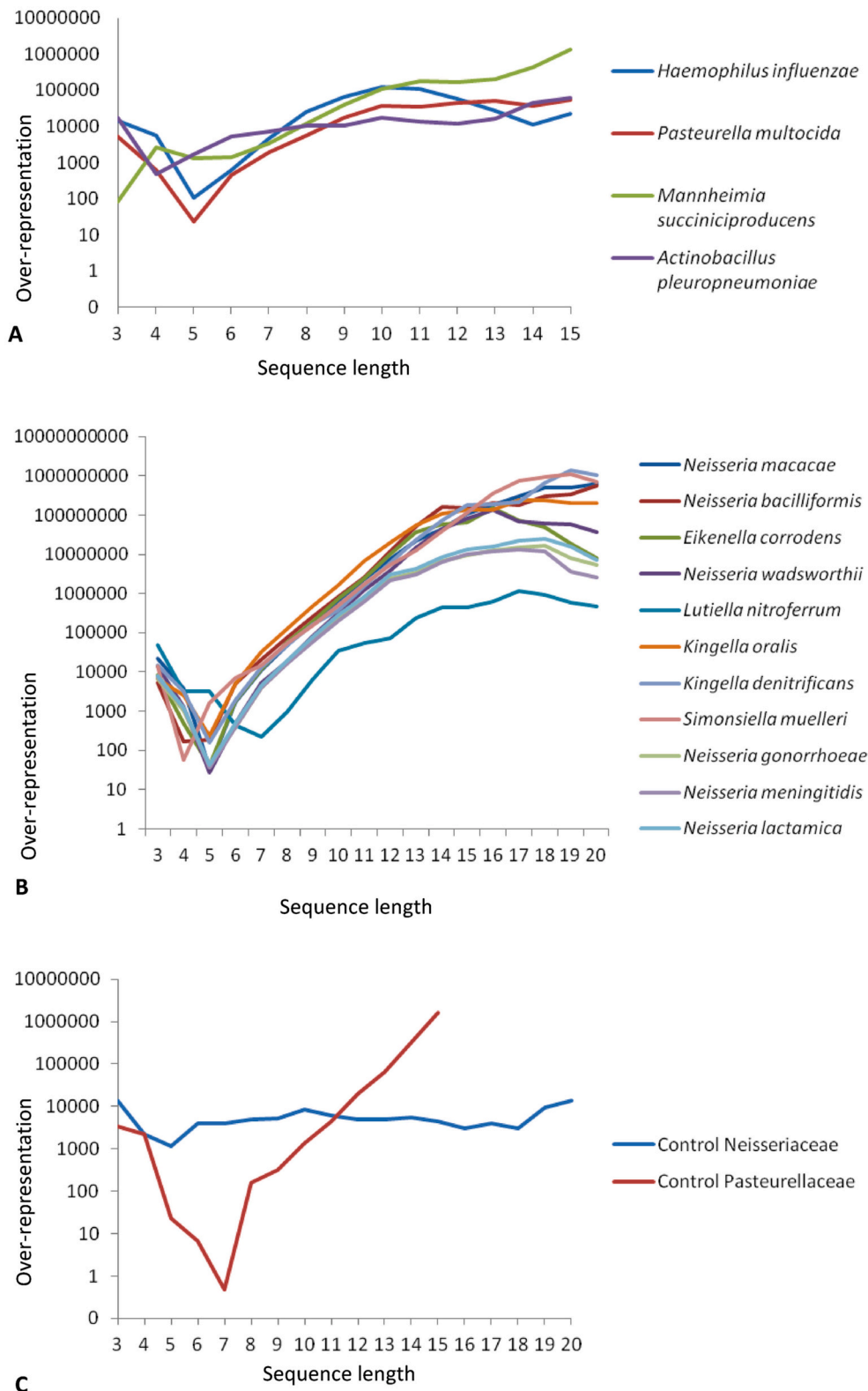
**Fig. 3.** Distribution of the over-representation levels (y-axis) of 3 to 20 base sequences (x-axis) that are contained in or that contain the DUES in the genomes of the Pasturellaceae (A), Neisseriaceae (B) and control species (C). X-axis: Length of the sequence containing or contained in the DUES, Y-axis: Sequence over-representation value calculated using the Chi-squared formula (see material and methods).

### 3.3. Phylogenetic relationships between DUESs and inference of their ancestral sequences

The 16S maximum likelihood phylogenetic tree of the different

Pasteurellales and Neisseriales strains and species for which 16S sequences were available in the database showed a clear separation between the members of both orders and a within-order and within-family arrangements in accordance with the taxonomical identifications of the

**Table 3**

The prevalence and full DUES sequences corresponding to each of the core sequences identified for the Neisseriaceae species in Table 2. Obs.: Observed number of DUESs in the genomic sequence, Exp.: Expected number of DUESs in the genomic sequence.

