An in vitro strategy for the selective isolation of anomalous DNA from prokaryotic genomes

M. W. J. van Passel, A. Bart, R. J. A. Waaijer1, A. C. M. Luyf1, A. H. C. van Kampen1 and A. van der Ende*

Department of Medical Microbiology and 1Bioinformatics Laboratory, Academic Medical Center, Amsterdam, the Netherlands

Received May 19, 2004; Revised and Accepted July 28, 2004

ABSTRACT

In sequenced genomes of prokaryotes, anomalous DNA (aDNA) can be recognized, among others, by atypical clustering of dinucleotides. We hypothesized that atypical clustering of hexameric endonuclease recognition sites in aDNA allows the specific isolation of anomalous sequences in vitro. Clustering of endonuclease recognition sites in aDNA regions of eight published prokaryotic genome sequences was demonstrated. In silico digestion of the Neisseria meningitidis MC58 genome, using four selected endonucleases, revealed that out of 27 of the small fragments predicted (<5 kb), 21 were located in known genomic islands. Of the 24 calculated fragments (>300 bp and <5 kb), 22 met our criteria for aDNA, i.e. a high dinucleotide dissimilarity and/or aberrant GC content. The four enzymes also allowed the identification of aDNA fragments from the related Z2491 strain. Similarly, the sequenced genomes of three strains of Escherichia coli assessed by in silico digestion using Xbal yielded strain-specific sets of fragments of anomalous composition. In vitro applicability of the method was demonstrated by using adaptor-linked PCR, yielding the predicted fragments from the N.meningitidis MC58 genome. In conclusion, this strategy allows the selective isolation of aDNA from prokaryotic genomes by a simple restriction digest–amplification–cloning–sequencing scheme.

INTRODUCTION

Horizontal gene transfer (HGT) was already identified in 1944 by the same experiment that demonstrated the transformation of non-virulent to virulent Streptococcus pneumoniae (1). The extent of HGT as an evolutionary phenomenon had not been addressed quantitatively on genomic scale until Lawrence and Ochman (2) calculated that ∼18% of the genome of Escherichia coli MG1665 was horizontally transferred since its divergence from the Salmonella lineage 100 million years ago. This identified HGT as a major factor in prokaryotic genome evolution. Recently, an extensive database of horizontally transferred genes based on complete bacterial and archaeal genomes has been made available (3).

The rationale behind the computational identification of horizontally transferred DNA is the genome hypothesis, which proposes that for a given prokaryotic genus genomic DNA is relatively constant in codon usage and GC content (4,5). In contrast, horizontally acquired anomalous DNA differs in codon usage and/or GC composition from the recipient genome and can therefore be identified when substantial sequence information is available.

An additional parameter in lateral genomics is based on oligonucleotide compositional extremes: the dinucleotide relative abundance values or genome signature ρ* (6,7). The genome signature is constant among members of a genus, but deviates substantially between members of different genera (8). When used for intragenomic comparisons, ρ* makes an excellent parameter for the identification of anomalous DNA regions. Aberrant dinucleotide frequencies in aDNA are then expressed as the genome dissimilarity δ*, being the average dinucleotide relative abundance difference between the aDNA region and the whole genome (6–8). Although the genome signature is capable of identifying clusters of alien genes and acquired pathogenicity-associated islands (PAI) with an atypical nucleotide composition, highly expressed regions such as ribosomal clusters can also display aberrant dinucleotide frequencies (8,9).

Till date, to our knowledge, no method exists that uses (one of) these parameters and enables the selective isolation of anomalous DNA sequences from a microbial genome in vitro. In order to develop such a technique, we investigated a special group of oligonucleotide composition extremes: the local overrepresentation in a genome of palindromic hexanucleotide sequences, specifically restriction endonuclease recognition sites, in aDNA regions. Like the genomic dinucleotide and trinucleotide frequencies (10,11), frequencies of restriction sites vary between the genomes of different microbial species (12). Avoidance of cognate recognition sequences is probably the operating mechanism (13,14). An HGT event between different organisms may introduce clusters of certain restriction sites in the recipient’s genome. Therefore, digestion of the chromosomal DNA with such a restriction endonuclease can produce a limited number of small restriction fragments, comprising potential anomalous DNA, which can be selectively amplified by adaptor-linked

*To whom correspondence should be addressed. Tel: +31 20 5664862; Fax: +31 20 6979271; Email: a.vanderende@amc.uva.nl
PCR [ALP (15)]. The resulting amplicons can be subsequently subcloned and identified by sequence analysis.

