FEBS Letters 522 (2002) 52–58 FEBS 26201

Transposon-like Correia elements: structure, distribution and genetic exchange between pathogenic *Neisseria* sp.

Nicolas Buisine[a,*], Christoph M. Tang[b], Ronald Chalmers[a,*]

[a] *Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK*
[b] *Department of Infectious Diseases, Imperial College of Sciences, Technology and Medicine, The Flowers Building, Armstrong Road, London SW7 2AZ, UK*

Received 22 March 2002; revised 8 May 2002; accepted 21 May 2002

First published online 6 June 2002

Edited by Takashi Gojobori

Abstract Correia elements are a prominent feature of all four *Neisseria* genome sequences. We report an in silico analysis of the structure and genomic distribution of these elements and some preliminary biochemical data. Correia elements fall into four major families, distinguished by a 50 bp internal deletion and five point mutations. The elements resemble a transposon with 25 bp inverted repeats and a TA duplication at the target site. Within the element there is a functional integration host factor binding site. The genomic distribution of Correia elements is essentially random except for some small Correia-less regions apparently acquired by horizontal transfer. Phylogenetic analysis suggests that their presence predates the divergence of *Neisseria meningitidis* and *Neisseria gonorrhoeae*.

Key words: Pathogen; Virulence; Polymorphism; Genetic exchange; Horizontal transfer; Integration host factor

1. Introduction

Neisseria meningitidis (N.m.) and *Neisseria gonorrhoeae* (N.g.) are important human pathogens, responsible to cerebrospinal meningitidis and gonorrhoeae, respectively. The impact of high recombinational rate on the adaptive evolution of pathogenic *Neisseria* is well illustrated by their complex population structure [1,2], the presence of mosaic genes, antigenic variation, and the occurrence of regions of the chromosome that appear to have been acquired horizontally, some of which contribute to the virulence of these bacteria [3,4].

Analysis of the genome sequence of N.m. Z2491 and MC58 revealed a high occurrence of repeated sequences [5,6], thought to be involved in intragenomic and intergenomic variation mechanisms [7,8]. Small repetitive elements (SREs), often called 'Correia elements', are repetitive sequences of unknown function, characteristic of neisserial genomes [9,10]. Surprisingly, despite this prevalence and the available genome sequences, the distribution and function of SREs have not been characterised in detail to date.

In this paper, we demonstrate that SREs are ancestral elements in pathogenic *Neisseria*, and their absence from a chromosomal region highlights elements which have probably been acquired by horizontal transfer. Furthermore, despite their small size, SREs have features characteristic of insertion sequences, and contain an integration host factor (IHF) binding and transcriptional start signals that may play a role in modulating gene expression.

2. Materials and methods

The source code for programs is available from NB on request. A custom BLAST output parser was used to collect information on Correia element position in the genome, strand location, size and level of similarity. The strand bias index (SBI) was calculated as follows: for each element found on the plus strand, a counter records one positive unit, whereas for each element on the minus strand, the counter records a negative unit. SBIs are then plotted as a regular map with a 5 kb window. A random distribution that produces equal numbers of Correia elements on each strand will have a SBI of zero. The larger the SBI, either positive or negative, the more the elements are located on one particular strand. In order to reconstruct full-length Correia elements, coordinates of BLAST highest-scoring segment pairs of a given size class were used to extract the corresponding sequence along with some flanking DNA sequence. This information was saved into a file and a simple Bash script was used to automate ClustalW alignments. Also used were Artemis and ACT, both available on the Sanger centre ftp server (http://www.sanger.ac.uk/Software/).

The DNA fragment used for the gel retardation assay corresponded to 80 bp from the central region of the Correia element located at positions 1 508 472–1 508 627 of the N.m. Z2491 genome. The element was cloned and the fragment was amplified by PCR using the primers 5′-GATATCGGATCCGCAGACAGTACAAATAG and 5′-GTTAACGGATCCTTAGCTCAAAGAGAAC. The control DNA fragment from Tn*10* and the protocol for the assay are described in [11]. Purified IHF protein was from *Escherichia coli* and was a gift from Howard Nash (NIH).

3. Results and discussion

3.1. Four families of Correia elements in N.m. and N.g.

While investigating a large duplication in the genome sequence of N.m. MC58, a BLAST search revealed hundreds of short repeats related to the 156 bp Correia element. This element was originally identified in N.g. 5019 during heteroduplex mapping experiments and shown by Southern blotting to be to be repeated at least 20 times [9,10]. When the 156 bp Correia element was used to search the MC58 genome sequence, more than 700 significant BLAST hits were obtained, ranging in size from 15 to 156 bp. Amongst these we found

*Corresponding author. Fax: (44)-1865-275297.
E-mail address: chalmers@bioch.ox.ac.uk (R. Chalmers).