| Species | Sequence | Obs. | Exp. | Obs./Exp. |
|---|---|---|---|---|
| *Neisseria macacae* | AAAAGGTCGTCTGAAAAC | 187 | 0,000069 | 2,710,145 |
| *Neisseria bacilliformis* | AGGCCGTCTGAAAA | 1511 | 0,013834 | 109,224 |
| *Eikenella corrodens* | AAAGGCTACCTGAAAA | 283 | 0,000522 | 542,146 |
| *Neisseria wadsworthii* | AATGCCTGTCTGAAAA | 410 | 0,001228 | 333,876 |
| *Kingella oralis* | AAGGCAGCCTGAAAACG | 129 | 0,000069 | 1,869,565 |
| *Kingella denitrificans* | AAAAAGCAGCCTGCACTTT | 126 | 0,000012 | 10,500,000 |
| *Simonsiella muelleri* | TTTTCAGGCTGCCTGAAAA | 156 | 0,000022 | 7,090,909 |
| *Neisseria gonorrhoeae* | AAAAATGCCGTCTGAAAC | 29 | 0,000051 | 568,628 |
| *Neisseria meningitidis* | AAAAATGCCGTCTGAAA | 54 | 0,000217 | 248,848 |
| *Neisseria lactamica* | AAAAATGCCGTCTGAAAC | 37 | 0,000055 | 672,727 |

bacteria included in the analysis (Fig. 4a). When it comes to the DUESs, the tree also shows a notorious concordance between the phylogenetic distribution of the species and the DUES variants that these harbor. Result of that, a notorious clustering of the different DUES variants can be observed both in the Pasteurellaceae and in the Neisseriaceae families. These results are even clearer if the phylogenetic tree is built using the 16S sequences only of the species found here to carry DUESs (Fig. 4b). That tree shows how the species of the Pasteurellacean family can be divided into five different clades; one, the largest, composed mainly of species harboring the *H. influenza*-like DUES and *H. parasuis* and *H. parainfluenzae*, that harbor the *A. pleuromonieae*-type DUESs, while the four others clades, with fewer species, show only species with the *A. pleuromonieae*-type DUESs. No DUES dendrogram can be built for the just two Pasteurellacean DUESs and the obvious consensus of both sequences is NAMRWGCGGTN (Fig. 4c). The Neisseriaceae family however shows dispersion of the different DUESs with the exception of the *N. gonorrhoea*- and *N. maccacae*-types that appear clustered in their respective clades (Fig. 4b). The *Neisseria bacilliformis*-type DUES, and despite its prevalence and clustering, appears in clearly separated branches. That DUES-type also appears at one extreme (see the possible origin) of a Neisseriacean DUESs dendrogram (Fig. 4d) and the inferred consensus sequence of the Neisseriacean DUESs seems to be NAAAAAGGCYGYCTGAAAAC (Fig. 4c).

In order to infer the ancestral sequence of the Pasteurellacean species, an analysis of the over-representation of each nucleotide along each of the 11 potential ancestral DUES positions (as inferred from the Pasteurellacean DUES consensus sequence) was carried out for all the current mismatched DUESs in the corresponding *H. influenza* and *A. pleuromoneae* genomes. As Table 4 shows, the sequences inferred based on the nucleotide over-representation values in each of the analyzed genomes are highly concordant with the real DUES, especially in the GCGG part. The result indicates that AAAAAGCGGT seems to be, or to contain, the ancestral Pasteurellacean DUES sequence.

## 4. Discussion

It is known that, apart from few exceptions (such as *Azotobacter vinelandii* [24], *Campylobacter coli* [25] and *Pseudomonas stutzeri* [26]), specificity of the DNA uptake by naturally competent bacteria depends on the presence in the extracellular DNA of defined sequences, called DNA Uptake Enhancing Sequence (DUES). It is also known that such bias is not strict; so that mismatched DUESs are also taken up—although less

efficiently (see [3]). Previous works demonstrated that mutations generate new "perfect" and mismatched DUESs that end-up accumulating in the corresponding genome due to the biased uptake by competent bacteria of the DNAs that contain these sequences [3,6]. The mechanism responsible for such accumulation was suggested to resemble a molecular drive that gradually "grows up" the preferred DUES in regions of the genome where these sequences would not disturb as to affect the cell's fitness [3,6].

Over-representation of specific short DNA sequences in the genome, understood as the difference between the expected and the observed numbers in a genome, could therefore be used in order to identify the actual sequence of the preferred DUES. The logic that I applied here is based on the expectation that adding an actual DUES nucleotide to the right position of a partial DUES would result in a more complete and, thus, more over-represented sequence in the genome, while adding a nucleotide that does not form part of the actual DUES to a partial or complete DUES would result in a less "perfect" DUES and, thus, less over-represented sequence. This way one can identify the actual DUES as the DUES-like sequence that looses over-representation in the genome if we take from it or add to it any nucleotide. As to the inferences on competence itself, it is legitimate to take the over-representation of the DUES as indicator both of the likelihood of a bacterium to be competent for uptake of specific DNAs, of the efficiency of the DNA uptake, and of the frequency of the competence and DNA uptake episodes (*i.e.*, of how competent is the bacterium). The logic here is that a bacterium that ceases to be competent would see the DUESs degenerate and disappear in parts or all its genome. Conversely, the more often a bacterium becomes competent, the more DNA it would take up, and the stronger the bias of its DNA uptake towards a DUES is, the more DUESs its genome would accumulate. With such logic in mind, and according to the results of the current work, competence for DUES-biased DNA uptake seems confined to species from the Pasteurellaceae and Neisseriaceae families; since over-representation of short DUES-like DNA sequences was not found in the sequenced genomes of any bacterium belonging to families other than Pasteurellaceae and Neisseriaceae, not even other Pasteurellales or Neisseriales families.