Clustering of restriction endonuclease recognition sites in diverse aDNA regions in prokaryotic genomes was illustrated by the in silico assessment of seven genome sequences of five different species. The restriction enzymes for which the hexameric recognition sites are underrepresented were identified for each genome, and restriction fragments between clustered sites, being <5 kb, were analysed for nucleotide composition concerning GC percentage and genomic dissimilarity.

Next, the restriction fragments of *Neisseria meningitidis* MC58 between 300 bp and 5 kb were analysed in silico for both GC content and genome signature compared to the genomic values. Also, the restriction fragments obtained with the selected restriction endonucleases from *N.meningitidis* MC58 and Z2491 strains were compared.

Finally, in order to demonstrate the applicability of this technique in vitro, ALP was performed on chromosomal DNA from strain MC58 digested by each of the selected restriction endonucleases. The resulting amplicons were sequenced to verify the predicted sequence composition.

**MATERIALS AND METHODS**

**Bacterial strain and growth conditions**

*N.meningitidis* MC58 is a serogroup B:15:P1.7,16 strain isolated from a case of invasive infection in the UK (16). This wild-type MC58 strain lacks the erythromycin resistance cassette insertion in the capsule gene locus in contrast to the wild-type MC58 strain (17). Neisseriae were grown on heated blood (chocolate) agar plates or in liquid Tryptic Soy Broth (DIFCO) medium at 37°C.

**Chromosomal DNA preparation and digestion**

Chromosomal DNA was isolated with the Puregene DNA isolation kit (Biozym). Restriction digests and subsequent heat inactivation were carried out according to the manufacturer’s instructions (Roche).

**Adaptor-linked PCR and DNA sequencing**

Adaptor-linked PCR was performed as described previously (18). The adaptor and linker sets are MP19 (5'-ACG TCG ACT ATC CAT GAA CAG ATC 3') and MP23 (5'-GAT CTG TTC ATG-3') for the Scal-digested genomic template, MP24 (5'-ACC GAC GTC GAC TAT CCA TGA ACA-3') and MP20 (5'-CTA GTG TTC ATG -3') for both the Nhel- and SpeI-digested chromosomal DNA and MP24 and MP23 for the BglIII-digested genomic template. PCR amplicons were purified by agarose gel extraction (Qiagen) and subcloned into a pCR2.1 vector (Invitrogen) according to the manufacturer’s instructions. *E.coli* DH5α was transformed by standard heat shock procedure. The constructed plasmids were isolated with the Wizard Kit (Promega). Inserts were sequenced using standard M13 primers or primer walking on vector or genomic DNA according to the manufacturer’s instruction (ABI). Sequences were analysed using the Staden Package (http://www.mrc-lmb.cam.ac.uk/pubseq/).

**Software**

The restriction site frequency tables from the various genomes were obtained from http://tools.neb.com/~posfai/DEPARTED. The in silico digestions of the various sequenced genomes (for accession numbers see Table 1) were performed using the Restriction Digest tool from The Institute for Genomic Research (TIGR) (http://www.tigr.org). In silico retrieval and identification of the restriction fragments was performed with the Position Search/Segment Retrieval tool from TIGR (http://www.tigr.org). The different genomes of *N.meningitidis* were compared using the Artemis Comparison Tool (ACT) (http://www.sanger.ac.uk).

**Data analysis**

Fragments were designated anomalous in GC composition if the GC content of the fragment is below the fifth or above the 95th percentile of the genomic GC content distribution, calculated with a window and step size identical to the fragment length (http://www.tigr.org).

