Genome Comparison In Silico in Neisseria Suggests Integration of Filamentous Bacteriophages by their Own Transposase

Mikihiko Kawai,1,2,3 Ikuo Uchiyama,4 and Ichizo Kobayashi1,2,5,*

Department of Medical Genome Sciences, Graduate School of Frontier Science, The University of Tokyo, Japan1, Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan2, Division of Pathology, Immunology and Microbiology, Graduate School of Medicine, The University of Tokyo, Japan3, Research Center for Computational Science, National Institutes of Natural Sciences, Nishigonoka 38, Myodaiji, Okazaki 444-8585, Japan1, and Graduate Program in Biophysics and Biochemistry, Graduate School of Science, The University of Tokyo, Japan5

(Received 24 August 2005; revised 24 November 2005; published online 22 February 2006)

Abstract

We have identified filamentous prophages, Nf (Neisserial filamentous phages), during an in silico genome comparison in Neisseria. Comparison of three genomes of Neisseria meningitidis and one of Neisseria gonorrhoeae revealed four subtypes of Nf. Eleven intact copies are located at different loci in the four genomes. Each intact copy of Nf is flanked by duplication of 5'-CT and, at its right end, carries a transposase homologue (pivNM/irg) of RNaseH/Retroviral integrase superfamily. The phylogeny of these putative transposases and that of phage-related proteins on Nfs are congruent. Following circularization of Nfs, a promoter-like sequence forms. The sequence at the junction of these predicted circular forms (5'-atCTtatat) was found in a related plasmid (pMU1) at a corresponding locus. Several structural variants of Nfs—partially inverted, internally deleted and truncated—were also identified. The partial inversion seems to be a product of site-specific recombination between two 5'-CTtat sequences that are in inverse orientation, one at its end and the other upstream of pivNM/irg. Formation of internally deleted variants probably proceeded through replicative transposition that also involved two 5'-CTtat sequences. We concluded that the PivNM/Irg transposase on Nfs integrated their circular forms into the chromosomal 5'-CT-containing sequences and probably mediated the above rearrangements.

Key words: transposase; filamentous bacteriophage; integration; prophage; genome comparison

1. Introduction

Genome comparison involving closely related organisms can be useful for understanding the mechanisms of genome rearrangements and for identifying elements that participate in them.1,2 During examination of a large chromosomal rearrangement that was seen when a comparison was made between two genomes of Neisseria meningitidis,3 we encountered filamentous prophages.

The filamentous bacteriophages have a single-stranded circular DNA genome.4 Some of them, such as M13, propagate in a double-stranded circular form in their host bacteria,4 whereas others, such as CTXφ of Vibrio cholerae, integrate themselves into the host chromosome.5 So far, their integration is known to be mediated by one of two types of tyrosine recombinases: integration into a dif-like site by the host-encoded XerC/D recombinase6 and integration into a tRNA gene by the phage-encoded recombinase.7 Quite different from these tyrosine recombinases in sequence and molecular mechanism are the pivNM/irg genes3,8–10 that belong to the Piv subfamily of the IS110/IS492 transposase family of RNaseH/Retroviral Integrase superfamily.9,11 The irg genes in Neisseria gonorrhoeae were regarded as the transposases of small insertion sequence (IS) elements,10 which were found linked with two homologues of filamentous phage proteins and were thought to be

* Communicated by Kenta Nakai

To whom correspondence should be addressed. Tel. +81-3-5449-5326, Fax. +81-3-5449-5422, E-mail: ikobaya@ims.u-tokyo.ac.jp


The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use, please contact journals.permissions@oxfordjournals.org
integrated into the host chromosome as part of the bacteriophage genome.\textsuperscript{10}

In the present work, we identified prophages of this family in four \textit{Neisseria} genomes and named this phage family \textit{Nf} (\textit{Neisseria} filamentous phages). We found that each intact copy of \textit{Nf} carries a \textit{preNM} irg homologue at their right end. Our further analysis strongly indicates that this phage family is integrated by a novel mechanism using its own PivNM/Irg transposase. This transposase is probably responsible for variously rearranged prophage genomes. Recently, Bille et al.\textsuperscript{12} published an experimental work that was in agreement with ours (see Section 3.6).

2. Materials and Methods

2.1. Accession numbers and gene names

Accession numbers of the complete genomes are as follows: \textit{N. meningitidis} serogroup A strain Z2491 (\textit{NmeA}) (NC\textunderscore 003116.1),\textsuperscript{8} \textit{N. meningitidis} serogroup B strain MC58 (\textit{NmeB}) (NC\textunderscore 003112.1),\textsuperscript{3} \textit{N. gonorrhoeae} strain FA1090 (\textit{Ngo}) (NC\textunderscore 002946.2) (http://www.genome.ou.edu/gono.html; http://www.ncbi.nlm.nih.gov/genomes/framik.cgi?db=genome&gi=635). The genome sequence of \textit{N. meningitidis} serogroup C strain FAM18 (\textit{Nmc}) was obtained from The Sanger Institute (ftp://ftp.sanger.ac.uk/pub/pathogens/nmc/). Accession numbers for the sequences of plasmids, filamentous phages and IS are as follows: pJS-B (NC\textunderscore 004758.1),\textsuperscript{13} pJTPS1 (NC\textunderscore 001399.1),\textsuperscript{14} pMU1 (NC\textunderscore 007093.1),\textsuperscript{15} M13 (NC\textunderscore 003287.2),\textsuperscript{16} Pf1 (NC\textunderscore 001331.1),\textsuperscript{17} P3 (NC\textunderscore 001418.1),\textsuperscript{18} VGJ\textsuperscript{3} (NC\textunderscore 004736.1),\textsuperscript{19} B5 (NC\textunderscore 003460.1),\textsuperscript{20} IS\textsuperscript{621}(AB097054-AB097056).\textsuperscript{9}

We used the entries of the complete genome sequences of \textit{NmeA}, \textit{NmeB} and \textit{Ngo} in the NCBI RefSeq (ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/) for gene names, amino acid sequences and coordinates. Genes of \textit{Ns} in \textit{Nmc} are predicted by GeneMark.hmm for Prokaryotes\textsuperscript{21} (Version 2.4; http://opal.biology.gatech.edu/GeneMark/gmhmmb_prok.cgi), with the nucleotide sequence of \textit{Ns} (from the left position to the right position, as shown in Table 1) as input sequences and \textit{N. meningitidis} as species.

