Comparative Genome Analysis of *Fusobacterium nucleatum*

Mia Yang Ang¹,²,†, Avirup Dutta¹,†, Wei Yee Wee¹,²,†, David Dymock³, Ian C. Paterson²,⁴, and Siew Woh Choo¹,²,⁵,*

¹Genome Informatics Research Laboratory, Centre for Research in Biotechnology for Agriculture (CEBAR), High Impact Research Building, University of Malaya, Kuala Lumpur, Malaysia
²Department of Oral and Craniofacial Sciences, Faculty of Dentistry, University of Malaya, Kuala Lumpur, Malaysia
³School of Oral & Dental Sciences, University of Bristol, Bristol, United Kingdom
⁴Oral Cancer Research and Coordinating Centre (OCRCC), Faculty of Dentistry, University of Malaya, Kuala Lumpur, Malaysia
⁵Genome Solutions Sdn Bhd, Suite 8, Innovation Incubator UM, Level 5, Research Management & Innovation Complex, University of Malaya, Kuala Lumpur, Malaysia

†These authors contributed equally to this work.
*Corresponding author: E-mail: l.choo@genomesolutions.com.my.

Accepted: July 31, 2016

Data deposition: This project has been deposited at NCBI Genbank under the accession AXUR00000000.

Abstract

*Fusobacterium nucleatum* is considered to be a key oral bacterium in recruiting periodontal pathogens into subgingival dental plaque. Currently *F. nucleatum* can be subdivided into five subspecies. Our previous genome analysis of *F. nucleatum* W1481 (referred to hereafter as W1481), isolated from an 8-mm periodontal pocket in a patient with chronic periodontitis, suggested the possibility of a new subspecies. To further investigate the biology and relationships of this possible subspecies with other known subspecies, we performed comparative analysis between W1481 and 35 genome sequences represented by the five known *Fusobacterium* subspecies. Our analyses suggest that W1481 is most likely a new *F. nucleatum* subspecies, supported by evidence from phylogenetic analyses and maximal unique match indices (MUMi). Interestingly, we found a horizontally transferred W1481-specific genomic island harboring the tripartite ATP-independent (TRAP)-like transporter genes, suggesting this bacterium might have a high-affinity transport system for the C4-dicarboxylates malate, succinate, and fumarate. Moreover, we found virulence genes in the W1481 genome that may provide a strong defense mechanism which might enable it to colonize and survive within the host by evading immune surveillance. This comparative study provides better understanding of *F. nucleatum* and the basis for future functional work on this important pathogen.

Key words: *Fusobacterium nucleatum*, comparative genomics, *F. nucleatum* W1481.

Introduction

The genus *Fusobacterium* can be distinguished by its Gram-negative, nonmotile and nonsporulating features, as described in the studies conducted by Bachmann and Gregor (1936). Cultures of *Fusobacterium* grow only under strictly anaerobic conditions and they are extremely susceptible to atmospheric oxygen (Spaulding and Rettger 1937b). *Fusobacterium* are normally present in the human oral cavity and throat (Spaulding and Rettger 1937a), however, recent studies have shown that the oropharynx, as well as the gastrointestinal and genitourinary tracts may also be populated with *Fusobacterium* spp. (Castellarin et al. 2012). *Fusobacterium* spp. can occasionally cause localized infections such as tonsillitis, para-tonsillar abscess or dental sepsis, invasive disease including septicemia, abscesses of the brain, liver and lung, or Lemierre’s syndrome, a septic thrombophlebitis of the internal jugular vein that is often associated with oropharyngeal infection (Huggan and Murdoch 2008; Kuppalli et al. 2012). The majority of research has focused on the role of *Fusobacterium* spp. in gingivitis and periodontitis, where *Fusobacterium nucleatum* (F. nucleatum) is the most frequently isolated species from dental plaque, where it is believed to play an important role in recruiting periodontal pathogens into subgingival...
dent, all genome sequences were uploaded to PanSeq (Laing et al. 2010) for alignment and the Single Nucleotide Polymorphism (SNPs) of the core/common genome regions used for DNA substitution model testing and also to generate the phylogenetic trees. The model selection was based on Bayesian information criterion (BIC). All the neighbor-joining phylogenetic trees were based on Kimura’s two parameter model, inferred with 1,000 bootstrap replications.