The δ* value for each restriction fragment was calculated as described earlier by Karlin and colleagues (7). In brief, the

**Table 1. Clustering of restriction enzyme recognition sites in anomalous DNA regions in sequenced genomes of various prokaryotes**

<table>
<thead>
<tr>
<th>Organism</th>
<th>Accession number (and reference)</th>
<th>Enzyme</th>
<th>Total number of fragments*</th>
<th>aDNA*</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Haemophilus influenzae</em> Rd20</td>
<td>NC000907 (29)</td>
<td>XmalI</td>
<td>7</td>
<td>7/7</td>
</tr>
<tr>
<td><em>E.coli</em> O157:H7 VT2</td>
<td>NC002695 (30)</td>
<td>Apal</td>
<td>7</td>
<td>7/7</td>
</tr>
<tr>
<td><em>E.coli</em> K-12</td>
<td>NC000913 (31)</td>
<td>AvrlI</td>
<td>5</td>
<td>5/5</td>
</tr>
<tr>
<td><em>E.coli</em> CFT073</td>
<td>NC004431 (32)</td>
<td>Xbal</td>
<td>3</td>
<td>2/3</td>
</tr>
<tr>
<td><em>Salmonella enterica</em> serovar Typhi CT18</td>
<td>NC003198 (33)</td>
<td>Xbal</td>
<td>2</td>
<td>1/1</td>
</tr>
<tr>
<td><em>Salmonella enterica</em> serovar Typhi CT18</td>
<td>NC003198 (33)</td>
<td>Xbal</td>
<td>7</td>
<td>4/5</td>
</tr>
<tr>
<td><em>Methanobacterium thermoautotrophicum</em> delta H</td>
<td>NC000916 (34)</td>
<td>Xbal</td>
<td>7</td>
<td>6/6</td>
</tr>
<tr>
<td><em>N.meningitidis</em> MC58</td>
<td>NC003112 (17)</td>
<td>SpeI</td>
<td>11</td>
<td>7/8</td>
</tr>
<tr>
<td><em>N.meningitidis</em> Z2491</td>
<td>NC003116 (35)</td>
<td>BglII</td>
<td>9</td>
<td>6/7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Scal</td>
<td>11</td>
<td>9/10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Spel</td>
<td>3</td>
<td>3/3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nhel</td>
<td>4</td>
<td>4/4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BglII</td>
<td>4</td>
<td>1/2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Scal</td>
<td>4</td>
<td>3/4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nhel</td>
<td>2</td>
<td>2/2</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td>67/74 (91%)</td>
<td></td>
</tr>
</tbody>
</table>

*a*All restriction fragments up to 5 kb are considered in this column.

*b*For aDNA composition calculations concerning GC percentage and genome dissimilarity, only the fragments between 300 bp and 5 kb were considered.
dinucleotide relative abundance values $p_{XY}$ are defined as the frequency of the dinucleotide XY divided by the product of the background frequencies of the individual nucleotides in the sequence and the reverse complement sequence [$p_{XY}^* = f_{XY}/(f_X \cdot f_Y)$]. $\delta^*$ is the average absolute dinucleotide relative abundance difference given by $\delta^*(f, g) = 1/16 \sum |p_{XY}^*(f) - p_{XY}^*(g)|$, where $p_{XY}^*(f)$ denotes the abundance values calculated for fragment $f$ and $p_{XY}^*(g)$ the abundance values calculated for the genome $g$. The $\delta^*$ of each fragment was compared to a distribution of $\delta^*$ values which we constructed for consecutive fragments of identical size obtained from the respective genome sequence. A fragment was scored positive for anomalous DNA composition if this $\delta^*$ value was above the 90th percentile of the $\delta^*$ distribution. For determination of genome signature dissimilarities, only sequences between 300 bp and 5 kb were considered. Five kilobase pairs is the median size of imported DNA in Neisseria (19), and it also represents a conservative size limit for technical convenience in amplification procedures. With restriction fragments <300 bp, computations of composition are not performed; this limit represents a conservative lower size limit previously used in studies identifying aDNA by codon usage (3,5).

