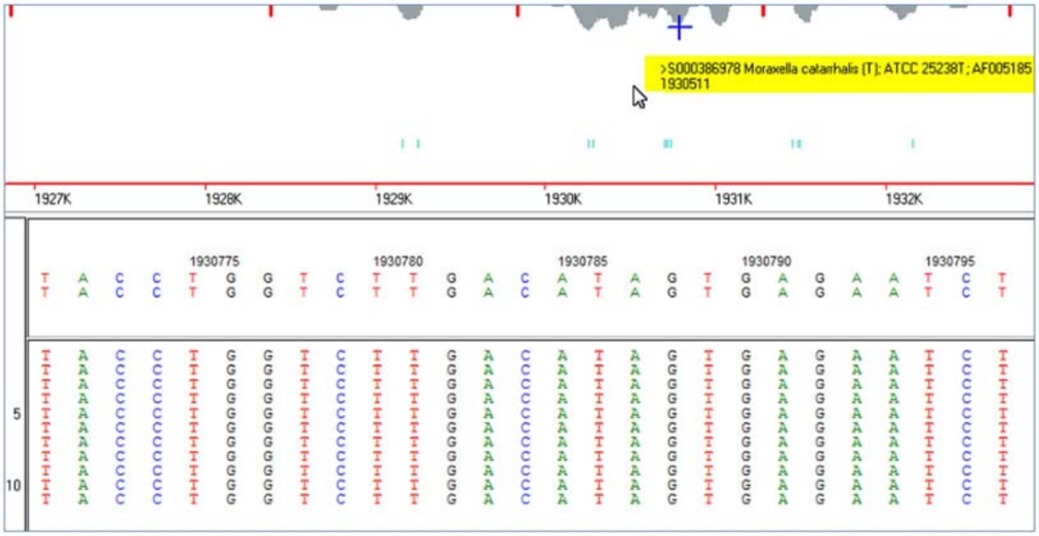
Using NextGENe® Software to Identify Bacteria in Metagenomic Samples by Simultaneous Alignment of 2nd Generation Sequencing Reads to Multiple 16S rRNA References

John McGuigan, Kevin LeVan, Ni Shouyong, Megan McCluskey, CS Jonathan Liu

Introduction

Early diagnosis and treatment of infections are more important than ever with the rise of antibacterial resistant "superbug" bacteria like Methicillin-resistant Staphylococcus aureus (MRSA). If the bacterial strain behind an infection is recognized earlier, treatment with effective antibiotics can begin earlier and recovery times are reduced. The usual procedure for diagnosing bacterial infections involves culturing samples on nonspecific media and then assaying the pure cultures. This approach can delay treatment for days or weeks and has limited sensitivity because clinical samples may have low concentrations of bacteria or may contain hard-to-culture strains. Faster methods such as real-time PCR testing [1] or more accurate methods such as specialized culture media [2] have been developed but both are limited to testing for a single type of bacteria. The same problems apply to identifying bacteria in environmental samples. Soil bacteria are notoriously difficult to isolate and some species need to be incubated for months to form visible colonies [3].

Using NextGENe to analyze sequence data from next generation sequencing systems (such as the Illumina and Ion Torrent) solves all of these problems. NextGENe is able to align sequence reads across many bacterial references at one time. Thus, accurate and rapid identification of even a small amount of bacteria is possible without being limited to testing for a single species or strain. The results are displayed in the sequence alignment viewer (figure 1) and are available for export in an expression report (figure 3). This method is also very useful for high throughput screening of antibacterial activity. Direct sequencing can be used to monitor the amount and identity of bacterial growth.



>S000386978 Moraxella catarrhalis (T); ATCC 25238T; AF005185
1930511

Figure 1: Alignment results. The identity of any species in the reference file is accessible in the alignment viewer or the expression report.

Methodology

A dataset (SRA001099) was downloaded from the NCBI short read archive (http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi). It is a sequencing library from the sputum of a cystic fibrosis (CF) patient containing 4,499 reads.

