Bacteroidales-specific antimicrobial genes influence the selection of the dominant fecal strain of Bacteroides vulgatus and Bacteroides uniformis from the gastrointestinal tract microbial community

Hyunmin Koo  
University of Alabama at Birmingham

Casey D. Morrow (caseym@uab.edu)  
University of Alabama at Birmingham

Research Article

Keywords:

Posted Date: November 22nd, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2276751/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

Background

Previous studies have used metagenomic sequencing to provide strain level resolution of the dominant fecal microbial community. Subsequent studies have shown that perturbations, such as antibiotics can result in the transient appearance of new fecal dominant strains that after a short time are replaced with the pre-antibiotic strains indicating the existence of a microbial community gastrointestinal tract (GIT) reservoir.

Results

To gain insight into the selection of dominant fecal microbial strains from the GIT, we have focused our studies on *Bacteroides vulgatus* and *Bacteroides uniformis*, which are known to be prominent in the human fecal microbial community. Using 3 longitudinal metagenomic sequencing data sets, we have analyzed sequence reads of individual fecal samples using BLAST+ to search for genes encoding *Bacteroidales*-specific antimicrobial proteins that have known functions to restrict species-specific replication of *B. uniformis* (BSAP-2) or *B. vulgatus* (BSAP-3). The BSAP gene, including 5’ and 3’ sequences, was then visualized using Integrative Genomics Viewer. Using the HMP data set, we found fecal samples from individuals had *B. vulgatus* or *B. uniformis* with either complete or deleted BSAP gene patterns that did not change over time. We also examined fecal samples from two separate longitudinal data sets of individuals that had been given either single or multiple antibiotics. The BSAP phenotype and gene pattern from most individuals given either single or multiple antibiotics recovered to be the same as the pre-antibiotic strain. However, in a few individuals, we found that the presence of the BSAP positive phenotype restricted the appearance of a BSAP negative phenotype during the recovery of the fecal dominant strain following antibiotic perturbation.

Conclusions

We demonstrate the dominant fecal *B. uniformis* or *B. vulgatus* strains contain a defined and stable BSAP phenotype. In some individuals, the BSAP positive phenotype dominated the BSAP negative phenotype during recovery from antibiotics, supporting the influence of the BSAP proteins in the selection from the GIT microbial community. The results of these studies provide a basis for targeting the *Bacteroides* BSAP genes to alter microbe-microbe and host-microbe interactions in efforts to modify functions of the *Bacteroides* spp. in the gut community.

Introduction

Numerous longitudinal studies have used metagenomic sequencing of longitudinal sampling coupled with new informatics approaches to provide resolution of the dominant fecal microbial community at the strain level that is unique to the individual and generally stable over time [1–4]. The stability of dominant fecal microbial strains can be influenced by gastrointestinal tract (GIT) perturbations, such as antibiotics,
that result in the transient appearance of new strains although this can vary between individuals [5–7]. A hallmark of the human GIT microbial community is that usually after a short time, the new fecal strains are replaced with the previous dominant strains supporting the existence of a complex GIT ecosystem with multiple microbial strains that compete for fecal dominance [5].

The *Bacteroides vulgatus* and *Bacteroides uniformis* are known to be prominent in the human fecal microbial community [2, 8]. We have previously shown that dominant fecal *B. vulgatus* or *B. uniformis* strains are unique to the individual and related to each other as determined from longitudinal strain tracking analysis [1, 4, 5, 9, 10]. In a recent study, Roodgar et al. showed that microbial strains found in the feces were probably stably maintained in GIT reservoirs defined by metabolic or spatial niches [11]. Previous studies have identified the *Bacteroidales*-specific antimicrobial proteins (BSAP) that have known functions to restrict species specific replication and colonization of *B. uniformis* and *B. vulgatus* [12, 13]. The presence of these microbe genes in an individual's *Bacteroides spp.* might then provide a competitive advantage to maintain that strain in the GIT ecosystem [14].

