Pyrosequencing to Reduce Time for Detection in Human Sepsis

MANUEL CABALLERO, CIV
JAMIE L. MYERS
HUI XIA

FINAL REPORT

May 2018

59th Medical Wing
Office of the Chief Scientist
2520 Ladd Street, BLDG. 3885
JBSA Lackland AFB, TX 78236-7517

DISTRIBUTION A. Approved for public release; distribution is unlimited.

DECLARATION OF INTEREST

The views expressed in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the Air Force, Department of Defense, nor the U.S. Government. This work was funded by Project Code Number AC12EM01. Authors are military service members, employees, or contractors of the US Government. This work was prepared as part of their official duties. Title 17 USC §105 provides that 'copyright protection under this title is not available for any work of the US Government.' Title 17 USC §101 defines a US Government work as a work prepared by a military service member, employee, or contractor of the US Government as part of that person's official duties.

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

Qualified requestors may obtain copies of this report from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

PYROSEQUENCING TO REDUCE TIME TO DETECTION IN HUMAN SEPSIS.

JOSEPH H. LYNCH, DAF
Medical Modernization Program Analyst
Integrated Clinical Medicine
59th Medical Wing-Science & Technology

CARLTON C. BRINKLEY, Ph.D., DAF
Dir., Diagnostics & Therapeutics Research
59th Medical Wing-Science & Technology

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings

Pyrosequencing to Reduce Time for Detection in Human Sepsis

Jamie L. Myers[1,2], Hui Xia[1], Manuel Y. Caballero[1]

[1]Center for Advanced Molecular Detection, Chief Scientist's Office, Science and Technology, 59th Medical Wing, US Air Force, JBSA-Lackland, San Antonio, Texas 78236
[2]Current Address: Hematology and Oncology Division, Department of Medicine, The University of Texas Health Science Center, San Antonio, Texas 78229

KEYWORDS

Pyrosequencing, sepsis, pathogen, PyroMark Q24 Advanced, BioMatrix µSeq Sepsis Diagnostic Sequencing kit

"The views expressed are those of the author's and do not reflect the official views or policy of the Department of Defense or its Components"

ABSTRACT

Sepsis results from systemic presence of infectious agents, and it involves dysregulated immune response to such infections. Because sepsis can progress to varying degrees of tissue and organ failure, with nearly 40 % mortality rate, it is imperative to implement effective treatment modalities as early as possible. To do that, rapid and precise identification of the sepsis-causing infectious agents is important. Although the traditional microbiological methods can identify the causative agents, they may require days to do so. Pyrosequencing is a technique suited for sequencing relatively short DNA molecules, which could be generated by PCR. This project's aim was to **1)** assess the utility of the BioMatrix µSeq Sepsis Diagnostic Sequencing kit, **2)** coupled with pyrosequencing using Qiagen's PyroMark Q24 Advanced System. The PyroMark Q24 Advanced was selected for this project based on its small footprint, user-friendly software and easy-to-use pyrosequencing protocol. The kit utilizes three primer sets, each specific for amplifying a bacterial 16S rRNA gene segment encompassing the hypervariable region V1, V2, or V3. The segments thus amplified are then sequenced using PyroMark Q24 Advanced. The combined approach is reputed to be useful for expedited identification of bacteria. Further, the approach is said to require only hours to identify the bacteria, unlike the traditional microbiological approaches that require culturing bacteria, and thus take days. To assess this combined approach, we used commercially available purified genomic DNAs of 31 bacteria. The DNAs were used singly or in combinations to assess the technique.
This testing and evaluation study has resulted in several findings: **1)** BioMatrix µSeq Sepsis Diagnostic Sequencing kit primers did not result in clean amplicons; the PCR reaction mixtures had unexpected products in the no template control, as evidenced by agarose gel electrophoresis. **2)** PCR with the same primers from Life Technologies did not result in unintended amplicons. **3)** The length of sequences generated did not approach the length expected from PyroMark Q24 Advanced. Moreover, the sequences generated had varying degrees of error and reliability. **4)** BLAST analysis performed using the sequences resulted in precise identification of the genus and species for some bacteria, but not some others. No sequence generated by pyrosequencing afforded definitive identification of any bacterium to subspecies or strain level. Given that the 16S rRNA gene hypervariable region sequences can be used to accurately identify the target bacteria, we think in this study the identification failures resulted from lack of sufficiently long sequences.

INTRODUCTION

Sepsis is a condition in which the body's defense mechanisms can lead to varying degrees of inflammatory response. The clinical and pathologic consequences of sepsis can range from mild and quite treatable to life-threatening. Depending on its severity, sepsis can lead to serious organ dysfunction, even failure, and ultimately death (Singer *et al*., 2016). Sepsis can result from the presence of pathogens in various parts of the body, as well as body fluids such as blood, urine, and lymph. Sepsis is often fatal when not diagnosed and treated early. However, there be can problems both in precisely diagnosing the condition and its causative agent. Imprecise diagnosis can result in implementation of improper and ineffective treatment regimes, potentially leading to fatal consequences (Gaieski *et al*., 2013; Peterson and Chase, 2017). The antibiotics prescribed, for example, may not work well or not at all if the causal agent is incorrectly identified and for which the antibiotic proves to be the wrong type. Not only that, the cost of treating sepsis is also enormous; in 2013, for example, the overall cost of sepsis treatment was estimated to be nearly $23 billion, making it among the most expensive conditions to treat (Torio and Moore, 2016).

High rate of morbidity and mortality associated with sepsis has been a persistent concern for the US Military, especially in the battlefield arenas, where precise, expeditious diagnoses and effective treatments are often not feasible. For example, it was recently reported that traumas that result from combat-related injuries have higher fatality rates, especially when the wounded do not receive fast and timely surgical and drug treatments (Ma *et al*., 2016).

The common diagnostic laboratory practices in use for identifying the causative agents of sepsis heavily rely on traditional microbiological and biochemical methods, which require culturing the organisms for precise identification, thereby extending the time from sample isolation to pathogen identification into many days. The difficulty in growing certain bacteria, especially when in very low numbers in the samples, further frustrate the need to identify the sepsis-causing pathogens expeditiously. Further, the patient samples can carry more than one pathogen. Thus, the need is not just to identify one or a few organisms, but to rapidly identify all pathogens in the samples to pin down the ones that cause sepsis. Clearly the need to develop more efficient and fast approaches and methods to simultaneously identify multiple sepsis bacteria is crucial, and it would have great usefulness for the military. The BioMatrix µSeq Sepsis Diagnostic Sequencing approach in combination with the Qiagen pyrosequencing machine Q24 Advanced is one such candidate approach (Motoshima *et al*., 2012; Chikamatsu *et al*., 2018). The overall goal of this work was to evaluate this system for expeditious and accurate identification of bacteria, and to see whether further applications of this system within the military would be warranted and feasible.

MATERIALS AND METHODS

Reagents. BioMatrix µSeq Sepsis Diagnostic Sequencing kit was purchased from BioMatrix Sciences (Rancho Santa Fe, CA). The BioMatrix kit primers were also ordered from Life Technologies (Grand Island, NY). The 31 purified bacterial genomic DNAs, listed in Table 1, were bought as separate preparations from ATCC (Gaithersburg, MD). The PCR master mix (DNA polymerase, dNTPs, buffer) was purchased from Promega (Madison, WI). All reagents for pyrosequencing on PyroMark Q24 Advanced were from the manufacturer of the instrument (Qiagen, Germantown, MD).

PCR. The DNA concentrations in the 31 commercial preparations were determined fluorometrically using Qubit (ThermoFisher Scientific, Waltham, MA). The preparations were then diluted to 20 ng/µL for PCR. The initial PCR work was carried out using the BioMatrix kit primers. PCR was also performed with the primers from Life Technologies, and the sequences of these primers were identical to those from BioMatrix. There were three sets of primers, each specific for amplifying a segment of the bacterial 16S rRNA gene containing the hypervariable region V1, V2, or V3 (Table 2). The reverse primer in each set was biotinylated at the 5' end by the suppliers. Each PCR reaction mixture contained 1X PCR Master Mix, 0.2 µM each of forward and reverse primers, and 1-30 ng of template DNA. The final reaction volume was 25 µL. PCR was performed on a ProFlex PCR thermocycler (ThermoFisher Scientific), and the cycling parameters were as follows: 94°C for 5 minutes, followed by 35 cycles of 94°C for 20 seconds, 54°C for 20 seconds, and 72°C for 30 seconds. At the end of 35 cycles, a final step of 72°C for 5 minutes was also included. Following amplification, the PCR mixtures were analyzed by agarose gel electrophoresis (2% E-Gel EX; ThermoFisher Scientific).

Pyrosequencing. PyroMark Q24 Advanced was used for pyrosequencing of PCR products. The sequencing procedure was as directed by the manufacturer (Qiagen). The PyroMark Q24 Advanced software was used to create run files with the appropriate sample and assay information. The dispensation order for all sequencing reactions was 25 (dNTP addition order, CTGA or TGCA). The biotinylated PCR products were bound to sepharose beads, and then the amplicons were separated into single-stranded DNA using the Vacuum Prep Workstation as directed by the manufacturer (Qiagen), leaving the biotinylated ssDNA strands bound to the beads. The forward primers As9b, V3S, and V1b were used for pyrosequencing. The primers were diluted in the PyroMark annealing buffer, and the final concentration of each primer was 0.375 µM. For pyrosequencing, the mixtures containing the biotinylated strands of amplicons and the sequencing primers were first heated to 80°C for 5 minutes, followed by application into the PyroMark Q24 Advanced sample portal.

Data analysis. Each file run was analyzed using the PyroMark Q24 Advanced software (version 3.0.0, build 21). To find the sequence matches, the final sequences determined by pyrosequencing were analyzed by BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi).

Table 1. List of bacteria for which the purified genomic DNA was used for this study.

No.	Bacteria	ATCC ID	Lot#
1	*Acinetobacter baumannii* strain AYE	BAA-1710D-5	59333495
2	*Aeromonas hydrophila* subsp. *hydrophila* ATCC 7966	7966D-5	57897823
3	*Bacteroides fragilis* strain VPI 2553	25285D-5	60613545
4	*Clostridium perfringens* strain NCTC 8237	13124D-5	61570257
5	*Enterobacter aerogenes* strain IFO 12010	15038D-5	59861894
6	*Enterobacter cloacae* subsp. *cloacae* strain CDC 442-68	13047D-5	27863845
7	*Enterococcus faecalis* strain V583	700802D-5	60628801
8	*Enterococcus faecium* strain MMC4	51559D-5	59679205
9	*Haemophilus influenzae* strain Rd	51907D	2662083
10	*Klebsiella oxytoca* strain LBM 90.11.033	700324D	3573171
11	*Legionella pneumophila* subsp. *pneumophila* strain Philadelphia-1	33152D-5	60360151
12	*Listeria monocytogenes* strain Li 23	19114D-5	57878011
13	*Mycobacterium avium* subsp. *paratuberculosis* strain K-10	BAA-968D-5	61460825
14	*Mycobacterium tuberculosis* strain X004439	BAA-2236D-2	61646488
15	*Neisseria meningitidis* serogroup B	53415D-5	62082502
16	*Pseudomonas aeruginosa* strain PAO1-LAC	47085D-5	62538828
17	*Serratia marcescens* strain CDC 3100-71	27137D-5	59679187
18	*Shigella flexneri* strain 24570	29903D-5	7502841
19	*Staphylococcus aureus* subsp. *aureus* strain TCH1516	BAA-1717D-5	61274435
20	*Staphylococcus epidermidis* FDA strain PCI 1200	12228D-5	59867137
21	*Staphylococcus haemolyticus* strain SM 131	29970D-5	57700713
22	*Staphylococcus hominis* subsp. *novobiosepticus* strain R22	700236D-5	58120539
23	*Staphylococcus lugdunensis* strain N860297	43809D	3082088
24	*Staphylococcus saprophyticus* subsp. *saprophyticus* strain NCTC 7292	15305D-5	58083812
25	*Staphylococcus schleiferi* subsp. *schleiferi* strain N850274	43808D-5	63756347
26	*Stenotrophomonas maltophilia* strain 810-2	13637D-5	57972904
27	*Streptococcus agalactiae* strain 2603 V/R	BAA-611D-5	61793995
28	*Streptococcus mitis* strain NCTC 12261	49456D-5	57968143
29	*Streptococcus mutans* Clarke	25175D-5	62923047
30	*Streptococcus pyogenes* strain SF370; M1 GAS	700294D-5	61246892
31	*Streptococcus sanguinis* strain SK36	BAA-1455D-5	57805007

Table 2. List of BioMatrix µSeq Sepsis Diagnostic Sequencing Kit primers used for PCR.

16S rRNA Hypervariable Region	Primer	Sequence	Expected Amplicon (bp)
V1	*Forward: V1b	5’ GYR TTA CTC ACC CGT YCG CCR C	114
	Reverse:Bio-pBR5	5’ [Biotin] GAA GAG TTT GAT CAT GGC TCA G	
V2	*Forward: As9b	5’ CGG CTG GCA CGK AGT TAG CC	200
	Reverse: Bio-As5	5’ [Biotin] ACA CGG YCC AGA CTC CTA C	
V3	*Forward: V3S	5’ GAC ARC CAT GCA SCA CCT	100
	Reverse: Bio-V3F	5’ [Biotin] GCA ACG CGA AGA ACC TT	

* These primers were also used for pyrosequencing.

RESULTS

The overall focus of this project was to evaluate the BioMatrix µSeq Sepsis Diagnostic Sequencing Kit for identification of the bacteria commonly associated with sepsis, but also found in various types of lesions, such as the skin wounds. The kit is devised to work well in combination with pyrosequencing on PyroMark Q24 Advanced instrument (Qiagen). The process comprises two main experimental techniques, performed sequentially: First, regular PCR using the kit primers, which can be done on any suitable thermocycler. Second, pyrosequencing of the unique PCR products thus generated, done on the PyroMark Q24 Advanced instrument. The sequences are then analyzed using sequence search tools to find the DNA sequences that match the newly determined sequences. The most common tool for such searches is BLAST (NCBI), which searches for matches in various types of databases in an all-encompassing random manner, as well as with the desirable search restrictions.

First, pilot experiments were performed using the BioMatrix kit, as described in the Materials and Methods section. These experiments aimed to assess the kit components to amplify the V2 hypervariable region of the 16S rRNA gene. The templates for the pilot experiments were purified genomic DNAs of *Streptococcus pyogenes* strain SF370 (M1 GAS) and *Shigella flexneri* strain 24570. The PCR products were analyzed by agarose gel electrophoresis.

The results from the pilot experiments show that the expected V2 region 200-bp segment amplified from both genomes (Figure 1A). However, the product was more robust when the template used was *S. flexneri* strain 24570 genome. The negative template control (NTC; PCR without any template DNA) gave an unexpected faint band that appeared to be of approximately the same size as the expected 200-bp product (Figure 1A). We reasoned that this band may have resulted from inadvertent contamination with one of the two templates, or it may be a random, nonspecific amplicon. We therefore repeated the NTC experiment. Two repeat experiments yielded the same unexpected band (Figure 1B, 1C). Together, these results suggested the possibility that the BioMatrix kit components may have contamination with an unknown template.

The experiments with the BioMatrix kit primers specific for the V1 and V3 hypervariable regions were performed in the same way as with the V2 region primers. The products were then analyzed by agarose gel electrophoresis. The expected amplicon for the V1 region primers is 114 bp and that for the V3 region 110 bp. The results in Figure 2 show that, unlike the results with the V2 region primers, the NTC PCR with V1 and V3 region primers did not result in any unexpected amplicons. Further, while both sets of primers gave robust amplicons for *Shigella flexneri* strain 24570, neither set amplified the expected amplicon when the genomic DNA of *Streptococcus pyogenes* strain SF370 (M1 GAS) was used as the template (Figure 2 A, B). However, the lack of amplification from the *S. pyogenes* DNA was not reproducible; it likely resulted from inadvertent absence of the target genome (see Figure 4).

As described above, the V2 region primers in the BioMatrix kit repeatedly resulted in an unexpected amplicon (Figure 1 A, B, C), suggesting the kit reagents to be the source of this unexpected amplification. We therefore decided to get the V2 primers custom-made by a different vendor (Life Technologies). To keep uniformity of the source, we also got the V1 and V3 primers from Life Technologies. A pilot negative template control experiment with the Life Technologies V1, V2, and V3 primers gave no unexpected amplicon (Figure 3). We then performed parallel experiments with all three sets of primers from both companies. These results clearly showed that whereas the BioMatrix kit primers resulted in nonspecific bands, the same primers from Life Technologies did not. Further, PCR done with the V1 and V3 region primers from Life Technologies, robustly amplified the respective target fragments from both

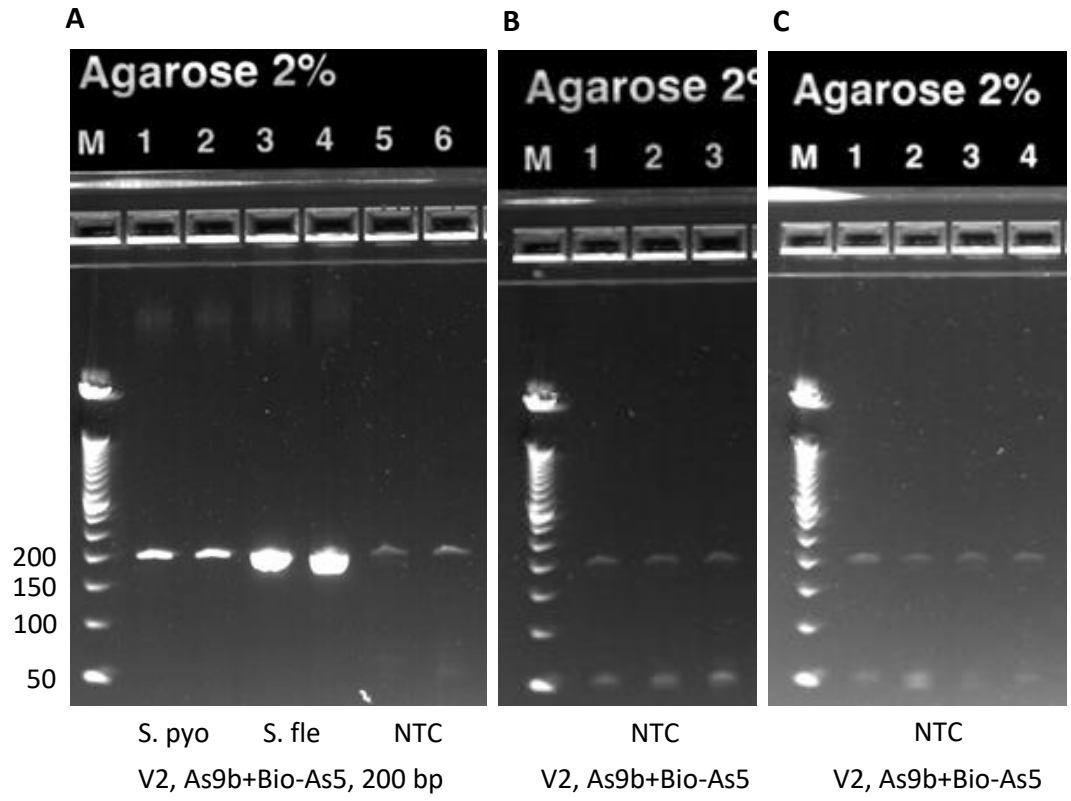
organisms (Figure 4 A, B, E, F). But the BioMatrix V1 and V3 primers also robustly amplified the target segments from the *Shigella flexneri* DNA, as well as from the *Streptococcus pyogenes* strain SF370 (M1 GAS) DNA (Figure 4 A, B, E, F). These results show that the absence of *S. pyogenes* amplicon in Figure 2 resulted from absence of the target DNA from the PCR reaction, a likely inadvertent omission. Strategies to mitigate any further omission of DNA template were put in place by having the lab techs repeat each other's work in case of a failed amplification result.