2.2. Bioinformatic analysis

The searches for macroscopic genome rearrangements and conserved gene clusters were performed by CGAT\textsuperscript{22} and MBGD\textsuperscript{23} (http://mbgd.genome.ad.jp/), respectively. CGAT was used to detect nucleotide sequence homology of a locus with the other parts of the genome. MBGD was used to detect homologous open reading frames (ORFs) and their neighbouring genes.

A homology search was performed by BLAST\textsuperscript{24} and fasta2 package.\textsuperscript{25} Multiple sequence alignments were constructed by ClustalW.\textsuperscript{26} Default parameters were used unless otherwise specified. A phylogenetic tree was constructed using ClustalW, NJplot\textsuperscript{27} and TreeViewPPC (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html).

3. Results and Discussion

3.1. Filamentous prophages in the genomes of \textit{Neisseria}

3.1.1. Nomenclature and terminology of \textit{Nf} prophages

In Table 1, we have listed all \textit{Nf} prophages that were found in the four \textit{Neisseria} genomes.

First, we describe the nomenclature of \textit{Nf} prophages. \textit{NF-G1(del)} may be used as an example of our nomenclature of \textit{Nf} prophages. Here (i) ‘\textit{Nf}’ stands for \textit{Neisseria} filamentous phage, (ii) ‘4’ stands for subtype (subtype 1–4, as classified in Section 3.1.3), (iii) ‘G’ stands for strain (A for \textit{NmeA}, B for \textit{NmeB}, C for \textit{NmcC} and G for \textit{Ngo}), (iv) ‘1’ stands for ordinal number within a subtype that is found within the same genome (a smaller number was given to the copy with smaller genome coordinates; for exceptions, see footnotes f–h of Table 1), and (v) ‘(del)’ stands for internal deletion. Partially inverted, truncated and internally deleted copies are designated as ‘(inv)’, ‘(tr)’ and ‘(del)’, respectively. The putative intact copies are not given such a description.

We followed the following terminology. ‘A copy of \textit{Nf}’ means a stretch of DNA sequence that encodes at least one gene that is homologous to the genes of ‘consensus, intact’ \textit{Ns}. By ‘the left/right of \textit{Ns}’, we mean the left/right of \textit{Ns}, as shown in Fig. 1A. ‘Intact’ copies are those that encode a potentially full set of homologues of phage genes that are oriented in the same direction and carry the expected left and right junction sequences (see Section 3.2 for a determination of the junctions). An only ‘partially inverted’ copy \textit{[NF-G2(inv)]} resembles the intact copies, except for partial inversion of its \textit{irg} gene (\textit{irg2}) region. ‘Truncated’ copies of \textit{Ns} are the copies with truncation of the right end or the left end. Three of the \textit{irg} copies (\textit{irg1}, 4 and 8) are not neighboured by the homologues of the above conserved genes.\textsuperscript{10} However, three regions containing these \textit{irg} genes have the expected left and right junctions (see Section 3.3.2), so we regarded them as ‘internally deleted’ copies of \textit{Ns} and named them \textit{NF-G1(del)}, \textit{NF-G4(del)} and \textit{NF-G8(del)} (Fig. 1A).

We used Courier font for nucleotide and amino acid sequences for clarity.

3.1.2. \textit{Nf} prophages in four \textit{Neisseria} genomes

A large chromosomal rearrangement is seen as a comparison between two genomes of \textit{N. meningitidis}, \textit{NmeA} and \textit{NmeB}.\textsuperscript{3} During examination of this rearrangement, we encountered homologues of filamentous phages. We also found several homologous prophages on the other loci. We named these prophages \textit{Ns} (for \textit{Neisseria} filamentous phages).
By investigating four sequenced *Neisseria* genomes (*Nme* A, *Nme* B, *Nme* C and *Ngo*), we identified 23 copies of Nf, including 11 intact, 1 partially inverted, 8 truncated and 3 internally deleted copies (Table 1). Two right-truncated homologues were located at the corresponding loci in three *N. meningitidis* genomes but not in the *N. gonorrhoeae* genome—that is, Nf2-A1(tr), Nf2-B1(tr) and Nf2-C1(tr) are at one corresponding locus and Nf2-A2(tr), Nf2-B2(tr) and Nf2-C2(tr) are at the other locus. So, there are 23 copies of Nf at 19 different loci in the four genomes. Notably, all the 11 intact and the 1 partially inverted copies are located at 12 different loci.