RESULTS

Local overrepresentation of hexameric restriction enzyme recognition sites in anomalous DNA regions of sequenced prokaryotic genomes

In order to identify clustered restriction endonuclease recognition sites in the sequenced prokaryotic genomes used in this study, we tested restriction enzymes of which the hexaprimromatic recognition sites are underrepresented in the genome sequences (http://tools.neb.com/~posfai/FINISHED). The tendency of the recognition sites to cluster in aDNA regions was assessed by analysing the sequence composition of the restriction fragments obtained in silico (Table 1 and supplementary Tables 1–8). Fragments <300 bp were not considered for their genome dissimilarity and GC composition values, because $\delta^*$ values of these small fragments are unreliable; this conservative minimal length is also used by Lawrence and Ochman and Garcia-Vallvé and co-workers (3,5). Nevertheless, many of these restriction fragments <300 bp are adjacent to the other fragments <5 kb in their respective genomes (as an example see N.meningitidis MC58 in Table 2 and Figure 1). The results showed that the eight analysed genome sequences did contain clusters of endonuclease restriction sites in aDNA regions (Table 1). The aggregated data showed 74 fragments (lengths between 300 bp and 5 kb) of which 67 (91%) were of anomalous composition.

Clustering of hexameric restriction enzyme recognition sites in the genome of N.meningitidis MC58 in different aDNA regions

Assessment of the occurrence of low-frequency restriction sites in the genome sequence of N.meningitidis MC58 revealed that many of the recognition sites of BglII, Nhel, Scal and SpeI clustered in the four regions known to contain large stretches of aDNA. These are also annotated as islands of horizontal transfer (IHT), thereby supporting the notion that in MC58 these recognition sites are relatively overrepresented in regions originating from horizontal transfer (3,8,17) (Figure 1, Table 2).

Of the 27 restriction fragments <5 kb, 21 were located within either of the clusters of anomalous genes described by Karlin (8). Nevertheless, many of these restriction fragments <300 bp are adjacent to the other fragments <5 kb in their respective genomes (as an example see N.meningitidis MC58 in Table 2 and Figure 1). The results showed that the eight analysed genome sequences did contain clusters of endonuclease restriction sites in aDNA regions (Table 1). The aggregated data showed 74 fragments (lengths between 300 bp and 5 kb) of which 67 (91%) were of anomalous composition.

Table 2. Restriction fragment numbers, lengths, $\delta^*$ and GC composition of fragments obtained after in silico digestion of the genome of N.meningitidis MC58 by four selected restriction enzymes

<table>
<thead>
<tr>
<th>Fragments*</th>
<th>Enzyme No.</th>
<th>Length (bp)</th>
<th>GC content</th>
<th>$\delta^*$</th>
<th>Genomic dissimilarity</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$&lt;100th$</td>
<td>$&gt;90th$</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>percentile</td>
<td>percentile</td>
</tr>
<tr>
<td>BglII</td>
<td>1</td>
<td>2996</td>
<td>47</td>
<td>136</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>2889</td>
<td>49</td>
<td>117</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2889</td>
<td>49</td>
<td>117</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>2654</td>
<td>46</td>
<td>123</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>2461</td>
<td>51</td>
<td>136</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>1194</td>
<td>55</td>
<td>125</td>
<td>–</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>477</td>
<td>34</td>
<td>+</td>
<td>218</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>75</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>21</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>Nhel</td>
<td>1</td>
<td>4723</td>
<td>43</td>
<td>99</td>
<td>+</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>4392</td>
<td>40</td>
<td>+</td>
<td>111</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>787</td>
<td>35</td>
<td>+</td>
<td>123</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>670</td>
<td>29</td>
<td>+</td>
<td>282</td>
</tr>
<tr>
<td>Scal</td>
<td>1</td>
<td>4824</td>
<td>43</td>
<td>+</td>
<td>121</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>4452</td>
<td>48</td>
<td>+</td>
<td>142</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>2496</td>
<td>48</td>
<td>+</td>
<td>132</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>2179</td>
<td>57</td>
<td>+</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>865</td>
<td>38</td>
<td>+</td>
<td>171</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>699</td>
<td>38</td>
<td>+</td>
<td>218</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>600</td>
<td>50</td>
<td>–</td>
<td>184</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>600</td>
<td>50</td>
<td>–</td>
<td>184</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>600</td>
<td>50</td>
<td>–</td>
<td>182</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>533</td>
<td>51</td>
<td>+</td>
<td>192</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>67</td>
<td>ND</td>
<td>ND</td>
<td>ND</td>
</tr>
<tr>
<td>SpeI</td>
<td>1</td>
<td>1672</td>
<td>36</td>
<td>+</td>
<td>132</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>579</td>
<td>35</td>
<td>+</td>
<td>212</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>470</td>
<td>24</td>
<td>+</td>
<td>274</td>
</tr>
</tbody>
</table>