Because the Life Technologies primers gave much cleaner results for all three target amplicons (V1, V2, V3), we carried out the rest of the work with these primers. The original purpose of pyrosequencing with the BioMatrix µSeq Sepsis Diagnostic Sequencing Kit had to be reconsidered due to contamination issues with their product. Therefore, the first purpose of this project was now to assess Life Technologies V1, V2, and V3 oligos for pyrosequencing. The amplicon sequences generated by pyrosequencing were analyzed by BLAST to identify the sequence matches in the database. Details of this analysis appear in the BLAST Analysis Results section.

Figure 1. PCR performed with BioMatrix µSeq Sepsis Diagnostic Sequencing Kit primers As9b and Bio-As5.

The *Streptococcus pyogenes* strain SF370 (M1 GAS) and *Shigella flexneri* 16S rRNA gene V2 regions were independently amplified in separate PCR reactions. The PCR mixtures were then analyzed on agarose gels.

Discussion. A) The expected 200 bp V2 region band is present for both organisms. The results show that the kit primers amplify the target fragment more robustly when the template DNA is *S. flexneri*. However, the no-template control (NTC) also has about the same size faint band of unknown identity. We considered the possibility that it may have resulted from contamination with one of the two template DNAs at CAMD, or the product reagents may have been contaminated at the supplier facilities. **B, C)** Additional experiments performed to see if the nonspecific NTC band in **A** is reproducible; clearly it is. Note that each of these experiments was performed using a freshly opened pouch of the BioMatrix kit reagents; this was done to avoid any carryover of contaminated pouch used for **A**. Together these results suggest that the source of nonspecific band is the kit mixture, not the two purified genomic DNAs.



M, 50-bp molecular weight marker ladder. 50, 100, 150, and 200 mark the band sizes in bp. The same ladder was used for all 3 gels. **A,** S. pyo, *Streptococcus pyogenes* strain SF370 (M1 GAS) (lanes 1-2); S. flex, *Shigella flexneri* (lanes 3-4); NTC, no template control (lanes 5-6). **B,** Second experiment no template control done with a freshly opened pouch of the BioMatrix reagents (lanes 1-3). **C,** Third experiment no template control done with another freshly opened pouch of reagents from BioMatrix (lanes 1-4).

Figure 2. PCR performed with BioMatrix µSeq Sepsis Diagnostic Sequencing Kit primers specific for the 16S rRNA gene hypervariable regions V1 and V3.

A, Gel analysis of products resulting from PCR with V1 region primers V1b and Bio-pBR5. **B,** Gel analysis of products resulting from PCR with V3 region primers V3S and Bio-V3F.

Results and Discussion. The results show that, unlike with the V2 region primers (Figure 1), PCR with the BioMatrix primers for the V1 and V3 regions did not result in any nonspecific or unexpected amplicons. Both sets of primers also amplified the expected bands from *Shigella flexneri* DNA; 114 bp for the V1 region and 100 bp for the V3 region. But neither primer set resulted in the expected amplicon for *Streptococcus pyogenes* strain SF370 (M1 GAS). Note, however, that this result was not reproducible, and indeed the BioMatrix primers did amplify the *S. pyogenes* V1 and V3 amplicons. The anomalous result shown in this figure evidently resulted from absence of any template DNA (See Figure 4).

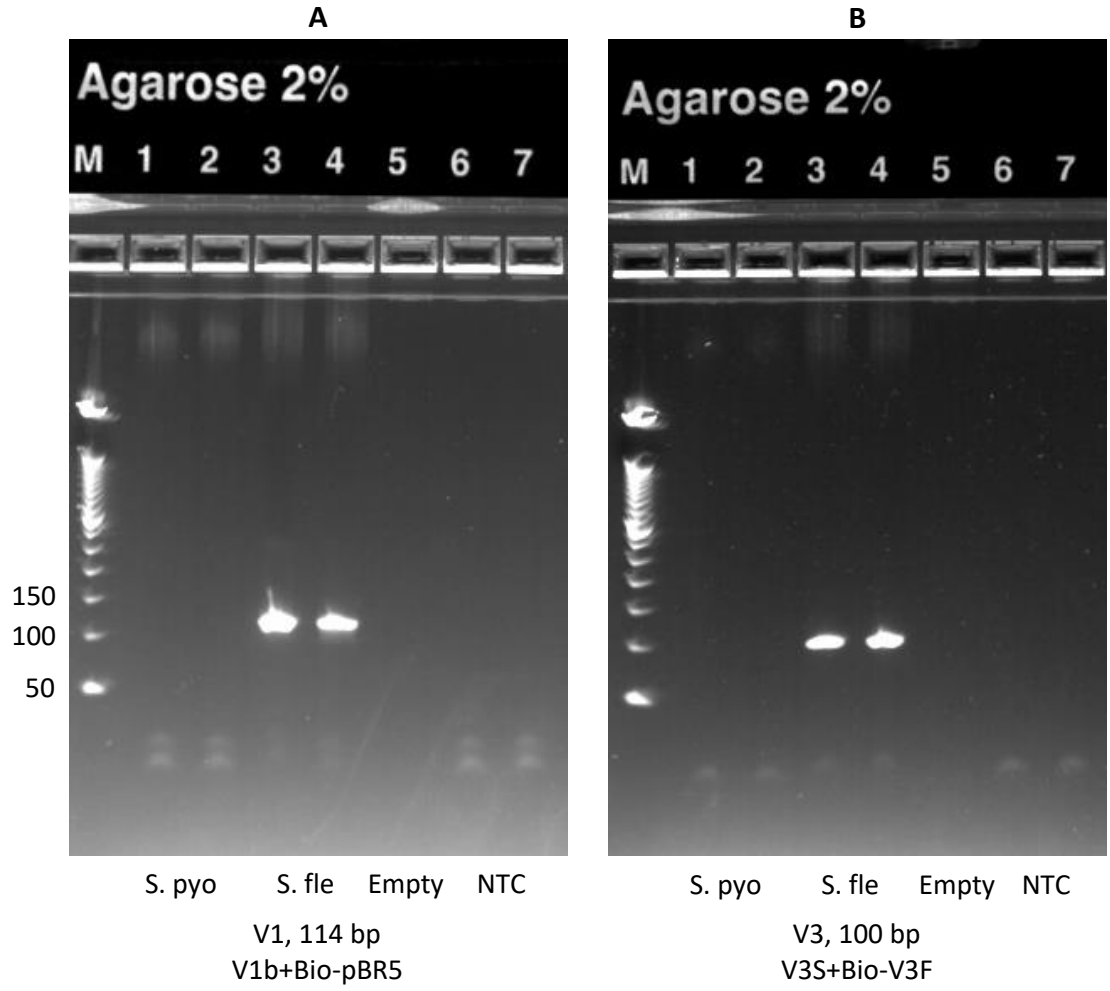


Figure 3. PCR and gel analysis of 16S rRNA gene V1, V2, and V3 regions using Life Technologies primers.

A, Gel analysis of products resulting from PCR with V2 region primers As9b/Bio-As5. **B,** Gel analysis of products resulting from PCR with V3 and V1 region primers V3S/Bio-V3F and V1b/Bio-pBR5.

Results and Discussion. PCR with the Life Technologies primers for the V1, V2 and V3 regions did not result in any nonspecific or unexpected amplicons. The three sets of primers also amplified the expected bands from *Streptococcus pyogenes* DNA; 200 bp for V2, 114 bp for the V1 region and 100 bp for the V3 region.

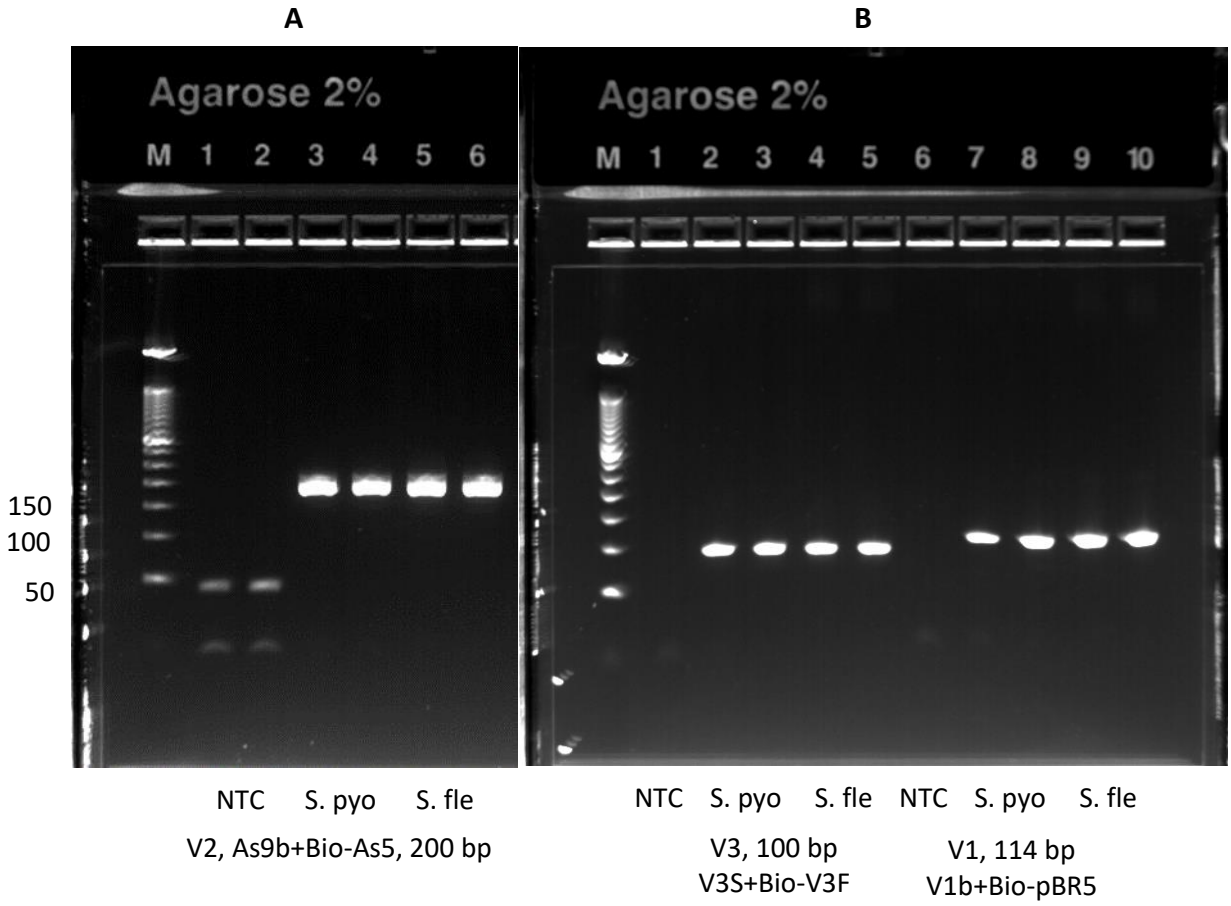
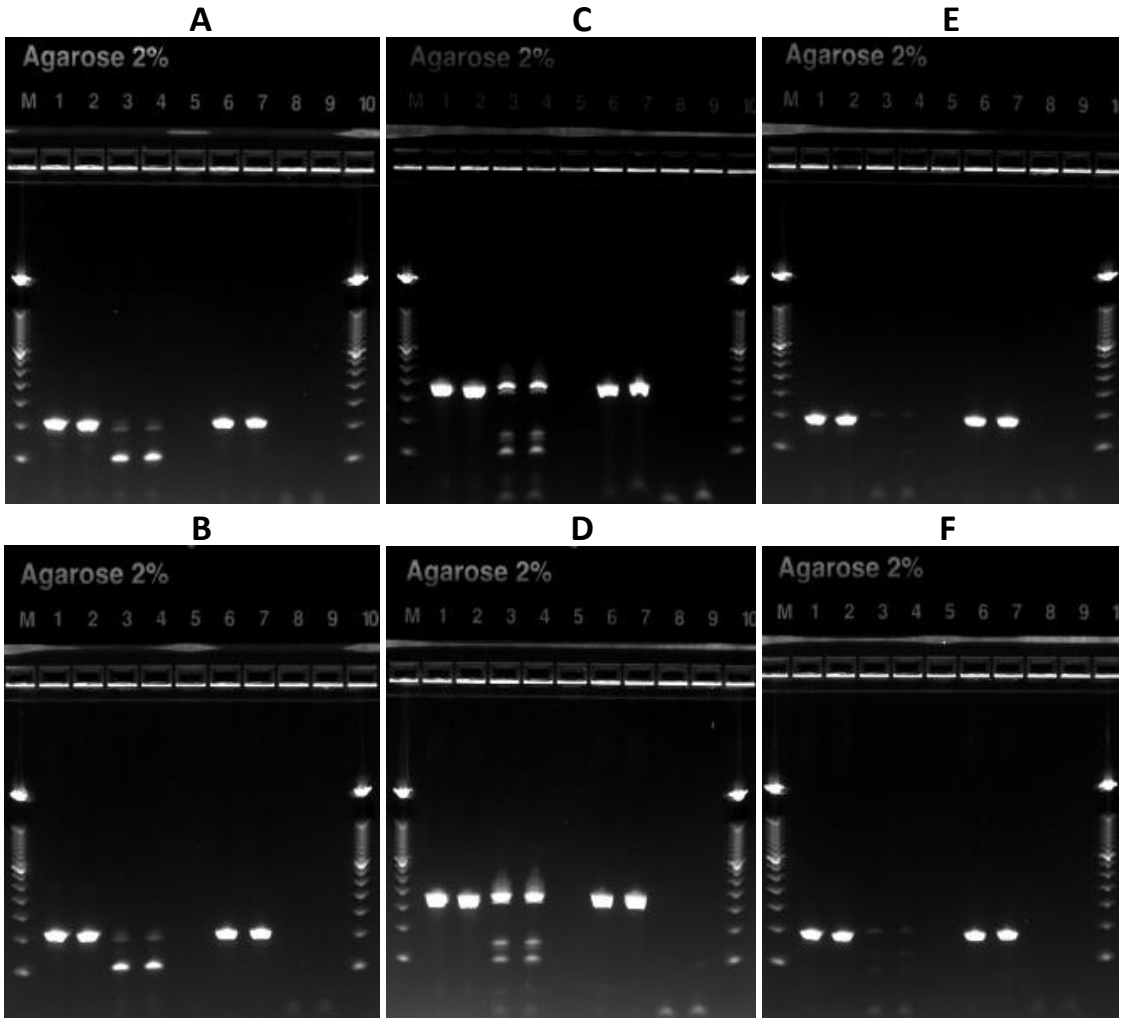


Figure 4. Parallel PCR and gel analysis of 16S rRNA gene V1, V2, and V3 regions using BioMatrix and Life Technologies primers.

A, B: V1 region analysis of *S. pyogenes* strain SF370 (M1 GAS) (A) and *S. flexneri* (B). **C, D:** V2 region analysis of *S. pyogenes* strain SF370 (M1 GAS) (C) and *S. flexneri* (D). **E, F:** V3 region analysis of *S. pyogenes* strain SF370 (M1 GAS) (E) and *S. flexneri* (F). In all panels, L1-2, segments amplified with BioMatrix primers; L6-7, segments amplified with Life Technologies primers; L3-4, NTC with BioMatrix primers; L8-9, NTC with Life Technologies primers; L5, empty; and M and L10, molecular weight marker ladder.

Results and Discussion. The primer sets from both companies strongly amplified the target segments of *S. pyogenes* strain SF370 (M1 GAS) and *S. flexneri* strain 24570 16S rRNA genes. However, all three BioMatrix primer sets resulted in unintended bands, which were the most noticeable for the V1 (A, B) and V2 (C, D) regions, but were also faintly visible for V3 (E, F). The same primers synthesized by Life Technologies resulted in no unintended bands. Together these results suggest that the appearance of nonspecific bands was not due to some intrinsic property of the primers *per se*, but rather that the preparations from BioMatrix were contaminated with some kind of DNA. We did not pursue these issues further. Instead, we decided to do the rest of the work with the Life Technologies primers. This work was done in February 2018 to retrospectively assess and verify whether the lack of *S. pyogenes* V1 and V3 target regions amplification with BioMatrix primers was fortuitous (Figure 2 A, B). As is clear from the results here, the lack of amplification was indeed accidental.



BLAST Analysis Results for Sequences Determined by Pyrosequencing

The salient BLAST analysis results for V1, V2, and V3 hypervariable region sequences for each bacterium follow. A short discussion also accompanies each set of results. The notes below are included to facilitate understanding of these results:

1. The primers used for pyrosequencing were V1b for the V1 region amplicon, As9b for the V2 region amplicon, and V3S for the V3 region amplicon.

2. The assignment of colors to sequence segments was by the PyroMark 24 Advanced pyrosequencing software. Based on the January 2016 PyroMark Q24 Advanced User Manual, the software assigned the yellow color for "check" and red "failed." The blue color signified "passed," and in this report it is shown as normal black. However, the BLAST analysis approach using these sequences was non-presumptive, and therefore all sequences were used to perform the analysis regardless of the color assignment by the software (see the Summary section).

3. To do the BLAST analysis, the V1, V2, and V3 amplicon sequences for each bacterium were first assembled into one sequence, but with nnnnn spacer separation. Thus, the order of these sequences in the single assembled sequence was V1nnnnnV2nnnnnV3. Initially, multiple searches were performed with a few sequences, even one by one using V1, V2, or V3 amplicon sequences. But the results were more consistent and reliable when the searches were performed using the assembled sequences, either as V1V2V3 or V1nnnnnV2nnnnnV3. This was done for all 31 bacteria listed in Table 1.

4. For each organism, separate BLAST searches were performed with the database setting as "nr/nt" or "16S ribosomal RNA (Bacteria and Archaea)." For each database setting, the "Program Selection" was either "Highly similar sequences (megablast)" or "Somewhat similar sequences (blastn)."

5. In this report, the alignments shown are only those resulting from the analysis of V1nnnnnV2nnnnnV3 sequence assemblies, with the settings "16S ribosomal RNA (Bacteria and Archaea)" and "Highly similar sequences (megablast)," unless stated otherwise for any alignment.

