Contents lists available at ScienceDirect

Journal of Microbiological Methods

journal homepage: www.elsevier.com/locate/jmicmeth

Rapid infectious disease identification by next-generation DNA sequencing

Jeremy E. Ellis, Dara S. Missan, Matthew Shabilla, Delyn Martinez, Stephen E. Fry *

Fry Laboratories, L.L.C., 15720 N. Greenway-Hayden Loop STE 3, Scottsdale, AZ 85260, United States

ARTICLE INFO

Article history:
Received 30 March 2016
Received in revised form 26 July 2016
Accepted 16 September 2016
Available online xxxx

Keywords:
Next generation DNA sequencing
Community profiling
Clinical NGS
NGS validation
Rapid infectious disease identification

ABSTRACT

Currently, there is a critical need to rapidly identify infectious organisms in clinical samples. Next-Generation Sequencing (NGS) could surmount the deficiencies of culture-based methods; however, there are no standardized, automated programs to process NGS data. To address this deficiency, we developed the Rapid Infectious Disease Identification (RIDI™) system. The system requires minimal guidance, which reduces operator errors. The system is compatible with the three major NGS platforms. It automatically interfaces with the sequencing system, detects their data format, configures the analysis type, applies appropriate quality control, and analyzes the results. Sequence information is characterized using both the NCBI database and RIDI™ specific databases. RIDI™ was designed to identify high probability sequence matches and more divergent matches that could represent different or novel species. We challenged the system using defined American Type Culture Collection (ATCC) reference standards of 27 species, both individually and in varying combinations. The system was able to rapidly detect known organisms in <12 h with multi-sample throughput. The system accurately identifies 99.5% of the DNA sequence reads at the genus-level and 75.3% at the species-level in reference standards. It has a limit of detection of 146 cells/ml in simulated clinical samples, and is also able to identify the components of polymicrobial samples with 16.9% discrepancy at the genus-level and 31.2% at the species-level. Thus, the system's effectiveness may exceed current methods, especially in situations where culture methods could produce false negatives or where rapid results would influence patient outcomes.

1. Introduction

Rapid identification of pathogenic organisms is a critical need in clinical settings. For patients with infections, such as septicemia, survival decreases hourly. Thus, the current standard of care is to administer broad-spectrum antibiotics until the infection is identified and then switching to more specific treatments (Faria et al., 2015; Perez et al., 2013). While this is an effective course of action for antibiotic susceptible bacteria, it is inadequate for antibiotic-resistant infections. In addition, this treatment regimen prolongs hospital stays and fosters antibiotic resistance. One study estimated that septicemia patients generated greater than US $40,000 per patient in direct hospital costs due to prolonged length of stay, and on average, early directed interventions reduced hospital costs to US $19,547 (Perez et al., 2013). Therefore, rapid pathogen identification would decrease both patient mortality and healthcare costs.

Currently, the gold standard for identifying bacteria is culturing methods. However, this method requires several days for positive identification of rapid-growing bacteria and even longer for fastidious or slow-growing organisms (Didelot et al., 2012). The positive predictive value of blood cultures can be constrained between 30% to >95% even when performed correctly (Afshari et al., 2012). Given the time-consuming nature and degree of variability inherent to culture methods, developing molecular methods for pathogen identification would be highly beneficial.

One of the most consistent molecular methods for identifying bacterial species is through next-generation sequencing (NGS) of the 16S rDNA gene sequencing. The 16S gene sequence organization is highly conserved in bacteria, which allows investigators to use universal primers to amplify the gene for downstream studies. The 16S gene, consisting of nine variable regions, may be effectively used to identify the bacterium via sequencing. By specifically targeting the variable regions more meaningful and enriched sequencing results per sample is obtained, thus decreasing the sequencing effort required for significant results and increasing the amount of samples that may be analyzed per instrument run (Claesson et al., 2010a). Previous efforts using NGS to target and identify organisms by variable region sequencing have not yielded reliable species-level identification due to target selection, technical limitations, and analysis methods (Junemann et al., 2012). As technology has advanced, NGS is becoming more accessible in a clinical

* Corresponding author.
E-mail address: medicaldirector@frylabs.com (S.E. Fry).

http://dx.doi.org/10.1016/j.mimet.2016.09.012
0167-7012/© 2016 Elsevier B.V. All rights reserved.

setting. Manufacturers have produced several relatively inexpensive benchtop sequencing models that have a low cost per sample, making NGS a feasible option for many hospitals. However, current methods require highly experienced personnel to both run the instrumentation and bioinformaticians to properly process the data, since there is currently no industry standard method for handling this information (Gullapalli et al., 2012). Successful implementation of NGS in a clinical setting requires a standardized data analysis pipeline that integrates across numerous NGS platforms (Table S1) which can automatically process and interpret the data and present the results in an easy-to-read format.

Given the industry requirements for a reproducible, automated data analysis pipeline for clinical NGS data, we developed the Rapid Infectious Disease Identification (RIDI™) system. The informatics pipeline was designed to become part of an NGS-based method for rapid clinical identification of bacteria. We strategically designed the software to meet a number of criteria for clinical implementation and experimentally challenged the system to determine its suitability. Overall, we developed an analysis pipeline that may be integrated into a clinical laboratory setting that is capable of highly accurate automated identification at the genus-level, acceptable species-level identification, and yields easy-to-read, actionable reports for use by clinicians.

2. Materials and methods

2.1. Informatics design and implementation

The RIDI™ system was designed to execute with minimal operator guidance, similar to other clinical laboratory systems, and to minimize errors during DNA sequence analysis. The software, which is compatible with data produced by all major benchtop sequencing systems, automatically and seamlessly interfaces with the Torrent Server (ThermoFisher Scientific), detects the data format, determines which analysis method to perform, and compiles the results into a clinician-friendly output. The RIDI™ informatics strategy is designed to identify both high probability sequence identification matches and more divergent sequence matches based on a percent identity threshold. This allows the system to differentiate between related species and detect novel microorganism species.

The sequence information generated from the patient sample is characterized using a combination of the publicly available NCBI databases (“nt” and “16SMicrobial”) and a RIDI™-specific database. The RIDI™ database consists of four distinct parts. The first part consists of NCBI database “correction” information in which incorrect results have been adjusted for accuracy. The second part includes information to filter the data and sort informative and uninformative results for the subsequent identification pipeline. The third component includes primer permutation rational expressions used for searching, classifying, and trimming the raw DNA reads, while the final portion classifies contaminating non-bacterial results to be eliminated from the analysis pipeline.