All the 11 intact copies and the 1 partially inverted copy of Nfs are /C24 8 kb in length (Table 1), with gene

<table>
<thead>
<tr>
<th>Straina</th>
<th>Subtype</th>
<th>Nameb</th>
<th>Length (bp)</th>
<th>Positionc</th>
<th>Strand</th>
<th>ORFs contained</th>
<th>pivNM/irg containedd</th>
<th>Locuse</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>NmeA 1</td>
<td>Nf1-A</td>
<td>8029</td>
<td>37.1872</td>
<td>17 45.200</td>
<td>+</td>
<td>NMA1792-1800</td>
<td>NMA1800</td>
<td></td>
<td>Right side truncated</td>
</tr>
<tr>
<td>NmeA 2</td>
<td>Nf2-A1(tr)</td>
<td>3555</td>
<td>1118.972</td>
<td>11.22.526</td>
<td>+</td>
<td>NMA1167-1173</td>
<td>–</td>
<td>I</td>
<td>Right side truncated</td>
</tr>
<tr>
<td>NmeA 3</td>
<td>Nf3-A</td>
<td>7996</td>
<td>7.69.557</td>
<td>7.61.562</td>
<td>–</td>
<td>NMA0783-0772</td>
<td>pivNM3(NMA0772)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NmeB 1</td>
<td>Nf1-B1</td>
<td>8695</td>
<td>16.00.454</td>
<td>16.09.148</td>
<td>+</td>
<td>NMB1542-1552</td>
<td>pivNM1-A(NMB1552)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NmeB 2</td>
<td>Nf2-B1(tr)</td>
<td>3551</td>
<td>9.88.033</td>
<td>9.91.583</td>
<td>+</td>
<td>NMB0971-0975</td>
<td>–</td>
<td>(I)</td>
<td>Right side truncated</td>
</tr>
<tr>
<td>NmeB 2</td>
<td>Nf2-B2(tr)</td>
<td>3552</td>
<td>5.05.188</td>
<td>5.01.637</td>
<td>–</td>
<td>NMB0485-0480</td>
<td>–</td>
<td>II</td>
<td>Right side truncated</td>
</tr>
<tr>
<td>NmeB 2</td>
<td>Nf2-B3</td>
<td>7649</td>
<td>18.25.043</td>
<td>18.32.693</td>
<td>+</td>
<td>NMB1742-1750</td>
<td>pivNM2(NMB1750)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NmeC 1</td>
<td>Nf1-C1</td>
<td>8046</td>
<td>33.529</td>
<td>25.248</td>
<td>–</td>
<td>Not annotated</td>
<td>[26.583-26.527]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NmeC 1</td>
<td>Nf1-C2</td>
<td>8046</td>
<td>2.79.961</td>
<td>2.88.006</td>
<td>+</td>
<td>Not annotated</td>
<td>[2.86.906-2.87.862]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NmeC 1</td>
<td>Nf1-C3</td>
<td>8010</td>
<td>17.43.242</td>
<td>17.51.251</td>
<td>+</td>
<td>Not annotated</td>
<td>[17.50.152-17.51.108]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NmeC 1</td>
<td>Nf1-C4</td>
<td>8032</td>
<td>18.93.109</td>
<td>19.01.140</td>
<td>–</td>
<td>Not annotated</td>
<td>[19.00.040-19.00.996]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NmeC 2</td>
<td>Nf2-C1(tr)</td>
<td>3556</td>
<td>9.52.056</td>
<td>9.55.611</td>
<td>+</td>
<td>Not annotated</td>
<td>–</td>
<td>I</td>
<td>Right side truncated</td>
</tr>
<tr>
<td>NmeC 2</td>
<td>Nf2-C2(tr)</td>
<td>3551</td>
<td>16.96.828</td>
<td>17.00.378</td>
<td>+</td>
<td>Not annotated</td>
<td>–</td>
<td>II</td>
<td>Right side truncated</td>
</tr>
<tr>
<td>Ngo 1</td>
<td>Nf1-G(tr)</td>
<td>2956</td>
<td>16.03.726</td>
<td>16.06.681</td>
<td>+</td>
<td>NGO1646-1648</td>
<td>irg7(NGO1648)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ngo 4</td>
<td>Nf4-G1(del)</td>
<td>1100</td>
<td>7.67.755</td>
<td>7.68.854</td>
<td>+</td>
<td>NGO00773</td>
<td>irg1(NGO00773)</td>
<td>An internal region deleted</td>
<td></td>
</tr>
<tr>
<td>Ngo 4</td>
<td>Nf4-G2(inv)</td>
<td>8176</td>
<td>10.88.351</td>
<td>10.80.176</td>
<td>–</td>
<td>NGO11147-1137</td>
<td>irg2(NGO11137)</td>
<td>A region containing irg2 is inverted</td>
<td></td>
</tr>
<tr>
<td>Ngo 4</td>
<td>Nf4-G3</td>
<td>7975</td>
<td>11.10.860</td>
<td>11.02.886</td>
<td>–</td>
<td>NGO1170-1164</td>
<td>irg3(NGO1164)</td>
<td>An internal region deleted</td>
<td></td>
</tr>
<tr>
<td>Ngo 4</td>
<td>Nf4-G4(del)</td>
<td>1174</td>
<td>11.48.077</td>
<td>11.46.904</td>
<td>–</td>
<td>NGO1200</td>
<td>irg4(NGO1200)</td>
<td>An internal region deleted</td>
<td></td>
</tr>
<tr>
<td>Ngo 4</td>
<td>Nf4-G5</td>
<td>7964</td>
<td>12.23.799</td>
<td>12.15.836</td>
<td>–</td>
<td>NGO1270-1262</td>
<td>irg5(NGO1262)</td>
<td>Left side truncated; head-to-head with Nf1-G(tr) separated by 73 bp</td>
<td></td>
</tr>
<tr>
<td>Ngo 4</td>
<td>Nf4-G6(tr)</td>
<td>4406</td>
<td>16.03.652</td>
<td>15.99.247</td>
<td>–</td>
<td>NGO1645-1641</td>
<td>irg6(NGO1641)</td>
<td>An internal region deleted</td>
<td></td>
</tr>
<tr>
<td>Ngo 4</td>
<td>Nf4-G8(del)b</td>
<td>1141</td>
<td>16.59.009</td>
<td>16.60.149</td>
<td>+</td>
<td>NGO1703</td>
<td>irg8(NGO1703)</td>
<td>An internal region deleted</td>
<td></td>
</tr>
</tbody>
</table>