For fragments <300 bp the GC percentage and $\delta^*$ were not determined (ND).
*Out of 25 fragments, 17 were located within one of the anomalous gene clusters A, B or C described by Karlin (8).
Adjacent in anomalous gene cluster A.
Adjacent in anomalous gene cluster C.
Present in anomalous gene cluster B.
Adjacent in anomalous gene cluster A.
Present in anomalous gene cluster C.
Present in anomalous gene cluster A.
Adjacent in anomalous gene cluster C.
Adjacent in anomalous gene cluster B.
Comparing the different restriction fragment patterns of the sequenced N. meningitidis strains Z2491 and MC58 in silico

The restriction fragment patterns obtained in silico from the two different N. meningitidis strains showed remarkable differences (Table 3). Various restriction fragments located in the annotated anomalous gene clusters or IHTs in N. meningitidis MC58 were not identified in N. meningitidis Z2491, consistent with the notion that these IHTs are absent in strain Z2491 (17). In addition, two anomalous restriction fragments from MC58 (MC58-ScaI-2179 and MC58-SpeI-470), which were not part of one of the previously mentioned IHTs, were located in aDNA regions only present in MC58. The MC58-ScaI-2179 fragment harboured ORF NMB1829, encoding a TonB-dependent receptor, and MC58-SpeI-470 contained a cluster of six open reading frames (ORFs). The latter showed a number of features typical for a PAI (20), such as an atypical GC content compared to the genome sequence and association with a transfer RNA (tRNA) gene (NMB1595) and an insertion sequence IS1106 (ORF NMB1601) at its boundaries. As the functions of these ORFs and their distribution in other pathogenic and non-pathogenic strains are unknown, this region does not formally qualify as PAI, although a heterologous origin is suspected.

Two fragments identified in silico in Z2491 were absent in the genome of MC58. Z2491-BglII-97 harboured a part of NMA0604, which encodes a hypothetical protein. The Z2491-ScaI-4101 fragment contained ORFs NMA0785 and NMA0786, encoding hypothetical proteins. Both NMA0785 and NMA0786 display an atypical GC composition and dinucleotide composition, and are described as putatively horizontally transferred by Garcia-Vallvé and colleagues (3).

Thus, the same set of four enzymes which was used to isolate anomalous sequences from the MC58 strain in silico identified aDNA fragments from the related strain Z2491. Similarly, the sequenced genomes of three strains of E. coli assessed by in vitro digestion using XbaI yielded strain-specific sets of fragments with anomalous composition (supplementary Tables 2–4). Unfortunately, due to sequence ambiguities in the E. coli EDL933 genome sequence, the δ* values of the XbaI restriction fragments from this strain could not be readily calculated, although the low GC percentage of these fragments compared to the E. coli genomic GC composition values suggest an anomalous nucleotide composition (supplementary Table 5).

Selective isolation of aDNA in vitro from N. meningitidis MC58 by adaptor-linked PCR

In order to validate that this strategy could be converted into an in vitro strategy with possible applications to unsequenced genomes, chromosomal DNA of strain MC58 was digested in vitro with BglII, NheI, ScaI or SpeI. The fragments obtained from each of the four digests were amplified by ALP. The amplicon pattern is very similar to the expected in silico restriction fragment patterns (Figure 2), albeit the minor differences observed. These can be explained by the possible inefficient amplification of large fragments (~4 kb) in the presence of smaller fragments. The resulting amplicons were subcloned and sequenced, verifying the sequences predicted by the in silico analysis (data not shown). This demonstrated the applicability of this method in vitro.