Abbreviations: IHF, integration host factor; SBI, strand bias index; SRE, simple repetitive element; N.m., *Neisseria meningitidis*; N.g., *Neisseria gonorrhoeae*

0014-5793/02/$22.00
PII: S0014-5793(02)02882-X

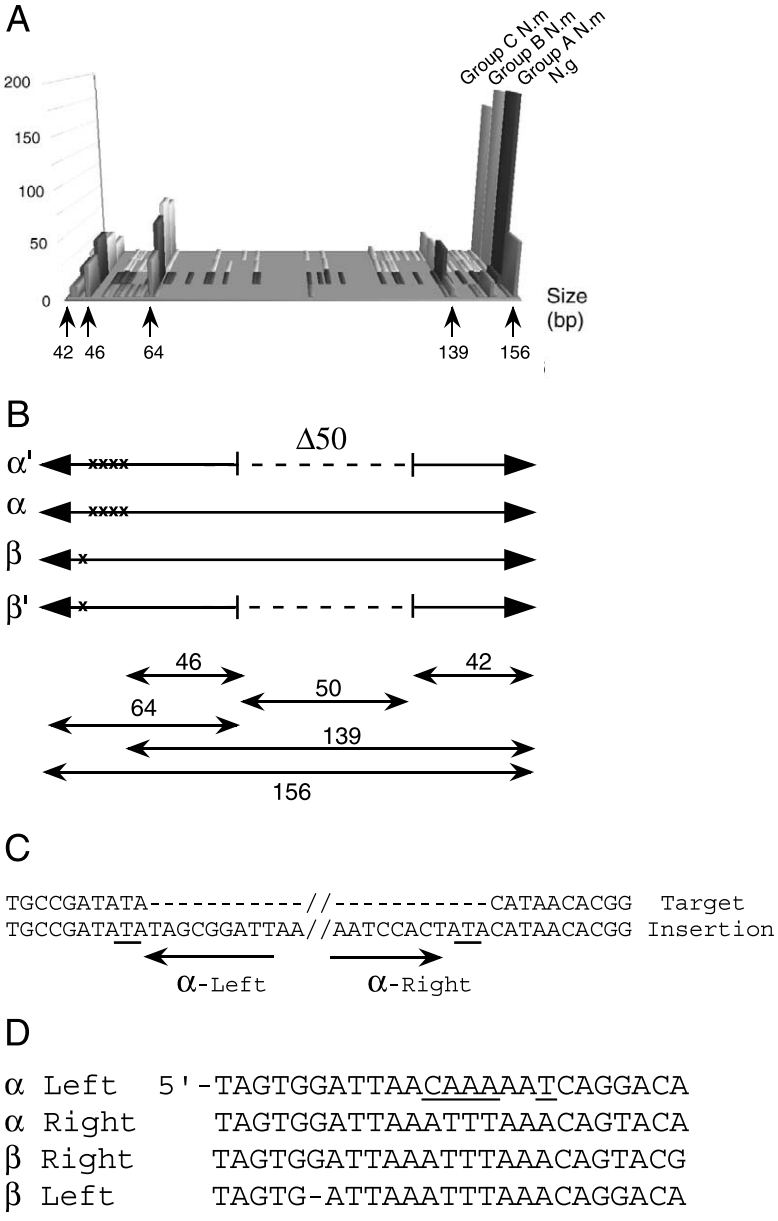

Fig. 1. Correia element sequences resemble transposons. A: A plot of the size distribution of BLAST hits to the canonical Correia element in four species of *Neisseria*. B: Structural organisation of four families of Correia elements. Deletions and point mutations that terminate extension of the BLAST hits in (A). The 'X' symbol represents four point mutations in the α family and a single nucleotide deletion in the β family. C: Example of an empty target site and the TA dinucleotide duplication on insertion. D: Sequence alignment of the terminal inverted repeat of the α and β families.

290 fragments of 42 bp or more with >94% sequence identity. A plot of the size distribution revealed that 70% of these fragments fell into five major classes of 42 bp, 46 bp, 64 bp, 139 bp and the full-length 156 bp Correia element (Fig. 1A). When the analysis was extended to the genome sequences of the N.m. group A strain Z2491 (see also [5]), the N.m. group C strain FAM18 and N.g., almost identical results were obtained with most BLAST hits falling into the five major classes (Fig. 1A).