I want to draw the attention to the fact that there was no difference, as to the density of the DUESs; between the finished and unfinished genomic sequences used for this work; meaning that the unfinished sequences considered here are representative samples of the corresponding full genomes—a fact also supported by the inferred proportion of the sequenced genomes (all the analyzed sequences were of over 1.8 Mb, for genome sizes that range between 2 and 6 Mb). Still, I cannot fully discard potential, although not likely, biases in the availability of genomes that might, although not likely, have led to the current results. For instance, I cannot discard that there could be a, or some, non pasteurellacean species that might harbor DUESs but whose genomes were not sequenced. Neither can I discard the possibility that sequencing was more biased towards DUES harboring species in the Neisseriaceae family than in the Pasturellaceae. Similarly, I cannot discard the possibility that bias in the sequenced species could be behind the differences between the Neisseriaceae and Pasturellaceae results. Still, with those cautious observations having been made, there are reasons to think that such biases are not likely to be the cause of the results reported in this work. The fact that the number of species analyzed here is similar between the two bacterial families speaks in favour of the lack of significant effect of any potential bias in the availability of the data on the results obtained here.

Sorting the analyzed species by density of the DUESs in their genomes should reflect the efficiency of their DNA uptake bias and the frequency of the episodes of their spontaneous competence for transformation. With that in mind, two direct aspects are to highlight regarding the results reported here on the species distribution of the DUESs. The first is the fact that the results show higher genomic densities of the DUESs in the Neisseriaceae species than in the Pasteurellaceae species, suggesting that bacteria of the first family have higher
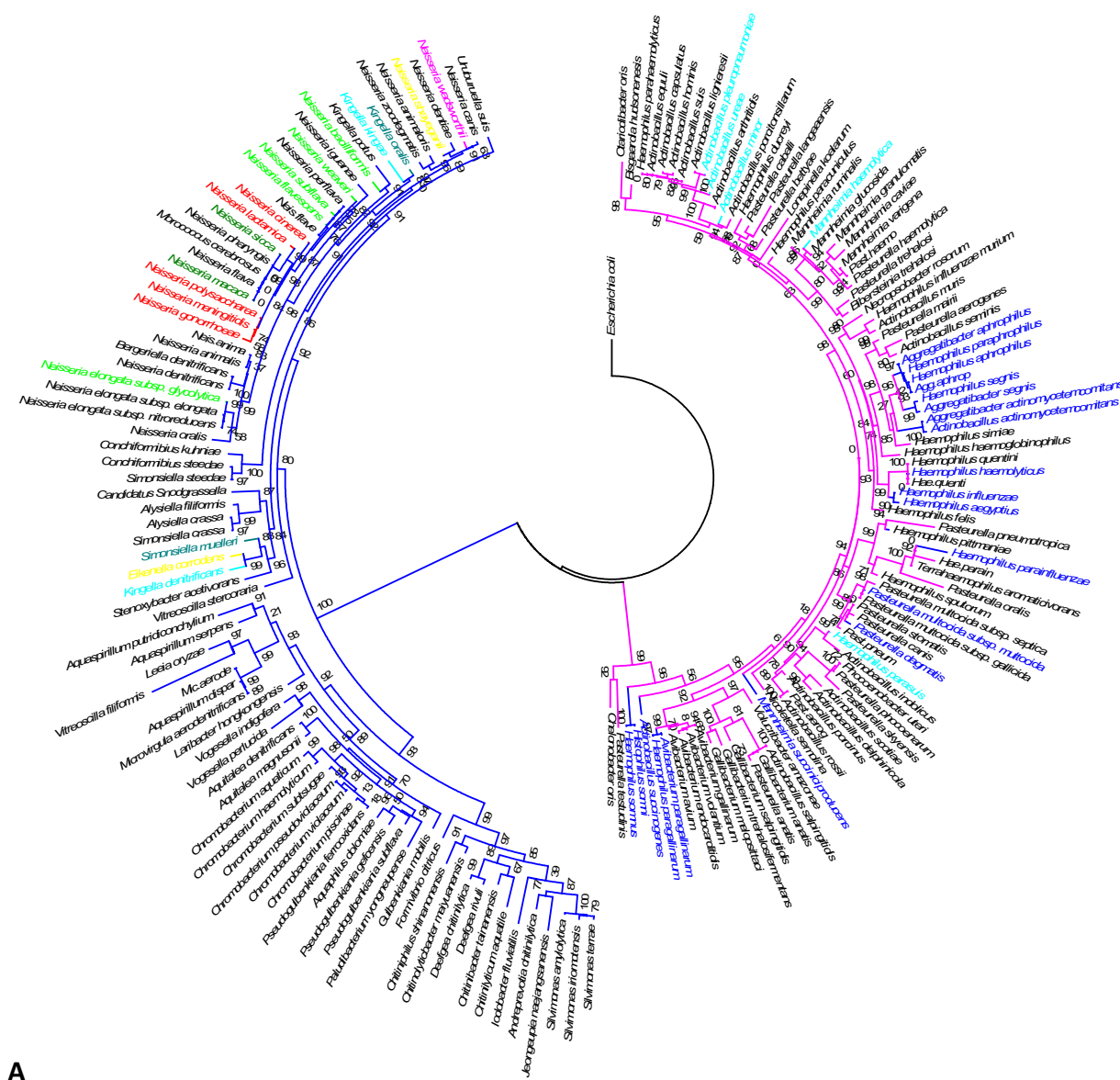
**Fig. 4.** Maximum likelihood phylogenetic tree of the 16 s ribosomal DNA of the Pasteurellales and Neisseriales species (a) and of the species whose genomes harbor DUESs(b)—in different colors in (a). Each color reflects the same DUES type and the brackets in (b) highlight clustering. In (c) is the dendrogram of the Neisseriaceae DUESs with Pseudogulbenkiania (Lutiella nitroferrum) as Neisseriales outgroup and in (d) are the alignments of these sequences.

