

# Metagenomic Studies of Viral and Bacterial Infections using Pyrosequencing Reads with NextGENe® Software

October 2009

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

## Introduction

Quick and accurate identification of viral and bacterial pathogens using next gen sequencing is valuable for both treatment and research (1). Early identification of an infectious agent can lead to a targeted treatment. Fast identification of novel pathogens- new viruses, bacteria, and other microorganisms- such as swine flu and Methicillin-resistant Staphylococcus aureus (MRSA) will speed up vaccine and new drug development. Drug efficacy studies are becoming easier through the use of metagenomic analysis of viral, bacterial and human sequences. The viral and bacterial concentration can be determined from the sequence reads after human background is subtracted using NextGENe software.

The traditional methods for detecting and identifying pathogens require culturing bacteria and viruses or detecting viral antigens. These procedures have several problems- they are costly in both time and money and there is a limit to their sensitivity because some viruses and bacteria are very difficult or impossible to culture, especially from small samples. Nucleic Acid Amplification tests (NATs) are a newer approach. Most involve PCR or Loop-mediated isothermal amplification (LAMP) PCR of DNA or RNA. Those that use multiplexed PCR or DNA microarrays offer a great advantage over the older methods because they are faster, more sensitive, and introduce less bias, but they are not without their own problems (2). The tests are limited by a relatively small number of candidate pathogens and while they may detect new strains, they aren't always useful for characterizing them.

Next generation sequencing technologies such as Roche GS FLX™, Illumina Genome Analyzer and Applied Biosystems SOLiDTM system make it possible to obtain millions of reads in a single run. NextGENe is able to quickly separate most host genome contamination from samples before aligning the remaining reads to bacterial or viral genomes for identification and characterization purposes. As seen in figure 1, NextGENe is able to align reads to several Bacteroides genomes at once.

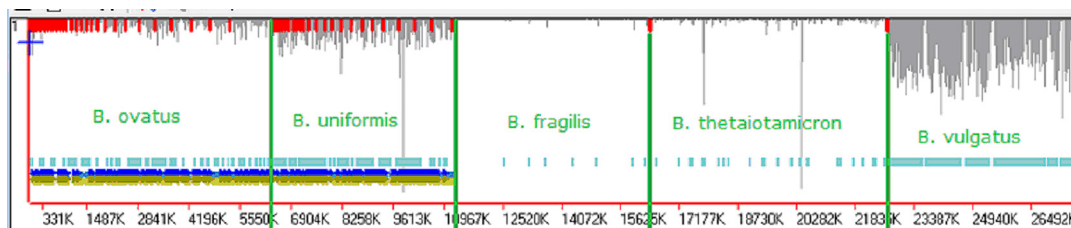


Figure 1: Metagenomic data aligned across five different Bacteroides genomes

## Methodology

Pyrosequencing data (SRA001127) was downloaded from the NCBI short read archive (<http://www.ncbi.nlm.nih.gov/Traces/sra>). The data consisted of two stool samples from the same patient- one taken when he was afflicted with diarrhea and one taken after treatment. The reads were first filtered for quality with NextGENe's format conversion tool. Reads were trimmed where more than 3 bases in a row had a quality score less than 16 and reads with a median score less than 20 or with less than 25 called bases were rejected, leaving 106,109 of 106,327 reads from the healthy sample and 96,679 of 96,941 reads in the diseased sample.

Another dataset (SRA002159) consisting of sequence data from three nasopharyngeal aspirates was analyzed using the same settings. The linker sequence (TGTGTTGGGTGTGTTTGG) was also trimmed. After quality filtering, 27,644 of 30,259 reads were left in sample one, 22,209 of 24,435 reads in sample two, and 19,209 of 21,087 reads in sample three.

Most human contamination was removed from the samples by aligning the reads to the whole human genome and then aligning the remaining reads to reference human fosmid data which helps to account for some of the gaps in the genome reference. The remaining reads were then aligned to various viral and bacterial genomes.