Each of the five major classes of BLAST hits, together with some flanking DNA sequences, was extracted from all four *Neisseria* genomes and aligned using ClustalW. The results revealed that the four classes of shorter than full-length BLAST hits are components of larger elements in which point mutations and/or a deletion terminated extension of the BLAST hit (Fig. 1B). The vast majority of the BLAST hits are therefore explained by the existence of four families of Correia elements: the α and β families are distinguished by four point mutations and a deletion in the left end of the elements, while the α′ and the β′ elements, respectively, are the result of identical 50 bp internal deletions (Fig. 1B).

More than 400 of the BLAST hits to sub-sections of the full-length element were not incorporated into the four families of Correia elements (above). A very few of these segments are longer than 42 bp (Fig. 1A) and are the result of point mutations and minor DNA rearrangements not present in the four major families of Correia elements. The remainder of the 400 odd BLAST hits not incorporated in the analysis were between 15 and 41 bp in length. These represented isolated

fragments of Correia elements which were not flanked by inverted repeats and could not therefore be part of a mobile element.

3.2. Terminal inverted repeats and target site duplication

The ends of all four families of Correia elements are defined by terminal inverted repeats, characteristic of transposons and insertion sequences. The precise end of the inverted repeat cannot be deduced from the DNA sequence alone as the four terminal bases form the palindrome, TATA. Transposon insertions are always flanked by a direct repeat of duplicated target sequences generated by the staggered overhang of target phosphate groups during insertion [12]. For Correia elements, it is therefore impossible to distinguish between the following three scenarios: 1, 23 bp inverted repeat with TATA target duplication; 2, 25 bp inverted repeat with TA duplication and 3, 27 bp inverted repeat with no target duplication (i.e. the Correia element is not a transposon).

To define the end of the Correia element unambiguously we searched for 'empty' target sites, i.e. sites occupied by a Correia element in one species of *Neisseria* but unoccupied in another. A total of 104 insertions were identified in the N.m. genome sequences that were absent from the corresponding loci in N.g. Comparison of the sequences revealed that in all cases a TA dinucleotide is duplicated on insertion (e.g. Fig. 1C). Thus, the full-length Correia element is 152 bp long with 25 bp terminal inverted repeats (Fig. 1D). The inverted repeats on the right hand end of the α and β families are identical except for a point mutation at bp 25. The left end of the α family has four point mutations, whereas the left end of the β family has a 1 bp deletion. We note that the terminal inverted repeat structure and the TA target site duplication have been described already in less detail [9,10,13]. However, the larger data set and more extensive alignments in the present analysis made the existence of the four families of Correia elements obvious.