RIDI™ software performs all data processing steps, including sequencing filtering, trimming, and editing. The bioinformatics workflow for quality control is as follows: (Faria et al., 2015) The system automatically detects the input sequence type (Illumina®, LifeTechnologies™, or PacBio®) and configures minimum quality threshold values used in sequence processing. (Perez et al., 2013) The system removes all sequences shorter than 100 bp. (Didelot et al., 2012) The primer sequences are removed from both the leading and trailing ends of the sequences. To achieve this, the algorithm evaluates the first 30 bp and last half of the sequence read for leading primer sequences within a single mutation distance. If full primer sequences are not detected, the software scans for fragments of the primer sequence to identify trimming boundaries. If the primer fragments are unable to be detected, the leading 30 bps of the raw sequence is conservatively removed to prevent potential cryptic primer-induced homology from being incorporated into the sequence. (Afshari et al., 2012) The leading and trailing 5 base pair (bp) windows are evaluated to determine the average Phred quality based on a system specific threshold values. If the average Phred score of the 5 bp window is not above the established threshold, then a single base is trimmed and average end Phred score is recalculated. This process continues iteratively until the average Phred quality meets the appropriate threshold value for the input sequencing system. (Claesson et al., 2010a) All sequences shorter than 75 bp after all trimming steps are discarded.

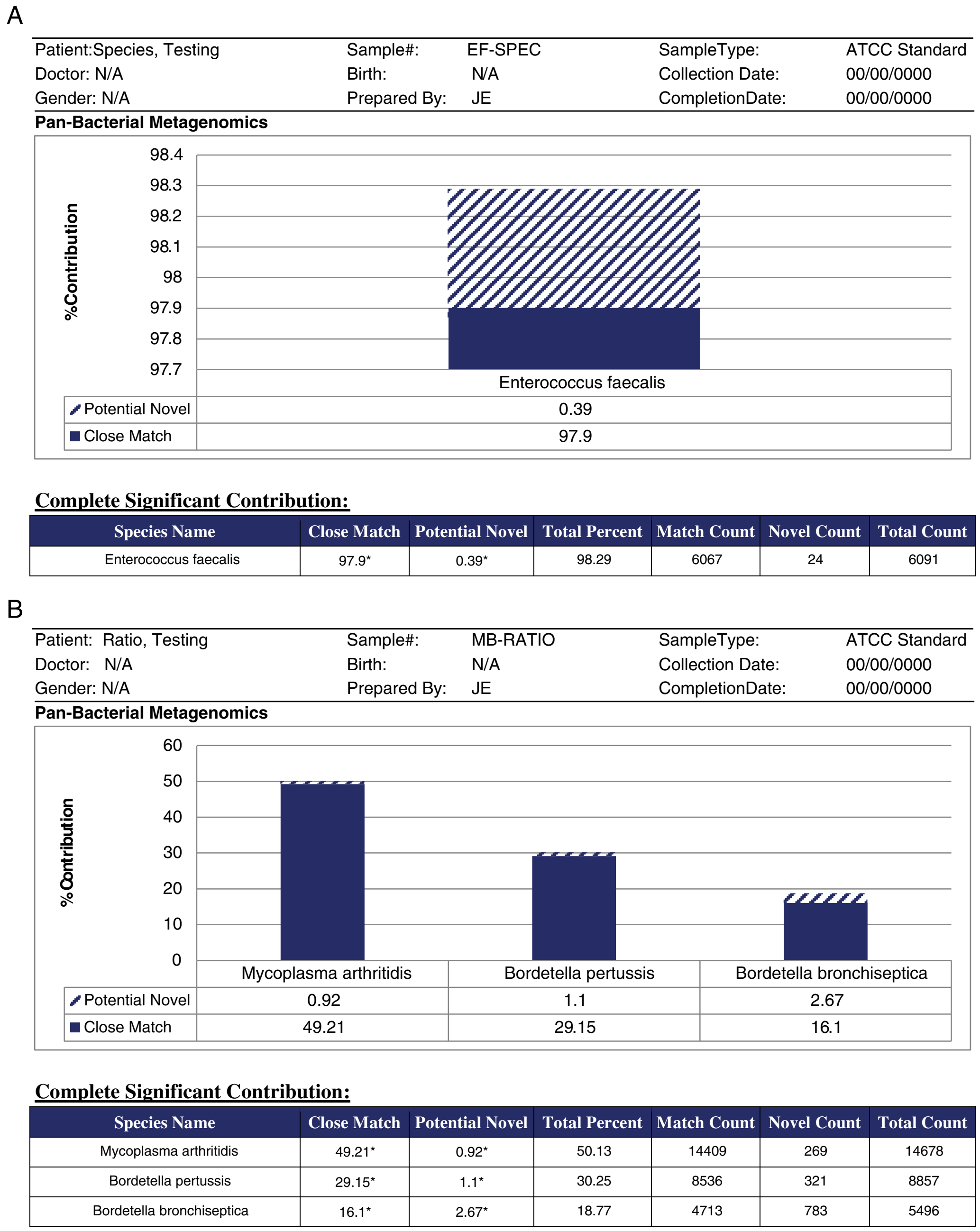
Once a set of trimmed reads has been obtained, the sequences are identified using a unique reductive BLAST strategy. This method uses a large initial BLAST window to quickly identify identical matches and remove them from the identification portion of the pipeline. The remaining sequences are queried with successively shorter BLAST windows (i.e., 65 bp, 40 bp, 25 bp, and 10 bp) for identification. Based on optimization trials (data not shown), a 65 bp BLAST search window results in sequence identification for 90.5% of the submitted sequences, with the subsequent 40, 25, and 10 bp search windows identifying an additional 8.2%, 1.2%, and the remaining 0.1%, respectively. Eliminating high homology sequences from small window BLAST searches reduces sequence identification time significantly, from hours to minutes. The software removes redundant sequences prior to the identification steps to reduce unnecessary processing and then adds the results back during results tabulation. Minimum sequences required to enable reliable organism calls was derived by subsampling positive result pools and deriving a predictive relationship between total BLAST results and the subsequent analyzed portion of sequences. Sequence matches with <97% identity are classified as novel reads unless representatives of the same indicated species are perfect 100% matches, then the threshold is lowered to 95% for that organism to account for sequencer induced variation. During the tabulation stage, the sum of reads for each species is tabulated and represented as a percentage to represent the distribution of organisms in the sample detected by sequencing. These results are then formatted into a convenient, easy-to-read report (Fig. 1). The output report consists of a bar graph and table representing the percent of close identity database-matched reads and potential novel reads (more divergent identity) as well as the percent contribution of each bacterial species out of the total sequence reads. As the software analysis progresses, all of the step-wise analysis variables, result lists, configurations, and decisions for each barcode set are archived to enable a full process audit. Upon completion of an analysis run a quality summary report is generated that may be reviewed by clinical laboratory staff for final approval and resulting. The final clinician friendly report enables easy recognition of potential bacterial pathogens, relative contribution ratios, and potential novel organisms at a glance. This report format intentionally differs from phylogenetic plots, representations of Operational Taxonomic Units (OTU), or principal component analysis plots commonly utilized in scientific research. In a clinical setting, a digital copy of the report could be transmitted via a Health Level-7 (HL7) interface to a variety of laboratory information systems (LIS).