## Procedure

1. Use the Format Conversion Tool to convert sample data to fasta format
  - a. Quality filtering can be performed simultaneously
2. Align to the preloaded human genome reference to remove contamination (figure 2)
  - a. “Matching Base Percentage” setting should be set near 85%
  - b. “Seed Length” can be increased to 35 in order to speed up the alignment because pyrosequencing reads are of sufficient length
3. Align the unmatched reads to reference human fosmid data to remove human sequences that are not part of the assembled human genome
  - a. Unmatched reads are saved as projectname\_unmatched.fasta in the “\projectname\projectname.files\sample\” directory
  - b. “Matching Base Percentage” should be set close to 85%
4. Perform a new alignment with the unmatched reads using a viral or bacterial reference or several references at once.
  - a. “Matching Base Percentage” should be set lower than before due to hypervariable regions in the bacterial genome, but no lower than 70%
5. Run an alignment for each species to be detected

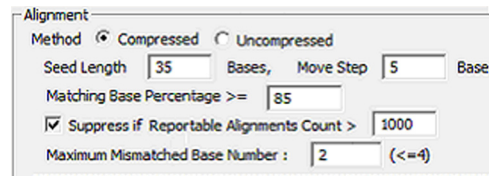


**Figure 2:** Whole Genome Alignment settings used to remove host DNA contamination

## Results

Each sample was first aligned to the human genome (build 36.1) with at least 85% base matching required for an alignment in order to remove most of the host reads and then aligned to a reference set of human fosmid data in order to account for some human DNA not found in the genome build. NextGENE aligns more human reads with 85% base matching than identifying the reads using BLAST with an expect-value cutoff of 10-40 (table 1). The remaining reads were then aligned to reference genomes of some common or expected pathogens with at least 85% base matching required. The results are summarized in table 2 where they are compared to the results of the previous study (3). Overall NextGENE was able to identify about twice as many reads as the previous study with the biggest increases from reads aligned to *B. ovatus* and *B. uniformis*.

|                         | NextGENE Analysis |       |           |       | BLAST Analysis |       |           |       |
|-------------------------|-------------------|-------|-----------|-------|----------------|-------|-----------|-------|
|                         | Sick              |       | Recovered |       | Sick           |       | Recovered |       |
|                         | Count             | %     | Count     | %     | Count          | %     | Count     | %     |
| <b>Total Bacteria</b>   | 16,420            | 79.2% | 15,948    | 99.3% | 8,506          | 74.2% | 7,985     | 99.0% |
| <b>Homo Sapiens</b>     | 4,316             | 20.8% | 108       | 0.7%  | 2,955          | 25.8% | 84        | 1.0%  |
| <b>Total Identified</b> | 20,736            |       | 16,056    |       | 11,461         |       | 8,069     |       |

**Table 1:** Removal of host genome contamination. Percentages are out of the total number of identified reads.

|                                   | NextGENE           |       |         |       | BLAST Analysis |       |         |       |
|-----------------------------------|--------------------|-------|---------|-------|----------------|-------|---------|-------|
|                                   | Sick               |       | Healthy |       | Sick           |       | Healthy |       |
|                                   | Count              | %     | Count   | %     | Count          | %     | Count   | %     |
| <b>Bacteroides vulgatus</b>       | 8,938              | 54.4% | 7,271   | 45.6% | 5,944          | 69.9% | 4,743   | 59.4% |
| <b>B. thetaiotaomicron</b>        | 2,026              | 12.3% | 2,574   | 16.1% | 767            | 9.0%  | 1,046   | 13.1% |
| <b>B. fragilis</b>                | 606                | 3.7%  | 789     | 4.9%  | 759            | 8.9%  | 842     | 10.5% |
| <b>B. ovatus*</b>                 | 2,796              | 17.0% | 3,501   | 22.0% | 48             | 0.6%  | 63      | 0.8%  |
| <b>B. uniformis*</b>              | 3,806              | 23.2% | 3,868   | 24.3% | 14             | 0.2%  | 8       | 0.1%  |
| <b>Parabacteroides distasonis</b> | 1,289              | 7.9%  | 2,012   | 12.6% | 818            | 9.6%  | 1,283   | 16.1% |
| <b>Campylobacter jejuni</b>       | 218                | 1.3%  | 19**    | 0.1%  | 156            | 1.8%  | 0       | 0.0%  |
| <b>Total Bacteroides</b>          | 15,217             | 92.7% | 14,316  | 89.8% | 7,532          | 88.5% | 6,702   | 83.9% |
| <b>Total time for analysis</b>    | Less than one hour |       |         |       | Hours to Days  |       |         |       |

**Table 2:** Summary of results from the detection of bacteria in stool samples.

\* These genomes are not complete and the increased number of identified reads between the two studies may be due to a difference in the references used.