susceptibility to become competent and/or more efficient DNA uptake (*i.e.*, they take up more DNA) than bacteria of the latter family. In such case, competent species of the Neisseriaceae could have more propensity to horizontal gene transfer and its consequences. Another result to highlight here is the higher diversity of the DUESs in the Neisseriaceae family than in the Pasteurellaceae. The Neisseriacean DUESs seem not only to differ in their nucleotidic composition but also in their sizes as well. In fact, while each of the two Pasteurellaceae DUESs are of 10 bp, the eight Neisseriaceae sequence types have sizes ranging between 14 and 19 bp. This result is also concordant with, and reinforces, the conclusion that the Neisseriaceae are more competent and/or more efficient at DNA uptake than the Pasteurellaceae. Still, and conversely, more interesting is probably the thought that; given that the differences in DUESs would hamper inter-species uptake of DNA, the results of this work suggest that inter-specific DNA-uptake, and thus cross-species movement of DNA, seems to be more frequent and efficient between Pasteurellaceae species than between Neisseriaceae species. The

situation seems therefore to reflect some sort of "compromise" between the efficiency of competence for transformation by DUES-biased DNA uptake and the diversification of the DUESs, whereby one can interpret that bacterial species of the "more competent" taxon have different DUESs in order to prevent excessive cross-species movement of DNA whereas species of the "less competent" taxon have less DUES types as their less efficient and/or frequent competence state would make the need for establishing barriers against uptake of DNA from other species insufficient for requiring (selecting for) the evolution of new DUES types. Of course, this is a purely evolutionary way of interpreting the results. A mechanistic interpretation of such result (*i.e.*, more frequent and more diverse DUESs in a family comparing to the other family) could be that there might be more DNA uptake and a higher mutability and/or lower stabilizing selection (and thus more diversification) of the molecule (protein) that binds the extracellular DUES in species of the Neisseriaceae family than in Pasteurellaceae species.

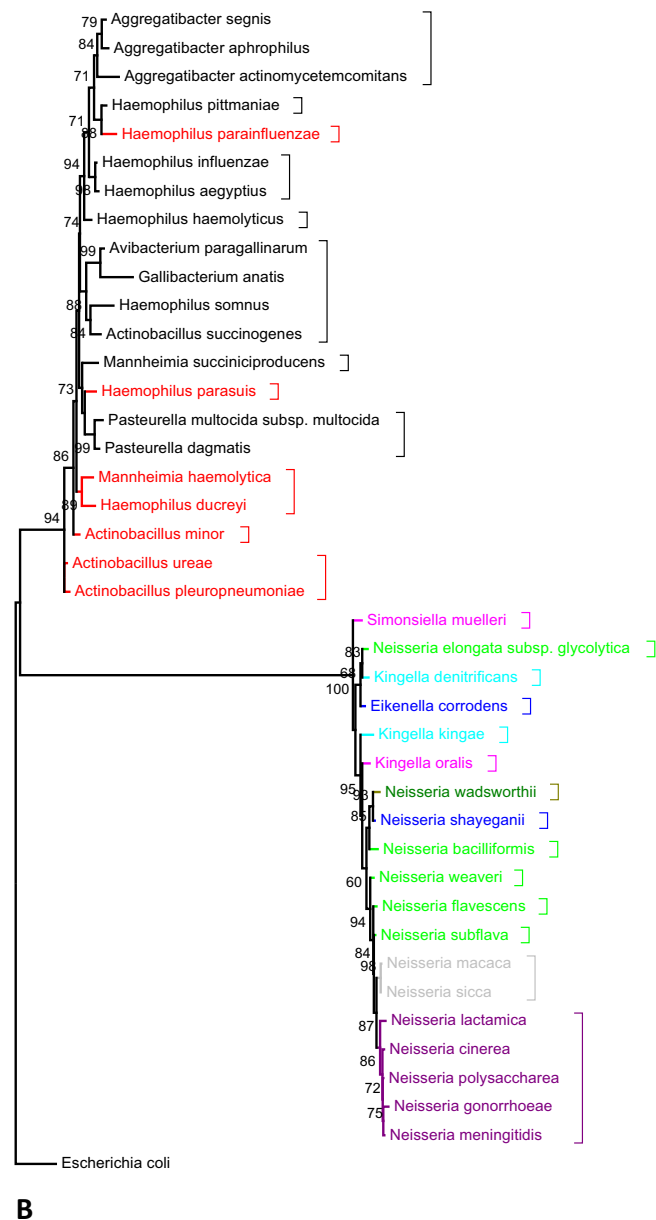¿Could there be a relation between the efficiency or frequency of the

**Fig. 4.** (*continued*).

DUES-biased DNA uptake events (competence) and the appearance of new DUES types?