Copies are sorted by strain, subtype and name.  
*a* *Nme* A: *N. meningitidis* Z2491 (serogroup A).  *Nme* B: *N. meningitidis* MC58 (serogroup B).  *Nme* C: *N. meningitidis* FAM18 (serogroup C).  *Ngo*: *N. gonorrhoeae* FA1090.  
*b*Nfs not specified as ‘(tr)’, ‘(del)’ nor ‘(inv)’ represent intact copies.  
*c*The genome coordinates indicate their left end (T next to the left direct repeat, CTtatat) and their right end (T next to the right direct repeat, aTCT).  
*d*pivNM of *Nme* C are expressed in genome coordinates annotated by ‘GeneMark.hmm for Prokaryote’.

Each number indicates corresponding locus. Each Nf not numbered is located at a unique locus. A rearranged locus is in parentheses.  
*f*No ordinal number is given to the single copy of the subtype in the genome.  
*g*The ordinal number of the Nf at the corresponding locus in *Nme* A (see locus column) is given.  
*h'S' stands for containing irg8.

By investigating four sequenced *Neisseria* genomes (*Nme* A, *Nme* B, *Nme* C and *Ngo*), we identified 23 copies of Nf, including 11 intact, 1 partially inverted, 8 truncated and 3 internally deleted copies (Table 1). Two right-truncated homologues were located at the corresponding loci in three *N. meningitidis* genomes but not in the *N. gonorrhoeae* genome—that is, Nf2-A1(tr), Nf2-B1(tr) and Nf2-C1(tr) are at one corresponding locus and Nf2-A2(tr), Nf2-B2(tr) and Nf2-C2(tr) are at the other locus. So, there are 23 copies of Nf at 19 different loci in the four genomes. Notably, all the 11 intact and the 1 partially inverted copies are located at 12 different loci. All the 11 intact copies and the 1 partially inverted copy of Nfs are ~8 kb in length (Table 1), with gene
Figure 1. Nf filamentous prophages on Neisseria chromosome, other filamentous bacteriophages and related plasmids. In the box are shown keys to gene families (the corresponding gene of M13 in parentheses) and DNA features (upper case letters, CT, indicate direct repeats flanking Nfs). Gene names are written above the arrows for colored genes. The alternative names for pivNM/irg are shown in parentheses. Note that members listed in (B) and (C) have been shown to be present in circular DNA form. (A) Nf prophages located at different loci in Neisseria genomes. The left half brackets indicate subtypes. Shown are Nfs located at different loci (Table 1) on four strains—Nme A Z2491, Nme B MC58, Nme C FAM18 and Ngo FA1090. Genes of Nf1-C1, C2, C3 and C4 are annotated by GeneMark.hmm for Prokaryotes. 21 Nf2-B1(tr) and Nf2-C1(tr) are not shown because they are located at the same locus and are quite similar to Nf2-A1(tr), as are Nf2-B2(tr) and Nf2-C2(tr) with Nf2-A2(tr) (Table 1). Nomenclature of Nf is described in Section 3.1.1 and the footnote for Table 1. Supplementary data is available at www.dnaresearch.oxfordjournals.org for details about colored genes and about the cg stem-loop region, of which the consensus sequence is 5’-cccccctnnnctaay-aggggggg-3’.

392 Prophages Integrated by their Transposase [Vol. 12,
organization being similar to that of filamentous phages (Fig. 1A versus B), except for the presence of a pivNM/irg transposase gene homologue at the right end (Fig. 1A).

As well as previously noticed homologues of the rstA protein of filamentous phage CTXφ (for phage DNA replication; corresponding to gII of M13) (NMA1792 and other brown arrows in Fig. 1A) and gI protein (for assembly) (NMA1799 and other light-blue arrows in Fig. 1A),10 we found other Nf genes corresponding to filamentous phage genes (gV, gVIII, gIII and gVI) by in silico research (Fig. 1; Supplementary Figure S1 is available at www.dnaresearch.oxfordjournals.org). In Fig. 1, these functionally corresponding, but not necessarily homologous, genes are shown with the same color.

Genes that are conserved among Nfs are oriented in the same direction (Fig. 1A), except for NGO1137 (irg2) of Nf4-G2(inv) (as discussed in Section 3.3.1). We assumed that their sense strand represents their single-strand (plus strand) within the presumed virion.

Plasmids that have genes that are homologous to those of Nf are also shown (Fig. 1C). pJS-B13 from N. meningitidis ET37, and pJTPS114 from Ralstonia solanacearum, do not carry a pivNM/irg homologue, whereas pMU115 from Eikenella corrodens does.

3.1.3. Classification of Nf into four subtypes We classified the Nfs into four subtypes, Nf1–Nf4, based on similarity of their three long ORFs: phage DNA replication protein homologues (corresponding to pII; homologous with rstA of CTXφ; brown arrows in Fig. 1A), adsorption protein homologues (corresponding to pIII; previously annotated as TspB) (yellow arrows in Fig. 1A) and pI homologues (light-blue arrows in Fig. 1A) [Fig. 2A, (1)–(3)], and by overall nucleotide sequence similarity (data not shown).