6. In the alignments, the term "Query" refers to the sequences generated by pyrosequencing and used for BLAST analysis. The term "Sbjct" refers to the database sequences that match the "Query" sequences to varying degrees.

7. The BLAST-generated terms "Range 1," "Range 2," and "Range 3" are not synonymous with alignments of V1, V2, and V3 amplicon segments, respectively. The amplicon segments should be recognized in "Query" sequences by the consecutive assembly and numbering scheme; that is, V1, then V2, and then V3.

8. Because each V1nnnnnV2nnnnnV3 format assembled sequence contains nnnnn, the maximum query coverage cannot be 100 % even if the match for all nucleotides is 100 %.

1. ***Acinetobacter baumannii* strain AYE**

   a. **V1 amplicon sequence**

   TAGGTCCGGT AGCAAGCTAC CT[T]CCCCGCC TCGACT[T]GCA TGTGTTA[A]GC TGCCGCCAGC [C]GT[T]CAATC

   b. **V2 amplicon sequence**

   GGTGCTTATT CTGCGAGTAA CGTCCACTAT CTCTAGGTAT TAACTAAAGT AGCCTCC[C]

   c. **V3 amplicon sequence**

   GTATCTAGAT TCCCGAAGGC ACCAATCCAT CTCTG[G]AA[A]G TTCTAGTATG TCAAGGCCAG GTAAGGTTC

   d. **BLAST Analysis Results and Discussion.** Two BLAST analyses were performed with search settings as "nr/nt" or "16S ribosomal RNA (Bacteria and Archaea)." Both approaches identified *A. baumannii* strains, placing them first in the BLAST generated table. The highest values for total score, query coverage, E-value, and identity for "nr/nt" search were 1486, 94 %, 3e-24, 99 %, respectively. These metrics were only for one bacterium, *A. baumannii* strain AR 0078 (complete genome); all other *A. baumannii* had lower values, reflecting varying degrees of sequence mismatches. The corresponding values for the 16S setting were 229, 61 %, 5e-28, and 99 %. Note that "nr/nt" BLAST search did not align the V1 sequence with any bacteria, except *A. baumannii* strain AR 0078 (complete genome), while the 16S rRNA database search didn't align the V1 sequence with any bacterium. The BLAST search results show that despite exclusion of V1 sequence, the analysis succeeded in identifying *A. baumannii*.

   The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Acinetobacter baumannii strain CIP 70.34 16S ribosomal RNA gene, partial sequence
Sequence ID: NR_116845.1 Length: 1528 Number of Matches: 3
Related Information

Range 1: 964 to 1033 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
119 bits(62)	4e-27()	69/70(99%)	1/70(1%)	Plus/Minus	

Features:

```
Query  138   GTATCTAGATTCCCGAAGGCACCAATCCATCTCTGGAAAGTT-CTAGTATGTCAAGGCCA  196
             |||||||||||||||||||||||||||||||||||||||||| ||||||||||||||||
Sbjct  1033  GTATCTAGATTCCCGAAGGCACCAATCCATCTCTGGAAAGTTTCTAGTATGTCAAGGCCA  974

Query  197   GGTAAGGTTC  206
             ||||||||||
Sbjct  973   GGTAAGGTTC  964
```

Range 2: 436 to 492 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
110 bits(57)	3e-24()	57/57(100%)	0/57(0%)	Plus/Minus	

Features:

```
Query  75   GGTGCTTATTCTGCGAGTAACGTCCACTATCTCTAGGTATTAACTAAAGTAGCCTCC  131
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  492  GGTGCTTATTCTGCGAGTAACGTCCACTATCTCTAGGTATTAACTAAAGTAGCCTCC  436
```

Range 3: 20 to 88 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #3

Score	Expect	Identities	Gaps	Strand	Frame
93.0 bits(48)	5e-19()	66/70(94%)	2/70(2%)	Plus/Minus	

Features:

```
Query  1    TAGGTCCGGTAGCAAGCTACCTTCCCCGCCTCGACTTGCATGTGTTAAGC-TGCCGCCAG  59
           ||||||||||||||||||||||||||  ||||||||||||||||||||| |||||||||
Sbjct  88   TAGGTCCGGTAGCAAGCTACCTTCCCCCGCTCGACTTGCATGTGTTAAGCCTGCCGCCAG  29

Query  60   CCGTTCAATC  69
           | ||||||||
Sbjct  28   C-GTTCAATC  20
```

2. *Aeromonas hydrophila* subsp. *hydrophila* ATCC 7966

a. V1 amplicon sequence

TCGCCGGCAA AAGATAGCAA GCATACTT[T]C CCTGCCTGCC [C]GCCTCCGCA CTTGCCATTG CTTGGTTTGA TGGCCTTGGC CCTGGCCCCA GGCCCGGC

b. V2 amplicon sequence

GGTGCTTCTT CTGCGAGTAA CGTCACAGTT GATACGTATT AGGCATCAAC CTTTCCTCCT C

c. V3 amplicon sequence

GTGTTCTGAT TCCGAAGGCA [A]CTCC[C]GCCA TCTCTGCAGG ATTGCCAGAC ATGTCAAGGC CAAGGGCTGA GGTTCTTC

d. <u>BLAST Analysis Results and Discussion.</u> Separate analyses were done with BLAST settings as "nr/nt" or "16S ribosomal RNA (Bacteria and Archaea)." With BLAST setting at "highly similar sequences," the tool used only the V2 sequence to generate alignments for both approaches. With the setting as "somewhat similar," the tool used V2 and V3 sequences, but still excluded the V1 sequence; this is the setting that was used for further searches. The "nr/nt" setting identified *A. hydrophila* as the best match, for which the highest values for total score, query coverage, E-value, and identity were 1946, 56 %, 1e-20, and 100 %, respectively. *A. hydrophila* strain KN-Mc-1R2

(complete genome) was the next best one, but had only one nucleotide mismatch. The sequence mismatches with the diverged more and more as the values for the alignment metrics decreased. Restricted alignment search for the 16S rRNA sequences generated a table with a number of *A. hydrophila* strains at the top of the list. The corresponding values for total score, query coverage, E-value, and identity were 209, 62 %, 2e-26, and 100 %. These results suggest that the pyrosequencing-generated sequences, even with red "failed" sequences, used for analysis could identify this bacterium through BLAST.

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Aeromonas hydrophila strain ATCC 7966 16S ribosomal RNA, partial sequence
Sequence ID: NR_119039.1 Length: 1460 Number of Matches: 3
Related Information

Range 1: 444 to 504 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
117 bits(61)	2e-26()	61/61(100%)	0/61(0%)	Plus/Minus	

Features:

```
Query  104  GGTGCTTCTTCTGCGAGTAACGTCACAGTTGATACGTATTAGGCATCAACCTTTCCTCCT  163
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  504  GGTGCTTCTTCTGCGAGTAACGTCACAGTTGATACGTATTAGGCATCAACCTTTCCTCCT  445

Query  164  C  164
            |
Sbjct  444  C  444
```

Range 2: 986 to 1045 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
60.3 bits(31)	5e-09()	59/63(94%)	4/63(6%)	Plus/Minus	

Features:

```
Query  170   GTGTTCTGATT-CCGAAGGCAACTCCCGCCATCTCTGCAGGATTGCCAGACATGTCAAGG  228
             |||||||||||| |||||||| ||||||| ||||||||||||||| |||||||||||||
Sbjct  1045  GTGTTCTGATTCCCGAAGGC-ACTCCCG-CATCTCTGCAGGATT-CCAGACATGTCAAGG  989

Query  229   CCA  231
             |||
Sbjct  988   CCA  986
```

Range 3: 72 to 101 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #3

Score	Expect	Identities	Gaps	Strand	Frame
31.5 bits(16)	2.2()	30/32(94%)	2/32(6%)	Plus/Minus	

Features:

```
Query  1    TCGCCGGCAAAAGATAGCAAGCATACTTTCCC  32
```

```
            ||||||||||||| ||||||||| |||||||||
Sbjct  101  TCGCCGGCAAAAG-TAGCAAGC-TACTTTCCC  72
```

3. *Bacteroides fragilis* strain VPI 2553

a. V1 amplicon sequence

TCTTTACCGA AGTAAATCGC TCAACTTGCA TGTGTTAGGC ACGCCGCCAG CGTTCATCCT GA

b. V2 amplicon sequence

GATCCTTATT CATATAATAC ATACAAAACA GTATACATAC TGCACTT[T]AT TCTTATATAA A[A]GAA

c. V3 amplicon sequence

GTCACCAATG TCCCCGAAGG GAACTCTCCG ATTAAGGAGA TGTCATTGGG ATGTCAAGCT TAGG[G]TAA

d. <u>BLAST Analysis Results and Discussion.</u> With the highest values of 234, 63 %, 6e-27, and 99 % for total score, query coverage, E-value, and identity, respectively, BLAST identified *Peptoclostridium difficile* as the best match, aligning V1 and V2 sequences. BLAST identified *Bacteroides fragilis* strain NCTC 9343 with the corresponding metrics of 121, 31 %, 2e-27, and 100 %, but aligning only the V2 sequence. Both alignments excluded V3. With the setting "Somewhat similar sequences (blastn)," BLAST listed *Bacteroides fragilis* strain NCTC 9343 with the total score, query coverage, E-value, and identity values of 164, 52 %, 1e-26, and 100 %, respectively. The alignments were for the V1 and V2 sequences; V3 was again excluded. These two alignments are shown below. Based on these results, the sequences generated by pyrosequencing could not have precisely identified *B. fragilis*, even with V1 and V3 having a "passed" sequences.

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Bacteroides fragilis strain NCTC 9343 16S ribosomal RNA, complete sequence
Sequence ID: NR_074784.2 Length: 1529 Number of Matches: 2
Related Information
Range 1: 439 to 503 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
118 bits(130)	1e-26()	65/65(100%)	0/65(0%)	Plus/Minus	

Features:

```
Query  68   GATCCTTATTCATATAATACATACAAAACAGTATACATACTGCACTTTATTCTTATATAA  127
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  503  GATCCTTATTCATATAATACATACAAAACAGTATACATACTGCACTTTATTCTTATATAA  444

Query  128  AAGAA  132
```

```
            |||||
Sbjct  443  AAGAA  439
```

Range 2: 24 to 66 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
46.4 bits(50)	6e-05()	36/43(84%)	0/43(0%)	Plus/Minus	

Features:

```
Query  20  CTCAACTTGCATGTGTTAGGCACGCCGCCAGCGTTCATCCTGA  62
           ||| |||||||||||||| ||  |  || |||||||||||||
Sbjct  66  CTCGACTTGCATGTGTTAAGCCTGTAGCTAGCGTTCATCCTGA  24
```

4. *Clostridium perfringens* strain NCTC 8237

a. V1 amplicon sequence

TAATCCTCTT CCGAAGAACA TCATCCCCTG CTGTTGTTAT CACGCCGCCG CCGCTCATTC C

b. V2 amplicon sequence

GTGGCTTCCT CCTTGGTACC GTCATTATCT TCCCCAAAGA CAGAGCTTTA CGATCCGAAA ACCATCATCA C

c. V3 amplicon sequence

GTCACCTTGT CCC[C]GAAGG[G] ATT[T]CCTCGA TTAAGAGTAA TGCAAGGGAT GTCAAGTGTA GGTAAGGTTC

d. <u>BLAST Analysis Results and Discussion.</u> For "nr/nt" search setting, BLAST generated a list of several *C. perfringens* strains (complete genomes). The total score, query coverage, E-value, and identity for all these bacteria were the same – 2515, 66 %, 2e-26, and 100 %, respectively. For the next lower total score of 251, the analysis listed many more *C. perfringens*. This search approach excluded the V1 sequence from alignment. Despite V1 exclusion, however, the analysis appears sufficient to precisely identify *C. perfringens*, but not any particular strain of it.

For the search setting "16S ribosomal RNA (Bacteria and Archaea)," BLAST identified *C. perfringens* as the top two matches, with the corresponding alignment metrics of 251, 66 %, 3e-30, 100 %. BLAST also listed four other species of *Clostridium*; however, the alignment metrics were much too low for these to be considered as candidate identifications. For example, the next lower set of values for these metrics was 111, 31 %, 1e-24, and 97%. Like the "nr/nt" search setting, the 16S restricted search approach too excluded the V1 sequence. Thus, the sequences used (V2 and V3) for BLAST were sufficient to precisely identify *C. perfringens*.

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Clostridium perfringens strain ATCC 13124 16S ribosomal RNA, complete sequence
Sequence ID: NR_121697.2 Length: 1513 Number of Matches: 2
Related Information
Range 1: 949 to 1018 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
130 bits(70)	3e-30()	70/70(100%)	0/70(0%)	Plus/Minus	

Features:

```
Query  143   GTCACCTTGTCCCCGAAGGGATTTCCTCGATTAAGAGTAATGCAAGGGATGTCAAGTGTA  202
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1018  GTCACCTTGTCCCCGAAGGGATTTCCTCGATTAAGAGTAATGCAAGGGATGTCAAGTGTA  959

Query  203   GGTAAGGTTC  212
             ||||||||||
Sbjct  958   GGTAAGGTTC  949
```

Range 2: 407 to 478 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
121 bits(65)	2e-27()	70/72(97%)	1/72(1%)	Plus/Minus	

Features:

```
Query  67    GTGGCTTCCTCCTTGG-TACCGTCATTATCTTCCCCAAAGACAGAGCTTTACGATCCGAA  125
             ||||||||||||||| ||||||||||||||||||||||||||||||||||||||||||||
Sbjct  478   GTGGCTTCCTCCTTGGGTACCGTCATTATCTTCCCCAAAGACAGAGCTTTACGATCCGAA  419

Query  126   AACCATCATCAC  137
             |||| |||||||
Sbjct  418   AACCTTCATCAC  407
```

5. *Enterobacter aerogenes* strain IFO 12010

a. V1 amplicon sequence

TCGTCACCCG AGAGCAAGCT CTCTGTGTCC CCTCTGTGTG CGCGCC

b. V2 amplicon sequence

GGTGCTTCTT CTGCGAGTAA CGTCAATCGC CAAGGTTATT AACCTTAATC GCCTTCCTCC TCGCATGAA

c. V3 amplicon sequence

GTCTCAGAGT TCCCGAAGGC ACCAAAGCAT CTCTGCTAAG TTCTCTGGAT GTCAAGAGTA GGTAA

d. <u>BLAST Analysis Results and Discussion.</u> For "nr/nt" search setting, the top four listings in the BLAST table were *Klebsiella aerogenes* strains; the total score, query coverage, E-value, and identity values were 2256, 85 % (70 % for one), 9e-24, and 100 %, respectively. At various lower

sets of these values, the organisms identified varied considerably; *Citrobacter spp*., *Vibrio spp*., *Morganella spp*., and many others. This list also had *Enterobacter cloacae* and *Enterobacter cancerogenus*, but no *Enterobacter aerogenes*. Also, BLAST used the V1, V2, and V3 sequences selectively for various alignments. The search was also performed with BLAST setting at "somewhat similar sequences," but this search gave similar results. The inevitable conclusion, therefore, is that these V1, V2, and V3 sequences generated by pyrosequencing did not afford precise identification of the target organism.

BLAST analysis with the restricted search "16S ribosomal RNA (Bacteria and Archaea)" identified four strains of *E. aerogenes* and one of *K. aerogenes* as the best matches. All five bacteria had the same values for total score, query coverage, E-value, and identity; 236, 70 %, 2e-27, and 100 %, respectively. With lower total scores, but the same other values, BLAST identified a number of different genera, species, and strains. Thus, as with the "nr/nt" search results, these results could not pinpoint the precise target organism based on the sequences used for analysis. Further, as with the "nr/nt" search, BLAST used the three sequences selectively for different alignments. Although the search done with the setting "somewhat similar sequences" aligned all three sequences, the identification results were essentially the same.

The overall conclusion from both results is that the sequences used here did not afford identification of the target bacterium.

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Enterobacter aerogenes strain JCM1235 16S ribosomal RNA gene, partial sequence
Sequence ID: NR_024643.1 Length: 1438 Number of Matches: 2
Related Information
Range 1: 947 to 1011 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
121 bits(65)	2e-27()	65/65(100%)	0/65(0%)	Plus/Minus	

Features:

```
Query  126   GTCTCAGAGTTCCCGAAGGCACCAAAGCATCTCTGCTAAGTTCTCTGGATGTCAAGAGTA  185
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1011  GTCTCAGAGTTCCCGAAGGCACCAAAGCATCTCTGCTAAGTTCTCTGGATGTCAAGAGTA  952

Query  186   GGTAA  190
             |||||
Sbjct  951   GGTAA  947
```

Range 2: 404 to 470 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
115 bits(62)	7e-26()	67/69(97%)	2/69(2%)	Plus/Minus	

Features:

```
Query  52    GGTGCTTCTTCTGCGAGTAACGTCAATCGCCAAGGTTATTAACCTTAATCGCCTTCCTCC  111
             ||||||||||||||||||||||||||||||||||||||||||||||| ||||||||||||
Sbjct  470   GGTGCTTCTTCTGCGAGTAACGTCAATCGCCAAGGTTATTAACCTTAA-CGCCTTCCTCC  412
```

```
Query  112  TCGCATGAA  120
            |||| ||||
Sbjct  411  TCGC-TGAA  404
```

6. *Enterobacter cloacae* subsp. *cloacae* strain CDC 442-68

a. V1 amplicon sequence

TCGTCACCCG AGAGCAAGCT CTCTGTGCTA CCGTTCGACT TGCATGTGTT AGGCCTGCCG CCA

b. V2 amplicon sequence

GGTGCTTCTT CTGCGG[G]TAA CGTCAATTGC TGCGGTTATT [T]AACCACAAC AACCTTCCCT TCCCCGCCTG AAAGTA

c. V3 amplicon sequence

GTCTCACAGT TCC[C]GAAGGC ACCAATCCAT CTCTGGAAAG TTCTGTGGAT GTCAAGACCA GGTAAGGTTC

d. <u>BLAST Analysis Results and Discussion.</u> BLAST placed *Enterobacter cloacae* strain Res2010EC27 chromosome (complete genome) on top of the search results table. The total score, query coverage, E-value, and identity level were 2820, 95 %, 2e-26, 100 %, respectively. However, many different genera, species, and strains followed, some with the same sequence identity as for *E. cloacae* in the segments aligned, some with only one nucleotide difference, and some with a few. These results made it impossible to identify the target bacterium with any reliability.

With the search setting "16S ribosomal RNA (Bacteria and Archaea)," and the corresponding alignment metrics of 356, 95 %, 3e-30, and 100 %, BLAST results listed six strains of *E. cloacae*. The total score for four was 356, for one 348, for another 346; the other values were the same for all six. The next bacterium in the list was *Salmonella enterica* subsp. *enterica* strain Ty2 (16S ribosomal RNA, partial sequence), which had a number of differences with the query sequences, sufficient to suggest that this is not a precise identification. Other bacteria that appeared with progressively lower alignment metrics had progressively greater sequence divergence with the query sequences, and therefore could not be considered candidate targets. These results suggest that although the query sequences used identified *E. cloacae* as the best match, the sequence homologies with *S. enterica* subsp. *enterica* were not sufficiently diverse, suggesting the target bacterium identification could be in doubt if the sample were unknown, which inevitable would be the case when the patient samples are used.