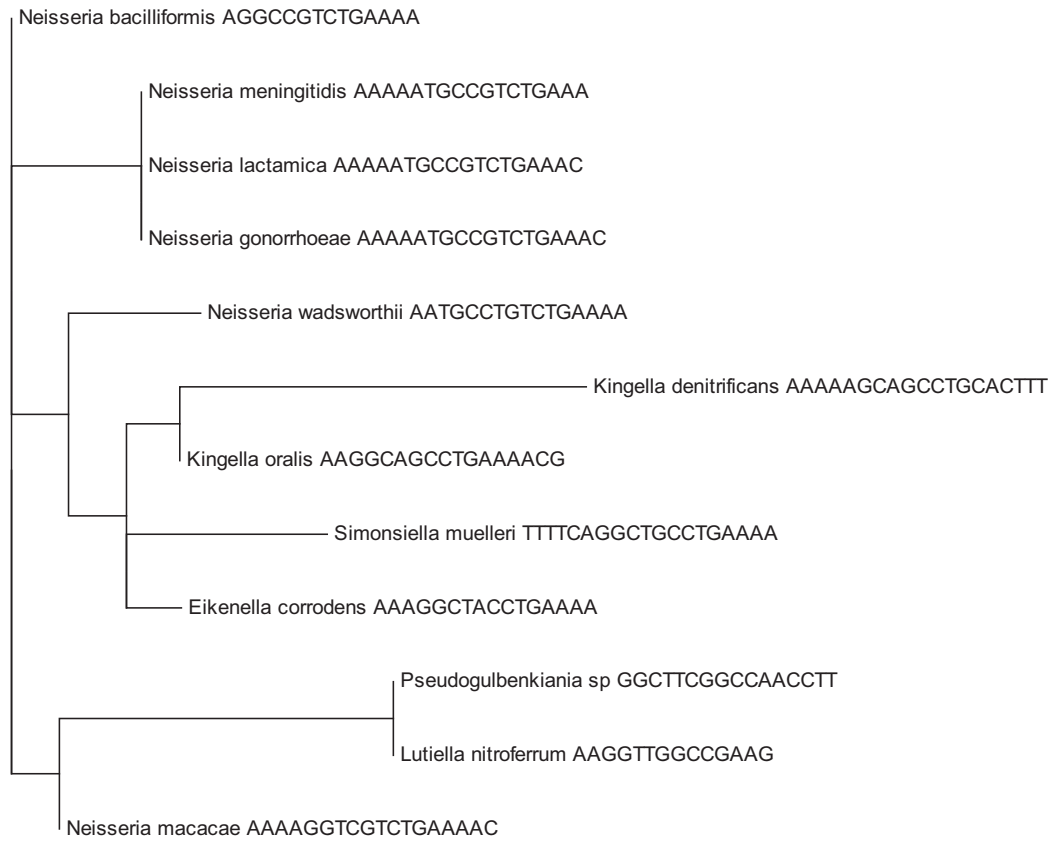The answer is very likely yes!

Considering as indicator the fact that *H. influenzae* shows growth-phase dependent $10^{-4}$ transformation frequency [27] while *N. gonorrhoeae* has growth-phase independent ~$2 \times 10^{-1}$ transformation frequency [28]; it is legitimate to conclude thus that the *in silico*-based data and interpretations reported here seem to suggest that while within-species uptake of DNA (uptake of conspecific DNA) seems more likely for the seemingly more competent Neisseriaceae species, cross-species uptake of DNA is more likely between the seemingly less competent Pasteurellaceae species. It is therefore possible that the appearance, maintenance and accumulation of new DUES types become selected for when the transformation efficiency and inter-species movement of DNA surpasses some threshold—DUESs could in such case be seen as barriers against excessive inter-specific DNA exchange. Such evolutionary dynamics would necessarily be based on mutability and selection of the protein that binds the extracellular DNA.

In line with the interpretation made above, the results show how the

*A. pleuromonieae*-type DUES-harboring species have two different DUES-like sequences with very similar densities, probably because of some higher flexibility in the specificity of the DNA uptake bias (*i.e.*, in the DNA binding by the receptor) of the competent cells of those bacteria.

The results show how not all the Pasteurellaceae species nor all the Neisseriaceae species have DUESs. Thus, either the DUES-lacking species have lost the DUESs, very likely after losing the capacity to become spontaneously competent for DNA uptake, or the DUES-harboring species have acquired these sequences while evolving DNA uptake bias concomitant or posterior to the evolution of competence. The fact that within each of these two DUES-harboring families, the genomes of most of the species do contain potential DUESs, means that biased DNA uptake is ancestral in each of these families—the probability of DNA uptake bias and DUESs being acquired independently (in parallel) in so many species is negligible. DUESs, and very probably competence before them, have therefore very likely been lost in some Pasteurellaceae and Neisseriaceae species and did not evolve independently in the so many competent and DUES-having bacteria.