The only one exception to a clear classification of sequence similarity was seen for the DNA replication protein homologue of Nf4-G2(inv). The branch of Nf4-G2(inv), which was located differently from those of the other Nf4 copies, can be explained by the assumption that Nf4-G2(inv) might have resulted from recombination between an Nf3-like copy and an Nf4 copy. In fact, the right 5942 bp part of Nf4-G2(inv) is similar to the corresponding part, if any, of the other Nf4 copies (data not shown), whereas its left part contains three regions: the left junction (see Section 3.2.2), the phage DNA replication protein (NGO1146) gene [Fig. 2A, (1)] and the cg stem-loop (Supplementary Figure S1E is available at www.dnaresearch.oxfordjournals.org), all of which are highly similar to those of Nf3-A but not to those of other Nfs.

3.2. Junctions of integrated Nf prophages and the host genome

Except for the right junction of Nf2 and Nf3, all the junctions were determined by alignment of copies of the same subtype, as detailed below (Sections 3.2.1–3.2.5).
To determine the right junctions of Nf2 and Nf3, we took advantage of the formation of a promoter-like sequence by joining the right and the left ends of Nfs, as described below (Section 3.2.6).

### 3.2.1. Sequences flanking Nfs

The dRS3 repeats constitute one family of repeated sequences in *Neisseria* and have a consensus sequence of `attcccnmnmnmgggaat.` Alignment of Nf1 copies (as shown in Fig. 3A) can be interpreted as Nfs being able to target 5'-CT in specific members of the dRS3 repeat family, `attccc(g/a)cCT(g/a)cgccg(g/a)aa(t/g),` and were integrated forming flanking direct repeats of 5'-CT. The consensus sequence that was targeted by Nf1 (5'-atccccgcCTgcggggaat) is the most abundant sequence of the dRS3 repeats (291 out of 672 copies in NmEA).

For the other three subtypes, we could not find any obvious similarity or pattern in the flanking sequences, although some sequences were observed several times—that is, `aaac` flanks the left of two Nf2 copies and two Nf4 copies and `ctcga` flanks the right of four Nf4 copies (Fig. 3B and C).

#### 3.2.2. Left junction

From a comparison of the left junction of three groups—Nf1 (Fig. 3A), Nf2 and Nf4 [Fig. 3B, (1)], and Nf3-A and Nf4-G2(inv) [Fig. 3B, (2)]—we concluded that the sequence 5'-CTtatat represents the left junction sequence for all of the Nfs examined [Fig. 3B, (3)].

#### 3.2.3. Right junction of Nf1 and Nf4

From a comparison of the right junction sequences of Nf1 (Fig. 3A) and Nf4 [Fig. 3C, (1)], we found that the right junction sequence that is common to Nf1 and Nf4 is 5'-atCTtatat (Fig. 3C, (2)). The "T" of the atCTtatat of Nf4 is the residue that is reported to be at the 3' end of aligned downstream regions of *irg* genes.

#### 3.2.4. Predicted circular form of Nf1 and Nf4

As filamentous phages have a circular phage genome, it is reasonable to assume that Nf1 and Nf4 would have a circular phage genome before integration. From the comparison of the common left and right junctions of Nf1 and Nf4 [Fig. 3D, (1)], the sequence at the junction of the hypothetical circular form is predicted to be 5'-atCTtatat [Fig. 3D, (2)].

Notably, we found that pMU1, which has the homologous gene set with NF (see Section 3.1.2 and Fig. 1C), has this predicted junction sequence (5'-atCTtatat) at the corresponding locus—that is, between the *pivNM*/*irg* homologue and the replication protein homologue [at position 3757–3749 bp; Fig. 3D, (3)].

#### 3.2.5. Assembly of a promoter-like sequence by circularization of Nf1 and Nf4

The sequence that is shared between the right end of Nf1 and Nf4 (5'-ttgaca, a typical −35 sequence of the −35 element of eubacterial promoters that is recognized by primary σ factors (boxed in solid, thick line in Fig. 3E). Because the predicted junction of the hypothetical circular form of Nf, atCTtatatntnt, contains a −10 promoter-like sequence (boxed with dotted line in Fig. 3E), we assumed that the junction region forms a promoter in the circular form.

The related plasmid pMU1 (see Fig. 1C) also carries this promoter-like sequence (Fig. 3E). The *pivNM*/*irg* homologue of pMU1 is most similar to the *pivNM*/*irg* of Nfs so far.

The transposase gene of IS621 is another close homologue of *pivNM*/*irg.* The left end of IS621 is CTgtatat and the right end is atCT (Fig. 3E), the former being similar to the left end of Nfs and the latter being identical to their right end. IS621 would also form a promoter-like sequence after assumed circularization (Fig. 3E).

The promoter-like sequence of the predicted circular form of Nfs and pMU1 could regulate the expression of phage replication gene, although some Nfs code another short ORF between the promoter-like sequence and this gene [e.g. Nf1-B1, Nf1-B2, Nf3-A and Nf4-G2(inv); Fig. 1A].

#### 3.2.6. Predicted right junction of Nf2 and Nf3

We cannot determine the right junction of Nf2 and Nf3 by alignment, as with Nf1 and Nf4, because Nf2-B3 and Nf3-A are the single examples of the Nf2 and Nf3 subtypes, respectively, that have an intact right junction (Fig. 1A).

Instead, we assumed that both Nf2-B3 and Nf3-A would have an atCT sequence at the right junction, as Nf1 and Nf4 copies. Because the right junction sequence atCT of Nf1 and Nf4 is located at 142 and 129 bp, respectively, downstream of the stop codon of *pivNM*/*irg* (Fig. 3E), we searched the range of 250 bp downstream of the stop codon of *pivNM* in Nf2-B3 and Nf3-A and found atCT at only one site in Nf2-B3 and at two sites in Nf3-A (Fig. 3F). Next, we looked for a potential −35 element upstream of these three candidate sequences and found a −35-like ttgann sequence in two of them (boxed with solid line in Fig. 3F). From these, we concluded that the right junction of Nf2-B3 and Nf3-A is atCT that is located 162 and 145 bp, respectively, downstream of the *pivNM* gene.