2.2. Challenge strategy

The RIDI™ system is designed to work across multiple NGS platforms (Table S1). For the verification experiments, we used the IonTorrent™ Personal Genome Machine (PGM™). This platform was selected based on the sequencing time required to obtain the target sequence read lengths. Furthermore, the PGM™ is one of a select few of NGS systems that is commonly used by clinical laboratories (Gullapalli et al., 2012).

To verify that the RIDI™ system effectively operates as a clinical tool, we designed a series of challenges testing the system performance, width of taxonomic detection, the limit of detection, effectiveness handling polymicrobial infections, and analytical performance. We obtained 27 bacterial standards from American Type Culture Collection (ATCC, Manassas, VA). NCBI's naming conventions are used in all

Please cite this article as: Ellis, J.E., et al., Rapid infectious disease identification by next-generation DNA sequencing, J. Microbiol. Methods (2016), http://dx.doi.org/10.1016/j.mimet.2016.09.012



A

Patient:Species, Testing	Sample#:	EF-SPEC	SampleType:	ATCC Standard
Doctor: N/A	Birth:	N/A	Collection Date:	00/00/0000
Gender: N/A	Prepared By:	JE	CompletionDate:	00/00/0000

Pan-Bacterial Metagenomics

	Enterococcus faecalis
▨ Potential Novel	0.39
■ Close Match	97.9

Complete Significant Contribution:

Species Name	Close Match	Potential Novel	Total Percent	Match Count	Novel Count	Total Count
Enterococcus faecalis	97.9*	0.39*	98.29	6067	24	6091

B

Patient: Ratio, Testing	Sample#:	MB-RATIO	SampleType:	ATCC Standard
Doctor: N/A	Birth:	N/A	Collection Date:	00/00/0000
Gender: N/A	Prepared By:	JE	CompletionDate:	00/00/0000

Pan-Bacterial Metagenomics

	Mycoplasma arthritidis	Bordetella pertussis	Bordetella bronchiseptica
▨ Potential Novel	0.92	1.1	2.67
■ Close Match	49.21	29.15	16.1

Complete Significant Contribution:

Species Name	Close Match	Potential Novel	Total Percent	Match Count	Novel Count	Total Count
Mycoplasma arthritidis	49.21*	0.92*	50.13	14409	269	14678
Bordetella pertussis	29.15*	1.1*	30.25	8536	321	8857
Bordetella bronchiseptica	16.1*	2.67*	18.77	4713	783	5496

Fig. 1. Reports generated by the RIDI™ system. (A) Single species report for *Enterococcus faecalis* demonstrating 97.90% of read identified as close matches, ≥97% identity, and 0.39% of reads displaying <97% identity. (B) Multispecies report for *Mycoplasma arthritidis* and *Bordetella pertussis*. Note the sum of the *Bordetella* species reflects the 50:50 input ratio.

cases, except when clinically accepted nomenclature differs from convention. In those cases, clinical nomenclature took precedence to improve clinical utility. A relevant example is our preferred medical nomenclature of *Clostridium difficile* over the proposed scientific classification of *Peptoclostridium difficile* (Yutin and Galperin, 2013). Additionally, utilization of these methods in a clinical setting would require parallel testing of control samples along with patient samples. It has been suggested that an approach demonstrating validation for each run be adopted with the inclusion of various controls. These controls include a genomic DNA extraction control, an amplification control, an inhibitor control, and contamination controls be implemented (Petti et al., 2008). Due to the multi-barcode capabilities of the assay these controls would be included and rotated throughout the various barcode positions for each run.

To assess the breadth of taxonomic detection, DNA samples from each bacterial species was evaluated independently. Polymicrobial conditions were tested with different genomic ratios of DNA from various species mixed prior to sequencing. To test the limit of detection, we performed 10-fold serial dilutions of blood spiked with live cultured *E. coli* cells (concentrations ranging from 2 to 20,000 cells/ml). From these dilutions total DNA was extracted using the QIAamp DSP DNA Blood Mini Kit (Qiagen, Hilden, Germany) according to the manufacturer's instructions. It is important to note that human DNA was not removed from the sample. Undiluted genomic DNA, keeping the putative amplification targets at maximal concentration, with concentrations, ranging from 1 ng/μl to 200 ng/μl, was used as PCR template to generate barcoded amplicons for sequencing. Amplification inhibition by concentrated DNA templates are controlled for by using an independent 18S PCR amplification which ensures no inhibitors are present. Analytical performance near the limit of detection was assessed by preparing positive extraction control DNA samples consisting of blood spiked with cultured *E. coli* (at 150 cells/ml) and unspiked blood samples. The subsequent samples were sequenced and analyzed by RIDI™.

For sequence library construction, an amplicon-based strategy was used to target two conserved 16S regions, variable regions 1 to 2 and 4 to 5, with low cycle PCR. Previous studies have shown that single "universal" primer pairs may limit the practical taxonomic universality of the subsequent assay (Kommedal et al., 2011). By targeting these two regions using a multiloci strategy, we achieve reliable detection across all bacterial species, while preserving the relative ratios of sequenced organisms in polymicrobial samples. Multiple sets of fusion PCR primer pairs were designed for species identification that included the primer target, IonTorrent™ adapter, linker, and barcode sequences (Table 1). The selected regions were chosen based on extensive experimentation whereby variable regions 3 to 2 and 3 to 4 failing to yield significant, reliable, and taxonomically informative sequence data (data not shown). The first primer set targets variable region 1 extending into variable region 2 in the 16S gene, termed Region 1. The second primer set targets variable region 5 and extends into variable region 4, termed Region 2. These regions are both compatible with sequencing and yield informatics resolution to assign genera to a very high degree and speciation at a clinically acceptable level, superior to single region-based methods. Recent independent studies have confirmed the finding that variable region-spanning amplicons improves identification resolution (Walters et al., 2016). The region spanning approach coupled with a strategy devoid of Operational Taxonomic Units (OTUs) reliably achieves species-level identification compared to previous attempts (Junemann et al., 2012).

A normalized ion sphere particle (ISP) library was generated by pooling equimolar amounts of the individually prepared amplicons, measured by the Qubit dsDNA HS Assay Kit (Life Technologies, Corp, Grand Island, NY), with either the Ion PGM™ Template OT2 400 Kit or the Hi-Q™ OT2 Kit (Life Technologies Corp). Amplicon matched negative amplification controls ensure that false positive amplicons do not occur by contaminated reagents during the amplification process. The quality of the ISP library was assessed using the Ion Sphere Quality Control Kit and the Qubit™ 2.0 Fluorometer (Life Technologies Corp). For the sequencing run, we used the Ion 314 Chip Kit v2 (Life Technologies Corp) according to the manufacturer's instructions with a few modifications using either the Ion PGM™ Sequencing 400 Kit or the Hi-Q™ Sequencing Kit (Life Technologies Corp). The total flow cycle number was increased to 900 cycles to promote read lengths longer than 400 bp. Previous research has shown that longer reads and associated read chemistry consistently produce higher quality sequences (Whiteley et al., 2012). In addition, the entire ISP library was loaded onto the chip using repeated loading steps to improve overall chip coverage and utilization. Even though test fragments are used as sequencing quality control measures of both the sequencing and ISP library steps, we have added a true analytical positive control as a more rigorous quality control metric.