The Pasteurellales and Neiseriales 16S phylogenetic tree highlights

Neisseria bacilliformis AGGCCGTCTGAAAA

Neisseria meningitidis AAAAATGCCGTCTGAAA

Neisseria lactamica AAAAATGCCGTCTGAAAC

Neisseria gonorrhoeae AAAAATGCCGTCTGAAAC

Neisseria wadsworthii AATGCCTGTCTGAAAA

Kingella denitrificans AAAAAGCAGCCTGCACTTT

Kingella oralis AAGGCAGCCTGAAAACG

Simonsiella muelleri TTTTCAGGCTGCCTGAAAA

Eikenella corrodens AAAGGCTACCTGAAAA

Pseudogulbenkiania sp GGCTTCGGCCAACCTT

Lutiella nitroferrum AAGGTTGGCCGAAG

Neisseria macacae AAAAGGTCGTCTGAAAAC

**C**

| *Neisseria macacae* | NNAAAAGGTCGTCTGAAAACNNN |
| *Neisseria bacilliformis* | NNNNNAGGCCGTCTGAAAANNNN |
| *Eikenella corrodens* | NNNAAAGGCTACCTGAAAANNNN |
| *Neisseria wadsworthii* | NNNAATGCCTGTCTGAAAANNNN |
| *Kingella oralis* | NNNNAAGGCAGCCTGAAAACGNN |
| *Kingella denitrificans* | NNNAAAAGCAGCCTGCA--CTTT |
| *Simonsiella muelleri* | TTTTCAGGCTGCCTGAAAANNNN |
| *Neisseria gonorrhoeae* | NAAAAATGCCGTCTGAAA-CNNN |
| *Neisseria meningitidis* | NAAAAATGCCGTCTGAAANNNNN |
| CONSENSUS | NAAAAAGGCYGYCTGAAAACNNN |

**D**

**Fig. 4.** (*continued*).

the absence of DUESs in non-neisseriacean and non-pasteurellacean species. It, together with the DUES-harboring species 16S tree, support the ancestral nature of the neisseriacean and pasteurellacean DUESs in their respective families; as these sequences appear scattered throughout the respective clades—*i.e.*, the Pasteurellaceae and the Neisseriaceae clades. Such distribution is more likely explained by common ancestry than by an unlikely independent evolution. Given the association of DUESs with biased uptake of DNA, we can infer that the biased uptake of DNA is ancestral in both bacterial families.

While the Pasteurellales branch of the tree shows only Pasteurellaceae species and, result of that, DUES-harboring species in all its clades, the Neisseriales tree shows subdivision into two large clades; one

containing DUES-harboring species, while the other completely lacks DUES-harboring bacteria. While the absence of 16S sequences from non-Pasteurellaceae species explains the Pasteurellales branch, the absence of DUESs in the genomes of non-Neisseriaceae species explains the Neisseriales branch. The scattering of the DUESs through the Pasteurellaceae and Neisseriaceae clades suggests that DNA uptake, or at least its bias towards DUESs, is not posterior to the diversification of any of these family. The lack of DUESs in the genomes of Pasteurellales and Neisseriales species other than Pasteurellaceae and Neisseriacea, if not an unlikely product of any probable sample size issue, would suggest that the emergence of the DNA uptake bias towards the DUESs might have been concomitant to, or just after, the split of these families from

**Table 4**

The possible ancestral DUES of the Pasteurellaceae species as inferred from the nucleotide over-representation values in the mismatched Flu-type and Pleu-type DUESs.

**Table 4.** The possible ancestral DUES of the Pasteurellaceae species as inferred from the nucleotide over-representation values in the mismatched Flu-type and Pleu-type DUESs.

| Position | Nucleotide | | | | Flu-type | | Infered ancestral | Pleu-type | | Nucleotide | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | G | C | T | Inferred | Real | | Real | inferred | A | G | C | T |
| 1 | 6,330 | 93,099 | 115,525 | 70,698 | C | A | **A** | A | A | 1,249 | -0,694 | -0,437 | -0,454 |
| 2 | 10,580 | 6,068 | 0,876 | 5,902 | A | A | **A** | A | G | 9,120 | 125,770 | 26,663 | 26,723 |
| 3 | 10,230 | 14,911 | 4,945 | 6,154 | G | A | **A** | C | A | 39,166 | 10,450 | 9,238 | 28,088 |
| 4 | 9,353 | 8,383 | 1,392 | 5,463 | A | G | **A** | A | A | 13,342 | 5,816 | 1,033 | -0,221 |
| 5 | 37,307 | 6,683 | 11,390 | 9,650 | A | T | **A** | A | A | 13,591 | 2,463 | -1,813 | -1,155 |
| 6 | 1,261 | 9,800 | -1,302 | 1,899 | G | G | **G** | G | G | 2,929 | 12,675 | 2,804 | 1,366 |
| 7 | 0,096 | 2,270 | 9,630 | 1,025 | C | C | **C** | C | C | -1,947 | -2,701 | 13,790 | -0,223 |
| 8 | 2,965 | 9,403 | 1,347 | 0,521 | G | G | **G** | G | G | 0,524 | 13,247 | 3,500 | -1,311 |
| 9 | 5,288 | 8,773 | 3,340 | 4,724 | G | G | **G** | G | G | 8,856 | 12,691 | -0,357 | 1,477 |
| 10 | 31,180 | 43,153 | 36,706 | 8,819 | C | T | **T** | T | T | 3,921 | 10,622 | 6,469 | 12,840 |
| +1 (11) | -0,401 | 0,030 | -0,515 | 0,701 | T | N | **N** | C or N | N | 6,176 | 4,126 | -0,454 | 6,967 |

their sister species (*i.e.*, the Pasteurellaceae ancestor and the Neisseriaceae ancestor evolved DNA uptake bias and DUESs).