DISCUSSION

A new parameter based on dinucleotide composition extremes has been introduced to identify genomic islands in complete genomes (6,7). The potential of this and other in silico methods to identify genomic islands is obviously limited to sequenced genomes. To our knowledge, no in vitro method exists which allows the selective isolation of aDNA from unsequenced genomes, except for subtractive hybridization strategies in which usually two related but different strains are compared (21–23). In order to develop an in vitro tool for the selective isolation of anomalous sequences from unsequenced genomes, we investigated whether clustering of
restriction enzyme recognition sites could lead to the preferential isolation of aDNA from various sequenced genomes.

We demonstrated clustering of genomically underrepresented restriction enzyme recognition sites in eight sequenced genomes of five prokaryotic species \textit{in silico}. We found that clustering of these recognition sites occurred predominantly in aDNA regions, including ribosomal loci, but also and more interestingly, in putative horizontally transferred loci which were described by Garcia-Vallvé and co-workers (3). However, some discrepancies between our data and their database exist, as the HGT database ignores non-coding sequences.

In the genome of \textit{N.meningitidis} MC58, the clustering of the four selected endonuclease recognition sites occurred in the three known IHTs (17), a recently described aDNA region (3), and also in smaller anomalous loci. Comparative analysis of the calculated restriction fragments from \textit{N.meningitidis} MC58 and Z2491 showed that similar putative horizontally acquired anomalous sequences could be isolated from both strains. Furthermore, aDNA confined to either one of these strains could be identified, suggesting that differences between strains can be identified and isolated. The strategy was validated by the \textit{in vitro} amplification of the predicted restriction fragments from the genome of \textit{N.meningitidis} MC58.

In this study, only a limited number of sequenced genomes was analysed to illustrate atypical clustering of endonuclease recognition sites in their respective aDNA regions. Theoretically, any prokaryotic genome may contain atypical clustering of endonuclease recognition sites. On the other hand, aDNA which is acquired via horizontal gene transfer is thought to adjust to the host’s nucleotide composition over time in a process called amelioration, the same mutational process that affects the entire genome (5). This implies that only aDNA resulting from evolutionary recent transfer events, in which the nucleotide content of the acquired DNA still differs substantially from the sequence composition of the host genome, can be adequately identified and isolated. Furthermore, in genomes of bacteria, such as \textit{Helicobacter pylori}, with an extreme plasticity due to high recombination rates, regions of aDNA may be rapidly obscured over time (24). Another potential limitation of our technique is that the restriction enzyme recognition sites may be methylated by