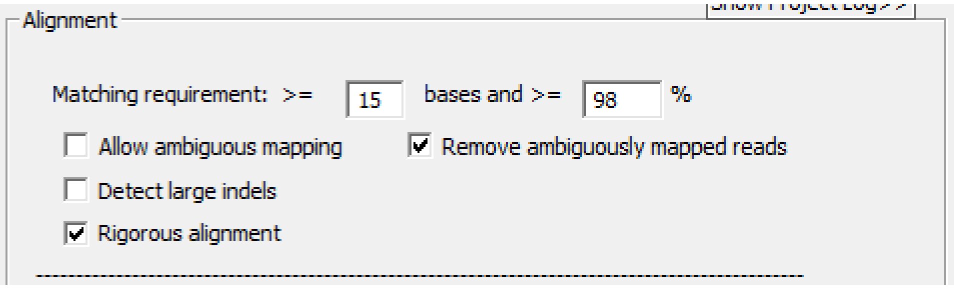
The data was converted to fasta format and filtered using NextGENe's format conversion tool. Reads with less than 60 called bases or with a median score less than 20 were rejected. Reads were trimmed where three or more consecutive bases with a score less than 16 were found. After processing there were 4,405 reads remaining. The remaining reads were aligned to a reference consisting of 5,164 bacterial 16S ribosomal RNA sequences downloaded from RDP database (https://rdp.cme.msu.edu/). All classified type-strain isolates were included in the reference.

A second dataset (SRA009753) was also examined. It consists of over 300,000 barcoded sequencing reads from eight different compost or soil samples. After DNA extraction from the community microbial samples PCR was used to amplify the 16S rRNA region. NextGENe's format conversion tool was also used to filter this dataset. After quality filtering the reads were sorted with NextGENe's Barcode Sorting tool which removed the barcode sequences and sorted the reads into separate files. The primer sequences were then removed with the format conversion tool. Each sample was aligned to the RDP 16s rRNA reference individually.

Error correction was used for both alignments in order to remove homopolymer errors. It works by parsing the reads into shorter keywords and comparing those keywords between reads in order to generate a consensus sequence. Keywords are created by dividing the reads where a nucleotide is repeated 3 or more times (a homopolymer). If aligned homopolymers vary in length the median number of bases is used in the consensus sequence.

Procedure

1. Filter and trim reads for quality using NextGENe's format conversion tool.
2. Align reads to the 16s rRNA reference. A high base matching percentage should be used because parts of the 16S rRNA are very similar between species (figure 2).
3. The expression report can be found under the "Report" menu on the toolbar in the sequence alignment viewer.



Alignment

Matching requirement: >= 15 bases and >= 98 %

□ Allow ambiguous mapping ☑ Remove ambiguously mapped reads

□ Detect large indels

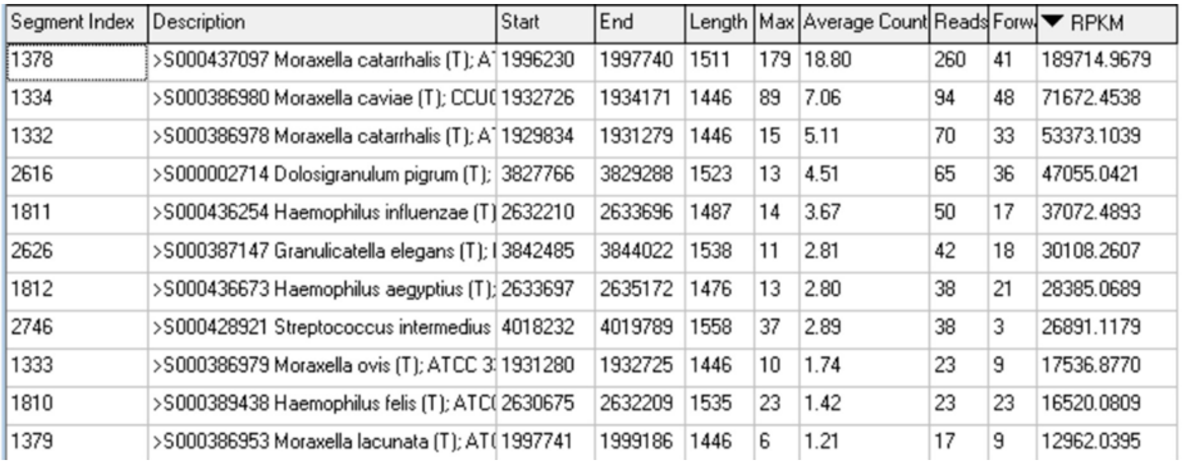
☑ Rigorous alignment

Figure 2: Alignment Settings

Results

Human Sputum Samples

In the first dataset 1,677 of the 4,405 reads were aligned to the references using 98% base matching. The expression report lists the number of reads aligned to each 16S rRNA reference and the results can be sorted based any one of several different measurements.