#### 3.2.7. Junction sequences common to four Nf subtypes

Altogether, we concluded that Nfs are flanked with CT at both ends. The common sequence at the left end of NF is CTatatat, and at that the right end is atCT, as seen in Fig. 3D, (1).

#### 3.3. Rearranged Nfs

#### 3.3.1. A partially inverted copy of Nf [Nf4-G2(inv)]

The *irg*2 gene of Nf4-G2(inv), a transposase homologue, is inverted relative to the other three genes.
and the remaining Nf4 homology (Fig. 1A). From a comparison with the intact copies of Nf4, the recombination sites of this partial inversion are identified as the 5 bp sequence (boxed in alignment of Fig. 4A). One recombination site, CTtat, is a part of the right junction CTTatat (an open triangle and underlined in Fig. 4A), which is composed of the right direct repeat (CT) and the flanking 5 bp host sequence (tatat; shown in italics). The other recombination site, ataAG (complement is CTTatat), is located between irg2 and...
Figure 4. Rearranged Nfs. Sequence alignments are shown for the relevant regions. A filled triangle, a filled circle and an inverted filled triangle indicate three conserved motifs of Nf, the left junction, the right junction, and a 5 bp sequence that is complementary to the left junction as in Figs 1 and 3. A black arrow indicates a \textit{pctNM/irg} homologue. (A) Nf with a partial inversion [Nf4-G2(inv)]. A box in thick line indicates the 5 bp junction sequence of the inversion. The right junction of Nf4-G2(inv), CTtatat (underlined and marked with an open triangle) is composed of the right direct repeat (CT) that is shared by all Nfs and the flanking host sequence (tatat; shown in italics). The dotted underline of the hypothetical parental form indicates the sequence that is identical to the predicted junction sequence of hypothetical circular Nf. (B) Nfs with an internal deletion. (1) Nf4-G4(del). Comparison with the left junction and the ataAG region of Nf4-G3 and Nf4-G5. Inverted arrows indicate inverted repeats. Dotted arrows indicate the homologous regions between the left end of intact Nf4 and the complement of Nf4-G4(del). (2) Nf4-G1(del) and Nf4-G8(del). Boxes in dashed line (ccct) and boxes in dotted line (tgcgt) indicate short repeats that are probably responsible for deletions that generated Nf4-G8(del) and Nf4-G1(del), respectively. (3) Possible steps of formation of Nf4-G4(del). An open circle with a dotted line indicates a hypothetical site that the transposase recognized. The thick arrow indicates a sequence that is involved in the replicative transposition event to form long inverted repeats. Boxes in dashed line with dots indicate short repeats that are probably responsible for deletion.
the other genes of Nf4-G2(inv) (Fig. 4A). This sequence is complementary to the 5’ bp sequence of the left common junction sequence of Nf, _CTtat_, and is located 10, 8 and 9 bp upstream of the start codon of _pioNM/_irg_ in Nf1, Nf2 and Nf4, respectively. Nf3 has a similar sequence (_taaAG_) at 10 bp upstream. The _taaAG_ sequence partially overlaps with the ribosome binding site _agagg_, according to Skaar et al. 

This inversion structure is perfectly consistent with conservative site-specific recombination involving the 5 bp sequences. The inversion was probably mediated by PivNM/Irg itself on an integrated form of a presumed parental Nf4-G2 (the second line of Fig. 4A) because the right junction (atCT) of the parental Nf4-G2 sequence, as well as the flanking host sequence (tatat) is the same as the predicted junction of the hypothetical circular form of Nfs, _CTtatatat_ (underlined with a dotted line in Fig. 4A). Alternatively, the inversion could have occurred on the circular form of an Nf4 member. If so, the partially inverted circular form may have been integrated into the chromosome by Irg2-mediated recombination between the junction sequence ..._ccCTtatatat... [Fig. 4A and 3B, (2)] and a chromosomal sequence ..._caCTtatagc... [Fig. 3B, (2) and 4A].

3.3.2. _Internally deleted copies of Nf_ The _irg_1, 4 and 8 genes are the only exceptions of the _pioNM/_irg_ homologues that are not neighboured by the homologues of phage genes at the left (Fig. 1A). The region downstream of these _irg_ genes to atCT, the right junction of Nfs, is highly homologous to that in the other Nf4 copies [Fig. 3C, (1)]. At a variable distance upstream from the _irg_ genes, we identified the left junction sequence of Nfs, _CTtatat_ [Fig. 4B, (1) and (2)]. We named these regions bound by these junction sequences: Nf4-G1(del), Nf4-G4(del) and Nf4-G8(del).

Indeed, the left 33 bp sequence of Nf4-G4(del) (solid, rightward arrow) aligned well with the left end of the intact Nf4 copies (Nf4-G3 and Nf4-G5) [Fig. 4B, (1)]. Notably, the complementary strand of the sequence just upstream of _irg_4 aligned longer with the left end of intact Nf4 [dotted arrows in Fig. 4B, (1)]. In Nf4-G4(del), the left end of the 33 bp sequence (_CTtatat..._) and the 34 bp sequence upstream of _irg_ [..._taaAG_; solid, leftward arrow in Nf4-G4(del) of Fig. 4B, (1)] form inverted repeats (33/34 bp match). Note that the right junction (ataaAG; boxed and indicated by an inverted triangle in Fig. 4B, (1)) is the same site as the recombination sequence for partial inversion (see Section 3.3.1 and Fig. 4A).