To test for the importance of the BSAP genes for the selection of the fecal dominant strain, we have determined the status of *B. vulgatus* or *B. uniformis* in fecal samples from studies that contained longitudinal metagenomic data sets from normal and antibiotic perturbed individuals. We have used Integrative Genomics Viewer (IGV) to visualize and extract the genomic region (5’ and 3’ sequences) containing BSAP genes to determine the status of the BSAP genes in the individuals. Our study establishes the existence and long-term stability of the BSAP phenotype in individuals. Additionally, analysis of the data set from two studies using different antibiotics supports a role for the BSAP genes in the change and recovery of dominant fecal strains following disruption of the GIT microbial ecosystem.

**Results**

Previous studies have reported the identification of the BSAP that are specific for *B. vulgatus* (BSAP-3) or *B. uniformis* (BSAP-2) [12, 13]. For BSAP-3, membrane attack perforin/complex protein sequences (MACPF) were identified in a gene cluster 16 from *B. vulgatus/B. dorei* that had activity to restrict growth in a species-specific manner of *B. vulgatus*. Similarly, in *B. uniformis*, MACPF was also identified that could restrict species-specific replication of *B. uniformis* which was named BSAP-2. To determine the presence of the gene encoding BSAP-3 or BSAP-2, the individual samples from 3 longitudinal data sets used in this study were used for the gene analysis based on the MACPF cluster 15 and 16 sequence reads using BLAST+ [15] and IGV tools [16].

We first analyzed the human microbiome (HMP) data set that is composed of paired samples taken at separate times (up to 1 year apart) [2, 17]. We selected 30 paired samples that we have previously shown to have related strains of *B. vulgatus* and *B. uniformis* over periods between 3 months and 1 year. To satisfy the parameters for WSS analysis to determine relatedness, the sequence reads that had a depth of greater than 3.5 with a minimum sequence coverage greater than 30% were used for further analysis [1]. We screened for the complete genes encoding BSAP-3 for *B. vulgatus* and BSAP-2 for *B. uniformis* in
each sample pair (Supplementary Table 1). We identified 11 sample pairs out of 30 for the complete gene encoding BSAP-3, and 7 of 28 samples for gene encoding BSAP-2 (2 sample pairs, S64, and S70 had WSS scores below the cutoff value for B. uniformis) (Figure 1). For 30 BSAP-3 and 28 BSAP-2 sample pairs, we found the same BSAP phenotype (either positive or negative) for time points taken at times that were different by 3-6 months.

To further characterize the BSAP negative phenotype, we used the IGV to display the 5’ and 3’ sequences bracketing the BSAP deletion [16] (Supplementary Figures 1A and 1B). We found that 4 of the 19 sample pairs for the BSAP-3 deleted phenotype were identical (S28, S31, S48, and S66) and 10 of the 19 sample pairs (S16, S17, S21, S23, S49, S60, S64, S65, S69, and S70) had a few nucleotide differences bracketing the deletion from the comparison of the two (a and b) time points (Figure 2A and 2B, and Supplemental Figure 1A). Similarly, for BSAP-2 from the HMP, we found 2 of the 20 (S21 and S66) BSAP-2 negative phenotypes had identical 5’ and 3’ sequences and 18 of the 20 sample pairs (S5, S10, S14, S16, S17, S19, S20, S23, S24, S28, S29, S31, S48, S51, S57, S60, S61, and S65) had only a few nucleotide differences 5’ and 3’ sequences bracketing the deletion from the comparison of the a and b time points (Figure 3A and 3B, and Supplemental Figure 1B). Thus, these results establish the existence of a reservoir of these strains for both B. vulgatus and B. uniformis in these individuals that is stable over an extended time (3-6 months) that is used in the selection of the dominant fecal strain.

From a further inspection of the BSAP negative phenotype sample sets with IGV, we found a subset of Individuals for BSAP-3 (S5, S14, S18, S62, and S63) and one individual (S49) for BSAP-2 that lacked a complete BSAP gene but did have partial (truncated) sequence fragments in the BSAP gene region. The BSAP-3 in individual S14 was particularly interesting since the B. vulgatus in the two samples were related as determined by WSS with the first sample containing a deleted BSAP-3 while the second, later sample had a partial BSAP-3 gene (Supplementary Figure 1A). Most importantly, we found that the pattern of the partial BSAP genes in these individuals was unique to that sample and thus differed between the two sample times.