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Enterobacter cloacae strain ATCC 13047 16S ribosomal RNA, complete sequence
Sequence ID: NR_102794.2 Length: 1543 Number of Matches: 3

Related Information
Range 1: 977 to 1046 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
130 bits(70)	3e-30()	70/70(100%)	0/70(0%)	Plus/Minus	

Features:

```
Query  150   GTCTCACAGTTCCCGAAGGCACCAATCCATCTCTGGAAAGTTCTGTGGATGTCAAGACCA  209
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1046  GTCTCACAGTTCCCGAAGGCACCAATCCATCTCTGGAAAGTTCTGTGGATGTCAAGACCA  987

Query  210   GGTAAGGTTC  219
             ||||||||||
Sbjct  986   GGTAAGGTTC  977
```

Range 2: 37 to 99 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
117 bits(63)	2e-26()	63/63(100%)	0/63(0%)	Plus/Minus	

Features:

```
Query  1   TCGTCACCCGAGAGCAAGCTCTCTGTGCTACCGTTCGACTTGCATGTGTTAGGCCTGCCG  60
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  99  TCGTCACCCGAGAGCAAGCTCTCTGTGCTACCGTTCGACTTGCATGTGTTAGGCCTGCCG  40

Query  61  CCA  63
           |||
Sbjct  39  CCA  37
```

Range 3: 432 to 502 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #3

Score	Expect	Identities	Gaps	Strand	Frame
108 bits(58)	1e-23()	71/76(93%)	5/76(6%)	Plus/Minus	

Features:

```
Query  69   GGTGCTTCTTCTGCGGGTAACGTCAATTGCTGCGGTTATTTAACCACAACAACCTTCCCT  128
           |||||||||||||||||||||||||||||||||||| |||||||||| |||||| |
Sbjct  502  GGTGCTTCTTCTGCGGGTAACGTCAATTGCTGCGGTTATT-AACCACAACA-CCTTCC-T  446

Query  129  TCCCCGCCTGAAAGTA  144
            |||||| ||||||||
Sbjct  445  -CCCCGC-TGAAAGTA  432
```

7. *Enterococcus faecalis* strain V583

a. V1 amplicon sequence

TCCTCTTTCC AATTGAGTGC AAGCACTCGG AGGAA[A]GAAG CAGTCTGACT [T]GCA TGTATT ATGGCAGCAG CCGCCA

b. V2 amplicon sequence

GTGGCTTTCT GGTTAGATAC CGTCAGGGAC GTTCAGTTAC TAACGTCCTT GTTCTTCTC

c. V3 amplicon sequence

GTCACTT[T]GT CCCGAAGGAA AGCTCTATCT CTAGAGTGGT CAAAGGATGT CAAGACCTGG TAAGG

d. BLAST Analysis Results and Discussion. With search setting at "nr/nt," the two best matches listed were *E. faecalis* strains ARO1/DG and FDAARGOS 338. The total score, query coverage, E-value, and identity level were 1265 (1250 for strain FDAARGOS 338), 95 %, 5e-22, and 98 %, respectively. At the same query coverage and identity of 97-98 %, the table listed many more *E. faecalis* strains. At lower metrics, other genera appeared, but the metrics were too low to consider them as reasonable target identities. Thus, these query sequences generated by pyrosequencing were sufficient to identify *E. faecalis*, although not any particular strain of this organism.

The "16S ribosomal RNA (Bacteria and Archaea)" setting BLAST results table listed four different strains of *E. faecalis*, and the corresponding alignment metrics for all four were the same; 315, 95 %, 1e-24, and 97 %. Following this, various genera and species were listed in the table. But the metrics were much lower; for example, the next lower query coverage and identity values were 59 % and 5e-23, respectively, and this was for *Carnobacterium viridans* strain MPL-11 (16S ribosomal RNA gene, partial sequence). We therefore did not consider and therefore these bacteria could not be considered as candidate identifications. Like the BLAST results from "nr/nt" setting, these results show that the only precise identification is *E. faecalis*, and that the query sequences generated by pyrosequencing were sufficient to identify this bacterium. However, the sequences did not identify a particular strain of the bacterium.

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Enterococcus faecalis strain JCM 5803 16S ribosomal RNA gene, partial sequence
Sequence ID: NR_040789.1 Length: 1517 Number of Matches: 3
Related Information
Range 1: 988 to 1054 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
111 bits(60)	1e-24()	65/67(97%)	2/67(2%)	Plus/Minus	

Features:

```
Query  146   GTCACTTTGT-CCCGAA-GGAAAGCTCTATCTCTAGAGTGGTCAAAGGATGTCAAGACCT  203
             |||||||||| |||||| ||||||||||||||||||||||||||||||||||||||||
Sbjct  1054  GTCACTTTGTCCCCGAAGGGAAAGCTCTATCTCTAGAGTGGTCAAAGGATGTCAAGACCT  995

Query  204   GGTAAGG  210
             |||||||
Sbjct  994   GGTAAGG  988
```

Range 2: 455 to 514 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
104 bits(56)	2e-22()	59/60(98%)	1/60(1%)	Plus/Minus	

Features:

```
Query  82    GTGGCTTTCTGGTTAGATACCGTCAGGG-ACGTTCAGTTACTAACGTCCTTGTTCTTCTC  140
            |||||||||||||||||||||||||||| ||||||||||||||||||||||||||||||
Sbjct  514   GTGGCTTTCTGGTTAGATACCGTCAGGGGACGTTCAGTTACTAACGTCCTTGTTCTTCTC  455
```

Range 3: 30 to 102 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #3

Score	Expect	Identities	Gaps	Strand	Frame
99.0 bits(53)	8e-21()	70/77(91%)	5/77(6%)	Plus/Minus	

Features:

```
Query  1     TCCTCTTTCCAATTGAGTGCAAGCACTCGG-AGGAAAGAAGCAGTCTGACTTGCATGTAT  59
            |||||||||||||||||||||||||||||| |||||||||| ||  ||||||||||||
Sbjct  102   TCCTCTTTCCAATTGAGTGCAAGCACTCGGGAGGAAAGAAGC-GTTCGACTTGCATGTAT  44

Query  60    TATGGCAGCAGCCGCCA  76
             || |||| | |||||||
Sbjct  43    TA-GGCA-C-GCCGCCA  30
```

8. *Enterococcus faecium* strain MMC4

a. V1 amplicon sequence

TCTCTTT[T]CC TGTGGAGCAG CTCCGGTGGA AAGAAGACGT CGACTGCATG TATTATGCGA CGCG

b. V2 amplicon sequence

GTGGCTTTCT GGTTAGATAC CGTCAAGGGA TGAACAGTTA CTCTCATCCT TGTTCTTCTC TAACAA

c. V3 amplicon sequence

GTCACTT[T]GC CCC[C]GAAGGG AAGCTCTATC TCTAGAGTGG TCAAAGGATG TCAAGACCTG GTAAGG

d. BLAST Analysis Results and Discussion. BLAST excluded the V1 sequence from any alignment, but used both V2 and V3, which is consistent with the software-judged unreliability of the V1 sequence. For the "nr/nt" search setting, many different strains of *E. faecium*, *E. hirae*, and *E. lactis* were listed. For the total score, query coverage, E-value, and identity level of 1442, 64 %, 3e-24, and 100 %, respectively, the top two on the list were *E. hirae* strain FDAARGOS 234 and *E. faecium* strain FDAARGOS 323 (both complete genomes). Given these results, we conclude the query sequences used here are insufficient to precisely identify the target bacterium, *E. faecium*.

For the search setting "16S ribosomal RNA (Bacteria and Archaea)," BLAST did not align the V1 sequence to any sequence in the database; it aligned only the V2 and V3 sequences. For the alignment metrics of 240, 64 %, 5e-28, and 100 % for total score, query coverage, E-value, and identity, respectively, BLAST table listed different species and strains of *Enterococcus*; *E. hirae*, *E. faecium*, *E. durans*, and some others. These results show that, like the "nr/nt" search, the restricted 16S rRNA gene database search failed to precisely identify the target bacterium, *E. faecium*.

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Enterococcus faecium strain JCM 5804 16S ribosomal RNA gene, partial sequence
Sequence ID: NR_112039.1 Length: 1523 Number of Matches: 2
Related Information
Range 1: 455 to 520 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
122 bits(66)	5e-28()	66/66(100%)	0/66(0%)	Plus/Minus	

Features:

```
Query  70   GTGGCTTTCTGGTTAGATACCGTCAAGGGATGAACAGTTACTCTCATCCTTGTTCTTCTC  129
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  520  GTGGCTTTCTGGTTAGATACCGTCAAGGGATGAACAGTTACTCTCATCCTTGTTCTTCTC  461

Query  130  TAACAA  135
           ||||||
Sbjct  460  TAACAA  455
```

Range 2: 996 to 1062 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
117 bits(63)	2e-26()	66/67(99%)	1/67(1%)	Plus/Minus	

Features:

```
Query  141   GTCACTTTGCCCCCGAAGGG-AAGCTCTATCTCTAGAGTGGTCAAAGGATGTCAAGACCT  199
           |||||||||||||||||||| ||||||||||||||||||||||||||||||||||||||||
Sbjct  1062  GTCACTTTGCCCCCGAAGGGGAAGCTCTATCTCTAGAGTGGTCAAAGGATGTCAAGACCT  1003

Query  200   GGTAAGG  206
           |||||||
Sbjct  1002  GGTAAGG  996
```

9. *Haemophilus influenzae* strain Rd

a. V1 amplicon sequence

TCGTCAGCAA GAAAGCAAGC TTCTCCTGCT ACCGTTCGAC TTGCATGTGT
TAATGCCTGC CGCC[C]AGCCG C

b. V2 amplicon sequence

GGTGCTTCTT CTGTATTTAA CGTCAATTTG AT[T]GTATCTA TTAATCAATC AATCAATCCA TTTCCCTTCCAATTCAATCC ATGCAAA

c. V3 amplicon sequence

GTCTCTAAGT TCCCGAAGGC ACAAGCTCAT CTCTGAGCTC TTCTTAGGAT GTCAAGAGTA GG

d. <u>BLAST Analysis Results and Discussion.</u> The "nr/nt" search analysis listed *H. influenzae* strain FDAARGOS 199 as the top match, with the match metrics of 1781, 79 %, 5e-22, and 100 % for total score, query coverage, E-value, and identity, respectively. For the next lower score of 1748 and the same values for the other metrics, BLAST listed 8 strains of *H. influenzae*. For 1743 and the same values for other metrics, it still listed *H. influenzae*. Then there was a significant drop in total score and query coverage, and the table had a long list of various species and strains of *Pasteurella*. Thus, the "nr/nt" search identified the target bacterium, *H. influenzae* accurately even with yellow "check" sequences in V1, V2, and V3.

The "16S ribosomal RNA (Bacteria and Archaea)," BLAST analysis placed and *H. influenzae* strain 680 and *H. aegyptius* strain CCUG 25716 as the top two bacteria identified with the V1, V2, and V3 sequences; the respective total score, query coverage, E-value, and identity were 291, 79 %, 9e-26, and 100 %. For the next lower total score of 284 and other metrics the same, BLAST listed *H. aegyptius* strain NCTC 8502 at the third position, with only two nucleotide differences in alignment. Then the values for the match metrics dropped precipitously, with a number of different genera and species listed, e.g., *Pasteurella spp*. and *Actinobacillus spp*. The results show that while the sequences identified *H. influenzae*, the identification was not unique, suggesting the 16S rRNA gene restricted search approach failed to identify the bacterium accurately. This is in contrast to the "nr/nt" approach, which did identify the bacterium.

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Haemophilus influenzae strain 680 16S ribosomal RNA, partial sequence
Sequence ID: NR_044682.2 Length: 1486 Number of Matches: 3
Related Information
Range 1: 981 to 1042 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
115 bits(62)	9e-26()	62/62(100%)	0/62(0%)	Plus/Minus	

Features:

```
Query  169   GTCTCTAAGTTCCCGAAGGCACAAGCTCATCTCTGAGCTCTTCTTAGGATGTCAAGAGTA  228
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1042  GTCTCTAAGTTCCCGAAGGCACAAGCTCATCTCTGAGCTCTTCTTAGGATGTCAAGAGTA  983

Query  229   GG  230
             ||
Sbjct  982   GG  981
```

Range 2: 34 to 100 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
108 bits(58)	2e-23()	66/69(96%)	3/69(4%)	Plus/Minus	

Features:

```
Query  1    TCGTCAGCAAGAAAGCAAGCTT-CTCCTGCTACCGTTCGACTTGCATGTGTTAATGCCTG  59
           |||||||||||||||||||||| ||||||||||||||||||||||||||||||| |||||
Sbjct  100  TCGTCAGCAAGAAAGCAAGCTTTCTCCTGCTACCGTTCGACTTGCATGTGTTAA-GCCTG  42

Query  60   CCGCCCAGC  68
            ||||| |||
Sbjct  41   CCGCC-AGC  34
```

Range 3: 454 to 503 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #3

Score	Expect	Identities	Gaps	Strand	Frame
67.6 bits(36)	3e-11()	48/53(91%)	4/53(7%)	Plus/Minus	

Features:

```
Query  77   GGTGCTTCTTCTGTATTTAACGTCAATTTGATTGTATCTATTAATCA-ATCAA  128
           |||||||||||||||||||||||||||||| ||  ||||||| || |||||
Sbjct  503  GGTGCTTCTTCTGTATTTAACGTCAATTTGAT-GTG-CTATTAA-CACATCAA  454
```

10. *Klebsiella oxytoca* strain LBM 90.11.033

a. V1 amplicon sequence

TCGTCACCCG AGAGCAAGCT CTCTGTGCTA C[C]GT[T]CGACT TGCATGTGTT ATGGCCTGCC GCCA

b. V2 amplicon sequence

GGTGCTTCTT CTGCGG[G]TAA CGTCAATGAA [A]TAAGGTTAT TAACCTCACT CCCTTCC[C]TC CCCGCTGAAA GTA

c. V3 amplicon sequence

GTCTCAGAGT TCCCGAAGGC ACCAAAGCAA TCTCTGCTAA GTTCTCTGGA TGATCAAGAA GTGAGGATGA A

d. <u>BLAST Analysis Results and Discussion.</u> Analysis conducted with database setting "nr/nt" identified a long list of various strains of *K. oxytoca*. The first match listed based on the highest total score, query coverage, E-value, and identity was *K. oxytoca* strain CAV 1335, complete genome; the respective values for the four metrics were 2553, 93 %, 3e-24, and 97 %, respectively. At lower values, the table still listed *K. oxytoca* strains. We therefore think that the

"nr/nt" approach correctly identified the target bacterium, *K. oxytoca*, although not down to the strain level.

The "16S ribosomal RNA (Bacteria and Archaea)," BLAST analysis results were similar, with three strains of *K. oxytoca* listed at the top of the match table. These had identical values for the total score, query coverage, E-value, and sequence identity; 332, 93 %, 5e-28, and 97 %, respectively. Other genera and species were listed with much lower metrics, and the sequence difference were much too great in comparison to the values for *K. oxytoca*. Thus, this approach too, like the "nr/nt" search approach, correctly identified the target bacterium, but not any particular strain of it, even with V1 and V3 having some red "failed" sequences.

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Klebsiella oxytoca strain ATCC 13182 16S ribosomal RNA gene, partial sequence
Sequence ID: NR_118853.1 Length: 1502 Number of Matches: 3
Related Information
Range 1: 425 to 495 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
122 bits(66)	5e-28()	71/73(97%)	2/73(2%)	Plus/Minus	

Features:

```
Query  70   GGTGCTTCTTCTGCGGGTAACGTCAATGAAATAAGGTTATTAACCTCACTCCCTTCCCTC  129
           ||||||||||||||||||||||||||||| ||||||||||||||||||||||||| ||
Sbjct  495  GGTGCTTCTTCTGCGGGTAACGTCAATGAA-TAAGGTTATTAACCTCACTCCCTTCC-TC  438

Query  130  CCCGCTGAAAGTA  142
           |||||||||||||
Sbjct  437  CCCGCTGAAAGTA  425
```

Range 2: 30 to 92 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
111 bits(60)	1e-24()	63/64(98%)	1/64(1%)	Plus/Minus	

Features:

```
Query  1    TCGTCACCCGAGAGCAAGCTCTCTGTGCTACCGTTCGACTTGCATGTGTTATGGCCTGCC  60
           ||||||||||||||||||||||||||||||||||||||||||||||||||| ||||||||
Sbjct  92   TCGTCACCCGAGAGCAAGCTCTCTGTGCTACCGTTCGACTTGCATGTGTTA-GGCCTGCC  34

Query  61   GCCA  64
           ||||
Sbjct  33   GCCA  30
```

Range 3: 975 to 1036 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #3

Score	Expect	Identities	Gaps	Strand	Frame
97.1 bits(52)	3e-20()	62/66(94%)	4/66(6%)	Plus/Minus	

Features:

```
Query  148   GTCTCAGAGTTCCCGAAGGCACCAAAGCAATCTCTGCTAAGTTCTCTGGATGATCAAGAA  207
             ||||||||||||||||||||||||||| |||||||||||||||||||| ||||||
Sbjct  1036  GTCTCAGAGTTCCCGAAGGCACCAAAGCA-TCTCTGCTAAGTTCTCTGGATG-TCAAGA-  980

Query  208   GTGAGG  213
             || |||
Sbjct  979   GT-AGG  975
```

11. *Legionella pneumophila* subsp. *pneumophila* strain Philadelphia-1

a. V1 amplicon sequence

TCGCCATCTG TCTAGCAAGC TAGACAATGC TGCCGTTCGA CTTGCATGTG TTAAGCA

b. V2 amplicon sequence

GGTGCTTCTT CTGTGGGTAA CGTCCAGTTA ATCAGCTCTT AACCTATCAA CCCTCCTCCC CACCTGAAAG

c. V3 amplicon sequence

GTATCAGTGT TCCCGAAGGC ACTAATGCAT CTCTGCAAAA TTCACTGTAT GTCAAGGG

d. <u>BLAST Analysis Results and Discussion.</u> BLAST analysis with the setting "nr/nt" essentially identified only one organism, *Legionella pneumophila*, although not a specific strain. The top match metrics were 1012, 94 %, 3e-24, and 99 % for total score, query coverage, E-value, and identity, respectively. The "nr/nt" setting search was therefore sufficient to identify the target bacterium.