The presence of *A. pleuromonieae*-type DUES-harboring species in the *H. influenzae*-type DUES-harboring species part of the tree and absence of *H. influenzae*-type DUES-harboring species in the *A. pleuromonieae*-type DUES-harboring species part of the tree seems to suggest that the *A. pleuromonieae*-type DUES could be the ancestral Pasteurellacean DUES. However, (i) the *A. pleuromonieae*-type DUES-harboring species part of the tree is small, (ii) there are only two DUES-types in the tree, (iii) only two species appear in a part of the tree that is not congruent with the DUES-type that these species harbor, and (iv) there is no clear clade of species harboring any of the two Pasteurellaceae DUES types. All these make it safer to consider that the 16S tree do not allow pinpointing with enough confidence any of the two Pasteurellacean DUESs as the ancestral one. For its part, the Neisseriaceae tree points towards either the *N. bacilloformis*-type or the *K. denitrificans*- or *E. corrodens*-types of DUES as possibly ancestral (due to them being the Neisseriaceae DUES-types that appear in separate parts of the tree). Not only the dispersion of the *N. bacilloformis*-type-harboring species in the tree is higher, also the unrooted dendrogram of the Neisseriaceae DUESs seems to support that DUES-type as the probable ancestral sequence for the Neisseriaceae. The consensus sequence of all the Neisseriaceae DUES types also shows clear resemblance to the *N. bacilloformis*-type DUES.

If the DNA uptake bias is ancestral, as concluded here, there must be an ancestral DUES for each family. The most likely ancestral DUES of the Neisseriaceae species having been inferred based on phylogeny and sequence similarities, phylogeny and sequence similarity results do not allow pinpointing which Pasteurellaceae DUES-type could be the ancestral one, and whether both have emerged from an extinct ancestral sequence. I therefore opted for a different strategy that is based on the logic that the ancestral sequence could have left its footprint in the pasteurellacean genomes no matter their current DUES. I looked for such footprint in each position of the DUES. When a position of the aligned sequences of both Pasteurellaceae DUES-types shows the same nucleotide, the decision is simple; that nucleotide is ancestral. However, there is discrepancy of nucleotides in some positions of the different pasteurellacean DUESs, I hence opted for a screening of the over-

representation levels of each of the four nucleotides in each of the ten positions of the DUESs and their mismatched forms. This way I could reveal the second most important (over-represented) nucleotide after the DUES nucleotide in each position of the pasteurellacean DUESs. In case the second most important nucleotide in a position of a DUES-type is the same as the most important nucleotide at the same position of the other DUES-type, that nucleotide is considered as the most likely ancestral one. The results of this way of looking for a footprint of an ancestral sequence suggest that the ancestral pasteurellacean DUES seems to be AAAAAGCGGTN. The results of this analysis also highlight the GCGG sequence as the most conserved core of the Pateurellacean DUES; just as *in silico* and wet lab. Experiments have demonstrated in [3]. Given that the logic and method used here could be applied to the analysis of other sequences (protein binding, transcription factor binding, repeated…), and hopping that it might be inspiring and helpful to other follow scientists, I have to highlight that its success depends on the compared species and sequences not being too distant (*i.e.*, divergence is expected to significantly erase possible old footprints in the genome).

To conclude, here I show that the DUESs, and hence the biased DNA uptake, seem confined to the Pateurellaceae and Neisseriaceae families. While most bacteria of these families harbor DUESs, not all do, suggesting that the species of these families are mainly competent and have DNA uptake bias towards DUES-containing DNAs. Within each family, the densities of DUESs are very different between species implying differences in the efficiency of the DNA uptake or its bias. One can infer that there seems to be a sort of compromise between the efficiency of the DNA uptake or its bias and between the varieties of DUESs that a group of bacteria show; so that the Pasteurellaceae, that have few DUES types, would have less efficient uptake and whereas the Neisseriaceae, that more DUES types, would have more efficient uptake and their diverse DUESs might very likely have evolved in order to prevent excessive cross species passage of DNA. One could speculate that this situation suggests the existence of thresholds for allowing cross-species transfer of DNA. Indeed, it is logical to expect mechanisms that limit horizontal DNA (gene) transfer, as recombination with distant DNAs could be harmful and excessive permeability to exogenous DNAs would ultimately erase species boundaries. Finally I suggest a way to detect footprints of

ancestral sequences in the genome. Applied here, that method allowed inferring the ancestral Pasteurellacean DUES—the Neisseriacean one being easily inferred based on sequence similarity and phylogeny.

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ygeno.2021.06.020.

## References

[1] F. Griffith, The significance of pneumococcal types, J. Hyg. 27 (1928) 113–159.

[2] M. Bakkali, Could DNA uptake be a side effect of bacterial adhesion and twitching motility? Arch. Microbiol. 195 (2013) 279–289, https://doi.org/10.1007/s00203-013-0870-1.