To determine the impact of the BSAP gene phenotype following perturbation and recovery of the gut ecosystem, we next analyzed a data set from a previous study that analyzed gut microbes following a single antibiotic treatment [5, 18]. This data set contained the collection of fecal samples from 6 individuals at day 0, day 7 and day 90 that were used as untreated controls (Supplemental Table 2). Analysis of the BSAP genes from these samples found that 3 of the 6 were BSAP-3 positive, while 2 of the 6 were positive for BSAP-2 (Supplementary Table 2 and Supplementary Figures 2A and 2B). The analysis from the control samples for the Raymond et al. study gave similar results with respect to the analysis of the sequences bracketing the BSAP deletions as that found for the analysis of the HMP data set (Supplementary Table 2 and Supplementary Figures 2A and 2B).

We next selected 17 individual sample sets from the Raymond et al. study that were given a treatment of cephalaxin with samples collected one day after treatment and 82 days after (a total of 90 days after pre sample). We had previously analyzed this data set using our strain tracking analysis to show recovery of
the pre-antibiotic strain at the day 90 samples for most individuals. Analysis of BSAP genes from the pre sample demonstrated that 3 of the 17 were BSAP-3 and 4 out of 17 were BSAP-2 positive (Figure 4). Although for most of the sample sets the BSAP phenotype did not change following antibiotics, we did find for BSAP-3 one individual (P15) was partial BSAP-3 at day 0 and following the antibiotics at day 7, but by day 90 was BSAP-3 positive with no strain change (Figure 4A, Supplementary Figures 3A, and Supplementary Table 2). In another individual (P17), we found a new strain on day 7 that was partial BSAP-3, but at day 90 the strain was related to the pre sample and was also BSAP-3 positive. Finally, one individual of note, P4, had a deleted BSAP-3 gene at days 0 and 7, with a B. vulgatus sub strain at day 90 with a complete BSAP-3 gene. This individual was the only example where we found a BSAP phenotype change from negative to positive without a strain change (Supplementary Figure 3A). The correlation of the dominant fecal strain in these two individuals with the presence of the BSAP positive phenotype supports that the BSAP-3 positive phenotype restricted the B. vulgatus with the BSAP-3 negative phenotype in the recovery of the GIT ecosystem following the cephalexin perturbation.

To further characterize the impact of the antibiotic on the BSAP phenotype negative samples, we next analyzed the sequence changes using IGV [16]. For the BSAP-3 phenotype negative samples, we found 8 samples in which the strains were related at Day 0 and Day 90 had identical or few nucleotide differences in the 5 and 3 regions bracketing the BSAP-3 deletion (P2, P3, P9, P13, P18, P19, P20, and P21) demonstrating the recovery of the pre antibiotic sub strain (Figure 4A, and Supplementary Figure 3A). For the BSAP-2 negative phenotype samples, 10 samples that had BSAP-2 phenotype negative for the Day 0 and Day 90 samples all had identical or few nucleotide differences in the 5’ and 3’ regions bracketing the deleted BSAP gene (P2, P9, P10, P11, P13, P15, P17, P18, P20, P21) (Figure 4B and Supplementary Figure 3B). Collectively, these results are consistent with the resiliency of the strain reservoir in the capacity to recover following perturbation of the GIT.

