Complete genome sequence of BS49 and draft genome sequence of BS34A, Bacillus subtilis strains carrying Tn916

Hilary P. Browne1,#, Seyed Yahya Anvar2,#, Jeroen Frank2, Trevor D. Lawley1, Adam P. Roberts3 and Wiep Klaas Smits4,*

1Host-Microbiota Interactions Laboratory, Wellcome Trust Sanger Institute, Hinxton, Cambridge, CB10 1SA, UK, 2Leiden Genome Technology Center, Human and Clinical Genetics, Leiden University Medical Center, 2300 RC, Leiden, The Netherlands, 3Department of Microbial Diseases, UCL Eastman Dental Institute, WC1X 8 LD London, UK and 4Department of Medical Microbiology, Leiden University Medical Center, 2300 RC, Leiden, The Netherlands

*Corresponding author: Department of Medical Microbiology, Leiden University Medical Center, PO Box 9600, 2300 RC, Leiden, The Netherlands. Tel: +31-71-526 1229, Fax: +31-71-526 6761; E-mail: w.k.smits@lumc.nl

#These authors contributed equally.

ABSTRACT

Bacillus subtilis strains BS49 and BS34A, both derived from a common ancestor, carry one or more copies of Tn916, an extremely common mobile genetic element capable of transfer to and from a broad range of microorganisms. Here, we report the complete genome sequence of BS49 and the draft genome sequence of BS34A, which have repeatedly been used as donors to transfer Tn916, Tn916 derivatives or oriT

Tn916-containing plasmids to clinically important pathogens.

Key words: mobile element; donor; conjugation; conjugative transposon

The BS49 and BS34A strains were originally obtained through introduction of Tn916 into Bacillus subtilis CU2189 (Christie et al., 1987) either by direct transformation of pAM120 plasmid (GenBank U49939.1) (Gawron-Burke and Clewell 1984) (Peter Mullany, personal communication) or by conjugation from Clostridium difficile strain FM12A (Roberts et al., 2003), respectively. Tn916 from strain FM12A was originally transferred from BS49 to C. difficile CD37 to give FM12A (Roberts et al., 2003; Brouwer et al., 2012). The BS49 strain contains multiple copies (our unpublished observations), whereas BS34A carries a single copy of Tn916 (Roberts et al., 2003).

Tn916 (reviewed in references Roberts and Mullany (2009, 2011)) has been used as a vector to perform (random) mutagenesis of recipient bacteria (Kathariou et al., 1990; Lin and Johnson 1991; Whetzel et al., 2003; Cookson et al., 2011; Mullany et al., 2012), to deliver cargo DNA (Rubens and Heggen 1988; Haraldsen and Sonenshein 2003; Roberts et al., 2003; McBride and Sonenshein 2011) or to mobilize plasmids containing the Tn916 origin of transfer, oriT

Tn916 (Francis et al., 2013; Mackin et al., 2013). BS49 and BS34A like their close relative B. subtilis subsp. subtilis demonstrate high levels of competence for genetic transformation. The genome sequences of strains BS49 and BS34A will be made available through Genbank.
Table 1. Variants in BS34A compared to BS49. Illumina HiSeq reads of strain BS34A were assembled against the complete genome sequence of strain BS49. Variants were called by Geneious R7 (http://www.geneious.com) using a minimal coverage of 25 and a frequency greater than 0.75. Variants at the ends of the Tn916 elements and immediately flanking the insertion sites were ignored.