Genomic Distance Calculation

Calculation of genomic distance was performed based on the number of maximal unique and exact matches (MUMs) of a minimal length shared by two genomes when being compared. For each genome pairs, MUMs were generated using MUMmer 3 software package (Delcher et al. 2003) using the parameters: -mum, -b, -c and -119. MUMmer package automatically detects matches that may not be unique while a Perl script developed by Marc and his colleague (Deloger et al. 2009) was used to trim overlapping MUMs and to calculate the MUM index (MUMi) values. MUMi varies between 0 for very similar and 1 for very distant genomes.

Whole-Genome Average Nucleotide Identity Analysis

To evaluate the genetic relatedness between W1481 and the other Fusobacterium genomes the average nucleotide identity (ANI) (Konstantinidis and Tiedje 2005) was calculated based on the method proposed by Goris et al. (2007). Two-way BLAST was chosen and only forward and reversed-matched orthologs were used in the calculations. For the robustness, the BLAST match have been set at least 50% identity at the nucleotide and amino acid level and a sequence coverage of at least 70%.

Amino Acid Identity Analysis

The calculation of average amino acid identity (AAI) was performed by the method as described by Konstantinidis and Tiedje (2005). The RAST annotated protein-coding sequences of the W1481 genome was used as the reference for comparison against other Fusobacterium genomes (also annotated using RAST), using the BLAST search to determine the conserved genes. The cut-off was set at ≥30% sequence identity and ≥70% sequence coverage at the amino acid level for the BLAST search. The average of the amino acid identity of all conserved genes between a pair of genomes was computed to measure the genetic relatedness between them.

Comparative Genomic Island Analysis

All genomes sequences were submitted to IslandViewer (Langille and Brinkman 2009) for genomic island (GI) prediction. IslandViewer implements a sequence composition based approaches derived from other validated GI prediction software packages, which were SIGI-HMM (Waack et al. 2006) and IslandPath-DIMOB (Hsiao et al. 2003). Both the composition based approach were shown to have a higher than 86% overall accuracy compared with other prediction software. IslandViewer also integrated IslandPick, which was developed
based on comparative genomics approach (Langille et al. 2008). To cluster these GIs for comparison across different *F. nucleatum* strains, the GI nucleotide sequences were clustered using BLASTClust (Altschul et al. 1990) with the thresholds of at least 50% sequence identity and 50% sequence coverage.

### Comparative Pathogenomic Analyses

Homology searches were performed on the protein sequences of 36 *Fusobacterium* strains against the Virulence Factors Database (VFDB) (Chen et al. 2005, 2012), by applying BLASTP of the BLAST software package (Altschul et al. 1997). The BLAST hits were processed using in-house Python script filtering and accepting only orthologs at the threshold of at least 50% sequence identity and 50% sequence completeness between the query and subject sequences. Instead of using the tabular comparison for pathogenomic composition provided in VFDB release of 2012, the data collected from VFDB was used to construct a graphical heat map representation using our in-house R Scripts, allowing for the comparison of the virulence gene profiles across all *Fusobacterium* strains.

### Pan-Genome Analyses

Functional ortholog clustering was performed using PGAP (Pan-genome Analysis Pipeline) (Zhao et al. 2012). The protein sequences of 21 *F. nucleatum* strains were used as the input file. The orthologs among these 21 strains were searched using the Gene Family (GF) method included in the PGAP pipeline. Protein sequences of each strain were clustered together and denoted with strain identifiers. BLASTALL program (Altschul et al. 1990) was performed among the protein sequences with the minimum score value of 50 and E-value at $1^{-8}$. The filtered BLAST results were clustered by MCL algorithm (Enright et al. 2002). In order to group the same genes into the same cluster, the global match region must have at least 70% of the longer gene protein sequence (coverage) and 40% sequence identity.

### Results and Discussion

#### Genome Overview and Annotation

The size of the sequenced genome of *F. nucleatum* W1481 was 2,477,971 bp, with a GC composition of 27.6%. About 2,163 coding sequences (CDSs) and 56 RNAs were predicted by RAST in W1481 (Ang et al. 2014). An illustration of genomic contents in the genome of W1481 is shown in figure 1A. The subsystem distribution statistic of W1481 based on the genome annotation performed by RAST pipeline is presented in figure 1B.