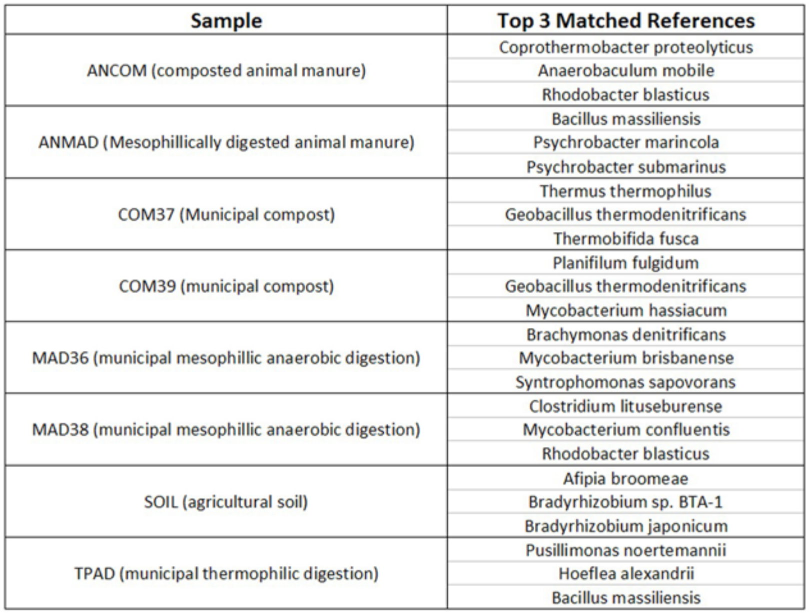
Segment Index	Description	Start	End	Length	Max	Average Count	Reads	Forw	▼ RPKM
1378	>S000437097 Moraxella catarrhalis (T); A	1996230	1997740	1511	179	18.80	260	41	189714.9679
1334	>S000386980 Moraxella caviae (T); CCU(	1932726	1934171	1446	89	7.06	94	48	71672.4538
1332	>S000386978 Moraxella catarrhalis (T); A	1929834	1931279	1446	15	5.11	70	33	53373.1039
2616	>S000002714 Dolosigranulum pigrum (T);	3827766	3829288	1523	13	4.51	65	36	47055.0421
1811	>S000436254 Haemophilus influenzae (T)	2632210	2633696	1487	14	3.67	50	17	37072.4893
2626	>S000387147 Granulicatella elegans (T);	3842485	3844022	1538	11	2.81	42	18	30108.2607
1812	>S000436673 Haemophilus aegyptius (T);	2633697	2635172	1476	13	2.80	38	21	28385.0689
2746	>S000428921 Streptococcus intermedius	4018232	4019789	1558	37	2.89	38	3	26891.1179
1333	>S000386979 Moraxella ovis (T); ATCC 3	1931280	1932725	1446	10	1.74	23	9	17536.8770
1810	>S000389438 Haemophilus felis (T); ATC(	2630675	2632209	1535	23	1.42	23	23	16520.0809
1379	>S000386953 Moraxella lacunata (T); AT(	1997741	1999186	1446	6	1.21	17	9	12962.0395

Figure 3: Expression Report

Error correction and alignment of the 4,405 reads took less than a minute. 42 different references had at least two reads aligned to them. 8 of the top 15 most-matched references (11.5% of matched reads) belonged to the Moraxella genus which is associated with lower respiratory infections in adults with chronic lung disease such as cystic fibrosis. .

Soil and Compost Samples

The number of aligned reads varied for each sample in the second dataset, but ranged from 639 to 6,651 (2.0% to 15.5%). Each alignment took less than 1 minute to finish. The top results are summarized in table 1. Most of the top matches were nitrogen-fixing, composting, or other soil bacteria.