<table>
<thead>
<tr>
<th>bp start in BS49</th>
<th>bp end in BS49</th>
<th>Nucleotide change</th>
<th>Codon change</th>
<th>Amino acid change</th>
<th>Poly-morphism</th>
<th>Location in BS49</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>786781</td>
<td>786782</td>
<td>+TA</td>
<td>Insertion</td>
<td>Intergenic region of BS49,08160 and BS49,08170</td>
<td>Insertion restores pseudogene BS49,08161 (BS34A,08161) corresponding to yetI of strain 168.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1152310</td>
<td>1152311</td>
<td>(A)7 -&gt; (A)8</td>
<td>Insertion</td>
<td>BS49_11890</td>
<td>Frameshift results in a fusion of BS49,11890 with BS49,11900/hcpE (restores pseudogene BS49,11891) (BS34A,11890/hcpE), corresponding to yisK of strain 168.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1706699</td>
<td>1706699</td>
<td>(A)8 -&gt; (A)7</td>
<td>Deletion</td>
<td>HslU (BS49,18020)</td>
<td>Truncated after 5 amino acids in BS49A (BS34A,17841 pseudogene). Possible alternative start codon annotated (BS34A,17840/hslU)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4188826</td>
<td>4188826</td>
<td>G -&gt; A</td>
<td>GCC -&gt; GTC</td>
<td>A -&gt; V</td>
<td>Transition</td>
<td>WallK (BS49,43870)</td>
<td>In PAS domain. Strain 168 has an alanine at this position, like BS49.</td>
</tr>
<tr>
<td>4189142</td>
<td>4189142</td>
<td>A -&gt; G</td>
<td>TTC -&gt; CTC</td>
<td>F -&gt; L</td>
<td>Transition</td>
<td>WallK (BS49,43870)</td>
<td>Between TM and HAMP domain. Strain 168 has a leucine at this position, like BS34A.</td>
</tr>
</tbody>
</table>

and BS34A will therefore contribute to the study of Tn916 conjugation requirements, including the possible role of host factors, thus facilitating the genetic manipulation of various bacterial strains.

Genomic DNA was isolated as described before (He et al., 2013) or using the Qiagen GenomicTip 500/G according to the manufacturer’s instructions. Libraries of BS49 and BS34A chromosomal DNA were prepared and sequenced using the Illumina Hi-Seq platform as described (Pettit et al., 2014) (EMBL ERS370048 and EMBL ERS370049). For single molecule real-time sequencing of strain BS49, a SMRTbell DNA template library with an insert size of ~20 kb was prepared according to the manufacturer’s specification. SMRT sequencing was carried out on the Pacific Biosciences RSII machine according to standard protocols (EMBL ERS550338). Sequencing reads were corrected using the HGAP pipeline (Chin et al., 2013) and de novo assembly was performed using Celera Assembler 8.1 (http://wgs-assembler.sourceforge.net). After manually curating the sequence of the Tn916 elements to ensure their accuracy, the chromosome was circularized using strain B. subtilis strain 168 (Genbank NC_000964.3) (Kunst et al., 1997) as a reference and the start of the genome was set at 409 bp upstream of the dnaA gene. The single contig was verified by aligning the Illumina reads using Geneious R7 (http://www.geneious.com) to correct for errors in homopolymeric stretches as a result of the consensus calling algorithm.

To generate the BS34A genome sequence, a consensus was generated in Geneious R7 from a reference alignment against the complete BS49 genome. A de novo assembly of the BS34A Illumina sequence data was carried out using Velvet (Zerbino and Birney 2008) to reconstruct the two empty target sites of Tn916 from strain BS49 and to determine the location and context of the single Tn916 element in strain BS34A. The element was subsequently manually inserted into the consensus BS34A sequence. This genome sequence should be considered an improved high-quality draft (Chain et al., 2009).

Following the assembly, Prokka (Seemann 2014) was used to generate annotations based on a database of trusted protein sequences derived from the genome sequence of strain 168. To identify modified bases, kinetic signals were processed for all
genomic positions after aligning sequencing reads to the final single chromosome sequence of strain BS49. The identification of sequence motifs was performed using the SMRT Portal. The output of SMRT Portal (available as Supporting Information) was converted into an Artemis-compatible TAB file using an in-house script. Raw genome sequence and annotations were merged and the Tn916 elements were manually annotated in Artemis (Rutherford et al., 2000).