BLAST search with the restricted database setting "16S ribosomal RNA (Bacteria and Archaea)" identified *L. pneumophila* strain JCM 7571 and *L. pneumophila* strain Philadelphia as the top two matches, with the best match metrics of 337, 94 %, 5e-28, and 99 % for total score, query coverage, E-value, and identity, respectively. The next organism listed with lower metrics was *L. anisa* strain ATCC 35297, but it had 4 nucleotide divergence in comparison to the *L. pneumophila* identity. Other genera, species, and strains had much lower metrics. We conclude that combined with the "nr/nt" search, this search results also correctly identified the bacterium.

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Legionella pneumophila strain JCM 7571 16S ribosomal RNA gene, partial sequence
Sequence ID: NR_113235.1 Length: 1466 Number of Matches: 3
Related Information
Range 1: 409 to 477 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
122 bits(66)	5e-28()	69/70(99%)	1/70(1%)	Plus/Minus	

Features:

```
Query  63   GGTGCTTCTTCTGTGGGTAACGTCCAGTTAATCAGCTCTTAACCTATCAACCCTCCTCCC  122
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  477  GGTGCTTCTTCTGTGGGTAACGTCCAGTTAATCAGCTCTTAACCTATCAACCCTCCTCCC  418

Query  123  CACCTGAAAG  132
            ||| ||||||
Sbjct  417  CAC-TGAAAG  409
```

Range 2: 962 to 1019 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
108 bits(58)	1e-23()	58/58(100%)	0/58(0%)	Plus/Minus	

Features:

```
Query  138   GTATCAGTGTTCCCGAAGGCACTAATGCATCTCTGCAAAATTCACTGTATGTCAAGGG  195
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1019  GTATCAGTGTTCCCGAAGGCACTAATGCATCTCTGCAAAATTCACTGTATGTCAAGGG  962
```

Range 3: 18 to 74 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #3

Score	Expect	Identities	Gaps	Strand	Frame
106 bits(57)	5e-23()	57/57(100%)	0/57(0%)	Plus/Minus	

Features:

```
Query  1   TCGCCATCTGTCTAGCAAGCTAGACAATGCTGCCGTTCGACTTGCATGTGTTAAGCA  57
           |||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  74  TCGCCATCTGTCTAGCAAGCTAGACAATGCTGCCGTTCGACTTGCATGTGTTAAGCA  18
```

12. *Listeria monocytogenes* strain Li 23

a. V1 amplicon sequence

TAACATTGGA AGAGCAAGCT CTTCCTCCGT TCGTTCGACT TGCATGTATT AGGCACGCCG CCA

b. V2 amplicon sequence

GTGGCTTTCT GGTTAGATAC CGTCAAGGGA CAAGCAGTTA CTCTTATCCT TGTTCTTCTC TAACAA

c. V3 amplicon sequence

GTCACTTTGT CCC[C]GAAGG[G] AAAGCTCTGT CTCCAGAGTG GTCAAAGGAT GTCAAGACCT GGTAA

d. <u>BLAST Analysis Results and Discussion.</u> The "nr/nt" BLAST search identified only *Listeria monocytogenes*, but different strains. The highest metrics were 2169, 95 %, 3e-24, and 100 % for total score, query coverage, E-value, and sequence identity, respectively. We conclude this search alone was sufficient to accurately identify the target bacterium, *L. monocytogenes*.

BLAST with the search setting at "16S ribosomal RNA (Bacteria and Archaea)" identified different species and strains of *Listeria*. The first organism listed was *Listeria innocua* strain ATCC 33090, with the highest metrics of 361, 95 %, 5e-28, and 100 % for total score, query coverage, E-value, and sequence identity, respectively. *L. monocytogenes* had 2 nucleotide difference with the query sequences in comparison to *L. innocua*. The table also listed other species and strains of *Listeria*, as well as other genera. These results show that in comparison to this restricted search, the "nr/nt" search approach for these sequences is the better one.

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Listeria monocytogenes strain NCTC 10357 16S ribosomal RNA, partial sequence
Sequence ID: NR_044823.1 Length: 1469 Number of Matches: 3
Related Information
Range 1: 986 to 1050 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
121 bits(65)	2e-27()	65/65(100%)	0/65(0%)	Plus/Minus	

Features:

```
Query  140   GTCACTTTGTCCCCGAAGGGAAAGCTCTGTCTCCAGAGTGGTCAAAGGATGTCAAGACCT  199
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1050  GTCACTTTGTCCCCGAAGGGAAAGCTCTGTCTCCAGAGTGGTCAAAGGATGTCAAGACCT  991

Query  200   GGTAA  204
             |||||
Sbjct  990   GGTAA  986
```

Range 2: 444 to 509 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
119 bits(64)	6e-27()	65/66(98%)	0/66(0%)	Plus/Minus	

Features:

```
Query  69    GTGGCTTTCTGGTTAGATACCGTCAAGGGACAAGCAGTTACTCTTATCCTTGTTCTTCTC  128
             ||||||||||||||||||||||||||||| ||||||||||||||||||||||||||||
Sbjct  509   GTGGCTTTCTGGTTAGATACCGTCAAGGGACNAGCAGTTACTCTTATCCTTGTTCTTCTC  450

Query  129   TAACAA  134
             ||||||
Sbjct  449   TAACAA  444
```

Range 3: 35 to 97 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #3

Score	Expect	Identities	Gaps	Strand	Frame
111 bits(60)	1e-24()	62/63(98%)	0/63(0%)	Plus/Minus	

Features:

```
Query  1   TAACATTGGAAGAGCAAGCTCTTCCTCCGTTCGTTCGACTTGCATGTATTAGGCACGCCG  60
           |||| ||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  97  TAACTTTGGAAGAGCAAGCTCTTCCTCCGTTCGTTCGACTTGCATGTATTAGGCACGCCG  38

Query  61  CCA  63
           |||
Sbjct  37  CCA  35
```

13. *Mycobacterium avium* subsp. *paratuberculosis* strain K-10

a. V1 amplicon sequence

TCGAGTACCT CCGAAGAGGC CTT[T]CCGTTC GACTTGCATG TGTTAAGCAC GCCGCCAGCG TTC

b. V2 amplicon sequence

GGTGCTTCTT CTCCACCTAC CGTCAATCCG AGAAAACC[C]A GGCA CCTTC[C]G TCGATGGGT GGAAAAGGAAGGGTTTTGAA

c. V3 amplicon sequence

GCACACAG[G]C CACAAGGAAC GCCTATCTCT AGACGCGTCC TGTGCATTGG TCAAAACCCC AAGGCATGAA AGGGA

d. <u>BLAST Analysis Results and Discussion.</u> At the "nr/nt" setting, the highest values for total score, query coverage, E-value, and sequence identity were 293, 83 %, 3e-26, and 100 %, respectively. For these values, BLAST listed *M. avium* strain DSM 44156, *M. avium* strain ATCC 25291, *M. bouchedurhonense* strain 4355387, *M. vulneris* strain NLA 000700772, and *M. colombinense* strain CIP 108962. *Mycobacterium avium* strain DSM 44156 as the first organism. All had the same metrics and sequence identity level. We therefore conclude that this search with the sequences generated by pyrosequencing failed to identify the target organism correctly.

The highest metrics for the search setting "16S ribosomal RNA (Bacteria and Archaea)" were 293, 83 %, 3e-26, and 100 %, respectively. For these values, BLAST listed several different species of *Mycobacterium*, including *M. avium*. Because they had the same sequence identity, we conclude that, like the "nr/nt" search setting, the sequences used for BLAST still failed to identify the target bacterium correctly.

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Mycobacterium avium strain DSM 44156 16S ribosomal RNA gene, partial sequence
Sequence ID: NR_025584.1 Length: 1472 Number of Matches: 3
Related Information

Range 1: 4 to 66 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
117 bits(63)	3e-26()	63/63(100%)	0/63(0%)	Plus/Minus	

Features:

```
Query  1    TCGAGTACCTCCGAAGAGGCCTTTCCGTTCGACTTGCATGTGTTAAGCACGCCGCCAGCG  60
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  66   TCGAGTACCTCCGAAGAGGCCTTTCCGTTCGACTTGCATGTGTTAAGCACGCCGCCAGCG  7

Query  61   TTC  63
           |||
Sbjct  6    TTC  4
```

Range 2: 400 to 463 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
89.8 bits(48)	6e-18()	62/68(91%)	4/68(5%)	Plus/Minus	

Features:

```
Query  69   GGTGCTTCTTCTCCACCTACCGTCAATCCGAGAAAACCCAGGCACCTTCCGTCGATGGGT  128
           ||||||||||||||||||||||||||||||||||| || |||||| |||||||| |
Sbjct  463  GGTGCTTCTTCTCCACCTACCGTCAATCCGAGAAAACCC-GG-ACCTTC-GTCGATGG-T  408

Query  129  GGAAAAGG  136
           | || |||
Sbjct  407  GAAAGAGG  400
```

Range 3: 948 to 1004 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #3

Score	Expect	Identities	Gaps	Strand	Frame
86.1 bits(46)	7e-17()	56/60(93%)	4/60(6%)	Plus/Minus	

Features:

```
Query  154   GCACACAGGCCACAAGG-AACGCCTATCTCTAGACGCGTCCTGTGCATTGGTCAAAACCC  212
           ||||||||||||||||| |||||||||||||||||||||||||||| | ||||| |||
Sbjct  1004  GCACACAGGCCACAAGGGAACGCCTATCTCTAGACGCGTCCTGTGCAT-G-TCAAA-CCC  948
```

14. *Mycobacterium tuberculosis* strain X004439

a. V1 amplicon sequence

TCGAGTATCT CCGAAGAGAC CTTTCCGTTC GACTTGCATG TGTTAAGCAC GCCGCCAGCG

b. V2 amplicon sequence

GGTGCTTCTT CTCCACCTAC CGTCAATCCG AGAGAACCCG GACCTTCGTC GATGGTGAAA GAGGTTTACA A

c. V3 amplicon sequence

GCACACAGGC CACAAGG[G]AA CGCCTATCTC TAGACGCGTC CTGTGCATGT CAAACCCAGG

d. <u>BLAST Analysis Results and Discussion.</u> With "nr/nt" search setting, essentially all organisms identified were various strains of *Mycobacterium tuberculosis*. The highest values for total score, query coverage, E-value, and identity were 356, 95 %, 5e-27, and 100 %, respectively. This search alone, therefore, was sufficient to accurately identify the target bacterium. It is also important to note that although the pyrosequencing software judged the sequences as "check" (yellow) or "failed" (red), they all proved sufficiently reliable to match *Mycobacterium species* sequences.

In contrast, the BLAST results with the "16S ribosomal RNA (Bacteria and Archaea)" setting were different: for the highest values for total score, query coverage, E-value, and identity of 356, 95 %, 8e-31, and 100 %, respectively, BLAST listed several *Mycobacterium* species and strains, *M. tuberculosis* strain H37Rv among them. However, the sequence alignments for these were identical, which shows that the restricted search approach failed to identify the target bacterium with these query sequences.

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Mycobacterium tuberculosis strain ATCC 27294 16S ribosomal RNA gene, partial sequence
Sequence ID: NR_116692.1 Length: 1454 Number of Matches: 3
Related Information
Range 1: 395 to 465 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
132 bits(71)	8e-31()	71/71(100%)	0/71(0%)	Plus/Minus	

Features:

```
Query  66   GGTGCTTCTTCTCCACCTACCGTCAATCCGAGAGAACCCGGACCTTCGTCGATGGTGAAA  125
            |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  465  GGTGCTTCTTCTCCACCTACCGTCAATCCGAGAGAACCCGGACCTTCGTCGATGGTGAAA  406

Query  126  GAGGTTTACAA  136
            ||||||||||
Sbjct  405  GAGGTTTACAA  395
```

Range 2: 7 to 66 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
111 bits(60)	1e-24()	60/60(100%)	0/60(0%)	Plus/Minus	

Features:

```
Query  1    TCGAGTATCTCCGAAGAGACCTTTCCGTTCGACTTGCATGTGTTAAGCACGCCGCCAGCG  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  66   TCGAGTATCTCCGAAGAGACCTTTCCGTTCGACTTGCATGTGTTAAGCACGCCGCCAGCG  7
```

Range 3: 947 to 1006 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #3

Score	Expect	Identities	Gaps	Strand	Frame
111 bits(60)	1e-24()	60/60(100%)	0/60(0%)	Plus/Minus	

Features:

```
Query  142   GCACACAGGCCACAAGGGAACGCCTATCTCTAGACGCGTCCTGTGCATGTCAAACCCAGG  201
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1006  GCACACAGGCCACAAGGGAACGCCTATCTCTAGACGCGTCCTGTGCATGTCAAACCCAGG  947
```

15. *Neisseria meningitidis* serogroup B

a. V1 amplicon sequence

TCGCCACCCG AGAAGCAAGC T[T]CTCTGTGC TGCCGTCCGA CT[T]GCATGTG TAA[A]GCATGC CG

b. V2 amplicon sequence

GGTGCTTATT CTTCAGGTAC CGTCATCAGC CGCTGATATT AGCAACAGCC TTTCTTCCCT GA

c. V3 amplicon sequence

GTGTTACGGC TCCCGAAGGC ACTCCTCCGT CTCCGGAGGA TTCCGTACAT GTCAAGACCA GG

d. <u>BLAST Analysis Results and Discussion.</u> BLAST identified *N. meningitidis* strain M1027 at the total score, query coverage, E-value, and identity level of 286, 82 %, 4e-25, and 100 %, respectively. But for a different set of these values – 321 (highest total score), 94 %, 2e-23, and 98 % - BLAST listed *N. cinerea* strain ATCC 14685. The major difference between the two alignments was that BLAST aligned the V1 sequence segment 1-37 at 100 % with *N. meningitidis*, but for *N. cinerea* it aligned 1-62 at 98 % (61/62). The reason for this is unclear. However, when we set the BLAST database at "nr/nt," most entries shown in the table were various strains of *N. meningitidis*, and these alignments also showed the same 1-62 segment identity at 100 %; thus, in this case "nr/nt" setting proved the better way to identify the target organism.

While the query sequences identified *N. meningitidis* as the most likely target, match differences with some other species were too close to this organism. Therefore, the identification should not be considered definitive.

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Neisseria meningitidis strain M1027 16S ribosomal RNA, partial sequence
Sequence ID: NR_104946.1 Length: 1415 Number of Matches: 3
Related Information
Range 1: 920 to 981 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
115 bits(62)	8e-26()	62/62(100%)	0/62(0%)	Plus/Minus	

Features:

```
Query  135  GTGTTACGGCTCCCGAAGGCACTCCTCCGTCTCCGGAGGATTCCGTACATGTCAAGACCA  194
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  981  GTGTTACGGCTCCCGAAGGCACTCCTCCGTCTCCGGAGGATTCCGTACATGTCAAGACCA  922

Query  195  GG  196
            ||
Sbjct  921  GG  920
```

Range 2: 378 to 440 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
110 bits(59)	4e-24()	62/63(98%)	1/63(1%)	Plus/Minus	

Features:

```
Query  68   GGTGCTTATTCTTCAGGTACCGTCATCAGCCGCTGATATTAGCAACAGCCTTT-CTTCCC  126
            |||||||||||||||||||||||||||||||||||||||||||||||||||| ||||||
Sbjct  440  GGTGCTTATTCTTCAGGTACCGTCATCAGCCGCTGATATTAGCAACAGCCTTTTCTTCCC  381

Query  127  TGA  129
            |||
Sbjct  380  TGA  378
```

Range 3: 1 to 37 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #3

Score	Expect	Identities	Gaps	Strand	Frame
69.4 bits(37)	6e-12()	37/37(100%)	0/37(0%)	Plus/Minus	

Features:

```
Query  1   TCGCCACCCGAGAAGCAAGCTTCTCTGTGCTGCCGTC  37
           |||||||||||||||||||||||||||||||||||||
Sbjct  37  TCGCCACCCGAGAAGCAAGCTTCTCTGTGCTGCCGTC  1
```

16. *Pseudomonas aeruginosa* strain PAO1-LAC

a. V1 amplicon sequence

TGAATCCAGG AGCAAGCTCC CTTCATCCGC TCGACTTGCA TGTGTTAGGC CTGCCGCCAG CG

b. V2 amplicon sequence

GGTGCTTATT CTGTTGGTAA CGTCAAAACA GCAAGGTATT AACTTACTGC CTTCCTCCCA ACTTAAAGTG CTTTA

c. V3 amplicon sequence

GTGTCTGAGT TCCCGAAGGC ACCAATCCAT CTCTGGAAAG TTCTCAGCAT GTCAAGG

d. **<u>BLAST Analysis Results and Discussion.</u>** Listed at the top of the BLAST generated table was *P. aeruginosa* isolate RW109. The BLAST setting was "nr/nt." The total score, query coverage, E-value, and identity were 1424, 95 %, 1e-27, and 99 %, respectively. For BLAST setting at "16S ribosomal RNA (Bacteria and Archaea)," the top three organisms were three different strains of *P. aeruginosa*, all with 356, 95 %, 2e-31, and 99 % for total score, query coverage, E-value, and identity, respectively. The results show that the query sequences identified the target bacterium as the most likely match.

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Pseudomonas aeruginosa strain ATCC 10145 16S ribosomal RNA gene, partial sequence
Sequence ID: NR_114471.1 Length: 1489 Number of Matches: 3
Related Information
Range 1: 415 to 490 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
134 bits(72)	2e-31()	75/76(99%)	1/76(1%)	Plus/Minus	

Features:

```
Query  68   GGTGCTTATTCTGTTGGTAACGTCAAAACAGCAAGGTATTAACTTACTGCC-TTCCTCCC  126
            |||||||||||||||||||||||||||||||||||||||||||||||||| ||||||||
Sbjct  490  GGTGCTTATTCTGTTGGTAACGTCAAAACAGCAAGGTATTAACTTACTGCCCTTCCTCCC  431

Query  127  AACTTAAAGTGCTTTA  142
            ||||||||||||||||
Sbjct  430  AACTTAAAGTGCTTTA  415
```

Range 2: 26 to 87 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
115 bits(62)	8e-26()	62/62(100%)	0/62(0%)	Plus/Minus	

Features:

```
Query  1   TGAATCCAGGAGCAAGCTCCCTTCATCCGCTCGACTTGCATGTGTTAGGCCTGCCGCCAG  60
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  87  TGAATCCAGGAGCAAGCTCCCTTCATCCGCTCGACTTGCATGTGTTAGGCCTGCCGCCAG  28

Query  61  CG  62
           ||
Sbjct  27  CG  26
```

Range 3: 975 to 1031 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #3

Score	Expect	Identities	Gaps	Strand	Frame

106 bits(57) 5e-23() 57/57(100%) 0/57(0%) Plus/Minus

Features:

```
Query  148   GTGTCTGAGTTCCCGAAGGCACCAATCCATCTCTGGAAAGTTCTCAGCATGTCAAGG  204
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1031  GTGTCTGAGTTCCCGAAGGCACCAATCCATCTCTGGAAAGTTCTCAGCATGTCAAGG  975
```

17. *Serratia marcescens* strain CDC 3100-71

a. V1 amplicon sequence

TCGTCACC[C]AGGAG CAAGCTCCCTGTGCTA CCGCTCGACT TGCATGTGTT AAGCCTGCC[C] GCC[C]AGC[C]G

b. V2 amplicon sequence

GGTGCTTCTT CTGCGAGTAA CGTCAATTGA TGAACGTATT AAGTCACCAC CTTCCTCCTC GC

c. V3 amplicon sequence

GTCTCAGAGT TCC[C]GAAGGC ACCAA[A]CATC TCTGATCTTG TAAGA

d. <u>BLAST Analysis Results and Discussion.</u> For "nr/nt" as the search database selection, BLAST listed most organisms as various strains of *S. marcescens*. These results suggest that this bacterium can be reliably identified with the sequences generated by pyrosequencing. For the "16S ribosomal RNA (Bacteria and Archaea)" search setting and the total score, query coverage, E-value, and identity values of 212, 69 %, 3e-24, and 98 %, respectively, BLAST listed top three bacteria as *S. marcescens* (three strains). Thus, the query sequences generated by pyrosequencing could identify *S. marcescens* as the best match; however, *S. nematodiphilia* alignments differed only by 3 nucleotides, making it impossible to definitively identify the target organism. Also note that with the setting "Somewhat similar sequences" all three segments were aligned. But that did not resolve the issue whether *S. marcescens* could be considered the definitive identification.