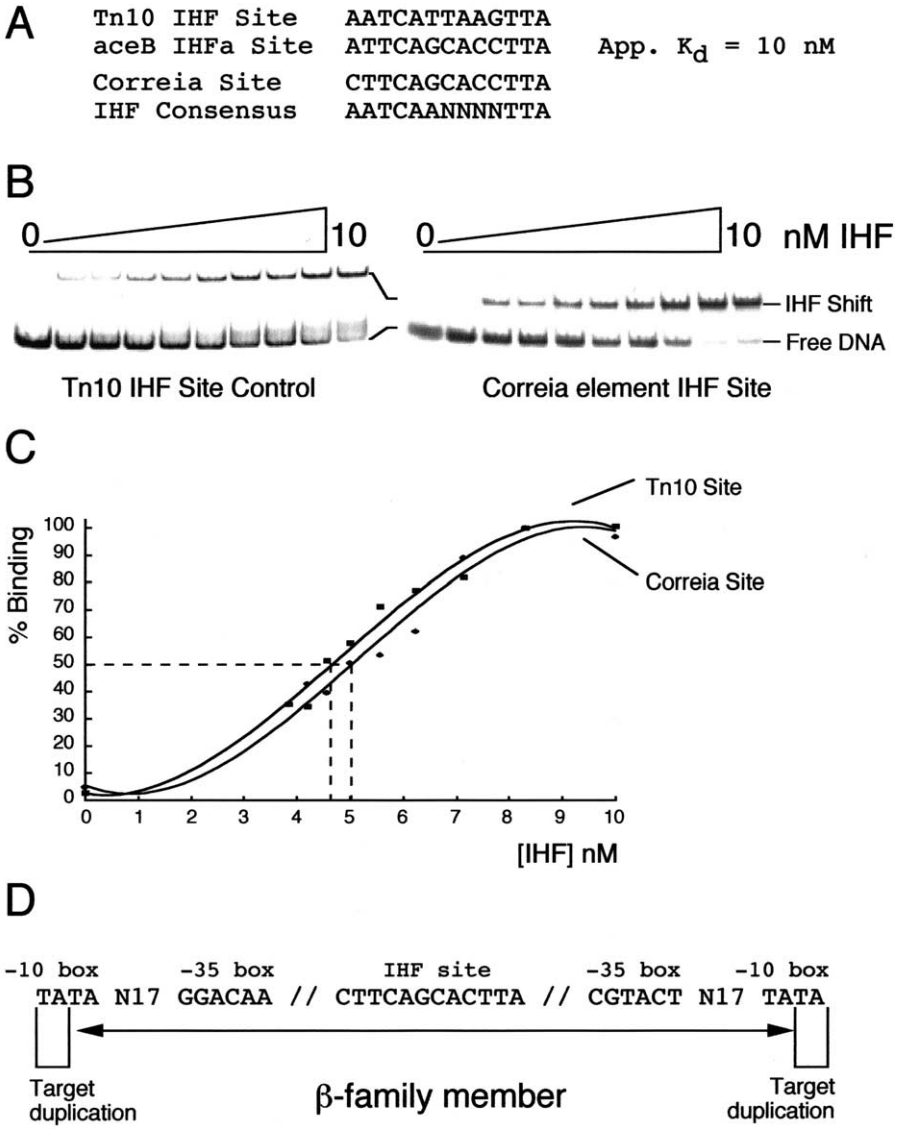

Fig. 2. Functional IHF binding site and sequence of predicted divergent promoters. A: The DNA sequence of Correia element IHF site is provided together with other selected sites. B: Gel shift assay for the function of the Correia IHF site. C: The gel from (B) was quantified using a phosphorimager and the data plotted to estimate the apparent K_d. D: The location of −10 and −35 boxes of divergent promoters predicted at the Correia element ends.