Our analysis of the 4.251.652 bp BS49 genome (EMBL PRJEB7328) yielded 4248 predicted coding sequences, 30 rRNAs, 87 tRNAs, 1 tmRNA and 94 miscellaneous RNAs. We failed to detect an over-represented motif associated with m4C modifications (n = 2035), but found a GAGCAG motif associated with m6A modifications (1255/2446; modified residue underlined).

The sequence of the 18 032 bp Tn916 elements (bps 1475100-1493131 and 3254988-3273019) (Genbank KM516885) showed that the elements in BS49 are identical to each other and highly similar to the element described for Enterococcus faecalis DS16 (GenBank U094221). The three single nucleotide variations in the Tn916 element of BS49 compared to the Enterococcus element were [nomenclature from Genbank U094221]: (1) the deletion of a guanine residue between ORF21 and ORF20 in the oriT region but not affecting the nick site (TGTTGTTG), (2) the insertion of a cytosine in ORF15, resulting in a frameshift and a mutant protein of 725 amino acids and (3) a G > T conversion in ORF9 resulting in a Q > K amino acid substitution.

For BS34A (EMBL PRJEB7328), we identified 4233 predicted coding sequences within the 4.233.615 bp genome sequence (the numbers of RNAs was the same as for BS49). The variations of the Tn916 element of BS34A (bps 1886552-1904583) compared to the E. faecalis element were the same as for BS49, suggesting that they arose either within the pAM120 plasmid or during the transformation that generated BS49.

We additionally verified the locations of Tn916 in the genome assembly using conventional PCRs for both BS34A and BS49 (text and Fig. S1, Supporting Information). The element was found to be inserted between BS34A_19270 and BS34A_19450 (yypP), BS49_15600/BS49_15780 (ykuC) and BS49_34490 (yufK)/BS49_34670 (yufl) in regions that are conserved in closely related 8. subtilis strains. As described for C. difficile (Mullaney et al., 2012), the insertion sites are characterized by stretches of A and T (Fig. S2A, Supporting Information). Interestingly, a sequence flanking the Tn916 element in BS34A shows homology to the ykyB-ykuc insertion site from BS49, and a sequence flanking Tn916 at that locus resembles the yufK-yufL insertion site (Fig. S2B, Supporting Information), suggesting that additional (chromosomal) sequence was carried over when Tn916 mobilized.

We compared the BS34A and BS49 genome sequences. Besides the Tn916 insertion sites, we identified five variants, including two non-conserved substitutions in the histidine kinase Whk (Table 1).

A progressiveMAUVE alignment (Darling et al., 2010) of the BS49 and 168 genome sequences showed a single large collinear block with >98% pairwise identity, demonstrating that strains BS49 and strain 168 are highly related. We then also aligned the illumina sequence reads from BS49A and BS49 to the genome of strain 168 using Geneious R7. The resulting list of variants confirms the variants identified above that distinguish the two strains, and in addition shows other mutations that are in common and are likely present in the ancestral strain CU2189 (Christie et al., 1987) (Table S1, Supporting Information). These include frameshifts in yafH, yesY, pksN and yubD [168 nomenclature]. Additionally, there are point mutations in for instance comp, sigI, sepF, ytpS/sftA and a G > A mutation in epsC that was shown to result in partial restoration of biofilm formation in strain 168 (McLoon et al., 2011). One or more of the mutations may underlie the difference between the auxotrophy of CU2189 (metB5 hisA1 thr-5) and strain 168 (trpC2).

SUPPLEMENTARY DATA
Supplementary data is available at FEMSLE online.

ACKNOWLEDGEMENTS
The authors thank AL Sonenshein and J Sorg for providing us with BS49 and the Sequence Analysis Support Core at the LUMC for assistance with the implementation of Prokka.

FUNDING
WKS is supported by a VENI fellowship of the Netherlands Organisation for Scientific Research and a Gisela Their Fellowship from the Leiden University Medical Center.

Conflict of interest statement. None declared.

REFERENCES