#### Phylogenetic Analysis Using Different Biomarker

Twenty-six (26) published *Fusobacterium* strains (including *F. nucleatum* strain W1481) were obtained from the National Center for Biotechnology Information (NCBI) database (Pruitt et al. 2007). Phylogenetic trees were reconstructed by maximum likelihood method using the Kimura 2-parameter (K2P) distance model. The 16S rRNA phylogenetic tree clearly clustered all the subspecies together, as previously shown (Manson McGuire et al. 2014). Based on the 16S rRNA phylogenetic tree (fig. 2A), W1481 was clustered together with *F. nucleatum* strains, indicating that it was indeed a *F. nucleatum* species. To further confirm the taxonomic position of W1481, 271,947 single nucleotide polymorphisms (SNPs) in the core genomes were concatenated and aligned. We reconstructed a more robust phylogenetic tree using the core-genome SNPs by using K2P distance model (fig. 2B). The core-genome SNP-based tree showed W1481 was clearly clustered together with *F. nucleatum* strains, further supporting that W1481 is *F. nucleatum*.

Interestingly, W1481 did not cluster with any of the known subspecies, pointing to the possibility that W1481 might be a novel subspecies of *F. nucleatum*.

#### Genomic Distance Based on Maximum Unique Matches Index (MUMi)

To further study the genetic relationship between W1481 and other subspecies of *F. nucleatum* at the subspecies level, we performed genomic distance-type index calculations. Genomic distance based on MUMi analysis was then performed by comparing two sets of data, in which the first set consisted of all the 36 genomes of *Fusobacterium* spp. and the second set consisted only of *F. nucleatum* genomes. The results of the MUMi analysis of the first set as summarized in figure 3A showed that W1481 was quite distant compared to other *Fusobacterium* strains/species with the range of values between 0.64 and 0.98. The closest strains to W1481 were all of *F. nucleatum*, reinforcing our view that this isolate is indeed *F. nucleatum*.

Although W1481 had the closest value (0.64) with *F. nucleatum* subsp. *animalis*, it was still unclear if this value was distant enough to support W1481 as a novel subspecies since we did not have a specific cut-off to define a subspecies. To define a cut-off value, we calculated the MUMi distance between all pairs of known *F. nucleatum* subspecies. To use a very stringent criterion, we chose the highest value (most distant) to separate a pair of two known subspecies as a cut-off to discriminate subspecies. Our MUMi analysis showed that the highest value between the subsp. *animalis* (*F. nucleatum* subsp. *animalis* 11_3_2) and subsp. *polymorphum* (*F. nucleatum* subsp. *polymorphum* ATCC10953), which was 0.6342 (fig. 3B). Therefore, we selected 0.6342 as a cut-off value to discriminate subspecies.

Interestingly, the MUMi distances between W1481 with any strain of known subspecies were all above the defined cut-off, supporting our view that W1481 might be a novel subspecies. W1481 was distantly separated from all known subspecies compared to the distances between any pair of
known subspecies (fig. 3A). Supplementary table S1, Supplementary Material online, shows the full sets of MUMi values generated by comparing all 36 strains of *Fusobacterium* species.

The MUMi results were further supported by the ANI and AAI analysis (using the same set of genomes with W1481 as the reference genome) which also showed that the closest strain to W1481, belonged to *F. nucleatum* subsp. *animalis* as shown in supplementary figure S1, Supplementary Material online.

**Pan-Genome and Genome Composition Analysis**

To further investigate the genomic structure and gene content of *F. nucleatum* W1481, we performed pan-genome analysis using the protein sequences of 21 *F.*

---

**Fig. 1.**—(A) Circular representation of the W1481 genome. The indication for each feature from the outermost layer: (a) Contigs (Dark Red); (b) Forward protein-coding genes (Orange); (c) Reverse protein-coding genes (Light Blue); (d) tRNAs (Light Green); (e) rRNAs (Dark Green); (f) Predicted Genomic Islands (Dark Blue); and (g) GC plot (Purple and Green). (B) Subcellular distribution as predicted by RAST.
nucleatum strains. All the genome sequences of *Fusobacterium* used in the study were also annotated using the RAST pipeline to ensure uniformity. The genome size of the strains analyzed ranged from 1.71 to 3.68 Mb with GC composition varying between 25.2% and 35.2%. An overview of all the 21 *Fusobacterium* genomes used in the present study is shown in table 1. The results showed that *F. nucleatum* pan-genome contained 6,666 gene clusters which were composed of 880 (13%) core gene clusters, 2,850 (43%) dispensable gene clusters and 2,936 (44%) strain-specific gene clusters. The accessory genes (dispensable and strain-specific genes) which cover 87% of the total *F. nucleatum* pan-genome suggested that this species contains a very diverse genomic structure.