3. Results

3.1. System performance

In summary, the PGM™ system is amenable to RIDI™ integration. Known organism standards were rapidly detected in <12 h, and the system was capable of multi-sample throughput. The RIDI™ software processes between 2000 and 8000 sequences/min depending on the organism and sample quality, averaging 5000 sequences/min. The informatics pipeline completed the analysis of standard barcoded sequencing runs between 10 and 30 min. On average, tens of thousands of sequences are yielded per barcode with several thousands in positive samples.

3.2. Width of taxonomic detection

We challenged the system using ATCC DNA reference standards of 27 species of interest or clinical relevance, both individually and in various combinations. The RIDI™ system was able to correctly match the individual reference standards to their appropriate identity with 99.52% ± 1.72% of the reads correctly identified at the genus-level and 75.29% ± 39.54% of the reads correctly identified at the species-level (Table 2). The reduction of the species identification rate compared to the genus identification results is primarily due to reduced species-level detection of a select few standards (Table 2, bolded), such as *Bartonella henselae* (0.09% identification) and *Borrelia burgdorferi* (0.73% identification); 19 of the 27 standards had species-level identification >90% of the obtained reads, while 26 of the 27 standards had genus-level identification >95% of the obtained sequences. Notably, these infrequent cases of species-level misidentifications are not a result of the amplicon-based strategy or the target regions, but are due to the presence of unresolved ambiguous identification calls whereby the software arbitrarily selects the alphabetically last result. Future iterations of the software analysis pipeline may directly address arbitrarily resolved ambiguous identification and resolve ambiguity in a quantitative manner. Ultimately these results support a universally broad range of taxonomic detection that enables useful taxonomic identification. No significant cross-barcode leakage or sequencing efficiency differences were observed while rotating positive standards through 12 total barcodes as anticipated.

3.3. Limit of detection (LOD)

The LOD was defined as the lowest concentration that provides a 95% detection rate and was determined using a dilution series of *E. coli*-spiked blood samples. The sequencing efficiency (defined as the number of bacterial reads over the total number of reads) varied greatly between the different replicates; however, DNA sequence reads for *Escherichia* genus, including *Shigella* species (Fukushima et al., 2002; Lan and Reeves, 2002), was detected in all samples; however, it was called by RIDI™ less frequently with decreasing cell concentration.

Table 1
Fusion primer design.

Region	Direction	Adapter (A for forward, P1 for reverse)	Barcode	Spacer	Target annealing	Target reference
V1-V2	V1/2-F	CCATCTCATCCCTGCGTGTCTCCGACTCAG	(N)*10	GAT	AGAGTTTGATCCTGGCTCAG	(Weisburg et al., 1991)
	V1/2-R	CCTCTCTATGGGCAGTCGGTGAT			CTGCTGCCTYCCGTA	(McKenna et al., 2008)
V3-V2	V3/2-F	CCATCTCATCCCTGCGTGTCTCCGACTCAG	(N)*10	GAT	ATTACCGCGGCTGCTGG	(Dethlefsen et al., 2008)
	V3/2-R	CCTCTCTATGGGCAGTCGGTGAT			AGYGGCGNACGGGTGAGTAA	(Sundquist et al., 2007)
V3-V4	V3/4-F	CCATCTCATCCCTGCGTGTCTCCGACTCAG	(N)*10	GAT	ACTCCTACGGRAGGCAGCAG	(Dethlefsen et al., 2008)
	V3/4-R	CCTCTCTATGGGCAGTCGGTGAT			TACNVGGGTATCTAATCC	(Claesson et al., 2010b)
V5-V4	V5/4-F	CCATCTCATCCCTGCGTGTCTCCGACTCAG	(N)*10	GAT	CCGTCAATTYYTTTRAGTTT	(Claesson et al., 2010b)
	V5/4-R	CCTCTCTATGGGCAGTCGGTGAT			AYTGGGYDTAAAGNG	(Claesson et al., 2010b)

Table 2
Genus and species-level identification rates of standard replicates.

Organisms tested		Genus-level identification			Species-level identification		
Species name	Replicates	Total %	R1 ID %[a]	R2 ID %[b]	Total %	R1 ID %[a]	R2 ID %[b]
Ralstonia solanacearum	41	99.80%	19.32%	80.61%	18.79%	99.91%	**0.04%**
Enterococcus faecalis	33	99.65%	79.36%	20.58%	94.76%	81.58%	18.36%
Haemophilus influenzae	18	99.86%	93.26%	6.61%	99.70%	93.35%	6.51%
Clostridium difficile	12	99.89%	19.40%	80.57%	99.87%	19.41%	80.57%
Bacillus cereus	8	99.52%	32.66%	67.27%	76.18%	34.43%	65.50%
Acinetobacter baumannii	7	99.84%	25.87%	74.04%	99.47%	25.85%	74.07%
Gemella haemolysans	7	99.98%	34.61%	65.29%	99.40%	34.77%	65.13%
Klebsiella pneumoniae	7	98.74%	46.19%	53.76%	98.66%	46.21%	53.75%
Legionella pneumophila	7	99.94%	37.63%	62.37%	86.87%	37.68%	62.32%
Parabacteroides distasonis	5	99.77%	12.16%	87.82%	99.55%	12.02%	87.95%
Bartonella henselae	4	99.97%	72.90%	26.84%	**0.09%**	100.00%	**0.00%**
Escherichia coli	4	96.65%	37.55%	62.27%	96.25%	37.23%	62.59%
Helicobacter pylori	4	99.95%	35.43%	64.57%	99.80%	57.08%	42.90%
Bacteroides vulgatus	3	99.97%	21.37%	78.59%	99.48%	21.45%	78.51%
Bordetella pertussis	3	99.55%	3.80%	96.20%	95.14%	**0.05%**	99.95%
Streptococcus sanguinis	3	99.98%	5.07%	94.88%	99.64%	5.09%	94.86%
Bartonella bacilliformis	2	99.99%	99.91%	**0.00%**	99.86%	99.91%	**0.00%**
Borrelia burgdorferi	2	99.89%	**0.72%**	99.26%	**0.73%**	77.78%	16.67%
Capnocytophaga gingivalis	2	100.00%	**0.03%**	99.94%	99.95%	**0.00%**	99.97%
Coxiella burnetii	2	87.11%	**0.76%**	99.05%	87.11%	**0.76%**	99.05%
Mycoplasma fermentans	2	99.86%	96.17%	3.79%	99.79%	96.21%	3.75%
Mycoplasma pneumoniae	2	96.97%	30.21%	69.62%	96.77%	30.15%	69.68%
Acholeplasma laidlawii	1	100.00%	69.62%	30.33%	99.48%	69.98%	29.98%
Akkermansia muciniphila	1	99.86%	71.84%	27.94%	99.86%	71.84%	27.94%
Bifidobacterium breve	1	99.78%	64.74%	34.97%	64.82%	99.56%	**0.00%**
Blautia producta	1	99.38%	20.20%	79.63%	89.92%	11.80%	88.00%
Mycoplasma arthritidis	1	99.91%	**0.39%**	99.59%	99.79%	**0.39%**	99.59%

[a] Region 1 percent identification, 16S rRNA variable region 1 spanning into variable region 2.
[b] Region 2 percent identification, 16S rRNA variable region 5 spanning into variable region 4.