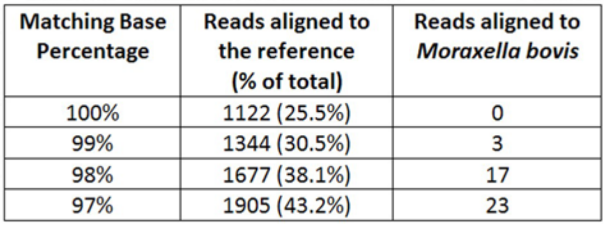
Sample	Top 3 Matched References
ANCOM (composted animal manure)	Coprothermobacter proteolyticus
	Anaerobaculum mobile
	Rhodobacter blasticus
ANMAD (Mesophillically digested animal manure)	Bacillus massiliensis
	Psychrobacter marincola
	Psychrobacter submarinus
COM37 (Municipal compost)	Thermus thermophilus
	Geobacillus thermodenitrificans
	Thermobifida fusca
COM39 (municipal compost)	Planifilum fulgidum
	Geobacillus thermodenitrificans
	Mycobacterium hassiacum
MAD36 (municipal mesophillic anaerobic digestion)	Brachymonas denitrificans
	Mycobacterium brisbanense
	Syntrophomonas sapovorans
MAD38 (municipal mesophillic anaerobic digestion)	Clostridium lituseburense
	Mycobacterium confluentis
	Rhodobacter blasticus
SOIL (agricultural soil)	Afipia broomeae
	Bradyrhizobium sp. BTA-1
	Bradyrhizobium japonicum
TPAD (municipal thermophilic digestion)	Pusillimonas noertemannii
	Hoeflea alexandrii
	Bacillus massiliensis

Table 1: The top three matched references for each sample in the second dataset.

Discussion

NextGENe's error correction using the condensation tool is used to correct homopolymer errors that are one of the biggest problems with semiconductor technology. After error correction the sample is rapidly aligned to the reference. The incorrect alignment of reads caused by highly similar sequences is avoided by requiring a very high base matching percentage and by rejecting reads that map perfectly to several sites in the reference. When references with only one aligned read are ignored in order to increase specificity, there are 42 bacterial species identified in the first dataset, close to the 36 identified in the original study [4].

Sensitivity and specificity can be adjusted by changing the required base matching percentage. 16s rRNA sequences are well conserved among bacteria, with those in the same genus often sharing greater than 96% similarity. Using 98 to 99% base matching is recommended in order to allow some species-level specificity while maintaining reasonable sensitivity. Table 2 shows the sensitivity (number of matched reads) and specificity (number of reads aligned to Moraxella bovis) at several base matching levels. The reads aligned to Moraxella bovis in the human sputum sample are assumed to be false positives because that species- unlike the others in the Moraxella genus- is normally found in cow eye infections.



Matching Base Percentage	Reads aligned to the reference (% of total)	Reads aligned to *Moraxella bovis*
100%	1122 (25.5%)	0
99%	1344 (30.5%)	3
98%	1677 (38.1%)	17
97%	1905 (43.2%)	23

Table 2: Comparison of sensitivity and specificity at different base matching percentages

NextGENe's barcode sorting tool makes it very easy to process samples that are barcoded like the second dataset. The barcode sequences are removed and the reads are grouped either based on automatic detection or based on a file defining the sequences. The analysis of two very different datasets demonstrates the utility of a large reference. Though the reads are compared to over 5000 different bacteria- including human pathogens and soil microbes- aligning over 40,000 reads takes less than a minute. As expected, mesophilic bacteria were found in the mesophilic digestion samples, soil bacteria were found in the soil sample, and thermophilic bacteria were found in the thermophilic digestion and compost samples all without culturing.

In addition to Roche GS FLX^{TM}, NextGENe is capable of processing data from the Illumina and Ion Torrent systems. The increased depth of coverage (10 million and 100 million reads per run, respectively) can provide even greater sensitivity so that bacteria present at very low concentrations can be detected.

References

1. Warren, D.K. et al. Detection of Methicillin-Resistant Staphylococcus aureus Directly from Nasal Swab Specimens by a Real-Time PCR Assay. J. Clin. Microbiol. 42, 5578-5581(2004).
2. Diederen, B. et al. Performance of CHROMagar MRSA Medium for Detection of Methicillin-Resistant Staphylococcus aureus. J. Clin. Microbiol. 43, 1925-1927(2005).
3. Davis, K.E.R., Joseph, S.J. & Janssen, P.H. Effects of Growth Medium, Inoculum Size, and Incubation Time on Culturability and Isolation of Soil Bacteria. Appl. Environ. Microbiol. 71, 826-834(2005).
4. Armougom, F. & Raoult, D. Use of pyrosequencing and DNA barcodes to monitor variations in Firmicutes and Bacteroidetes communities in the gut microbiota of obese humans. BMC Genomics 9, 576(2008).