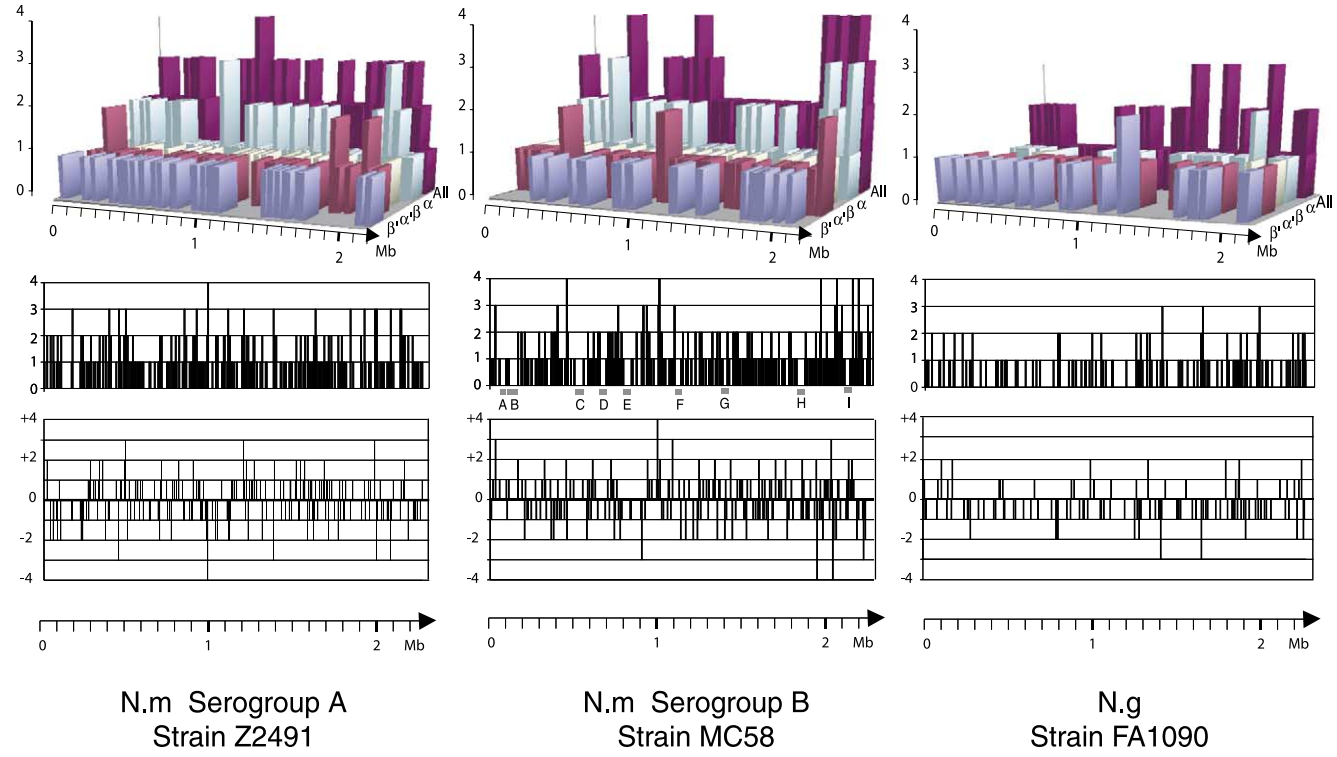

Fig. 3. Genomic distribution and SBI for Correia elements in three strains of *Neisseria*. The top row of panels plots the genomic distribution of the four families of Correia elements using a 5 kb window. The middle row of panels plots the genomic distribution of all Correia elements in the respective strains. The bottom row of panels plots the SBI (see text) of all Correia elements in the respective strains.

3.3. Correia elements contain an IHF binding site

Close examination of the full-length Correia element revealed a DNA sequence identical except for a single point mutation to the IHFa binding site in the *E. coli aceB* promoter (Fig. 2A) [14]. This site is present in the α and β families of Correia elements but is absent from the deletion derivatives, α′ and β′. The IHF site has three point mutations with respect to the IHF consensus (Fig. 2A) and a gel shift assay was therefore performed to confirm whether this site is functional (Fig. 2B). The well characterised IHF binding site from the outside end of Tn*10* was included as a control. The apparent K_d for both sites was estimated to be approximately 5 nM from the plot of the gel shift data (Fig. 2C). No IHF-dependent gel shift was detected for an α′ family member at high IHF concentrations (data not shown). We therefore conclude that the IHF binding site is located at the predicted position within the 50 bp deletion present in the α′ and β′ family members.

In *E. coli* the concentration of free IHF available to bind a given specific site is 15–35 nM [15]. If the concentration in *Neisseria* is at least 5 nM, we can conclude that the Correia element IHF site will probably be occupied because of the functional conservation of IHF in Gram-negative bacteria. Indeed, IHF function is conserved to such an extent that the individual subunits of the IHF heterodimer from N.g. complement the respective subunits from *E. coli* [16].

3.4. Divergent promoters in the Correia element ends

Further analysis of the DNA sequence predicts that there may be divergent promoters directing transcription out from both ends of the Correia element (Fig. 2D). It is common for transposons to influence transcription of nearby genes [17]. Tn*10*, for example, has a constitutive promoter driving transcription out into the target DNA [18]. Alternatively, if there is a −35 box hexamer in the end of the transposon, and the site of insertion places it the correct distance from a cryptic −10 box in the target DNA, a new promoter will be created [17]. In this regard Correia elements have a high potential for creating functional promoters because the terminal TA dinucleotide together with the duplication of the target TA dinucleotide on insertion always creates a perfect TATA box at each end of the element (Fig. 2D). Previously, promoter activity has been detected from the left end of a β′ family member [19]. However, experiments will be required to determine whether both ends of the α and β families are functional as promoters in vivo and whether the absence of the IHF site in the α′ and β′ family members affects regulation of transcription. We also note that the −35 box at the left end of the β family member shown in Fig. 2D matches five out of the six bases in the −35 box of the IHF-regulated $ilvp_G$ promoter [20].

3.5. Transposition of Correia elements

Duplication of the TA dinucleotide target site demonstrates that Correia elements spread by a transposition mechanism (above). In bacteria, transposons typically occupy 1–4% of the genome with any particular family of elements represented by one or a few copies [12]. Transposases are always poorly expressed and are sometimes tightly regulated to prevent a positive feedback of transposition events by *trans*-activation of elements as the transposase gene copy number rises, e.g. [21]. Correia elements are clearly not autonomous and the uncou-

pling of the element from the cognate transposase is likely to be critical for genome stability in the face of so many insertions.

Direct transposition of the 152 bp Correia elements may seem unlikely because of the requirement to synapse the ends of the transposon prior to the first step of the reaction. However, the presence of the IHF site in the middle of the element provides a mechanism to facilitate synapsis. The centre of the 180° IHF-induced bend [22] is at bp 78 of the canonical 152 bp Correia element. This location, almost exactly in the middle of the element, will therefore bring the ends of the element into an almost-perfect alignment.

3.6. Genomic distribution and strand bias of Correia elements

The distributions of the four families of Correia elements were plotted using a 5 kb window (Fig. 3, top row of panels). Group C N.m. was not included because the genome sequence has not been assembled yet. All four families of elements are evenly distributed throughout the genomes, although N.g. has significantly fewer elements than either strain of N.m. When all four families of elements are combined in a single plot (Fig. 3, middle row of panels), the distribution remains essentially random with several small but obvious gaps (see below).