In order to know more about the genomic structure of *F. nucleatum*, we further predicted the pan-genome size by extrapolation through calculation of the gene clusters and core clusters for N genomes, where N is the number of *F. nucleatum* genomes (*N* = 1, 2, 3 ..., 19, 20, 21). To avoid bias, the pan-genome size and core genome for each of the permutations of genome comparisons was predicted. Results show that the pan-genome size of *F. nucleatum* increased with the increase in the number of genomes. Our mathematical functional model (*Y* = 1260.48511544134...
FIG. 3.—(A) MUMi analysis between W1481 and other Fusobacterium spp. The initial result showed that W1481 has the closest value (0.64) when compared to *F. nucleatum* subsp. *animalis*. (B) Comparison between the MUMi values among W1481 and the groups of five existing subspecies of *F. nucleatum*. 

Comparative Genome Analysis of *Fusobacterium nucleatum*
The MUMi results of each subspecies are summarized and compared. The upper value of the box and whisker box showed the highest MUMi value, whereas the lower value showed the lowest MUMi value while the value in middle showed the average MUMi value when we performed comparison. A clear separation of W1481 from the rest of the known F. nucleatum subspecies is able to be performed at the threshold of 0.6342. (Note: FNA denotes subsp. animalis; FNV denotes subsp. vincentii; FNN denotes subsp. nucleatum; FNP denotes subsp. polymorphum and FNF denotes subsp. fusiforme).