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Serratia marcescens strain DSM 30121 16S ribosomal RNA gene, partial sequence
Sequence ID: NR_041980.1 Length: 1505 Number of Matches: 2
Related Information
Range 1: 421 to 483 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
110 bits(59)	3e-24()	62/63(98%)	1/63(1%)	Plus/Minus	

Features:

```
Query  75    GGTGCTTCTTCTGCGAGTAACGTCAATTGATGAACGTATTAAG-TCACCACCTTCCTCCT  133
             |||||||||||||||||||||||||||||||||||||||||| ||||||||||||||||
Sbjct  483   GGTGCTTCTTCTGCGAGTAACGTCAATTGATGAACGTATTAAGCTCACCACCTTCCTCCT  424
```

```
Query  134  CGC  136
            |||
Sbjct  423  CGC  421
```

Range 2: 14 to 80GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
102 bits(55)	6e-22()	65/69(94%)	4/69(5%)	Plus/Minus	

Features:

```
Query  1   TCGTCACCCA-GGAGCAAGCT-CCCTGTGCTACCGCTCGACTTGCATGTGTTAAGCCTGC  58
           |||||||||| |||||||||| ||||||||||||||||||||||||||||||||||||
Sbjct  80  TCGTCACCCAGGGAGCAAGCTCCCCTGTGCTACCGCTCGACTTGCATGTGTTAAGCCTGC  21

Query  59  CCGCCCAGC  67
           | ||| |||
Sbjct  20  C-GCC-AGC  14
```

18. *Shigella flexneri* strain 24570

a. V1 amplicon sequence

TCGTCAGCGA AACAGCAAGC GCTTCCTGTT ACCG TTCGAC TTGCATGTGT TATGCTGCCG CCAGCC

b. V2 amplicon sequence

GGTGCTTCTT CTGCGGGTAA CGTCAATGAG CAA[A]G[G]ATGA TTAATATTAT CATCCCTTCC CCTTCCCCCCTGCCCCTGGC AAAGGAATA

c. V3 amplicon sequence

GTCTCACGGT TCC[C]GAAGGC ACAT[T]CTCAT CTCTGAAA[A]C TTCCGTGGAT GTCAAGACCA GGTAAGG

d. <u>BLAST Analysis Results and Discussion.</u> With database set at "nr/nt," the analysis listed the first 8 organisms as *Shigella flexneri*, but different strains; the total score, query coverage, E-value, and identity being the same for all – 2130, 82 %, 9e-25, 100 %, respectively. With the database set at "16S ribosomal RNA (Bacteria and Archaea)," BLAST listed *S. flexneri* strain ATCC 29903 as the first identified organism. The total score, query coverage, E-value, and identity were 306, 82 %, 2e-28, and 100 %, respectively. However, BLAST also showed exactly the same alignments for *Escherichia albertii* strain Albert 19982, except one base mismatch, and this was the same level of match as for *Shigella sonnei* strain CECT 4887. With somewhat lower total scores, Blast identified other bacteria as well, among them *Escherichia coli* strain U 5/41, *Escherichia fergusonii* strain ATCC 35469, *Escherichia coli* strain NBRC 102203 and *Shigella boydii* strain P288.

Together, these results suggest that the "nr/nt" setting BLAST search could identify the target bacterium with greater reliability than the restricted 16S rRNA gene search, which found matches with other bacteria as well, with small differences. Both V1 and V2 have some red "failed" sequences which can have some negative effect using BLAST for a specific target.

BLAST aligned all of V1, except the last C, and despite the whole sequence being yellow or red; 64/67 = 96 %. Clearly, the color assignment by the software as an index of sequence reliability was wrong, and therefore the color assignments should not be used as the defining parameters for which parts of the sequences could be considered reliable and then used for BLAST analysis.

BLAST aligned V2 sequence up to CCCTTCC, despite the yellow and red; 54/60 = 90 %. Again, the software assignment of yellow and red was imprecise.

BLAST aligned all of V3, despite the yellow; 67/67 = 100 %

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Shigella flexneri strain ATCC 29903 16S ribosomal RNA gene, partial sequence
Sequence ID: NR_026331.1 Length: 1488 Number of Matches: 3
Related Information
Range 1: 960 to 1026 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
124 bits(67)	2e-28()	67/67(100%)	0/67(0%)	Plus/Minus	

Features:

```
Query  166   GTCTCACGGTTCCCGAAGGCACATTCTCATCTCTGAAAACTTCCGTGGATGTCAAGACCA  225
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1026  GTCTCACGGTTCCCGAAGGCACATTCTCATCTCTGAAAACTTCCGTGGATGTCAAGACCA  967

Query  226   GGTAAGG  232
             |||||||
Sbjct  966   GGTAAGG  960
```

Range 2: 16 to 82 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
106 bits(57)	6e-23()	64/67(96%)	2/67(2%)	Plus/Minus	

Features:

```
Query  1    TCGTCAGCGAAACAGCAAGC-GCTTCCTGTTACCGTTCGACTTGCATGTGTTATGC-TGC  58
            |||||||||||||||||||| |||||||||||||||||||||||||||||| || |||
Sbjct  82   TCGTCAGCGAAACAGCAAGCTGCTTCCTGTTACCGTTCGACTTGCATGTGTTAGGCCTGC  23

Query  59   CGCCAGC  65
            |||||||
Sbjct  22   CGCCAGC  16
```

Range 3: 430 to 485 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #3

Score	Expect	Identities	Gaps	Strand	Frame
75.0 bits(40)	2e-13()	54/60(90%)	4/60(6%)	Plus/Minus	

Features:

```
Query  72   GGTGCTTCTTCTGCGGGTAACGTCAATGAGCAAAGGATGATTAATATTATCATcccttcc  131
            ||||||||||||||||||||||||||||||||||| | |||||  ||| | ||||||||
Sbjct  485  GGTGCTTCTTCTGCGGGTAACGTCAATGAGCAAAGG-T-ATTAACTTTA-C-TCCCTTCC  430
```

19. *Staphylococcus aureus* subsp. *aureus* strain TCH1516

a. V1 amplicon sequence

TAA[A]CATCAG A[A]GAAGCAAG CT[T]CTCGTC[C] GT[T]CGCTCGA CTTGCATGTA TTAGGCACGC CGCCCA

b. V2 amplicon sequence

GTGGCTTTCT GATTAGGTAC CGTCAAGATG TGCACAGTTA CTTACACATA TGTTCTTCCC TAA

c. V3 amplicon sequence

GTCACTTTGT CCCCGAAGGG AAGGCTCTAT CTCTAGAGTT GTCAAAGGAT GTCAAGATT[T] GGTAAGGTC

d. <u>BLAST Analysis Results and Discussion.</u> Search with the database setting as "nr/nt" put *S. aureus* subsp. *aureus* strain Tager 104 on top of the list, with the match metrics of 2398, 93 %, 4e-23, and 97 %, respectively. But with somewhat lower corresponding values, BLAST also listed many different strains of *S. aureus*. Search with the database setting at "16S ribosomal RNA (Bacteria and Archaea)" listed the first three organisms as *S. aureus* strain S33 R, *S. aureus* strain ATCC 12600, and *S. aureus* strain NBRC 100910. The total score, query coverage, E-value, and identity level were the same for all three; 343, 92 %, 2e-26, and 100 %, respectively. At lower total scores but the same other values, BLAST listed some other species, but they had at least 1 bp alignment difference. Note that the total scores are much higher when the target alignment DNA is the whole genome, as is the case here, even when the same size segments are aligned as for the 16S rRNA gene database setting. The reason is that for the whole genomes, these sequences are aligned at more than one site, and that is because the 16S rRNA gene is a multi-copy gene.

Although the sequences generated by pyrosequencing identified S. aureus as the top choice, it is impossible to consider this definitive. The reason is that some other organisms had near identical matches. *S. simiae*, for example, differed only by two nucleotides. Therefore the sequences were not sufficient to definitively identify the target organism.

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Staphylococcus aureus strain ATCC 12600 16S ribosomal RNA gene, partial sequence

Sequence ID: NR_115606.1 Length: 1476 Number of Matches: 3
Related Information
Range 1: 424 to 486 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
117 bits(63)	2e-26()	63/63(100%)	0/63(0%)	Plus/Minus	

Features:

```
Query  72   GTGGCTTTCTGATTAGGTACCGTCAAGATGTGCACAGTTACTTACACATATGTTCTTCCC  131
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  486  GTGGCTTTCTGATTAGGTACCGTCAAGATGTGCACAGTTACTTACACATATGTTCTTCCC  427

Query  132  TAA  134
            |||
Sbjct  426  TAA  424
```

Range 2: 961 to 1030 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
117 bits(63)	2e-26()	68/70(97%)	2/70(2%)	Plus/Minus	

Features:

```
Query  140   GTCACTTTGT-CCCCGAA-GGGAAGGCTCTATCTCTAGAGTTGTCAAAGGATGTCAAGAT  197
            |||||||||| ||||||| ||||||||||||||||||||||||||||||||||||||||
Sbjct  1030  GTCACTTTGTCCCCCGAAGGGGAAGGCTCTATCTCTAGAGTTGTCAAAGGATGTCAAGAT  971

Query  198   TTGGTAAGGT  207
            ||||||||||
Sbjct  970   TTGGTAAGGT  961
```

Range 3: 13 to 73 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #3

Score	Expect	Identities	Gaps	Strand	Frame
108 bits(58)	1e-23()	61/62(98%)	1/62(1%)	Plus/Minus	

Features:

```
Query  3   AACATCAGAAGAAGCAAGCTTCTCGTCCGTTCGCTCGACTTGCATGTATTAGGCACGCCG  62
           |||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  73  AACATCAG-AGAAGCAAGCTTCTCGTCCGTTCGCTCGACTTGCATGTATTAGGCACGCCG  15

Query  63  CC  64
           ||
Sbjct  14  CC  13
```

20. *Staphylococcus epidermidis* FDA strain PCI 1200

a. V1 amplicon sequence

TAACGTCAGA GGAGCAAGCT CCTCGTCTGT [T]CGCTCGACT [T]GCATGTATT AGGCACGCC

b. V2 amplicon sequence

GTGGCTTTCT GATTAGGTAC CGTCAAGACG TGCATAGTTA CTTACACATT TGTTCTTCCC TAA

c. V3 amplicon sequence

GTCACTCTGT CCCCGAAGG[G] AAAACTCTAT CTCTAGAGGG [G]TGCAGAGGA TGGTCAAGAA TTTGGGTTGA AGG

d. <u>BLAST Analysis Results and Discussion.</u> When the search was performed with database setting of "nr/nt," BLAST analysis listed *S. epidermidis* ATCC 12228 (complete genome) at the top of the table, with the total score, query coverage, E-value, and identity values of 1354, 59 %, 1e-22, and 100 %, respectively. At lower total scores, it listed many other strains of *S. epidermidis*, but also some other *S. spp.*, e.g., *S. capitis* strain ISLP22 with the same query coverage, E-value, and identity. Search with the database setting at "16S ribosomal RNA (Bacteria and Archaea)" listed seven *Staphylococcus spp.*, among them *S. epidermidis* strains Fussel and NBRC 100911. The other five were *S. capitis* strains. All seven had the same total score, query coverage, E-value, and identity; 227, 59 %, 2e-26, 100 %, respectively. Note that BLAST excluded the V3 sequence from all alignments displayed when the setting was "Highly similar sequences," but it included all three segments when the setting was "Somewhat similar sequences." However, the latter did not improve the reliability of identifications. These results suggest that the sequences generated by pyrosequencing and used for BLAST here could not reliably pinpoint the match solely to *S. epidermidis.*

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Staphylococcus epidermidis strain NBRC 100911 16S ribosomal RNA gene, partial sequence
Sequence ID: NR_113957.1 Length: 1476 Number of Matches: 2
Related Information
Range 1: 422 to 484 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
117 bits(63)	2e-26()	63/63(100%)	0/63(0%)	Plus/Minus	

Features:

```
Query  65   GTGGCTTTCTGATTAGGTACCGTCAAGACGTGCATAGTTACTTACACATTTGTTCTTCCC  124
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  484  GTGGCTTTCTGATTAGGTACCGTCAAGACGTGCATAGTTACTTACACATTTGTTCTTCCC  425

Query  125  TAA  127
           |||
Sbjct  424  TAA  422
```

Range 2: 14 to 72 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
110 bits(59)	4e-24()	59/59(100%)	0/59(0%)	Plus/Minus	

Features:

```
Query  1   TAACGTCAGAGGAGCAAGCTCCTCGTCTGTTCGCTCGACTTGCATGTATTAGGCACGCC  59
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  72  TAACGTCAGAGGAGCAAGCTCCTCGTCTGTTCGCTCGACTTGCATGTATTAGGCACGCC  14
```

21. *Staphylococcus haemolyticus* strain SM 131

a. V1 amplicon sequence

TAACGTCAAA GGAGCAAGCT CCTTGTCTGT TCGCTCGACT TGCATGTATT AGGCACGCCG

b. V2 amplicon sequence

GTGGCTTTCT GATTAGGTAC CGTCAAGACG TGCATAGTTA CTTACACGTA TGTTCTTCCC TAA

c. V3 amplicon sequence

GTCACTT[T]GT CCCC[C]GAAGG G[G]AAG[G]CTCT ATCTCTAGAG TTGTCAAAGG ATGTCAAGAT TTGGTAA

d. <u>BLAST Analysis Results and Discussion.</u> With the database setting as "nr/nt," the top organism listed in the BLAST analysis table was *S. hemolyticus* strain SGAir0252 (complete genome), for which the total score, query coverage, E-value, and identity were 2084, 95 %, 8e-25, 100 %, respectively. However, for many different species (e.g., *S. aureus, S. cohnii, S. simulans*) and strains that followed this listing in the table with lower total scores, but the same other values, differed only slightly, sometimes by only one nucleotide. Thus, these sequences generated by pyrosequencing could not be used to precisely identify the target organism, *S. haemolyticus*.

With the database search set at "16S ribosomal RNA (Bacteria and Archaea)," BLAST identified *S. haemolyticus* strains 2416 and SM 131, with a difference of only one nucleotide in the alignments. The respective values for total score, query coverage, E-value, and identity were 350 (341 for SM 131), 95 %, 2e-27, and 99 %. Restricting the search to the 16S rRNA database, therefore, appeared to identify this organism with greater reliability than with the "nr/nt" search. But the differences for some other species were not appreciable, and therefore the identification could not be considered definitive.

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Staphylococcus haemolyticus strain JCM 2416 16S ribosomal RNA gene, partial sequence
Sequence ID: NR_113345.1 Length: 1473 Number of Matches: 3
Related Information
Range 1: 962 to 1028 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
121 bits(65)	2e-27()	66/67(99%)	0/67(0%)	Plus/Minus	

Features:

```
Query  134   GTCACTTTGTCCCCCGAAGGGGAAGGCTCTATCTCTAGAGTTGTCAAAGGATGTCAAGAT  193
             |||||||||||||||||||||| ||||||||||||||||||||||||||||||||||||
Sbjct  1028  GTCACTTTGTCCCCCGAAGGGGAARGCTCTATCTCTAGAGTTGTCAAAGGATGTCAAGAT  969

Query  194   TTGGTAA  200
             |||||||
Sbjct  968   TTGGTAA  962
```

Range 2: 422 to 484 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
117 bits(63)	2e-26()	63/63(100%)	0/63(0%)	Plus/Minus	

Features:

```
Query  66    GTGGCTTTCTGATTAGGTACCGTCAAGACGTGCATAGTTACTTACACGTATGTTCTTCCC  125
             |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  484   GTGGCTTTCTGATTAGGTACCGTCAAGACGTGCATAGTTACTTACACGTATGTTCTTCCC  425

Query  126   TAA  128
             |||
Sbjct  424   TAA  422
```

Range 3: 13 to 72 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #3

Score	Expect	Identities	Gaps	Strand	Frame
111 bits(60)	1e-24()	60/60(100%)	0/60(0%)	Plus/Minus	

Features:

```
Query  1     TAACGTCAAAGGAGCAAGCTCCTTGTCTGTTCGCTCGACTTGCATGTATTAGGCACGCCG  60
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  72    TAACGTCAAAGGAGCAAGCTCCTTGTCTGTTCGCTCGACTTGCATGTATTAGGCACGCCG  13
```

22. *Staphylococcus hominis* subsp. *novobiosepticus* strain R22

a. V1 amplicon sequence

TAACGTCAAA GGAGCAAGCT CCTCGTCTGT TGCTCACCTT GCATGTATTA GGCACGCCGC CA

b. V2 amplicon sequence

GTGGCTTTCT GATTAGGTAC CGTCAAGACG TGCACAGTTA CTTACACGTT [T]GTTCTT[**T**]CC C

c. V3 amplicon sequence

GTCACTTTGT CCCCGAAGGG AAACTTCTAT CTCTAGAAGG GTCAAAGGAT GTCAAGATTT GGTAAGGTTC T

d. **BLAST Analysis Results and Discussion.** For "nr/nt" database setting, BLAST identified six bacteria with identical alignments, except one nucleotide difference for some, and with the values 332, 95 %, 2e-25, and 99 % for total score, query coverage, E-value, and identity, respectively. All were listed as "uncultured bacterium," various clones. At total score of 328 and the same other values, the analysis listed three different strains of *S. hominis* subsp. *hominis*, with only one nucleotide alignment difference. At progressively lower values for the four metrics, BLAST further identified many different strains of *S. hominis*, but these had more than one nucleotide mismatches with the query sequences. With the database search set at "16S ribosomal RNA (Bacteria and Archaea)," the first two bacteria listed in the table were *S. hominis* strain DM 122 and subsp. *novobiosepticus* strain GTC 1228. The total score, query coverage, E-value, and identity metrics for these two were 326 (229 for the latter), 95 %, 2e-27, and 97 %, respectively.