\begin{table}
\centering
\begin{tabular}{|l|l|l|l|l|l|l|}
\hline
Enzyme & Size Coordinates & Strain MC58 & Strain Z2491 & GI in strain MC58 & Presence in the other strain & \\
& Strain Z2491 & & & & & \\
\hline
BglII & — & 4213 & — & 1179114–1183327 & — & Dispersed in MC58, with inversions and loss of restriction fragment \hline
2996 & — & 525422–528418 & — & A & Absent in Z2491 \hline
2889 & — & 1863020–1865909 & — & C & Largely present in Z2491 \hline
2889 & — & 522533–525422 & — & A & Largely present in Z2491 \hline
2654 & — & 1860366–1863020 & — & C & Largely absent in Z2491 \hline
2461 & 2449 & 726967–729428 & 875412–877861 & — & Similar sequences \hline
1194 & — & 1859172–1860366 & — & C & Largely present in Z2491 \hline
477 & — & 614379–614856 & — & — & Absent in Z2491 \hline
— & 97 & 578316–578413 & — & — & Absent in MC58 \hline
75 & 75 & 542107–542182 & 688499–688574 & A & Similar sequences \hline
21 & — & 1444731–1444752 & — & B & Absent in Z2491 \hline
NheI & 4723 & — & 505027–509750 & A & Largely absent in Z2491 \hline
4392 & — & 1834407–1838799 & — & C & Absent in Z2491 \hline
787 & 776 & 2231113–2231900 & 299657–300433 & X & Similar sequences \hline
670 & 670 & 543262–543932 & 689470–690140 & A & Similar sequences \hline
Scal & 4824 & — & 526521–531345 & — & A & Largely absent in Z2491 \hline
4452 & — & 1598–510650 & — & A & Absent in Z2491 \hline
— & 4101 & — & 769714–773815 & — & Similar sequences, but polymorphism at the recognition site \hline
— & 3181 & — & 1928315–1931496 & — & Similar sequences, but polymorphism at the recognition site \hline
2496 & — & 1007047–1009543 & — & — & Largely absent in Z2491 \hline
2179 & — & 1925725–1927904 & — & — & Largely absent in Z2491 \hline
865 & 865 & 1815665–1816530 & 1927450–1928315 & C & Similar sequences \hline
699 & 699 & 1814966–1815665 & 1927651–1927450 & C & Similar sequences \hline
600 & — & 1447767–1448367 & — & B & Similar to Z2491-Scal-4101 \hline
600 & — & 1447167–1447767 & — & B & Similar to Z2491-Scal-4101 \hline
600 & — & 616671–617271 & — & — & Similar to Z2491-Scal-4101 \hline
533 & — & 1446567–1447100 & — & B & Similar to Z2491-Scal-4101 \hline
67 & — & 1447100–1447167 & — & B & Similar to Z2491-Scal-4101 \hline
Scal & 1672 & — & 2223993–2225665 & — & X & Dispersed in MC58, with inversions and loss of restriction fragment \hline
— & 4101 & — & — & — & Absent in Z2491 \hline
— & 1443931–1444510 & — & B & Absent in Z2491 \hline
579 & — & 1443931–1444510 & — & B & Absent in Z2491 \hline
470 & — & 1659795–1660265 & — & — & Absent in Z2491 \hline
\hline
\end{tabular}
\caption{Restriction fragments, coordinates, of the different tested \textit{N.meningitidis} strains, indicating the absence or presence of the different fragments in the other strain (GI refers to the different genomic island as depicted in Figure 1)}
\end{table}
restriction-modification (RM) systems. For example, \textit{H.pylori} contains many RM enzymes, rendering the genome resistant to their activity (25).

Only hexapalindromic recognition sites of endonucleases have been tested; we did not examine other recognition sites (such as non-palindromic recognition sites). The cores of the restriction sites of the selected restriction enzymes persistently consist of the genomically underrepresented tetranucleotides previously described by Karlin and co-workers (10). Genomic underrepresentation of tetrapalindromes may be due to structural defects caused by these sequences or special functional roles associated with these sequences (10). Whether genomic aDNA regions, with these tetranucleotides overrepresented, predominantly originate form donor organisms in which these tetranucleotides are less associated with structural defects or special functional roles, remains unclear.

The restriction enzymes, for which the recognition sites were often found to cluster atypically in the genomes assessed in this study, such as SpeI, XbaI, AvrII and NheI, are also commonly used for genotyping by pulsed-field gel electrophoresis (PFGE) (26). For example, to identify an \textit{E.coli} O157 outbreak cluster, Tsuji and co-workers (27) performed PFGE with the XbaI enzyme. A higher prevalence of the recognition sites of these enzymes in aDNA regions, such as horizontally transferred genes, may partly explain the high differentiating capacity of PFGE when performed with these enzymes. Insertion of horizontally acquired DNA harbouring these sites in a higher frequency than the recipient genome will result in the introduction of novel small fragments, which are usually not visualized by PFGE. However, the large fragment in which the region of aDNA is inserted will disappear from the PFGE pattern.

As the genome signature is conserved between closely related species (28), this technique may enable the selective isolation of aDNA from novel outbreak strains in the population of a pathogenic species of which a representative complete genome sequence is available, illustrated by the identification of the different anomalous fragments in the two \textit{Neisseria} strains. It would be of interest to test different neisserial genoclusters for anomalous sequences with this novel strategy. In conclusion, the strategy presented in this study allows the selective isolation of anomalous sequences from prokaryotic genomes by a simple restriction digest–amplification–cloning–sequencing scheme. This simple technique can have major practical applications in studying horizontal gene transfer.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at NAR Online.

ACKNOWLEDGEMENTS

We would like to thank Drs Mark Achtman and Christina Vandenbroucke-Grauls for critically reading the manuscript.

REFERENCES