Table 1
Overview of the genome statistics

<table>
<thead>
<tr>
<th>#</th>
<th>Strain name</th>
<th>Strain status</th>
<th>Genome size (Mb)</th>
<th>GC content (%)</th>
<th>Number of contigs</th>
<th>Contig N50</th>
<th>Number of ORFs</th>
<th>Number of rRNAs</th>
<th>Number of tRNAs</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><em>Fusobacterium</em> nucleatum W1481</td>
<td>I</td>
<td>2.48</td>
<td>27.5</td>
<td>186</td>
<td>7,109</td>
<td>2,219</td>
<td>46</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td><em>F. nucleatum</em> subsp. <em>animalis</em> 11_3_2</td>
<td>I</td>
<td>2.7</td>
<td>26.9</td>
<td>81</td>
<td>101,807</td>
<td>2,725</td>
<td>47</td>
<td>14</td>
</tr>
<tr>
<td>3</td>
<td><em>F. nucleatum</em> subsp. <em>animalis</em> 21_1A</td>
<td>I</td>
<td>2.15</td>
<td>27</td>
<td>2</td>
<td>1,755,614</td>
<td>2,083</td>
<td>46</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td><em>F. nucleatum</em> subsp. <em>animalis</em> 3_1_33</td>
<td>I</td>
<td>2.29</td>
<td>27</td>
<td>18</td>
<td>266,619</td>
<td>2,135</td>
<td>44</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td><em>F. nucleatum</em> subsp. <em>animalis</em> 4_8</td>
<td>C</td>
<td>2.26</td>
<td>27.1</td>
<td>1</td>
<td>–</td>
<td>2,202</td>
<td>56</td>
<td>15</td>
</tr>
<tr>
<td>6</td>
<td><em>F. nucleatum</em> subsp. <em>animalis</em> 7_1</td>
<td>I</td>
<td>2.47</td>
<td>26.8</td>
<td>95</td>
<td>45,444</td>
<td>2,477</td>
<td>42</td>
<td>2</td>
</tr>
<tr>
<td>7</td>
<td><em>F. nucleatum</em> subsp. <em>animalis</em> ATCC 51191</td>
<td>I</td>
<td>2.48</td>
<td>25.2</td>
<td>465</td>
<td>4,679</td>
<td>2,443</td>
<td>42</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td><em>F. nucleatum</em> subsp. <em>animalis</em> ChDC F324</td>
<td>I</td>
<td>2.27</td>
<td>26.9</td>
<td>123</td>
<td>39,283</td>
<td>2,210</td>
<td>35</td>
<td>0</td>
</tr>
<tr>
<td>9</td>
<td><em>F. nucleatum</em> subsp. <em>animalis</em> D11</td>
<td>I</td>
<td>2.34</td>
<td>26.9</td>
<td>355</td>
<td>11,476</td>
<td>2,387</td>
<td>44</td>
<td>4</td>
</tr>
<tr>
<td>10</td>
<td><em>F. nucleatum</em> subsp. <em>animalis</em> F0419</td>
<td>I</td>
<td>2.44</td>
<td>27</td>
<td>5</td>
<td>2,257,469</td>
<td>2,345</td>
<td>39</td>
<td>13</td>
</tr>
<tr>
<td>11</td>
<td><em>F. nucleatum</em> subsp. <em>nucleatum</em> ATCC 23726</td>
<td>I</td>
<td>2.24</td>
<td>27</td>
<td>67</td>
<td>69,905</td>
<td>2,143</td>
<td>43</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td><em>F. nucleatum</em> subsp. <em>nucleatum</em> ATCC 25586</td>
<td>C</td>
<td>2.17</td>
<td>27.2</td>
<td>1</td>
<td>–</td>
<td>2,102</td>
<td>47</td>
<td>15</td>
</tr>
<tr>
<td>13</td>
<td><em>F. nucleatum</em> subsp. <em>nucleatum</em> ChDC F316</td>
<td>I</td>
<td>2.16</td>
<td>27</td>
<td>83</td>
<td>68,085</td>
<td>2,112</td>
<td>32</td>
<td>3</td>
</tr>
<tr>
<td>14</td>
<td><em>F. nucleatum</em> subsp. <em>polymorphum</em> ATCC 10953</td>
<td>C</td>
<td>2.43</td>
<td>26.8</td>
<td>1</td>
<td>–</td>
<td>2,390</td>
<td>45</td>
<td>10</td>
</tr>
<tr>
<td>15</td>
<td><em>F. nucleatum</em> subsp. <em>polymorphum</em> F0401</td>
<td>I</td>
<td>2.51</td>
<td>27</td>
<td>2</td>
<td>2,507,721</td>
<td>2,533</td>
<td>51</td>
<td>19</td>
</tr>
<tr>
<td>16</td>
<td><em>F. nucleatum</em> subsp. <em>vincentii</em> 3_1_27</td>
<td>I</td>
<td>2.2</td>
<td>27</td>
<td>77</td>
<td>111,961</td>
<td>2,115</td>
<td>46</td>
<td>13</td>
</tr>
<tr>
<td>17</td>
<td><em>F. nucleatum</em> subsp. <em>vincentii</em> 3_1_36A2</td>
<td>C</td>
<td>2.27</td>
<td>27.1</td>
<td>1</td>
<td>–</td>
<td>2,189</td>
<td>43</td>
<td>3</td>
</tr>
<tr>
<td>18</td>
<td><em>F. nucleatum</em> subsp. <em>vincentii</em> 4_1_13</td>
<td>I</td>
<td>2.25</td>
<td>26.9</td>
<td>50</td>
<td>106,163</td>
<td>2,207</td>
<td>43</td>
<td>5</td>
</tr>
<tr>
<td>19</td>
<td><em>F. nucleatum</em> subsp. <em>vincentii</em> ATCC 49256</td>
<td>I</td>
<td>2.12</td>
<td>27.3</td>
<td>302</td>
<td>14,248</td>
<td>2,321</td>
<td>44</td>
<td>12</td>
</tr>
<tr>
<td>20</td>
<td><em>F. nucleatum</em> subsp. <em>vincentii</em> ChDC F8</td>
<td>I</td>
<td>1.99</td>
<td>26.9</td>
<td>186</td>
<td>23,738</td>
<td>1,833</td>
<td>28</td>
<td>1</td>
</tr>
<tr>
<td>21</td>
<td><em>F. nucleatum</em> subsp. <em>fusiforme</em> ATCC 51190</td>
<td>I</td>
<td>1.84</td>
<td>27.2</td>
<td>198</td>
<td>23,781</td>
<td>1,778</td>
<td>45</td>
<td>11</td>
</tr>
</tbody>
</table>

Note.—C, complete genome; I, incomplete genome.
Spectinomycin 9-O-adenylyltransferase and a number of multi-drug resistance proteins and efflux pumps such as multi-drug resistance protein 2, permease of the drug/metabolite transporter (DMT) superfamily, multi-antimicrobial extrusion protein (Na(+)/drug antiporter), MATE family of MDR efflux pumps and Na(+)-driven multidrug efflux pump. We also found the presence of hemolysin and putative large exoprotein involved in heme utilization or adhesion of ShlA/HecA/FhaA family in the W1481 genome.