[3] M. Bakkali, Genome dynamics of short oligonucleotides: the example of bacterial DNA uptake enhancing sequences, PLoS One 2 (2007), e741, https://doi.org/10.1371/journal.pone.0000741.

[4] J.L. Berry, A. Cehovin, M.A. McDowell, S.M. Lea, V. Pelicic, Functional analysis of the interdependence between DNA uptake sequence and its cognate ComP receptor during natural transformation in Neisseria species, PLoS Genet. 9 (2014) e1004014, https://doi.org/10.1371/journal.pgen.1004014. PGENETICS-D-13-00964 [pii].

[5] A. Cehovin, et al., Specific DNA recognition mediated by a type IV pilin, Proc. Natl. Acad. Sci. U. S. A. 110 (2013) 3065–3070, doi:1218832110 [pii], https://doi.org/10.1073/pnas.1218832110.

[6] M. Bakkali, T.Y. Chen, H.C. Lee, R.J. Redfield, Evolutionary stability of DNA uptake signal sequences in the Pasteurellaceae, Proc. Natl. Acad. Sci. U. S. A. 101 (2004) 4513–4518.

[7] S.D. Goodman, J.J. Scocca, Identification and arrangement of the DNA sequence recognized in specific transformation of Neisseria gonorrhoeae, Proc. Natl. Acad. Sci. U. S. A. 85 (1988) 6982–6986.

[8] C. Elkins, C.E. Thomas, H.S. Seifert, P.F. Sparling, Species-specific uptake of DNA by gonococci is mediated by a 10-base-pair sequence, J. Bacteriol. 173 (1991) 3911–3913.

[9] O.H. Ambur, S.A. Frye, T. Tonjum, New functional identity for the DNA uptake sequence in transformation and its presence in transcriptional terminators, J. Bacteriol. 189 (2007) 2077–2085, doi:JB.01408-06 [pii], https://doi.org/10.1128/JB.01408-06.

[10] P.M. Duffin, H.S. Seifert, DNA uptake sequence-mediated enhancement of transformation in Neisseria gonorrhoeae is strain dependent, J. Bacteriol. 192 (2010) 4436–4444, doi:JB.00442-10 [pii], https://doi.org/10.1128/JB.00442-10.

[11] P.M. Duffin, H.S. Seifert, Genetic transformation of Neisseria gonorrhoeae shows a strand preference, FEMS Microbiol. Lett. 334 (2012) 44–48, https://doi.org/10.1111/j.1574-6968.2012.02612.x.

[12] S.A. Frye, M. Nilsen, T. Tonjum, O.H. Ambur, Dialects of the DNA uptake sequence in Neisseriaceae, PLoS Genet. 9 (2013) e1003458, https://doi.org/10.1371/journal.pgen.1003458. PGENETICS-D-12-01142 [pii].

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

[14] S.H. Goodgal, M.A. Mitchell, Sequence and uptake specificity of cloned sonicated fragments of Haemophilus influenzae DNA, J. Bacteriol. 172 (1990) 5924–5928.

[15] R.D. Fleischmann, et al., Whole-genome random sequencing and assembly of Haemophilus influenzae Rd, Science 269 (1995) 496–512.

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

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

[18] S. Karlin, J. Mrazek, A.M. Campbell, Frequent oligonucleotides and peptides of the Haemophilus influenzae genome, Nucleic Acids Res. 24 (1996) 4263–4272.

[19] R.J. Redfield, et al., Evolution of competence and DNA uptake specificity in the Pasteurellaceae, BMC Evol. Biol. 6 (2006) 82.

[20] K. Katoh, G. Asimenos, H. Toh, Multiple alignment of DNA sequences with MAFFT, Methods Mol. Biol. 537 (2009) 39–64, https://doi.org/10.1007/978-1-59745-251-9_3.

[21] K. Katoh, K. Misawa, K. Kuma, T. Miyata, MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform, Nucleic Acids Res. 30 (2002) 3059–3066.

[22] S. Kumar, G. Stecher, K. Tamura, MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets, Mol. Biol. Evol. 33 (2016) 1870–1874, https://doi.org/10.1093/molbev/msw054.

[23] T.A. Hall, BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT, Nucleic Acids Symp. Ser. 41 (1999) 95–98.

[24] J.L. Doran, W.J. Page, Heat sensitivity of *Azotobacter vinelandii* genetic transformation, J. Bacteriol. 155 (1983) 159–168.

[25] Y. Wang, D.E. Taylor, Natural transformation in campylobacter species, J. Bacteriol. 172 (1990) 949–955.

[26] C.A. Carlson, L.S. Pierson, J.J. Rosen, J.L. Ingraham, Pseudomonas stutzeri and related species undergo natural transformation, J. Bacteriol. 153 (1983) 93–99.

[27] E.M. Wise Jr., S.P. Alexander, M. Powers, Adenosine 3′:5′-cyclic monophosphate as a regulator of bacterial transformation, Proc. Natl. Acad. Sci. U. S. A. 70 (1973) 471–474, https://doi.org/10.1073/pnas.70.2.471.

[28] H.L. Hamilton, J.P. Dillard, Natural transformation of Neisseria gonorrhoeae: from DNA donation to homologous recombination, Mol. Microbiol. 59 (2006) 376–385.