To investigate the impact of a greater perturbation on the strain reservoir, we characterized the strain stability of individuals that had been treated with a suppressive antibiotics cocktail consisting of 3 antibiotics (meropenem, gentamicin, and vancomycin) [5, 19]. In this longitudinal study, we found that in some instances the treatment resulted in a strain change from the pre antibiotic strain. Overall, 2 of 10 individuals showed BSAP-3 positive and 3 of 10 showed BSAP-2 positive (Figure 5, Supplementary Figure 4, and Supplementary Table 3). We found several examples where there was a change in an individual from BSAP negative to the BSAP positive phenotype. Individual P1 had a strain change early after antibiotics that were BSAP-3 negative, but the same samples at later times (42 and 180 days) were the same strain as the pre-strain and were both also BSAP-3 positive (Figure 5A). Another individual (P4) that was a partial BSAP gene on day 0 had a strain change with the new strains at day 42 and 180 that were both BSAP-3 positive (Figure 5A). Similarly, in this individual, for B. uniformis there was also a new strain at days 42 and 180 that were both BSAP-2 positive (Figure 5A and 5B). Finally, for P11 BSAP-2 where there was no change in the strain of B. uniformis but the BSAP-2 phenotype changed from partial to positive (Figure 5B).
IGV analysis of the BSAP-3 negative samples revealed a complex pattern, most probably due to the strain change in individuals that occurred following the 3 antibiotics. For BSAP-3, we found that only 1 of 10 samples (P1) had positive BSAP-3 at days 0, 42, and 180 (Figure 5A). We did find two individuals, P11 and P12 where the *B. vulgatus* strain at Day 0 was related to the strain at Day 180. In the case of P11, the 5 and 3 regions at Day 0 and Day 180 fecal dominant sub strains were different, whereas the 5 and 3 sequences from P12 were the same at Day 0 and 180 (Figure 5A and Supplementary Figure 4A). For *B uniformis*, individuals P3 and P9 had related strains at Day 0 and 180 and both the 5 and 3 sequences were the same (Figure 5B and Supplementary Figure 4B).

**Discussion**

In this study, we have analyzed for the presence of genes encoding BSAP proteins in fecal *B. vulgatus* or *B. uniformis* strains from individuals with normal and perturbed GIT ecosystems. Our analysis combines the identification of BSAP genes (BSAP-2 and BSAP-3) with the use of BLAST and IGV tools to provide a comprehensive analysis of the fecal dominant *Bacteroides* spp.. strains. From the analysis of normal individuals (HMP data set), we found the BSAP phenotype was stable in an individual over times up to 6 months. The results from the two antibiotic data sets show disruption and recovery of the GIT from antibiotics results in a fecal dominant strain with a BSAP phenotype corresponding, most of the time, to the pre-antibiotic strain. Our studies support that the fecal dominant microbial strain is most probably maintained and selected from a stable microbial strain reservoir within the GIT and that in some individuals could be influenced by the BSAP phenotype.

Microbes are known to encode a variety of proteins that have the potential to provide a competitive advantage in a complex ecosystem [20–23]. Previous studies have shown the *Bacteroidales*-specific antimicrobial genes function in a species-specific fashion to restrict the replication and subsequent colonization [12, 13]. The analysis of the BSAP genes then provides opportunities to investigate species-species interactions. For example, the presence of the BSAP genes (positive phenotype) would provide a selective advantage over the *Bacteroidales* microbe without BSAP (negative phenotype) in the GIT. To further define the presence or absence of the BSAP gene in the population, we first focused on the well-characterized HMP data set that contained paired samples from individuals taken at two different times 3 months to 1 year apart [2, 17]. Using WSS analysis, we had previously determined that for the samples collected, the strains of *B. vulgatus* and *B. uniformis* of an individual were related to each other over this period [1]. Consistent with previous studies, we found most individuals either had a BSAP positive or negative phenotype that is consistent with previous studies that described the BSAP genes [12, 13]. However, our studies also demonstrate from examination of the longitudinal samples where we found that the BSAP phenotypes were stable over time in the unperturbed, normal GIT microbial ecosystem. Interestingly, we found BSAP positive in only a small subset of individuals.