W1481 showed the presence of ATP-dependent oligopeptide transporter OppABCDF, a member of the ATP-Binding Cassette (ABC) Superfamily of transporters (Pearce et al. 1992). OppABCDF have been reported to transport oligopeptides up to five amino acids in length but not free amino acids as shown by studies based on binding affinity and competition assays (Guyer et al. 1986). The system is involved in the uptake of oligopeptides, as well as recycling of cell wall peptides (Hiles et al. 1987). OppA is the periplasmic substrate-binding component that binds oligopeptides (Tame et al. 1994), OppB and OppC are the membrane components of the ABC transporter, whereas OppD and OppF are the ATP-binding components of the ABC transporter (Pearce et al. 1992). Another ABC transport system present in W1481 is the nik operon which is responsible for uptake of nickel ions, an essential nutrient participating in various cellular processes. The first five genes of the nik operon (nikABCDE) were identified in W1481, however, nikR the DNA-binding protein repressing the transcription of nikABCDE (Navarro et al. 1993) was not detected in W1481.

We also found two putative genes, ltrA (encoding low temperature requirement protein A). These ltrA genes were not present in the other F. nucleatum strains. ltrA genes are essential for bacterial growth at low temperature (Zheng and Kathariou 1995). We suggest that the presence of these genes in W1481 might allow it to survive in lower temperatures as compared to other F. nucleatum strains thus enabling it to survive in more diverse environments in term of temperature.

Genomic Island Prediction

Genomic islands were predicted for the genome of W1481 together with the other 20 F. nucleatum genomes by the IslandViewer 3 software. A total of 91 nonredundant GIs were predicted in all F. nucleatum genomes used in this study (supplementary table S2, Supplementary Material online). Of the 91 putative GIs, two GIs (GI41 and GI73) were conserved across all F. nucleatum genomes. Inside GI43, there are two genes which encode for pyruvate formate-lyase enzyme. Insertion of these GI to the F. nucleatum genomes allow catalyzation of the reversible conversion of pyruvate and coenzyme-A into formate and acetyl-CoA (Becker et al. 1999). For the W1481 genome, we found 23
GIs, of which three GIs (GI1, GI7, and GI10) were only present in W1481, and not in other F. nucleatum strains (supplementary table S2, Supplementary Material online). We found the presence of putative genes code for adenosylcobinamide amidohydrolase, which is required in the salvage pathway of cobinamide (Woodson and Escalante-Semerena 2004), and the components of cobalt chelatase (CobN, ChlI/ChlID) involved in B12 (cobalamin) biosynthesis (Rodionov et al. 2003) in the GI1. We have also found L(+)-tartrate dehydratase alpha subunit gene and L(+)-tartrate dehydratase beta subunit gene in another W1481 specific GI (GI7). These genes are involved in the catabolism of lactate and succinate (Do et al. 2015). Thus, GI1 and GI7 may allow W1481 to synthesize or catabolize cobinamide, lactate and succinate.

In addition, we also observed some interesting genes and features in the GIs shared between W1481 and some of the F. nucleatum strains. For example, we found that the horizontally transferred GI3, which was shared by W1481 with the strains of subsp. animalis (11_3_2), subsp. fusiforme (ATCC51190) and subsp. nucleatum (ATCC23726), carries the tripartite ATP-independent (TRAP)-like transporter genes, encoding for TRAP-type C4-dicarboxylate transport system-periplasmic component, TRAP-type C4-dicarboxylate transport system-small permease component and TRAP-type C4-dicarboxylate transport system-large permease component. These TRAP transporters are a high-affinity transport system for the C4-dicarboxylates malate, succinate, and fumarate (Forward et al. 1997). Strikingly, the GI4 which was shared by W1481 with the strains of subsp. polymorphum (ATCC10953 & F401) and three strains of subsp. animalis (F0419, D11 & 7_1), contain a cluster of eight putative CRISPRs genes which may provide the host with acquired and heritable resistance to foreign DNA (Horvath and Barrangou 2010; Marraffini and Sontheimer 2010; Jore et al. 2012). The distribution of the GIs among the strains indicates that W1481 share a number of GIs with the strains of subsp. animalis, vincentii and polymorphum.