Together these results show that the sequences generated identified *S. hominis* with reasonable reliability. But the sequences generated by pyrosequencing were too short to precisely identify the strain.

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Staphylococcus hominis strain DM 122 16S ribosomal RNA, partial sequence
Sequence ID: NR_036956.1 Length: 1544 Number of Matches: 3
Related Information
Range 1: 976 to 1046 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
121 bits(65)	2e-27()	70/72(97%)	2/72(2%)	Plus/Minus	

Features:

```
Query  134   GTCACTTTGT-CCCCGAAGGGAAACTTCTATCTCTAGAAGGGTCAAAGGATGTCAAGATT  192
             |||||||||| ||||||| ||||||||||||||||||||||||||||||||||||||||
Sbjct  1046  GTCACTTTGTCCCCCGAA-GGAAACTTCTATCTCTAGAAGGGTCAAAGGATGTCAAGATT  988

Query  193   TGGTAAGGTTCT  204
             ||||||||||||
Sbjct  987   TGGTAAGGTTCT  976
```

Range 2: 445 to 504 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
106 bits(57)	5e-23()	60/61(98%)	1/61(1%)	Plus/Minus	

Features:

```
Query  68    GTGGCTTTCTGATTAGGTACCGTCAAGACGTGCACAGTTACTTACACGTTTGTTCTTTCC  127
             |||||||||||||||||||||||||||||||||||||||||||||||||||||||||| ||
Sbjct  504   GTGGCTTTCTGATTAGGTACCGTCAAGACGTGCACAGTTACTTACACGTTTGTTCTT-CC  446
```

```
Query  128  C  128
            |
Sbjct  445  C  445
```

Range 3: 30 to 92 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #3

Score	Expect	Identities	Gaps	Strand	Frame
99.0 bits(53)	8e-21()	60/63(95%)	1/63(1%)	Plus/Minus	

Features:

```
Query  1   TAACGTCAAAGGAGCAAGCTCCTCGTCTGTT-GCTCACCTTGCATGTATTAGGCACGCCG  59
           ||||||||||||||||||||||||||||||| ||||  ||||||||||||||||||||
Sbjct  92  TAACGTCAAAGGAGCAAGCTCCTCGTCTGTTCGCTCGACTTGCATGTATTAGGCACGCCG  33

Query  60  CCA  62
           |||
Sbjct  32  CCA  30
```

23. *Staphylococcus lugdunensis* strain N860297

a. V1 amplicon sequence

TAACGTCAAA GGAGCAAGCT CCTTATCTGT TCGCTCGACT TGCATGTATT AGGCACGCCG

b. V2 amplicon sequence

GTGGCTTTCT GATTAGGTAC CGTCAAGACG TGCACAGTTA CTTACACGTT TGTTCTTCCC TAATAA

c. V3 amplicon sequence

GTCACTTTGT CCCCCGAAGG G[G]AAGACTCT ATCTCTAGAG CGGTCAAAGG ATGTCAAGAT TTGGTAA

d. <u>BLAST Analysis Results and Discussion.</u> For "nr/nt" database setting, the query coverage, E-value, and identity values 95 %, 8e-25, and 100 %, respectively, the analysis identified four strains of *S. lugdunensis* as the top four. The strains were FDAARGOS 381, FDAARGOS 377, FDAARGOS 143, and FDAARGOS 222. The respective total scores for these were 2151, 2136, 2028, and 1793, all subject sequences in alignments were listed as complete genome. The alignment differences were one to a few nucleotides. For lower metrics for the four parameters, the analysis still listed a large number of *Staphylococcus spp*. When the search setting was "16S ribosomal RNA (Bacteria and Archaea)," the best match was with *S. lugdunensis* strain ATCC 43809. The metrics for this match were 359, 95 %, 1e-28, and 100 % for total score, query coverage, E-value, and identity, respectively. For lower metrics, the analysis listed some other species, e.g., *S. pasteuri, S. nepalensis, and S. auricularis*, and these showed at least 2 nucleotide divergence.

Overall, BLAST placed *S. lugdunensis* as the first choice in the table. But some other species were within 2-3 nucleotide differences, and therefore the sequences generated by pyrosequencing here could not be considered sufficient to definitively identify the bacterium.

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Staphylococcus lugdunensis strain ATCC 43809 16S ribosomal RNA gene, partial sequence
Sequence ID: NR_024668.1Length: 1492 Number of Matches: 3
Related Information
Range 1: 982 to 1048GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
124 bits(67)	1e-28()	67/67(100%)	0/67(0%)	Plus/Minus	

Features:

```
Query  137   GTCACTTTGTCCCCCGAAGGGGAAGACTCTATCTCTAGAGCGGTCAAAGGATGTCAAGAT  196
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1048  GTCACTTTGTCCCCCGAAGGGGAAGACTCTATCTCTAGAGCGGTCAAAGGATGTCAAGAT  989

Query  197   TTGGTAA  203
             |||||||
Sbjct  988   TTGGTAA  982
```

Range 2: 439 to 504GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
122 bits(66)	5e-28()	66/66(100%)	0/66(0%)	Plus/Minus	

Features:

```
Query  66    GTGGCTTTCTGATTAGGTACCGTCAAGACGTGCACAGTTACTTACACGTTTGTTCTTCCC  125
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  504   GTGGCTTTCTGATTAGGTACCGTCAAGACGTGCACAGTTACTTACACGTTTGTTCTTCCC  445

Query  126   TAATAA  131
             ||||||
Sbjct  444   TAATAA  439
```

Range 3: 33 to 92GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #3

Score	Expect	Identities	Gaps	Strand	Frame
111 bits(60)	1e-24()	60/60(100%)	0/60(0%)	Plus/Minus	

Features:

```
Query  1   TAACGTCAAAGGAGCAAGCTCCTTATCTGTTCGCTCGACTTGCATGTATTAGGCACGCCG  60
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  92  TAACGTCAAAGGAGCAAGCTCCTTATCTGTTCGCTCGACTTGCATGTATTAGGCACGCCG  33
```

24. *Staphylococcus saprophyticus* subsp. *saprophyticus* strain NCTC 7292

a. V1 amplicon sequence

TAACGTCAAA GGAGCAAGCT CCTTATCTGT TCGCTCGACT TGCATGTATT AGGCACGCCG

b. V2 amplicon sequence

GTGGCTTTCT GATTAGGTAC CGTCAAGACG TGCACAGTTA CTTACACTTT GTTCTTCCCT AATAA

c. V3 amplicon sequence

GTCACTTTGT CCCCGAAGGG AAGGCTCTAT CTCTAGAGTT TTCAAAGGAT GTCAAGATTT GGTAAGG

d. <u>BLAST Analysis Results and Discussion.</u> With database search setting "nr/nt," and the total score, query coverage, E-value, and identity metrics 2170, 91 %, 4e-23, and 99 %, respectively, the best match was *Staphylococcus spp.* AntiMn-1 (complete genome). With lower total scores, but only one to a few nucleotide alignment differences, BLAST identified *S. succinus*, *S. cohnii*, and *S. aureus*. In the "16S ribosomal RNA (Bacteria and Archaea)" database search, four strains of *S. saprophyticus* were on the top of the list – ATCC 15305, JCR2427, NBRC 102446, and subsp. *saprophyticus* ATCC 15305. The total score, query coverage, E-value, and identity for all four were 343, 95 %, 8e-26, and 98 %, respectively.

Although the analysis placed *S. saprophyticus* first in the list of identified bacteria, the query sequence differences with some other species and genera were not sufficiently diverse to consider *S. saprophyticus* as the definitive identification.

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Staphylococcus saprophyticus strain ATCC 15305 16S ribosomal RNA gene, partial sequence
Sequence ID: NR_115607.1Length: 1477Number of Matches: 3
Related Information
Range 1: 421 to 486GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
115 bits(62)	8e-26()	65/66(98%)	1/66(1%)	Plus/Minus	

Features:

```
Query  66   GTGGCTTTCTGATTAGGTACCGTCAAGACGTGCACAGTTACTTACAC-TTTGTTCTTCCC  124
           |||||||||||||||||||||||||||||||||||||||||||||| ||||||||||||
Sbjct  486  GTGGCTTTCTGATTAGGTACCGTCAAGACGTGCACAGTTACTTACACGTTTGTTCTTCCC  427

Query  125  TAATAA  130
            ||||||
Sbjct  426  TAATAA  421
```

Range 2: 962 to 1030GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
115 bits(62)	8e-26()	67/69(97%)	2/69(2%)	Plus/Minus	

Features:

```
Query  136   GTCACTTTGT-CCCCGAA-GGGAAGGCTCTATCTCTAGAGTTTTCAAAGGATGTCAAGAT  193
             |||||||||| ||||||| ||||||||||||||||||||||||||||||||||||||||
Sbjct  1030  GTCACTTTGTCCCCCGAAGGGGAAGGCTCTATCTCTAGAGTTTTCAAAGGATGTCAAGAT  971

Query  194   TTGGTAAGG  202
             |||||||||
Sbjct  970   TTGGTAAGG  962
```

Range 3: 15 to 74GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #3

Score	Expect	Identities	Gaps	Strand	Frame
111 bits(60)	1e-24()	60/60(100%)	0/60(0%)	Plus/Minus	

Features:

```
Query  1   TAACGTCAAAGGAGCAAGCTCCTTATCTGTTCGCTCGACTTGCATGTATTAGGCACGCCG  60
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  74  TAACGTCAAAGGAGCAAGCTCCTTATCTGTTCGCTCGACTTGCATGTATTAGGCACGCCG  15
```

25. *Staphylococcus schleiferi* subsp. *schleiferi* strain N850274

a. V1 amplicon sequence

TAACTTCAAA G[G]AGCAAGCT CCTCGTCCGT TCGCTCGACT TGCATGTATT AGGCACGCCG CC[C]AGC[C]

b. V2 amplicon sequence

GTGGCTT[T]CT GGTTAGGTAC CGTCAAGACG TGCACAGTTA CTTACACAAT TTGTTTCTTT CCCTTCAA

c. V3 amplicon sequence

GTCACTTTGT CCTCCGAAGA GGAAAACTCT ATCTCTAGAG CGGTCAAAGG ATGTCAAGAT TTGGTAA

d. <u>BLAST Analysis Results and Discussion.</u> For "nr/nt" database search, the top three matches were *S. schleiferi* strain 1360-13 (complete genome), *S. schleiferi* strain 2142-05, and *S. schleiferi* strain 2317-03. The match metrics for all three were 2003, 92 %, 8e-25, and 100 % for total score, query coverage, E-value, and identity, respectively. For lower total scores but the same other values, several different species of *Staphylococcus*, as well as strains of *S. schleiferi* appeared. These showed from one or two to many nucleotide differences with the strain listed as the best match. With search set at "16S ribosomal RNA (Bacteria and Archaea)," and total score, query coverage, E-value, and identity of 333, 92 %, 1e-28, and 100 %, respectively, BLAST identified *S. schleiferi* strain DSM4807 and *S. schleiferi* subsp. *coagulans* strain GA211 as the best matches.

But then for lower total scores, even if for the same other values, the alignment differences of two or more nucleotides emerged. Based on these search results, we conclude that the query sequences generated by pyrosequencing were not of sufficient length and quality to precisely identify *Staphylococcus schleiferi*.

The alignments shown below are based on the settings “Highly similar sequences (megablast)” and “16S ribosomal RNA (Bacteria and Archaea).”

Staphylococcus schleiferi strain DSM 4807 16S ribosomal RNA gene, partial sequence
Sequence ID: NR_037009.1Length: 1527Number of Matches: 3
Related Information
Range 1: 962 to 1028GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
124 bits(67)	1e-28()	67/67(100%)	0/67(0%)	Plus/Minus	

Features:

```
Query  146   GTCACTTTGTCCTCCGAAGAGGAAAACTCTATCTCTAGAGCGGTCAAAGGATGTCAAGAT  205
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1028  GTCACTTTGTCCTCCGAAGAGGAAAACTCTATCTCTAGAGCGGTCAAAGGATGTCAAGAT  969

Query  206   TTGGTAA  212
             |||||||
Sbjct  968   TTGGTAA  962
```

Range 2: 8 to 72GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
115 bits(62)	8e-26()	65/66(98%)	1/66(1%)	Plus/Minus	

Features:

```
Query  1    TAACTTCAAAGGAGCAAGCTCCTCGTCCGTTCGCTCGACTTGCATGTATTAGGCACGCCG  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  72   TAACTTCAAAGGAGCAAGCTCCTCGTCCGTTCGCTCGACTTGCATGTATTAGGCACGCCG  13

Query  61   CCCAGC  66
            || |||
Sbjct  12   CC-AGC  8
```

Range 3: 424 to 484GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #3

Score	Expect	Identities	Gaps	Strand	Frame
93.5 bits(50)	4e-19()	60/64(94%)	3/64(4%)	Plus/Minus	

Features:

```
Query  73   GTGGCTTTCTGGTTAGGTACCGTCAAGACGTGCACAGTTACTTACACAATTTGTTTCTTT  132
            |||||||||||||||| |||||||||||||||||||||||||||||| |||||| |||
Sbjct  484  GTGGCTTTCTGGTTAGGTGCCGTCAAGACGTGCACAGTTACTTACACA-TTTGTT-CTT-  428

Query  133  CCCT  136
            ||||
```

```
Sbjct  427  CCCT  424
```

26. *Stenotrophomonas maltophilia* strain 810-2

a. V1 amplicon sequence

TCGCCACCCA GAGAGCAAGC TCTCCTGTGC TGCCGTTCGA CTTGCATGTG TTAGGCCTAC CG

b. V2 amplicon sequence

GGTGCTTATT CTTTGGGTAC CGTCATCCCA ACCGGGTATT AACCAGCTGG ATTTCTTTCC CAACAAA[A]GG GCTTTACAAC CGA

c. V3 amplicon sequence

GTGTTCGAGT TCCCGAAGGC ACCAATCCAT CTCTGGAAAG TTCTCGACAT GTCAAGGCCA

d. <u>BLAST Analysis Results and Discussion.</u> The best match in the "nr/nt" database was *Stenotrophomonas maltophilia* strain NCTC 10257, identified with total score, query coverage, E-value, and identity of 1512, 94 %, 1e-32, and 100 %, respectively. After that, sequence differences in the aligned segments emerged, from five mismatches for the next one down in the table to many with progressively lower metrics. For the search setting "16S ribosomal RNA (Bacteria and Archaea)," the top four matches were *S. maltophilia* strains ATCC 13637, LMG958, IAM 12423, and NBRC 14161. The match metrics for all four were 378, 94 %, 2e-36, and 100 %, respectively. For the next lower score of 368 but the same other values, a difference of 3 nucleotides emerged. And the differences were greater for lower and lower metrics. With slightly lower metrics, resulting from 2-nucleotide difference in alignments, BLAST also identified two strains of *Stenotrophomonas pavanii*. We think this difference is too close to conclude that the query sequences used here for BLAST search proved sufficient to identify the target organism, *Stenotrophomonas maltophilia*.

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Stenotrophomonas maltophilia strain ATCC 13637 16S ribosomal RNA gene, partial sequence
Sequence ID: NR_112030.1 Length: 1467 Number of Matches: 3
Related Information
Range 1: 397 to 477 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
150 bits(81)	2e-36()	81/81(100%)	0/81(0%)	Plus/Minus	

Features:

```
Query  68   GGTGCTTATTCTTTGGGTACCGTCATCCCAACCGGGTATTAACCAGCTGGATTTCTTTCC  127
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  477  GGTGCTTATTCTTTGGGTACCGTCATCCCAACCGGGTATTAACCAGCTGGATTTCTTTCC  418
```

```
Query  128  CAACAAAAGGGCTTTACAACC  148
            ||||||||||||||||||||
Sbjct  417  CAACAAAAGGGCTTTACAACC  397
```

Range 2: 13 to 74 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
115 bits(62)	9e-26()	62/62(100%)	0/62(0%)	Plus/Minus	

Features:

```
Query  1   TCGCCACCCAGAGAGCAAGCTCTCCTGTGCTGCCGTTCGACTTGCATGTGTTAGGCCTAC  60
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  74  TCGCCACCCAGAGAGCAAGCTCTCCTGTGCTGCCGTTCGACTTGCATGTGTTAGGCCTAC  15

Query  61  CG  62
           ||
Sbjct  14  CG  13
```

Range 3: 960 to 1019 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #3

Score	Expect	Identities	Gaps	Strand	Frame
111 bits(60)	1e-24()	60/60(100%)	0/60(0%)	Plus/Minus	

Features:

```
Query  156   GTGTTCGAGTTCCCGAAGGCACCAATCCATCTCTGGAAAGTTCTCGACATGTCAAGGCCA  215
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1019  GTGTTCGAGTTCCCGAAGGCACCAATCCATCTCTGGAAAGTTCTCGACATGTCAAGGCCA  960
```

27. *Streptococcus agalactiae* strain 2603 V/R

a. V1 amplicon sequence

TCATCAGTCT AGTGTAAACA CCAAACCTCA GCGTCTACTG CTGTTTAGAC GCGCC

b. V2 amplicon sequence

GTCCCTTTCT GGTTAGTTAC CGTCACTTGG TAGATTTCCA CTCCTACCAA CGTCT[T]CTCT A

c. V3 amplicon sequence

GTCACTTCTG CTCCGAAGAG AAAGCCATAT CTCTAGGCCG G[G]TCAGAAGG AATGGTCAAG AACCTGGTGA A

d. <u>BLAST Analysis Results and Discussion.</u> BLAST did not align the V3 segment sequence for the setting "Highly similar sequences," but aligned all three sequences for the setting "Somewhat similar sequences." The best two matches in the "nr/nt" database were *S. agalactiae* strain FDAARGOS 254 and *Streptococcus spp.* group B strain FDAARGOS 229. The metrics for these matches were 1243, 58 %, 1e-18, and 97 % for total score, query coverage, E-value, and identity,

respectively. Both of these strains showed identical alignments with the query sequences. With search setting of "16S ribosomal RNA (Bacteria and Archaea)," there were four best matches, all with total score, query coverage, E-value, and identity of 177, 58 %, 2e-22, and 97 %, respectively. The four matches were actually *S. agalactiae* strains JCM 5671 and ATCC 13813, each listed twice. Only four other matches listed in the table, and these were two strains of *S. pyogenes* and 2 other species of *Streptococcus*. The match metrics for these were significantly lower. These results show that the V1 and V2 sequences generated by pyrosequencing were sufficient to identify *S. agalactiae*, even with some red "failed" sequences in V1.