Next, the SBI of the Correia element distribution was calculated using a 5 kb window (Section 2). The SBI quantifies the orientation of Correia element insertions with respect to each other and the genome sequence as a whole. For example, if two elements in the same window have the same orientation the strand bias is either +2 or −2, but if they have opposite orientations the strand bias is zero. There was no significant strand bias of insertions on a genome wide scale (Fig. 3, bottom row of panels). However, at the local level there is a bias for insertions in the same orientation. For example, in MC58 there are seven windows that each contain four Correia elements (Fig. 3, middle panel). Of these, three have the maximum strand bias value of four (Fig. 3, bottom panel).

Local strand biases have implications for homologous recombination and DNA repair where the orientation of repeat sequences dictates the relative rates of deletions and inversions. In *E. coli* the primary function of the homologous recombination proteins is to repair double strand breaks and stalled replication forks. Reciprocal homologous exchanges are rare in *E. coli* but may be more frequent in *Neisseria* sp. in which these functions are required for natural transformation. Reciprocal exchange between inverted repeats produces inversions, whereas exchange between direct repeats causes deletion of the intervening DNA. The prevalence of directly repeated Correia elements in the neisserial genome sequences suggests that there are mechanisms that preclude intramolecular homologous exchange.

3.7. Correia-less regions of the genome correspond with islands of horizontal transfer (IHTs) in MC58

The genomic distribution map for Correia elements in the group B strain of N.m. contains several small but obvious gaps (designated as regions A–I in Fig. 3, middle panel). Some of these correspond to previously described IHTs and critical determinants of pathogenicity [6]. IHT-A (NMB0066–0074, 0091–0100), IHT-B (NMB0498–0521) and IHT-C (NMB1746–1775) correspond to Correia-less regions A (44 kb), C (54 kb) and H (50 kb), respectively. Note that these Correia-less regions are much larger than the corresponding IHTs. This can be attributed to the low average density of Correia elements (approximately 1 per 10 kb) in the N.m. genome. We therefore ignored Correia-less regions less than about 30 kb.

There is a growing appreciation of the importance of horizontal transfer as a mechanism of environmental adaptation, especially in pathogenic bacteria. The main tools for identification of such regions are GC content and dinucleotide signature. However, these may be misleading if these parameters are identical in the donor and recipient. Another more recently developed tool for whole genome analysis is wavelet theory. This has been used to identify regions of horizontal transfer in various bacteria including N.m. [23].

To determine whether the absence of Correia elements is a good diagnostic test of idiosyncratic regions of the genome we examined the remaining Correia-less regions in more detail. The unusual nature of region B (57 kb) was immediately apparent as it harboured the ribosomal DNA. Region D (37 kb) contains genes for various housekeeping functions and an interrupted 10 kb section of DNA with 33% GC content. Region E (50 kb) contains two sections (9 and 6 kb) of very low GC content (33%); the first section contains several restriction/modification genes, whereas the second contains hypothetical proteins of unknown function. Region F (38 kb) appears to be a defective Mu-like prophage. Region G (29 kb) has a 12 kb tract of DNA with 33% GC content. Some of the annotated genes in this region appear to be duplicated and disrupted copies of a frpA/C-like gene. Finally, region I (37 kb) contains a tract of DNA with 40% GC (NMB2007–NMB2017) along with housekeeping genes, genes for hypothetical proteins and genes for the lipopolysaccharide coat.

All of the Correia-less regions examined contained tracts of DNA that were clearly idiosyncratic to the genome as a whole. They appear to harbour specific classes of genes often acquired by horizontal transfer, i.e. prophage, restriction systems and pathogenicity genes. Other common features of these regions are a high density of hypothetical genes of unknown function, duplications and DNA with low GC content. Since Correia elements insert at a TA dinucleotide, low GC content should favour insertions. It therefore appears that these regions were acquired after the spread of Correia elements in the genome. We conclude that the absence of Correia elements is an excellent diagnostic for 'interesting' regions of the genome. This may be a useful tool in the comparison of pathogenic and commensal strains and could be implemented using standard hybridisation protocols.