Comparative Pathogenomic Analysis

To better understand the pathogenic potential of W1481, we performed a comparative pathogenomic analysis of W1481 along with the strains of the five subspecies of F. nucleatum (fig. 6). W1481 showed the presence of a number of virulence genes which were shared by all or most of the other strains under study mostly related to lipopolysaccharide layer of the bacterial cell surface, such as galE (UDP-glucose 4-epimerase), galU (glucose-1-phosphate uridylyltransferase), gtaB (UTP-glucose-1-phosphate uridylyltransferase), acpxL (acyl carrier protein), kdsA (2-dehydro-3-deoxyphosphoorcinate aldolase), lpxA (UDP-N-acetylglucosamine acyltransferase), gmhA/lpcA (phosphoheptose isomerase) and eno (phosphopyruvate hydratase). W1481 also shared the presence of groEL a 60-kDa chaperonin, htpB (Hsp60, 60K heat shock protein), hemL, and phtA (phosphatase acetal, a component of the acyl carrier protein biosynthesis pathway). W1481 also showed the presence of groEL a 60-kDa chaperonin, htpB (Hsp60, 60K heat shock protein), hemL, and phtA (phosphatase acetal, a component of the acyl carrier protein biosynthesis pathway).
One interesting observation was the presence of \textit{ddhABC} genes which are related to CDP-glucose pathways. These genes were present in W1481 and only one other strain of \textit{F. nucleatum} subsp. \textit{vincentii} ATCC49256. \textit{ddhABC} genes are involved in lipopolysaccharide O-antigen biosynthesis which have been reported to play a crucial role in effective colonization of host tissues, and evasion of the host immune system by gram negative bacteria (Lerouge and Vanderleyden 2002; Skurnik and Bengoechea 2003). None of the other strains under study showed the presence of these genes. W1481 also showed the presence of Putative O-antigen ligase WaaL, which is responsible for the ligation of O-antigen onto the core region of the lipid A-core block a crucial step in the lipopolysaccharide biosynthetic pathway (Han et al. 2012).

Another interesting set of genes were that of neuramic acid synthetase (\textit{neuB} and \textit{neuB2}) reported to be catalyzing the last step of the de novo sialic acid biosynthetic pathway, which is a major element of bacterial surface structure (Feng et al. 2012). Again these were present in only W1481 and two strains of \textit{F. nucleatum} subsp. \textit{vincentii} (ATCC49256 and 3_1_27) and one other strain, \textit{F. nucleatum} subsp. \textit{polyumphum} ATCC10953. It has been suggested that bacterial pathogens employ the strategy of “molecular mimicry” by incorporating sialic acid on their cell-surface to disguise themselves as host cells to circumvent and/or counteract the host’s immune responses (Severi et al. 2007). W1481 also possess Alpha-2-macroglobulin (A2M) which are located in the periplasm and are believed to trap external proteases through a covalent interaction with an activated thioester thus providing protection to the cell. Studies in \textit{Salmonella enterica} ser. \textit{typhimurium} have shown that A2M mimics the innate immune system of eukaryotes (Wong and Dessen 2014). The comparative analysis results suggest that W1481 might be slightly more effective in colonization and resisting the host immune system.

**Conclusion**

The results of phylogenetic and MUMi distance analyses all suggest the possibility that W1481 is likely a novel subspecies of \textit{F. nucleatum}. W1481 has several unique horizontally transferred GIs, indicating the HGT may play an important role in the evolution of this bacterium. Comparative analysis suggests that W1481 is well equipped to defend itself, colonize and survive within the host evading its immune system, although experimental verification is required to confirm these observations. This comparative in silico study provides novel insights into the genome of W1481 and lays the foundation for future experimental studies.

**Supplementary Material**

Supplementary figure S1 and tables S1 and S2 are available at Genome Biology and Evolution online (http://www.gbe.oxfordjournals.org/).

**Acknowledgments**

\textit{Fusobacterium nucleatum} W1481 was originally isolated by William Wade (currently at Barts and the London School of Medicine and Dentistry, Queen Mary, University of London). We would like to thank all members of Genome Informatics Research Group (GiRG) in contributing to this research. The work was supported by University of Malaya and Ministry of Education, Malaysia under the High Impact Research (HIR) grant (grant number: UM.C/625/2/HIRMOHE/CHAN-08) and UMRG grant from University of Malaya and Ministry of Education, Malaysia (grant number: RG541-13HTM).

**Literature Cited**


Gharbia SE, Shah HN. 1990. Identification of \textit{Fusobacterium} species by the GISA and by