Another putative deletion derivative, Nf4-G8(del), aligned very well with Nf4-G4(del) with a deletion [Fig. 4B, (2)]. Nf4-G1(del) aligned well with these two deletion derivatives, [Nf4-G4(del) and Nf4-G8(del)], and with the intact Nf4s (Nf4-G3 and Nf4-G5) with a deletion [Fig. 4B, (2)].

Formation of Nf4-G8(del) is readily explained by illegitimate recombination involving short direct repeats, _ccct_ (boxed in dotted line), in an Nf4-G4(del)-type parent [Fig. 4B, (2)]. Deletions stimulated by palindromic sequences were shown to occur between short direct repeats in the laboratory. 

Similarly, Nf4-G1(del) would be formed through illegitimate recombination involving short direct repeats, _tgct_ (boxed in dotted line), from these types of deletion derivatives [Nf4-G4(del) and Nf4-G8(del)] or from the intact Nf4 copies [Fig. 4B, (2)]. As the parent, we prefer the structure of Nf4-G4(del), or similar, with long inverted repeats, as the inverted repeats will stimulate and select this type of deletion.

Formation of Nf4-G4(del) from intact Nf4 copies can be explained by the following replicative transposition model. (i) A copy of a region of the left terminus (longer than 33 bp) is inserted into the left of the _ataaAG_ sequence upstream of the transposase gene. This is a reaction that is similar to replicative transposition. The transposase may have recognized a secondary site within Nf as in one-ended transposition. (ii) The resulting long inverted repeats make the entire region unstable and cause or select an illegitimate recombination event involving a short sequence identity (_agagg_). This mechanism is the same as we proposed for the other two deletion derivatives. This model can provide a uniform explanation to the formation of the three deletion derivatives.

Although these three copies are probably deletion derivatives of intact Nfs, the possibility remains that they are integrated into these sites after the deletion event. Unfortunately, we do not know how much of the Nf sequence is necessary _in cis_ for integration or whether these three copies can express a functional transposase. The deletion in Nf4-G1(del) apparently removed the first residue of the start codon of the _irg_ gene [Fig. 4B, (2)] and resulted in the start codon of NGO0772 ( _irg1_ ) to be 33 bp upstream from the left end of Nf4-G1(del).

3.4. _Simple insertion revealed by comparison with another genome without Nf_

Comparison of Nf1-A integrated into NgoA with the corresponding unoccupied locus of Ngo (Fig. 5A) made it clear that the target sequence of Nf1-A integration is CT. Unfortunately, with the other intact copies of Nf1, such an obvious assignment was not possible because of the variability of the dRS3 repeats.

Comparison of one of the internally deleted copies, Nf4-G1(del), in Ngo with the corresponding unoccupied locus in NgoA strongly indicated that it also targeted CT (Fig. 5B). According to the comparison between Ngo and NgoA, the target sequence of NgoA-G8(del) also seems to be CT, although the two base pairs that are 3’ to CT are different from those in the corresponding unoccupied locus. For the other copies, comparison is not
3.5. Alternative models for Nf integration

In some of the above arguments (see Section 3.3), we implicitly assumed that PivNM/Irg transposase that is present on the Nf phage genome is likely to have mediated integration of the phage genome into the chromosome. In this section, we will list the alternative models and provide evidence against them.

3.5.1. Four possible models The first model we have been pursuing (A) is that the Nf phage was integrated by its own pivNM/irg transposase (transposase-as-integrase model). Three other models are as follows: (B) Nf without pivNM/irg was integrated by some unspecified mechanism and later the pivNM/irg was inserted as IS (phage-then-IS model); (C) pivNM/irg was inserted into the Nf phage genome and later the phage genome with pivNM/irg was integrated by some unspecified mechanism unrelated to pivNM/irg (phage-with-IS model) and (D) the pivNM/irg was inserted into the chromosome as IS and later, by some unspecified mechanism, Nf without pivNM/irg was inserted between the left junction of the IS and the start codon of pivNM/irg (IS-then-phage model). An idea similar to the phage-with-IS model was presented earlier.10

3.5.2. Evidences for or/and against the models Below, the transposase-as-integrase model (model A) is discussed in comparison with the other three models (models B–D).

(i) Models B–D provide no explanation as to how the phage was integrated into the chromosome. On the other hand, model A is straightforward by explaining that pivNM/irg transposase acts as the integrase for integration of Nf.

(ii) Tight linkage of phage-related genes of the Nf and pivNM/irg gene is a straightforward consequence of models A and C, but is not a straightforward consequence of models B and D. The apparently solitary irg genes [Nf4-G1(del), Nf4-G4(del) and Nf4-G8(del)] can be explained by deletion from an intact copy of Nf as detailed above (see Section 3.3.2). This supports model A. These deletion events may have happened after the integration of Nf, although the possibility that the deletion derivatives were integrated as simple IS cannot be excluded (see Section 3.3.2). Thus, we cannot exclude model D from this argument, but it is difficult to explain, by model D, why Nfs without pivNM/irg specifically target the fixed point between the left junction of ISpivNM/irg and the start codon of pivNM/irg and why they are not integrated to the other loci of Neisseria genomes.

(iii) Similarity of the junction sequences of Nf and IS621, the second closest homologue of pivNM/irg of Nf (see Section 3.2.5, supports the idea that both junctions of Nf were derived from the activity of PivNM/Irg and supports models A and D but not the other two models.