Based on Probit analysis, the LOD for *E. coli* using our DNA extraction method was approximately 146 cells/ml (Table 3).

3.4. Polymicrobial profiling

We then tested the system handling of polymicrobial samples. First, we analyzed *Bordetella pertussis* combined with *Enterococcus faecalis*, *Bartonella bacilliformis*, *Mycoplasma arthritidis*, or *B. burgdorferi* in varying genomic ratios (Table 4). Underperformance was observed for some amplicons further illustrating the need for a multiloci strategy (Table 4, bolded). Some combinations, such as those with *M. arthritidis* and *B. burgdorferi*, had very small genus-level discrepancies (0.38%–3.06%), whereas others, such as *E. faecalis*, were more significant (19.52%–54.30%). Notably, *E. faecalis* Region 1 detection performed poorly at lower genomic ratios, and the percent discrepancy increased as the amount of *E. faecalis* DNA decreased. Second, we analyzed organisms from the same genus to assess possible chimera or cross-genus identification (*Mycoplasma pneumoniae* with *M. arthritidis* and *B. bacilliformis* with *B. henselae*). Shared genus organisms present some difficulty in unambiguously assigning identity of the individual reads; therefore, the predominant species observed during genus-level identification was designated as the correct species calls for that set. All other species identified of that shared genus were designated to be the non-dominating species. The percent discrepancies for these experiments were relatively low, ranging from 4.58%–18.95%. Overall, the percent discrepancy between the actual genomic ratios and the ratios generated by the system was 16.88% ± 17.39% at the genus-level and 31.18% ± 12.33% at the species-level.

Table 3
Probit analysis of *E. coli* of spiked blood samples.

Cells/ml	% genus ID	Sequencing efficiency	log10	Probit	% detection
20,000	87.40%	33.79%	4.3	14.0	100.00%
20,000	81.49%	5.94%			
2000	86.18%	29.68%	3.3	10.6	100.00%
2000	78.00%	9.29%			
200	78.57%	27.08%	2.3	7.1	98.27%
200	0.00%	8.89%			
20	73.41%	25.32%	1.3	3.7	9.19%
20	0.00%	18.70%			
20	0.00%	9.04%			
2	0.00%	21.17%	0.3	0.2	0.00%
2	0.00%	9.71%			

3.5. Analytical performance

The performance of a clinical assay is frequently characterized by sensitivity, specificity, positive predictive value, and negative predictive value. These analytical performance metrics may be derived by comparing the expected outcome of an assay to the actual test results of the assay. A total of 187 control samples were analyzed, 33 *E. coli* spiked blood samples and 154 unspiked blood samples. All 33 spiked samples yielded a positive detection of the expected *Escherichia*/*Shigella* genus, while the 154 unspiked samples did not yield a positive bacterial identification call. These results support promising performance characteristics much greater than a clinically acceptable 95% sensitivity, specificity, positive predictive value, and negative predictive value. Additional replicates are required to resolve the relatively rare false positive and false negative events.

4. Conclusion

Currently, clinicians use culture techniques to identify bacterial pathogens; however, these methods are time consuming even for rapidly growing species and have multiple points of failure (Didelot et al., 2012). Given the recent decrease in NGS hardware costs and increases in throughput, identifying bacteria using sequencing methods represents a viable alternative. For this technique to be integrated successfully into clinical use there should be a user-friendly framework in place in which samples can be easily converted into an interpretable result without requiring extensive intervention or bioinformatics expertise. Thus, we aimed to implement software that integrates with existing NGS

Please cite this article as: Ellis, J.E., et al., Rapid infectious disease identification by next-generation DNA sequencing, J. Microbiol. Methods (2016), http://dx.doi.org/10.1016/j.mimet.2016.09.012

Table 4
Genus and species-level identification rates of mixed populations.

Organisms tested		Genus-level identification				Species-level identification			
Species name	Genomic ratio	Total %	% discr[a]	R1 ID %[b]	R2 ID %[c]	Total %	% discr[a]	R1 ID %[b]	R2 ID %[c]
Enterococcus faecalis	79.21%	98.61%	19.52%	75.21%	24.73%	96.04%	18.23%	77.24%	22.70%
Bordetella pertussis	20.79%	1.16%		**0.00%**	100.00%	1.16%		**0.00%**	100.00%
Enterococcus faecalis	29.74%	84.00%	54.30%	**0.00%**	99.71%	73.94%	49.27%	**0.00%**	99.67%
Bordetella pertussis	70.26%	15.93%		**0.00%**	100.00%	15.93%		**0.00%**	100.00%
Enterococcus faecalis	55.94%	94.59%	38.75%	38.40%	61.47%	87.27%	35.10%	41.61%	58.30%
Bordetella pertussis	44.06%	5.21%		**0.00%**	99.51%	5.18%		**0.00%**	99.51%
Enterococcus faecalis	55.94%	96.66%	40.76%	59.04%	40.91%	92.17%	38.54%	61.92%	38.03%
Bordetella pertussis	44.06%	3.25%		**0.00%**	100.00%	3.20%		**0.00%**	100.00%
Bartonella bacilliformis	50.00%	78.30%	28.55%	73.11%	25.94%	77.79%	31.51%	73.14%	25.91%
Bordetella pertussis	50.00%	21.20%		30.02%	69.76%	14.77%		**0.00%**	99.92%
Bordetella pertussis	50.00%	52.98%	3.06%	64.10%	35.88%	19.00%	17.13%	**0.56%**	99.44%
Mycoplasma arthritidis	50.00%	46.86%		46.10%	53.90%	46.75%		46.18%	53.82%
Bordetella pertussis	50.00%	49.53%	0.38%	38.62%	61.38%	30.25%	9.94%	**0.26%**	99.73%
Mycoplasma arthritidis	50.00%	50.28%		20.37%	79.61%	50.13%		20.42%	79.55%
Bordetella pertussis	50.00%	49.52%	2.94%	28.86%	70.77%	34.90%	32.52%	**0.08%**	99.47%
Borrelia burgdorferi	50.00%	44.60%		**0.19%**	99.78%	**0.07%**		40.00%	40.00%
Bordetella pertussis	95.67%	91.50%	3.83%	74.13%	25.84%	23.60%	38.18%	**0.58%**	99.36%
Borrelia burgdorferi	4.33%	7.83%		1.50%	98.25%	0.05%		87.50%	12.50%
Bordetella pertussis	95.67%	94.35%	1.01%	81.88%	18.09%	17.07%	41.43%	**0.67%**	99.28%
Borrelia burgdorferi	4.33%	5.03%		3.58%	96.04%	**0.08%**		87.50%	12.50%
Mycoplasma pneumoniae	25.09%	13.48%	11.61%	69.91%	30.05%	–	–	–	–
Mycoplasma arthritidis	74.91%	86.52%		56.42%	43.58%	–		–	–
Mycoplasma pneumoniae	50.12%	54.70%	4.58%	98.59%	1.41%	–	–	–	–
Mycoplasma arthritidis	49.88%	45.30%		78.67%	21.21%	–		–	–
Bartonella bacilliformis	79.78%	71.48%	8.11%	55.59%	44.34%	–	–	–	–
Bartonella henselae	20.22%	28.15%		95.11%	4.70%	–		–	–
Bartonella bacilliformis	56.80%	37.68%	18.95%	55.08%	44.92%	–	–	–	–
Bartonella henselae	43.20%	61.97%		94.99%	4.93%	–		–	–