3.8. Phylogenetic analysis suggests genetic exchange between Correia element copies

The α′ and β′ families of Correia elements differ from the α and β families because of an identical 50 bp internal deletion. This raised the question of whether the deletion arose early in the evolutionary history of the elements and was spread by transposition, or whether the deletion derivatives are produced by ongoing genetic exchange. To investigate this question we searched for cases in which different family members are present at identical locations in different strain backgrounds. The *rpmJ* genes in Z2491 and MC58 have β and β′ Correia element insertions at precisely the same nucleotide (Fig. 4A). This undoubtedly represents a single transposition event as the probability of two elements inserting at the same TA dinucleotide is very small. An even more informative in-

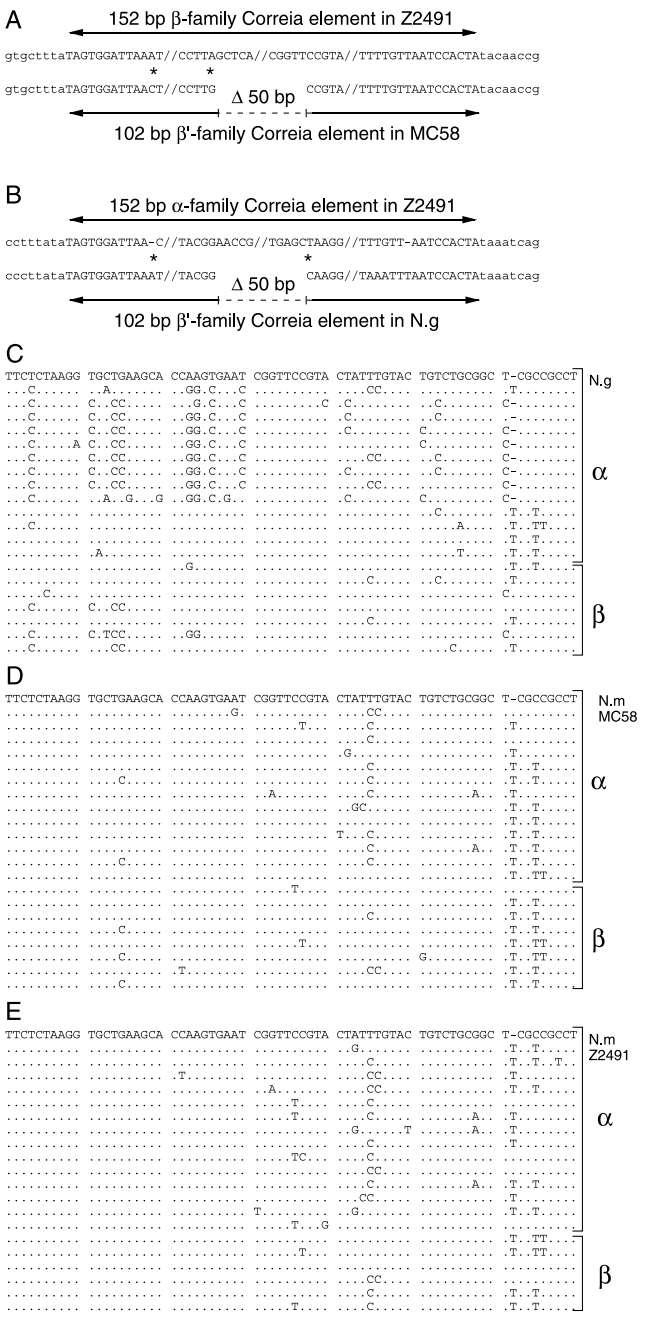

Fig. 4. Evidence for the level of genetic exchange between Correia elements in different strains. A: DNA sequence of β and β′ family members found at exactly the same location in Z2491 and MC58. B: DNA sequence of α and β′ family members found at exactly the same location in Z2491 and N.g. C: DNA sequence alignment of the central region of representative Correia elements from the α and β families in N.g. (D and E) as in (C) but from N.m. MC58 and Z2491, respectively.

sertion is present in the *sspA* genes of N.g. and Z2491 (Fig. 4B). In Z2491 there is an α family member which has been converted to a β′ in N.g. by acquisition of the internal deletion and the point mutations in the left end of the element.

To further investigate genetic exchange between different copies of the Correia elements in MC58, a phylogenetic tree was constructed (109 informative sites). Bootstrap analysis (1000 tests) did not support any of the nodes (values less than 30%). In addition, no consensus tree was found after parsimony analysis. Similar results were obtained for Correia elements in N.g. (89 informative sites). If the four families of Correia elements arose by accumulation of point mutations and spread through the genome by transposition, it would predict a phylogenetic tree with four branches. The absence of a phylogenetic tree supports genetic exchange amongst the elements. Indeed, this view is supported by visual inspection of the sequence alignments where it is apparent that the elements are connected by a network of mutations (Fig. 4C–E). In this network, individual or linked groups of mutations are not correlated with the different families of elements but are shared within and between the α and β families in an apparently random way. We can therefore conclude that there has been a high level of genetic exchange among the Correia elements within strains.