(iv) The closest homologue of pivNM/irg is the one that is coded on pMU1. pMU1 has the pivNM/irg homologue, along with other homologues of Nf phage genes (see Section 3.1.2 above and Fig. 1C). pMU1 encodes pivNM/irg and has an atCTtatat sequence, which is exactly the predicted junction sequence for the hypothetical circular form of Nf at the corresponding locus (see Section 3.2.4 above and Fig. 3E). This form resembles an intermediate form of transposition of several IS families, such as the IS3 family,33,34 Nf and pMU1, along with IS621, have the assembled promoter-like sequence (see Section 3.2.5 and Fig. 3). IS192 of the same IS110/IS492 family as pivNM/irg takes a circular form with the junction consisting of a single copy of the direct repeats and with a promoter-like sequence.35 These points argue for the possibility that Nfs are likely to have a circular form together with pivNM/irg and that the predicted junction sequence ‘CT’ is similar to that of pMU1 and support models A and D but not the other two models.

(v) Each of the subtypes, Nf1–Nf4, carries a distinct type of pivNM/irg gene (pivNM1/irg7 for Nf1, pivNM2 for Nf2, pivNM3 for Nf3 and irg1-6, 8 for Nf4), as revealed by phylogenetic analyses (Fig. 2A compared with 2B). This means that the copies of one subtype of Nf are linked with one specific subtype of pivNM/irg and vice versa; this is compatible with models A and C. On the other hand, in models B and D, there is no obvious reason to assume that copies of the same subtype of Nf carry pivNM/irg genes of the same type.
(vi) Amino acid sequences of phage-related genes of Nf2 and Nf4 are similar [Fig. 2A, (1)–(3)], but those of their transposases, pivNM2 and irg, are not similar to each other (Fig. 2B). Nucleotide sequences of the central regions of Nf2 and Nf4 are also similar but those of pivNM2 and irg are diverged (data not shown). This might support models B and D, which indicate that there are different origins for phages and pivNM/irg. However, if we assume that the pivNM/irg of each subtype was somehow inserted independently into Nf (Nf without pivNM/irg), models A and C would be supported. pJS-B and pJTPS1 have similar gene organizations to that of Nf, but do not encode the pivNM/irg homologue nor have an atCTtatat junction sequence (Fig. 1C). Independent insertion of pivNM/irg into the circular form of Nf without pivNM/irg (like pJS-B) might have occurred during evolution. The mechanism of this insertion might be different from that of Nf integration into bacterial chromosomes because Nf integration is marked by flanking CT repeats, whereas we could not find such duplication for pivNM/irg insertion into hypothetical Nf without pivNM/irg, although we cannot exclude the possibility that loss of the CT sequence was selected.

From the above observations and considerations, we concluded that the most likely mechanism is the transposase-as-integrase model (model A). The circular form of Nf before integration into (and after excision, if any, from) the chromosome would have an atCTtatat sequence at the junction and recombine with the target CT of the chromosome by the activity of the PivNM/Irg transposase. Although this reaction mechanism is similar to that of site-specific recombinase, Piv family proteins do not have amino acid motifs that are conserved among the tyrosine- or serine-recombinase families, but are similar to DDE transposases.

This mechanism is distinct from the two currently known mechanisms for filamentous phage integration, integration into a dif-like site mediated by XerC/D site-specific tyrosine-recombinase and integration into tRNA probably mediated by an integrase of the tyrosine-recombinase family.

3.6. Comparison with the work by Bille et al.\textsuperscript{12}

After this work was completed, Bille et al.\textsuperscript{12} published an experimental work of which some results are in agreement with those that we described above. Their results related to ours are listed below [(i)–(vi)], whereas the findings that are specific to our work are also listed [(a)–(e)].

(i) Discovery of filamentous phages (corresponding to seven intact copies of Nf1, according to our naming) in the genomes of *Neisseria* based on observations of their low G+C content, homology of the first ORF to bacteriophage replication proteins and the similarity to the order and size of ORFs that are encoded by other filamentous phages. (ii) Those phages (Nf1) are inserted into dRS3. (iii) Extrachromosomal circular single-stranded DNA of one prophage (Nf1-A, according to our naming) was detected by polymerase chain reaction (PCR). (iv) Junction of the circular form was sequenced. (v) The region containing the pivNM gene was contained in the circular form, as judged from the length of PCR products (supporting our result that pivNM/irg genes are integral part of Nfs). (vi) The inactivation of four genes—NMA1792, 1797 and 1799 on Nf1-A and pilQ on a locus of the host chromosome—had predicted effects on extracellular and cytoplasmic circular DNA forms.

Our findings that are not overlapping with theirs are as follows: (a) we found and characterized not only the Nf1 subtype, but also three other subtypes (Nf2, Nf3 and Nf4). (b) The left and right sequences common to all the four Nf subtypes are CTtatat and CTCT, respectively. (c) Our analyses strongly indicate that the integration of Nf is mediated by its own PivNM/Irg transposase (transposase-as-integrase model). (d) We identified and characterized structural variants of Nfs. We were able to explain their origin in terms of aberrant activity of this transposase. (e) We also added further bioinformatic evidence that ORFs of Nfs correspond to proteins of filamentous phages.

Note that their work\textsuperscript{12} further supports the transposase-as-integrase model. Our results proved the power of bioinformatic approach through genome comparison for detailed and thorough analysis of mobile elements and genome rearrangements.

**Acknowledgements:** We thank Noriko Takahashi and Takeshi Tsuru for helpful suggestions on the manuscript. This work was supported by grants from MEXT of the Japanese government to I.K. (Genome Biology, Genome Homeostasis, DNA Repair, Kiban-evolution, Kiban-genome, 21COE; genome language).

**Supplementary Data:** Supplementary data are available online at www.dnaresearch.oxfordjournals.org.

**References**