To further characterize these BSAP phenotypes, we used the IGV tool to visualize the BSAP genes and surrounding 5’ and 3’ regions. For those samples that were BSAP positive, all had complete BSAP genes although we did find in some instances regions 5’ and 3’ had discrete deletions. In contrast, using the IGV
analysis of the BSAP-3 and BSAP-2 negative phenotypes we found for most of the negative phenotype samples that deletions were outside of the BSAP gene and that the 5' and 3' regions surrounding the deleted gene that were identical (or only a few nucleotides different) at two different time points. Surprisingly, we found different individuals that had shared sequence deletions (Figs. 2 and 3). The sharing between different individuals suggests a common mechanism for the generation of deletions. One possibility would be the deletions might be a result of a transposition event to delete the BSAP gene and surrounding DNA regions, although we did not find the consensus sequence for a known Bacteroides conjugative transposon (CTnDOT) [24, 25]. However, a previous study has reported that the Bacteroides genome region containing the BSAP gene was included in a region of integrative and conjugative elements which included numerous, yet undefined, putative transposases [26, 27]. We also found a few individuals with partial BSAP gene patterns were different between the sampling time points. Although the origin of these BSAP gene patterns is unknown, one possibility is that they occurred during a non-typical transposition process that left remnants of the BSAP gene. Since transposition is enhanced during environmental stress (e.g. nutritional changes or even disease), further understanding of the dynamics of the appearance of these partial BSAP phenotypes in an individual over time will be needed to determine if a correlation exists with microbial ecosystems that were or are currently under stress [28–30].

The analysis of the impact of antibiotics on the BSAP phenotypes provides information on the impact of a defined perturbation on the microbial strain community. Overall, the result from both antibiotic studies demonstrates the stability of the BSAP phenotype as evidenced by no change pre and post antibiotics. Indeed, for most individuals, there was no strain change occurred and from the IGV analysis, there were also no changes in the surrounding regions of the BSAP negative phenotype supporting that the fecal dominant strain is most probably selected from a stable strain reservoir which is not eliminated by the antibiotics. We did find a few instances where the BSAP phenotype was changed during recovery from the antibiotic. We found for individual P17 that the cephalexin treatment resulted in the appearance of a new strain at Day 7 that was replaced by the original strain on day 90 that had a positive BSAP-3 phenotype. Individual P4 of the cephalexin study, had a deleted BSAP-3 gene at days 0 and 7, with a B. vulgatus sub strain, however had a complete BSAP-3 gene at day 90. Similarly, for individual P1 of the Palleja et al. samples, the BSAP-3 phenotype changed from positive to negative from day 0 to day 7 post antibiotics. However, the fecal dominance of the Day 7 new strain was replaced by the pre antibiotic BSAP-3 positive strain at days 42 and 180, supporting the dominance of the BSAP positive phenotype in the GIT strain reservoir used as a source for the selection of the fecal dominant strain.

**Conclusions**

The results of our study using the three previously published longitudinal data sets containing metagenomic analysis provide new insights into the GIT ecosystem that contains the Bacteroides strain community. Specifically, the analysis of the stability of the BSAP phenotype under normal conditions and the recovery of the phenotype following perturbation support a role for the BSAP proteins in the microbe-microbe interactions that occur during the selection of the dominant fecal Bacteroides strain.
The results of our analysis also have implications for efforts to modify the functions of microbial communities, especially the *Bacteroides*, since these microbes are generally the most predominant in the human GIT [2, 8]. Since the BSAP phenotype for most individuals was either BSAP positive or negative that was stable and resilient after perturbation with antibiotics, our studies provide a foundation for the application of genome modification technology targeted to the BSAP genes to alter the *Bacteroides* microbial community microbe-microbe and host-microbe interactions to change the functions in the human GIT ecosystem [8, 31].

**Materials And Methods**

**Ethics statement.**

We obtained the publicly available original sequence files from the European Nucleotide Archive (accession numbers: PRJEB8094 for Raymond et al., ERP022986 for Palleja et al.) and https://portal.hmpdacc.org/ for the HMP data set. All participants’ samples were collected, sequenced, and deposited by Raymond et al., Palleja et al., and HMP. The sequence data that we used in this study was fully anonymized before being made publicly available.

**Publicly available data sets used in this study.**

In this study, we used 3 publicly available metagenomic data sets for healthy individuals 1) from the NIH Human Microbiome Project (HMP) [17], 2) pre and post treated with a single antibiotic (cefprozil) [18], and 3) pre and post treated with 3 antibiotics (meropenem, gentamicin, and vancomycin) [19]. For Raymond et al., fecal samples were collected from 18 individuals at three different time points: pre-treatment (Day 0), end of antibiotic treatment (Day 7), and 3 months post-treatment (Day 90), and we selected all individuals’ samples to run the analysis. We have also additionally selected 6 individual fecal samples that did not receive antibiotic treatment as controls. For the HMP data set, 30 individual samples that were previously used to establish our WSS analysis were selected to run the analysis [1, 17]. For the Palleja data set, fecal samples from 12 individuals were collected at five different time points: pre-treatment (Day 0), immediately after antibiotics treatment (Day 4), and three post-treatment time points (Day 8, 42, and 180). From this data set, we have selected all 12 individual samples at four different time points (pre and post treatment) to run the analysis.