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Streptococcus agalactiae strain JCM 5671 16S ribosomal RNA gene, partial sequence
Sequence ID: NR_113262.1 Length: 1471 Number of Matches: 2
Related Information
Range 1: 422 to 484 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
104 bits(56)	2e-22()	61/63(97%)	2/63(3%)	Plus/Minus	

Features:

```
Query  61   GTCCCTTTCTGGTTAGTTACCGTCACTTGGTAGATTT-CCACTCCTACCAACGT-CTTCT  118
            |||||||||||||||||||||||||||||||||||| ||||||||||||||| |||||
Sbjct  484  GTCCCTTTCTGGTTAGTTACCGTCACTTGGTAGATTTTCCACTCCTACCAACGTTCTTCT  425

Query  119  CTA  121
            |||
Sbjct  424  CTA  422
```

Range 2: 14 to 72 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
73.1 bits(39)	5e-13()	53/59(90%)	4/59(6%)	Plus/Minus	

Features:

```
Query  1    TCATCAGTCTAGTGTAAACACCAAACCTCAGCGT-CTACT-GC-TGT-TTAGACGCGCC  55
            ||||||||||||||||||||||||||||||||| ||||| || ||| |||| | ||||
Sbjct  72   TCATCAGTCTAGTGTAAACACCAAACCTCAGCGTTCTACTTGCATGTATTAGGCACGCC  14
```

28. *Streptococcus mitis* strain NCTC 12261

a. V1 amplicon sequence

TCATCCGGAC G[G]AAG[G]ACGA AGACTCCTCC [C]TTCCTACGC GTCTACTTGC TATGTATTAT

b. V2 amplicon sequence

GTCCCTTTCT GGTAAGATAC CGTCACAGTG TGAACTTTCC ACTCTCACAC TCGTTCTTCT

c. V3 amplicon sequence

GTCACCTCTG TCCCGAAGGA AAACTCTATC TCTAGAGCGG TCAGAGGGAT GTCAAGACCT GG

d. <u>BLAST Analysis Results and Discussion.</u> BLAST did not align the V1 segment sequence for the setting "highly similar sequences). With the database setting as "nr/nt," BLAST analysis results listed several different species and strains of *Streptococcus*, with *S. mitis* being at the top of the list. However, many different species and strains of *Streptococcus* had the same metrics; 227, 63 %, 8e-26, and 100 % for total score, query coverage, E-value, and identity, respectively. This made it impossible to specifically pinpoint the target organism.

With search setting of "16S ribosomal RNA (Bacteria and Archaea)," the results were essentially the same as for "nr/nt" setting; the search generated a table that listed different species and strains of *Streptococcus* with identical alignments and the same metrics for total score, query coverage, E-value, and identity; 227, 63 %, 8e-26, 100 %, respectively. These included *S. mitis*, *S. pneumoniae*, *S. oralis*, and *S. infantis*. As for the "nr/nt" search setting, the V2+V3 sequences therefore proved insufficient to precisely identify *Streptococcus mitis*.

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Streptococcus mitis strain NCTC12261 16S ribosomal RNA, partial sequence
Sequence ID: NR_115560.1 Length: 1403 Number of Matches: 2
Related Information
Range 1: 982 to 1043 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
119 bits(62)	4e-27()	62/62(100%)	0/62(0%)	Plus/Minus	

Features:

```
Query  131   GTCACCTCTGTCCCGAAGGAAAACTCTATCTCTAGAGCGGTCAGAGGGATGTCAAGACCT  190
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1043  GTCACCTCTGTCCCGAAGGAAAACTCTATCTCTAGAGCGGTCAGAGGGATGTCAAGACCT  984

Query  191   GG  192
             ||
Sbjct  983   GG  982
```

Range 2: 443 to 502 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
116 bits(60)	6e-26()	60/60(100%)	0/60(0%)	Plus/Minus	

Features:

```
Query  66    GTCCCTTTCTGGTAAGATACCGTCACAGTGTGAACTTTCCACTCTCACACTCGTTCTTCT  125
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  502   GTCCCTTTCTGGTAAGATACCGTCACAGTGTGAACTTTCCACTCTCACACTCGTTCTTCT  443
```

29. *Streptococcus mutans* Clarke

a. V1 amplicon sequence

TCAA[A]GAAAA [A]CA[A]CGGTGT GCAAGCACAG TGTGT[T]CCTT GCGTCCCTCT TTTAGACCCC

b. V2 amplicon sequence

GTCCCTT[T]CT GGTAAGCTAC CGTCACTGTG TGAACTTTCC ACTCTCACAC ACGTTCTTGA

c. V3 amplicon sequence

GTCTCCGATG TACCGAAGTA ACTTCCTATC TCTAAGAATA GCATCGGATG TCAAGACC

d. <u>BLAST Analysis Results and Discussion.</u> BLAST excluded the V1 sequence from the search. However, the V2+V3 sequence retained for search mostly identified various strains of *S. mutans*, but also some "uncultured" bacteria, which may be *S. mutans*; no other genus or species was listed in the table generated by the search. The highest values for total score, query coverage, E-value, and identity were 1073, 62 %, 5e-21, and 100 %, respectively. For the search setting as "16S ribosomal RNA (Bacteria and Archaea)," the results were similar to those for the "nr/nt" setting, except that some other species of *Streptococcus* were also listed. Still, most were *S. mutans* strains. The corresponding highest metric were 214, 62 %, 1e-24, and 100 %. These results suggest that the V2+V3 sequences generated by pyrosequencing could identify the target organism, *S. mutans* even with red "failed" sequences on V2.

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Streptococcus mutans strain NCTC 10449 16S ribosomal RNA, partial sequence
Sequence ID: NR_114726.1 Length: 1512 Number of Matches: 3
Related Information
Range 1: 433 to 492 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
116 bits(60)	6e-26()	60/60(100%)	0/60(0%)	Plus/Minus	

Features:

```
Query  66   GTCCCTTTCTGGTAAGCTACCGTCACTGTGTGAACTTTCCACTCTCACACACGTTCTTGA  125
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  492  GTCCCTTTCTGGTAAGCTACCGTCACTGTGTGAACTTTCCACTCTCACACACGTTCTTGA  433
```

Range 2: 975 to 1033 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame

98.7 bits(51)	9e-21()	58/59(98%)	1/59(1%)	Plus/Minus

Features:

```
Query  131   GTCTCCGATGTACCGAAGTAACTTCCTATCTCTAAGAATAGCATCGG-ATGTCAAGACC  188
             ||||||||||||||||||||||||||||||||||||||||||||| ||||||||||||
Sbjct  1033  GTCTCCGATGTACCGAAGTAACTTCCTATCTCTAAGAATAGCATCGGGATGTCAAGACC  975
```

Range 3: 32 to 70 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #3

Score	Expect	Identities	Gaps	Strand	Frame
54.5 bits(28)	2e-07()	37/39(95%)	1/39(2%)	Plus/Minus	

Features:

```
Query  14  ACGGTGTGCAAGCACAGTGTGTTCC-TTGCGTCCCTCTT  51
           |||||||||||||||||||||||| ||||||||| |||
Sbjct  70  ACGGTGTGCAAGCACAGTGTGTTCCCTTGCGTCCCACTT  32
```

30. *Streptococcus pyogenes* strain SF370; M1 GAS

a. V1 amplicon sequence (multiple PCR and pyrosequencing experiments were done)

TCCCTTT[T]G[G] CCAAATTGCC C[C]AATGGGCC ATTGGGCCAA ATGGGCC[C]AA AGGCCCCAAA GGCCAATTGC

CATTCCAATT GCCATGGCCT TGCCTTCATT CCATTGCCAT TGGCATTGAT TTAATTGGCA ATGCAGCCAG GCCCAGGCCC AGGCC

b. V2 amplicon sequence

GTCCCTTTCT GGTTAGTTAC CGTCACTTG[G] TGGATTT[T]CC ACTCCCACC[C] ATCATTCTT[T] CTCTAACAAA CAGA

c. V3 amplicon sequence

GTCACCGATG TACCGAAGTA AAACATCTAT CTTCCTTATG AAGGCAGGGG CCATGGCCGG GGCAATGGGG ATTGGCAAATGGCAAAGCCC A

d. <u>BLAST Analysis Results and Discussion.</u> BLAST did not align V1 and V3 sequences, but it did align the entire V2 sequence. When the database search setting was "nr/nt," most of the bacteria listed in the table were *S. pyogenes* strains, but some were listed as "uncultured" bacteria, and of these some had the same sequence identity as *S. pyogenes*, suggesting they may be *S. pyogenes* isolates. The "nr/nt" search identified *S. pyogenes* successfully.

A special case was that of *Bacillus licheniformis* strain SR-05-02, which had the same match metrics as *S. pyogenes* strains. Because *Streptococcus* and *Bacillus* have markedly different lineages in systematics, we reasoned that *B. licheniformis* may be an erroneous label for the sequence entered under this designation (Accession # KC821514.1; 1306 bp; 16S rRNA gene). To

determine whether this notion held, we did an independent BLAST analysis of this 1306-bp sequence. And indeed we found that this sequence is specific for *Streptococcus*, not *Bacillus*; all entries in the table generated by this search were various species and strains of *Streptococcus*, including *S. pyogenes*. Clearly the 1306 bp sequence (Accession # KC821514.1) does not belong to *B. licheniformis*, and therefore the entry is erroneous.

For the database search setting "16S ribosomal RNA (Bacteria and Archaea)," BLAST identified only *S. pyogenes*, *S. agalactiae*, *S. loxodontisalivarius*, and *S. saliviloxodontae*. The highest metrics were for the *S. pyogenes* strains I-273 and JCM5674; 117, 22 %, 4e-26, and 96 % for total score, query coverage, E-value, and identity, respectively. The values for the other two species were considerably lower. These results suggest that the V2 sequence, despite any doubts placed on it by the pyrosequencing software (yellow, "check;" red, "failed"), was sufficient to precisely identify *S. pyogenes*.

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Streptococcus pyogenes strain I-273 16S ribosomal RNA gene, partial sequence
Sequence ID: NR_028598.1 Length: 1393 Number of Matches: 1
Related Information
Range 1: 396 to 466 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
117 bits(63)	4e-26()	71/74(96%)	3/74(4%)	Plus/Minus	

Features:

```
Query  161  GTCCCTTTCTGGTTAGTTACCGTCACTTGGTGGATTTTCCACTCCCACCCATCATTCTTT  220
            |||||||||||||||||||||||||||||||||||||||||||||||||| |||||||||
Sbjct  466  GTCCCTTTCTGGTTAGTTACCGTCACTTGGTGGATTTTCCACTCCCACC-ATCATTCTT-  409

Query  221  CTCTAACAAACAGA  234
            ||||||||| ||||
Sbjct  408  CTCTAACAA-CAGA  396
```

31. *Streptococcus sanguinis* strain SK36

a. V1 amplicon sequence

TCATCCAAGA AGAGCAAGCT CCTCTCTTCA GCGTTCTACT TGCATGTATT AGGCACGCCG CCAGC[C]G

b. V2 amplicon sequence

GTCCCTT[T]CT GGTAAGATAC CGTCACAGTG TGAACTTTCC ACTCTCACAC CCGTTCTTCT [T]C

c. V3 amplicon sequence

GTCACCTCTG TCCCGAAGGA AAA[A]CATCTA TCTCTAGAGC GGT CAGAAGG GAATGGTTCA AAGAACCCTG G

d. **BLAST Analysis Results and Discussion.** BLAST excluded the V3 sequence from alignments, but used the essentially the entire V1+V2 sequences. The search analysis performed under the setting "nr/nt" produced a table that listed four strains of *Streptococcus thermophiles* at the highest metrics of 1343, 58 %, 1e-23, and 100 % for total score, query coverage, E-value, and identity, respectively. All four had identical alignments. Indeed, many alignments with lower total scores but the same other values were still identical to the alignments for those that had total score of 1343. Under this search, none of the bacteria listed were *S. sanguinis*.

With the search setting "16S ribosomal RNA (Bacteria and Archaea)," BLAST found different species and strains of *Streptococcus*, and of these the first two listed were *S. sanguinis* strains SK1 and JCM 5708. Both of these had the same metrics; 233, 59 %, 2e-27, and 100 % for total score, query coverage, E-value, and identity, respectively. But with progressively lower values for these parameters, the sequence divergences emerged. These results suggest that *S. sanguinis* identification was reliable, which is unlike the results generated when the setting was "nr/nt."

The alignments shown below are based on the settings "Highly similar sequences (megablast)" and "16S ribosomal RNA (Bacteria and Archaea)."

Streptococcus sanguinis SK1 16S ribosomal RNA gene, partial sequence
Sequence ID: NR_024841.1 Length: 1460 Number of Matches: 2
Related Information
Range 1: 22 to 86 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #1

Score	Expect	Identities	Gaps	Strand	Frame
121 bits(65)	2e-27()	65/65(100%)	0/65(0%)	Plus/Minus	

Features:

```
Query  1   TCATCCAAGAAGAGCAAGCTCCTCTCTTCAGCGTTCTACTTGCATGTATTAGGCACGCCG  60
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  86  TCATCCAAGAAGAGCAAGCTCCTCTCTTCAGCGTTCTACTTGCATGTATTAGGCACGCCG  27

Query  61  CCAGC  65
           |||||
Sbjct  26  CCAGC  22
```

Range 2: 439 to 498 GenBankGraphics Next Match Previous Match First Match

Alignment statistics for match #2

Score	Expect	Identities	Gaps	Strand	Frame
111 bits(60)	1e-24()	60/60(100%)	0/60(0%)	Plus/Minus	

Features:

```
Query  73   GTCCCTTTCTGGTAAGATACCGTCACAGTGTGAACTTTCCACTCTCACACCCGTTCTTCT  132
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  498  GTCCCTTTCTGGTAAGATACCGTCACAGTGTGAACTTTCCACTCTCACACCCGTTCTTCT  439
```

SUMMARY

1. Because sepsis and the resulting high mortality rate (up to 40 %) is a serious problem both in the general population and the military (Singer *et al*., 2016; Ma *et al*., 2016; Johnston *et al*., 2013), timely identification of the sepsis causing agents is essential for proper treatment. The overall aim of this work was to assess whether the combined approach of amplifying certain hypervariable region segments of the bacterial 16S rRNA gene followed by sequencing by PyroMark Q24 Advanced could rapidly and accurately identify certain sepsis bacteria.

2. The number of bacteria to be tested and listed in the proposal is 63. Of these, 31 were tested as a batch. Because the results with these were not definitive, the rest of the bacteria were not tested.

3. The work reported here was for the first phase of the proposed work; namely, to amplify the target fragments, sequence them, and then do BLAST analysis to find the identity targets.

4. Phase two of the proposed work aimed to develop protocols for multiple or mixed population amplicon preparations and sequencing. Given that the work with 31 bacteria did not yield uniform, reliable results, this phase was not attempted.

5. Another approach was that more than one amplicon could be sequenced at the same time. This mostly failed; when two or more amplicons were combined for pyrosequencing, the process typically happened for only one of them (results not shown).

6. Genomic DNA of two bacteria, *Streptococcus pyogenes* and *Shigella flexneri*, was used in the pilot studies to initially evaluate the capacity of V1, V2, and V3 sets of primers to amplify the corresponding target segments of the 16S rRNA gene. Whereas the BioMatrix kit primers resulted in unintended bands of unknown nature in the NTC, the same primers synthesize by Life Technologies gave clean results; that is, there were no unintended amplicons (Figs. 1, 2, 3 and 4). This is an important finding of the study because a central part of the overall goal of the proposed work was to evaluate the utility of BioMatrix primers coupled with pyrosequencing in rapid identification of sepsis bacteria. However, this finding does not imply that the BioMatrix primers *per se* have any intrinsic flaw; after all, the same primers synthesized by Life Technologies worked very well, giving no unexpected bands. The most likely cause of the unexpected bands is contamination of the BioMatrix primers with some DNA template. It appears that the source of such contamination was at the Company's premises because several different pouches of the sets of primers gave the same unintended bands.

7. Although the expected V1, V2, and V3 amplicons were readily obtained by PCR, the pyrosequencing with PyroMark 24 suffered from two problems: **One**, the sequences generated were much shorter than the size of the respective amplicons. **Two**, the sequences had varying degrees of errors.

8. As mentioned in the Results section (explanatory note 2), the entire V1, V2, and V3 sequences for each bacterium were used for BLAST analysis, and the software color assignments, though noted, were not used as the metrics for sequence reliability. The reason is that such calls were not uniformly reliable. A clear example of this is the set of sequences for *Shigella flexneri* strain 24570. For example, the software placed yellow or red calls on all of the V1 sequence, and yet BLAST aligned all of this sequence with the target organism sequences, except the last C (see item 18 for details).

Mycobacterium tuberculosis strain X004439 is another such example; all three sequences were judged yellow or red by the software; however, BLAST found 100 % target identity for all three of the sequences (listed #14). *Haemophilus influenzae* strain Rd sequences were yet another example of this pattern of yellow and red color assignment unreliability.

The blue ("pass") didn't always prove reliable. For example, *Acinetobacter baumannii* strain AYE V1 sequence was mostly blue ("pass"; black in this report) or yellow ("check"). The "nr/nt" BLAST aligned this sequence only with one *A. baumannii* strain, while the 16S restricted search altogether excluded it.

9. The main problems were that the sequences generated by pyrosequencing were not sufficiently long and reliable to afford identification with pinpoint accuracy. However, we note that sufficiently long and accurately determined sequences of the hypervariable regions of the 16S rRNA gene afford precise identification of the bacteria, even the strains.

It should be noted, however, that the results reported in this study by no means suggest that the hypervariable region sequences of the bacterial 16S rRNA gene cannot be employed to accurately identify the target bacteria. Indeed the approach is a proven one. Clearly, here the identification problems resulted because the sequences generated by pyrosequencing were too short and many had errors.

10. The overall conclusion of this study is that the combined PCR and PyroMark Q24 Advanced pyrosequencing to accurately identify the sepsis bacteria did not work uniformly. In most cases it gave confusing results which is not a good fit with a general clinical laboratory.

REFERENCES

1. Singer M *et al*. The Third International Consensus Definitions for Sepsis and Septic Shock. *JAMA* 2016; 315(8): 801-810.
2. Gaieski DF *et al*. Benchmarking the incidence and mortality of severe sepsis in the United States. *Crit Care Med* 2013; 41(5): 1167-1174.
3. Peterson LKN and Chase K. Pitfalls in the treatment of sepsis. *Emerg Med Clin N Am* 2017; 35(1): 185-198.
4. Torio CM and Moore BJ. National inpatient hospital costs: the most expensive conditions by payer, 2013: statistical brief #204. Healthcare Cost and Utilization Project (HCUP) Statistical Briefs [Internet]. Rockville (MD): Agency for Healthcare Research and Quality (US); 2016.
5. Ma XY *et al*. Early prevention of trauma-related infection/sepsis. *Mil Med Res* 2016; 3(33).
6. Motoshima M *et al*. Identification of bacteria directly from positive blood culture samples by DNA pyrosequencing of the 16S rRNA gene. *J Med Microbiol* 2012; (61): 1556-1562.
7. Chikamatsu K *et al*. Evaluation of PyroMark Q24 pyrosequencing as a method for the identification of mycobacteria. *Diagn Microbiol Infect Dis* 2018; (90): 35-39.
8. https://blast.ncbi.nlm.nih.gov/Blast.cgi
9. Johnston AM *et al*. Sepsis management in the deployed field hospital. *J R Army Med Corps* 2013; 159(3): 175-180.