[a] Percent discrepancy.
[b] Region 1% identification.
[c] Region 2% identification.

technologies to automatically process bacterial sequencing data and produce clear, understandable, and relevant reports.

When designing this system, we devised an a priori list of key features for our software to include for it to be considered successful in this regard. Thus, the RIDI™ system was designed with the following conditions and requirements in mind. To meet our requirements the system should (i) be "open" to permit additions and refinements of the reference databases, (ii) use industry-standard analysis methods in a traceable format, (iii) use commercially available NGS hardware with minimal modifications, (iv) permit additional modular improvements to adapt to technological advances, (v) be as automated as possible to minimize operator error, particularly with respect to the informatics, (vi) generate clear, concise, actionable, and comparable to current clinical laboratory reports, (vii) reliably generate relevant taxonomic assignments for the detected organisms, (viii) have a sample-to-result time less than the time of current culture-based methods, (ix) have clinically relevant detection sensitivity, (x) have the capability to process polymicrobial samples and reasonably represent relative contribution ratios. Requirements (i) through (vi) were integrated at the pipeline design stage. However, the remaining requirements had to be verified experimentally.

To address requirements (vii) and (viii), generating reliable taxonomic outputs and producing results in <24 h, respectively, we tested the system's ability to identify species from a diverse array of monoculture samples. Our system had >99% identification at the genus-level and 75% identification at the species-level. Future software improvements will include the removal of arbitrary result selection, thus improving the overall sequence identification rate. Future efforts will focus on resolving the few cases of ambiguity to reduce incorrect species calls resulting from arbitrary factors. Such improvements will be incorporated into future versions of the software. While there are minor inefficiencies with species-level classification, organism genera are not only accurately identified, but in the majority of cases, the primary result correctly identifies the target species. When compared with culture methods, our results are comparable or significantly surpass current methodologies (Guo et al., 2014; Matsuda et al., 2012), and we can identify organisms in <12 h regardless of the species and their growth rate in culture. From a practical standpoint, the achieved genus-level identification coupled with the reported species is sufficient to make informed treatment decisions in a clinical setting. For example, a study investigating the frequency of low-level bacteremia in children observed that genus-only-level bacterial identification, which this method exceeds, in 128 patients resulted in initiation of treatment in 5 patients and change of treatment in 83 patients. In addition, the antibiotic spectra was reduced in 59.4% of cases, resulting in a 53.9% cost reduction (Kellogg et al., 2000). Given the speed and accuracy of the RIDI™ system, patient outcome improvement and cost is expected to be correspondingly improved.

Next we tested the LOD of this method to address requirement (ix), that the system has clinically relevant detection levels. Based on *E. coli*-spiked blood samples, the LOD of the RIDI™ system was 146.2 cells/ml blood. There is some uncertainty regarding how to define clinically relevant infections in the literature. The clinical relevance of an infection varies by the age of the patients and the infecting species. For example, depending on the study, clinically relevant bacteremia has been defined as having a bacterial density <1 CFU/ml to >100 CFU/ml (Yagupsky and Nolte, 1990). In one study investigating *E. coli* septicemia in neonates, one-third of the patients had bacterial loads >1000 CFU/ml (Dietzman et al., 1978). Numerous studies have found bacteria counts in excess of 100 CFU/ml for encapsulated bacterial species (Yagupsky and Nolte, 1990). Another consideration when comparing these methods is the loss of organisms during culture methods. Bacterial counts can be skewed lower using culture methods due to loss of bacteria during plating, variability in media, less than optimal growing conditions, and the presence of non-viable cells (Didelot et al., 2012). In addition to the ambiguous definition of a clinically relevant infection, there is a mismatch in the output metrics by these clinical tests. Culture methods report findings as CFU/ml, whereas molecular techniques (PCR, NGS, etc.)

often report cells/ml or genomes/ml. Bacteria that exist as single cells in vivo, one CFU frequently is equivalent to a single viable cell. However, many bacteria naturally congregate into clusters of cells. For example, *S. aureus* naturally forms clusters of 5–20 cells (Haaber et al., 2012), and each cluster would represent a single CFU by culture methods. Thus, the cell count in a given sample may be significantly higher than a CFU count recovered for the same sample. One study calculated the detection rates of culture methods and estimated that anaerobic cultures would require 14.9 CFU/ml to achieve a 95% detection rate and aerobic cultures would require 45 CFU/ml (Kellogg et al., 2000). Depending on the cell to CFU conversion, the RIDI™ system approaches or exceeds the sensitivity of culture methods. While there should be a push toward increased sensitivity, this effort can and does increase false positive risk. In a study investigating the clinical significance of positive blood cultures in infants, the authors calculated a 40% false positive rate for culture methods (Sabui et al., 1999). Thus, at an average adult load of 30 CFU/ml, patients with sepsis may be indistinguishable from non-sepsis patients by culture methods. Even though we aim for the most sensitive assay possible using NGS methods, this goal must be tempered with minimizing false positives. Based on all of these factors the RIDI™ system fulfills our requirement for clinically relevant levels of infection.