**Analysis of Bacteroides BSAP genes.**

Before running the analysis, quality control steps were included removing any human reference genome (hg19) using bowtie2 (version 2.3.4.3) with default parameters [32], and filtering low quality reads (sliding window of 50 bases having a QScore < 20) using Trimomatic (version 0.36) [33]. Each fecal metagenomics sample was used to align with *Bacteroides vulgatus* CL09T03C04 for BSAP-3 and *Bacteroides uniformis* CL03T00C23 for BSAP-2 using Burrows-Wheeler aligner program (BWA; version 0.7.13) BWA tool [34]. Aligned reads from each reference genome were then sorted and indexed using SAMtools (version 0.1.19) [35]. The resultant bam file was converted to FASTQ format using BEDTools
Each converted FASTQ file was then assembled using MEGAHIT and the resultant contig file was selected for BLASTX search against the BSAP-2 and BSAP-3 sequence reads using BLAST+ [15, 36]. To visualize aligned Bacteriocin genes (BSAP-2 and BSAP-3) for paired samples for each individual, the VCF file generated for *B. vulgatus* and *B. uniformis* of each sample was uploaded to Integrative Genomics Viewer (IGV) and aligned to their reference genome [16]. For the BSAP-3, 2,178,016 to 2,183,973 regions of *B. vulgatus* were selected to display 5’ and 3’ sequences that also include BSAP-3 genes (location: 2,178,616 to 2,180,073). For the BSAP-2, 1,381,372 to 1,392,000 regions of *B. uniformis* were selected to show 5’ and 3’ sequences that also includes BSAP-2 genes (location: 1,383,872-1,385,398). To further characterize the BSAP negative phenotype, nucleotide sequences of the selected 5’ and 3’ regions were extracted using IGV and then uploaded to Jalview for visualization [16, 37].

To determine the relatedness of strains, individual paired samples were additionally used for the Window-based single-nucleotide variant (SNV) similarity (WSS) analysis which was previously developed based on the Human Microbiome Project (HMP) data set [1]. The resultant WSS score was used to compare against cut-off value that was previously established in our previous study (related strain pair: WSS score > cut-off; unrelated strain pair: WSS score < cut-off) [1, 38].

**Declarations**

**Data availability.**

The original sequencing data set of the stool samples used in this study were downloaded from the European Nucleotide Archive (accession numbers: PRJEB8094 for Raymond et al., ERP022986 for Palleja et al.) and https://portal.hmpdacc.org/ for the HMP data set.

**Ethics approval and consent to participate.**

Not applicable. We obtained the publicly available original sequence files from the European Nucleotide Archive (accession numbers: PRJEB8094 for Raymond et al., ERP022986 for Palleja et al.) and https://portal.hmpdacc.org/ for the HMP data set. All participants’ samples were collected, sequenced, and deposited by Raymond et al., Palleja et al., and HMP. The sequence data that we used in this study was fully anonymized before being made publicly available. The individual sources of the sequence data contain all relevant methods of collection, experimental protocols and informed consent considerations.

**Consent for publication.**

Not applicable.

**Competing interests.**

The authors declare no competing interests.

**Acknowledgments.**
We thank the UAB Information Technology Research Computing group for providing the high-performance computing support necessary for bioinformatics analyses. We also thank Heersink School of Medicine for supporting this study.

**Funding**

The authors received no specific funding for this work.

**Author contributions.**

H.K. and C.D.M. conceived the study. H.K. contributed bioinformatics analyses on sequencing data sets. H.K., and C.D.M. wrote the manuscript. All authors approved the final manuscript.