The most likely mechanisms driving the genetic exchange are homologous recombination and gene conversion. However, homologous recombination generates concomitant deletions and inversions depending on the relative orientation of the participating sites, e.g. [12]. As this would tend to scramble the genome, we conclude that the most likely mechanism for genetic exchange amongst Correia elements is gene conversion associated with double strand break repair and the recovery of replication forks stalled at DNA lesions.

3.9. *Correia element invasion appears to predate the divergence of N.m. and N.g. which are genetically isolated*

N.g., N.m. Z2491 and MC58 have more than 100 Correia elements in common, i.e. elements inserted at identical target sites. This raises the question of whether these insertions were present before the divergence of N.g. and N.m., or whether they are the products of ongoing genetic exchange. Inspection of the network of mutations in each of the three strains provides a clear answer (Fig. 4C–E). In the N.m. strains Z2491 and MC58 the networks of mutations are almost identical. In N.g., however, the network is entirely different (compare Fig. 4C with Fig. 4D,E). This indicates that the elements in N.m. are engaged in an ongoing process of genetic exchange between strains. N.g., on the other hand, appears to be isolated from the community of genetic exchange inhabited by the N.m. strains. If N.g. is indeed isolated, then Correia elements must have invaded the genome before the divergence of N.g. and N.m.

Genetic exchange between pathogenic microorganism is an important factor in determining levels of virulence. The network of mutations in Correia elements is therefore a potentially useful marker to survey the extent of horizontal exchange between members of the *Neisseria* community.

Acknowledgements: We thank Sven Sewitz for technical help. R.C. is a Royal Society University Research Fellow. This work was funded by grants from The Wellcome Trust to R.C.

References

[1] Smith, N.H., Holmes, E.C., Donovan, G.M., Carpenter, G.A. and Spratt, B.G. (1999) Mol. Biol. Evol. 16, 773–783.
[2] Feil, E.J. and Spratt, B.G. (2001) Annu. Rev. Microbiol. 55, 561–590.
[3] Maiden, M.C. (1998) Clin. Infect. Dis. 27 (Suppl. 1), S12–S20.
[4] Kriz, P., Giorgini, D., Musilek, M., Larribe, M. and Taha, M.K. (1999) Res. Microbiol. 150, 273–280.
[5] Parkhill, J. et al. (2000) Nature 404, 502–506.
[6] Tettelin, H. et al. (2000) Science 287, 1809–1815.

[7] Barten, R. and Meyer, T.F. (2001) Mol. Gen. Genet. 264, 691–701.
[8] Howell-Adams, B. and Seifert, H.S. (2000) Mol. Microbiol. 37, 1146–1158.
[9] Correia, F.F., Inouye, S. and Inouye, M. (1986) J. Bacteriol. 167, 1009–1015.
[10] Correia, F.F., Inouye, S. and Inouye, M. (1988) J. Biol. Chem. 263, 12194–12198.
[11] Crellin, P. and Chalmers, R. (2001) EMBO J. 20, 3882–3891.
[12] Chalmers, R. and Blot, M. (1999) in: Organization of the Prokaryotic Genome (Charlebois, R.L., Ed.), pp. 151–169, American Society for Microbiology, Washington, DC.
[13] Mazzone, M., De Gregorio, E., Lavitola, A., Pagliarulo, C., Alifano, P. and Di Nocera, P.P. (2001) Gene 278, 211–222.
[14] Resnik, E., Pan, B., Ramani, N., Freundlich, M. and LaPorte, D.C. (1996) J. Bacteriol. 178, 2715–2717.
[15] Yang, S.W. and Nash, H.A. (1995) EMBO J. 14, 6292–6300.
[16] Hill, S.A., Belland, R.J. and Wilson, J. (1998) Gene 215, 303–310.
[17] Mahillon, J. and Chandler, M. (1998) Microbiol. Mol. Biol. Rev. 62, 725–774.
[18] Simons, R.W., Hoopes, B.C., McClure, W.R. and Kleckner, N. (1983) Cell 34, 673–682.
[19] Black, C.G., Fyfe, J.A. and Davies, J.K. (1995) J. Bacteriol. 177, 1952–1958.
[20] Parekh, B.S. and Hatfield, G.W. (1996) Proc. Natl. Acad. Sci. USA 93, 1173–1177.
[21] Kleckner, N. (1989) in: Mobile DNA (Berg, D.E. and Howe, M.M., Eds.), pp. 227–268, American Society for Microbiology, Washington, DC.
[22] Rice, P.A., Yang, S., Mizuuchi, K. and Nash, H.A. (1996) Cell 87, 1295–1306.
[23] Lio, P. and Vannucci, M. (2000) Bioinformatics 16, 932–940.