Requirement (x) relates to the system's ability to handle polymicrobial samples and determine the relative ratios of the organisms in the sample. Overall, we obtained approximately 17% discrepancy at the genus-level and an acceptable 31% discrepancy at the species-level. Close inspection of the data reveals that the largest discrepancies result in samples that had identification failures in Region 1 and is most apparent in samples containing *E. faecalis* (Table 4). The percent discrepancy increased as the amount of *E. faecalis* contribution decreased. This is not a surprising result, because there is a well-known issue of primer affinity variation among different species (Kommedal et al., 2011), which can contribute to discrepancies seen in the relative ratios of bacterial species. This example also highlights the importance of using a multi-loci strategy when identifying bacteria using NGS. If a single-loci strategy was used exclusively, contribution by *E. faecalis* may have been irreparably over-estimated relative to the other contributing organisms. Typically, bacterial community distributions in samples are determined based on dilution series culturing techniques, which is highly imprecise and error prone (McKee et al., 1985; Babaahmady et al., 1998; Zhu et al., 2004). There is some precedent for 16S sequencing to determine bacterial abundance in blood samples; however, this technique varied highly depending on the sample processing method (Faria et al., 2015). Any improvement in bacteria population characterization in samples would be considered beneficial in a clinical setting.

One of the key features of the RIDI™ system is the generation of a final report (Fig. 1). The goal was to create an easy-to-read report similar to culture-based identification results, while also providing additional clinically relevant information unique to the system. This new information is encapsulated in the Close Match and Potential Novel report metrics. There is a non-trivial debate regarding the definition of a bacterial species, with suggested values ranging from 97%–99.5% identity to be considered a single species (Schlaberg et al., 2012). Provided that the RIDI™ system analyzes a relatively small portion of the bacterial genome, a portion of the 16S rRNA gene, and that the DNA is derived from clinical samples that may not sequence efficiently, we selected a relatively conservative definition for novelty. We defined a close match to be a sequence read with ≥97% identity with the reference sequence or ≥95% identity if a 100% match to that same organism was previously observed. If the sequence diverges by >3% or 5%, respectively, it is considered a potentially "novel" organism. Potentially, divergent, or unexpected organisms may impact treatment decisions in the clinic; therefore, providing this information early in the course of treatment is considered a valuable result. Unlike 16S sequencing-based assays, culture-based methods are not equipped to identify novel, unexpected, or atypical organisms. Thus, the clarity provided by the RIDI™ report offers unique and unbiased clinically relevant information that clinicians currently do not have routine access. For example, a physician may treat a potentially novel infection as they would a known related species but monitor the patient closely throughout therapy or widen the antibiotic coverage to ensure that the treatment regime is successful.

In conclusion, we have developed an automated bioinformatics software analysis system that easily integrates into existing benchtop NGS pipelines to perform rapid identification and detection of bacterial pathogens in clinical samples while presenting the results in an easy-to-read format. The RIDI™ system meets the initial development criteria and clinically acceptable performance parameters. The RIDI™ informatics pipeline has proven cross-platform compatibility with the Illumina®, IonTorrent™, and the PacBio® benchtop sequencer formats (Table S1). Ultimately, the system is expected to exceed current methods where culture may yield a false negative result or where time-to-results influences patient outcome. Adoption of rapid identification sequence-based methods as an additional diagnostics tool may be of great value to both clinical microbiologists and physicians. Future studies include a non-interventional comparative study to demonstrate efficacy when compared to existing culture-based methods in speed, genus/species typing, and potential treatment outcomes. Bioinformatics improvements are underway to improve species-level identification rates in the infrequent instances where the sequence results are ambiguous with the goal of exceeding a sequence-by-sequence species identification rate >90%.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.mimet.2016.09.012.

Acknowledgements

This study has been funded by Fry Laboratories, LLC.

References

Afshari, A., Schrenzel, J., Ieven, M., Harbarth, S., 2012. Bench-to-bedside review: Rapid molecular diagnostics for bloodstream infection–a new frontier? Crit. Care 16 (3), 222. http://dx.doi.org/10.1186/cc11202 (PubMed PMID: 22647543; PMCID: 3580598).

Babaahmady, K.G., Challacombe, S.J., Marsh, P.D., Newman, H.N., 1998. Ecological study of Streptococcus Mutans, Streptococcus Sobrinus and lactobacillus spp. at sub-sites from approximal dental plaque from children. Caries Res. 32 (1), 51–58 Epub 1998/01/23. PubMed PMID: 9438572.

Claesson, M.J., Wang, Q., O'Sullivan, O., Greene-Diniz, R., Cole, J.R., Ross, R.P., O'Toole, P.W., 2010a. Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. Nucleic Acids Res. 38 (22), e200. http://dx.doi.org/10.1093/nar/gkq873.

Claesson, M.J., Wang, Q., O'Sullivan, O., Greene-Diniz, R., Cole, J.R., Ross, R.P., O'Toole, P.W., 2010b. Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. Nucleic Acids Res. 38 (22), e200. http://dx.doi.org/10.1093/nar/gkq873 PubMed PMID: 20880993; PMCID: 3001100.

Dethlefsen, L., Huse, S., Sogin, M.L., Relman, D.A., 2008. The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. PLoS Biol. 6 (11), e280. http://dx.doi.org/10.1371/journal.pbio.0060280 PubMed PMID: 19018661; PMCID: 2586385.

Didelot, X., Bowden, R., Wilson, D.J., Peto, T.E.A., Crook, D.W., 2012. Transforming clinical microbiology with bacterial genome sequencing. Nat. Rev. Genet. 13 (9), 601–612.

Dietzman, D.E., Fischer, G.W., Schoenknecht, F.D., 1978. Neonatal *Escherichia coli* septicemia–bacterial counts in blood. J. Pediatr. 85 (1), 128–130. http://dx.doi.org/10.1016/S0022-3476(74)80308-2.

Faria, M., Conly, J.M., Surette, M.G., 2015. The development and application of a molecular community profiling strategy to identify polymicrobial bacterial DNA in the whole blood of septic patients. BMC Microbiol. 15 (1), 215. http://dx.doi.org/10.1186/s12866-015-0557-7 (Epub 2015/10/18. PubMed PMID: 26474751; PMCID: Pmc4609058).

Fukushima, M., Kakinuma, K., Kawaguchi, R., 2002. Phylogenetic analysis of Salmonella, Shigella, and *Escherichia coli* strains on the basis of the gyrB gene sequence. J. Clin. Microbiol. 40 (8), 2779–2785 PubMed PMID: 12149329; PMCID: 120687.

Gullapalli, R.R., Desai, K.V., Santana-Santos, L., Kant, J.A., Becich, M.J., 2012. Next generation sequencing in clinical medicine: challenges and lessons for pathology and biomedical informatics. Journal of Pathology Informatics 3, 40. http://dx.doi.org/10.4103/2153-3539.103013 PubMed PMID: 23248761; PMCID: 3519097.

Guo, L., Ye, L., Zhao, Q., Ma, Y., Yang, J., Luo, Y., 2014. Comparative study of MALDI-TOF MS and VITEK 2 in bacteria identification. Journal of Thoracic Disease 6 (5), 534–538. http://dx.doi.org/10.3978/j.issn.2072-1439.2014.02.18 PubMed PMID: PMC4015025.