**References**


Figures
### Figure 1

**BSAP summary for HMP data set.**

The number of individual pairs that (A) both pairs were BSAP-2 and BSAP-3 positive and (B) both pairs were BSAP-2 and BSAP-3 negative.

<table>
<thead>
<tr>
<th>Samples</th>
<th>BSAP-3</th>
<th>No.</th>
<th>BSAP-2</th>
<th>No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>a</td>
<td>positive</td>
<td>11</td>
<td>positive</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>positive</td>
<td></td>
<td>positive</td>
</tr>
<tr>
<td>B</td>
<td>a</td>
<td>negative</td>
<td>19</td>
<td>negative</td>
</tr>
<tr>
<td></td>
<td>b</td>
<td>negative</td>
<td></td>
<td>negative</td>
</tr>
<tr>
<td></td>
<td>BSAP-3</td>
<td>BSAP-3</td>
<td>BSAP-3</td>
<td>BSAP-3</td>
</tr>
<tr>
<td>-------</td>
<td>------------------------</td>
<td>------------------------</td>
<td>------------------------</td>
<td>------------------------</td>
</tr>
<tr>
<td>S28a</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
</tr>
<tr>
<td>S28b</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
</tr>
<tr>
<td>S31a</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
</tr>
<tr>
<td>S31b</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
</tr>
<tr>
<td>S66a</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
</tr>
<tr>
<td>S66b</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
</tr>
<tr>
<td>S16a</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
</tr>
<tr>
<td>S16b</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
</tr>
<tr>
<td>S21a</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
</tr>
<tr>
<td>S21b</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
<td>TTTTTAAGAAATTAGGAAT</td>
</tr>
</tbody>
</table>

**Figure 2**

**BSAP-3 gene alignment for HMP data set.**

Individual pairs that both pairs had BSAP-3 negative were selected to extract aligned sequence reads that included the BSAP-3 gene along with 5’ and 3’ of the BSAP-3 gene (20 base pairs for each 5’ and 3’ end). (A) Both individual pairs had exacted 5’ and 3’ sequence reads; (B) minor differences (less than 5 nucleotides) were observed between individual pairs; (C) major differences (more than 5 nucleotides) were detected between individual pairs.
BSAP-2 gene alignment for HMP data set.

Individual pairs that both pairs had BSAP-2 negative were selected to extract aligned sequence reads that included the BSAP-2 gene along with 5’ and 3’ of the BSAP-2 gene (20 base pairs for each 5’ and 3’ end). (A) Both individual pairs had exacted 5’ and 3’ sequence reads; (B) minor differences (less than 5 nucleotides) were observed between individual pairs; (C) major differences (more than 5 nucleotides) were detected between individual pairs. “Y” and “S” are degenerate nucleotide code: “Y” indicates C/T and “S” indicates C/G (https://people.bath.ac.uk/jm2219/biology/degenerate.htm).
Gene analysis that includes WSS and BLAST was conducted for the Raymond et al. data set. For each sample, strain relatedness was determined by WSS analysis, and the presence/absence of (A) BSAP-3 and (B) BSAP-2 genes were determined by BLAST analysis. The “pos” indicates BSAP-3/BSAP-2 gene was observed; “partial” indicates sequence fragment of BSAP-3/BSAP-2 gene was observed; “neg” indicates BSAP-3/BSAP-2 gene was not observed; “No score” indicates that no score was observed from WSS analysis.
Gene analysis that includes WSS and BLAST was conducted for the Palleja et al. data set. For each sample, strain relatedness was determined by WSS analysis and the presence/absence of the (A) BSAP-3 and (B) BSAP-2 genes were determined by BLAST analysis. “R” indicates related strain, “NR” indicates not related strain observed, “NS” indicates no WSS score observed from WSS analysis. The “pos” indicates BSAP-3/BSAP-2 gene was observed; “partial” indicates sequence fragment of BSAP-3/BSAP-2 gene was observed; “neg” indicates BSAP-3/BSAP-2 gene was not observed; “No score” indicates that no score was observed from WSS analysis.

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.
• SupplementaryFiguresR1.pdf
• SupplementaryTables.pdf