Haaber, J., Cohn, M.T., Frees, D., Andersen, T.J., Ingmer, H., 2012. Planktonic aggregates of *Staphylococcus aureus* protect against common antibiotics. PLoS One 7 (7), e41075.

http://dx.doi.org/10.1371/journal.pone.0041075 Epub 2012/07/21. PubMed PMID: 22815921; PMCID: Pmc3399816.

Junemann, S., Prior, K., Szczepanowski, R., Harks, I., Ehmke, B., Goesmann, A., Stoye, J., Harmsen, D., 2012. Bacterial community shift in treated periodontitis patients revealed by ion torrent 16S rRNA gene amplicon sequencing. PLoS One 7 (8), e41606. http://dx.doi.org/10.1371/journal.pone.0041606 (PubMed PMID: 22870235; PMCID: 3411582).

Kellogg, J.A., Manzella, J.P., Bankert, D.A., 2000. Frequency of low-level bacteremia in children from birth to fifteen years of age. J. Clin. Microbiol. 38 (6), 2181–2185 PubMed PMID: PMC86758.

Kommedal, Ø., Lekang, K., Langeland, N., Wiker, H.G., 2011. Characterization of polybacterial clinical samples using a set of group-specific broad-range primers targeting the 16S rRNA gene followed by DNA sequencing and RipSeq analysis. J. Med. Microbiol. 60 (Pt 7), 927–936. http://dx.doi.org/10.1099/jmm.0.028373-0 PubMed PMID: PMC3168215.

Lan, R., Reeves, P.R., 2002. *Escherichia coli* in disguise: molecular origins of Shigella. Microbes and infection/Institut Pasteur 4 (11), 1125–1132 PubMed PMID: 12361912.

Matsuda, N., Matsuda, M., Notake, S., Yokokawa, H., Kawamura, Y., Hiramatsu, K., Kikuchi, K., 2012. Evaluation of a simple protein extraction method for species identification of clinically relevant staphylococci by matrix-assisted laser desorption ionization-time of flight mass spectrometry. J. Clin. Microbiol. 50 (12), 3862–3866. http://dx.doi.org/10.1128/JCM.01512-12 PubMed PMID: 22993187; PMCID: 3502947.

McKee, A.S., McDermid, A.S., Ellwood, D.C., Marsh, P.D., 1985. The establishment of reproducible, complex communities of oral bacteria in the chemostat using defined inocula. J. Appl. Bacteriol. 59 (3), 263–275 Epub 1985/09/01. PubMed PMID: 3932293.

McKenna, P., Hoffmann, C., Minkah, N., Aye, P.P., Lackner, A., Liu, Z., Lozupone, C.A., Hamady, M., Knight, R., Bushman, F.D., 2008. The macaque gut microbiome in health, lentiviral infection, and chronic enterocolitis. PLoS Pathog. 4 (2), e20. http://dx.doi.org/10.1371/journal.ppat.0040020 PubMed PMID: 18248093; PMCID: 2222957.

Perez, K.K., Olsen, R.J., Musick, W.L., Cernoch, P.L., Davis, J.R., Land, G.A., Peterson, L.E., Musser, J.M., 2013. Integrating rapid pathogen identification and antimicrobial stewardship significantly decreases hospital costs. Arch. Pathol. Lab. Med. 137 (9), 1247–1254. http://dx.doi.org/10.5858/arpa.2012-0651-OA (PubMed PMID: 23216247).

Petti, C.A., Bosshard, P.P., Brandt, M.E., Clarridge, J.E., Feldblyum, T.V., Foxall, P., Furtado, M.R., Pace, N., Procop, G., 2008. Interpretive Criteria for Identification of Bacteria and Fungi by DNA Target Sequencing; Approved Guideline: Clinical and Laboratory Standards Institute (CLSI).

Sabui, T., Tudehope, D.I., Tilse, M., 1999. Clinical significance of quantitative blood cultures in newborn infants. J. Paediatr. Child Health 35 (6), 578–581 Epub 2000/01/15. PubMed PMID: 10634987.

Schlaberg, R., Simmon, K.E., Fisher, M.A., 2012. A systematic approach for discovering novel, clinically relevant bacteria. Emerg. Infect. Dis. 18 (3), 422–430. http://dx.doi.org/10.3201/eid1803.111481 PubMed PMID: PMC3309591.

Sundquist, A., Bigdeli, S., Jalili, R., Druzin, M.L., Waller, S., Pullen, K.M., El-Sayed, Y.Y., Taslimi, M.M., Batzoglou, S., Ronaghi, M., 2007. Bacterial flora-typing with targeted, chip-based pyrosequencing. BMC Microbiol. 7, 108. http://dx.doi.org/10.1186/1471-2180-7-108 PubMed PMID: 18047683; PMCID: 2244631.

Walters, W., Hyde, E.R., Berg-Lyons, D., Ackermann, G., Humphrey, G., Parada, A., Gilbert, J.A., Jansson, J.K., Caporaso, J.G., Fuhrman, J.A., Apprill, A., Knight, R., 2016. Improved bacterial 16S rRNA gene (V4 and V4–5) and fungal internal transcribed spacer marker gene primers for microbial community surveys. mSystems 1 (1). http://dx.doi.org/10.1128/mSystems.00009-15.

Weisburg, W.G., Barns, S.M., Pelletier, D.A., Lane, D.J., 1991. 16S ribosomal DNA amplification for phylogenetic study. J. Bacteriol. 173 (2), 697–703 PubMed PMID: 1987160; PMCID: 207061.

Whiteley, A.S., Jenkins, S., Waite, I., Kresoje, N., Payne, H., Mullan, B., Allcock, R., O'Donnell, A., 2012. Microbial 16S rRNA ion tag and community metagenome sequencing using the ion torrent (PGM) platform. J. Microbiol. Methods 91 (1), 80–88. http://dx.doi.org/10.1016/j.mimet.2012.07.008.

Yagupsky, P., Nolte, F.S., 1990. Quantitative aspects of septicemia. Clin. Microbiol. Rev. 3 (3), 269–279 PubMed PMID: PMC358159.

Yutin, N., Galperin, M.Y., 2013. A genomic update on clostridial phylogeny: Gram-negative spore formers and other misplaced clostridia. Environ. Microbiol. 15 (10), 2631–2641. http://dx.doi.org/10.1111/1462-2920.12173 PubMed PMID: 23834245; PMCID: 4056668.

Zhu, Q., Quivey, R.G., Berger, A.J., 2004. Measurement of bacterial concentration fractions in polymicrobial mixtures by Raman microspectroscopy. J. Biomed. Opt. 9 (6), 1182–1186. http://dx.doi.org/10.1117/1.1803844 Epub 2004/12/01. PubMed PMID: 15568938.