\*\*See discussion section

NextGENe was also able to remove host genome contamination from pyrosequencing data collected from nasopharyngeal aspirates, as seen in table 3. NextGENe was then used to identify reads aligning to viral genomes and was successful in detecting two different strains of Influenza A in three different samples to confirm the previous study (4) while ruling out the presence of several dozen other viruses (table 4). All three samples showed reads aligning to both H1N1 and H3N2 Influenza A, but the reads that align to multiple strains can be ruled out in order to make a strain identification. NextGENe was also able to detect some E. coli contamination in each sample with the most found in sample 1.

|                         | Sample 1 | Sample 2 | Sample 3 |
|-------------------------|----------|----------|----------|
| Number of Sequences     | 30259    | 24435    | 21087    |
| After quality filtering | 27644    | 22209    | 19209    |
| After genome alignment  | 9973     | 7007     | 5727     |
| After fosmid alignment  | 9406     | 6647     | 5350     |
| % Human Contamination   | 68.9%    | 72.8%    | 74.6%    |

**Table 3:** Removal of host genome contamination from pyrosequencing data collected from nasopharyngeal aspirates.

|                         | Sample 1 | Sample 2 | Sample 3 |
|-------------------------|----------|----------|----------|
| Influenza A (H1N1)      | 103      | 15       | 19       |
| Influenza A (H3N2)      | 315      | 1        | 81       |
| Only H1N1               | 1        | 14       | 0        |
| Only H3N2               | 199      | 0        | 56       |
| Influenza A (5 strains) | 366      | 16       | 90       |
| Influenza B             | 0        | 0        | 0        |
| Influenza C             | 0        | 0        | 0        |
| Parainfluenza           | 0        | 0        | 0        |
| Enteroviruses           | 0        | 0        | 0        |
| Adenoviruses            | 0        | 0        | 0        |
| Bocaviruses             | 0        | 0        | 0        |
| Rhinoviruses            | 0        | 0        | 0        |
| E. coli                 | 67       | 2        | 7        |

**Table 4:** Identification of pyrosequencing reads as specific viral strains.

## Discussion

NextGENe has several features that make it a valuable tool in metagenomic studies of viruses and bacteria. Thanks to its unique Burrows Wheeler Transform alignment method for whole genomes, NextGENe is able to align over 100,000 reads of 454 sample reads to the whole human genome in about 20 minutes using a single core on a computer with 8 GB of RAM. When whole genome alignment is combined with alignment to fosmid data, most of the interfering human sequences can be removed in less than 30 minutes. NextGENe is also very fast when aligning to smaller genomes- alignments to 7 bacterial genomes at once took less than 90 seconds and alignments to over two dozen viral genomes took less than 10 seconds. Hundreds of bacterial genomes and thousands of viral genomes are combined and indexed so that they can be tested for all at once on a timescale similar to whole human genome alignment (less than 30 minutes for this sample size).

In some metagenomic studies such as the ones described earlier (3, 4) reads are identified using BLAST. This process is very time consuming- even if each read takes one second to identify, the datasets used here (roughly 280,000 reads) would take over 75 hours to process. Another advantage for NextGENe is the ability to align reads that would otherwise have resulted in a “no significant similarity found” result with BLAST. NextGENe is also an improvement over BLAST searches because it aligns reads to the best match rather than returning multiple possible alignments. The software provides an option to reject ambiguously aligned reads (those that align to multiple sites) which is a problem when aligning to several bacterial genomes that can have similar sequences. There were 19 reads found aligning to *C. jejuni* in the healthy sample when there was no *C. jejuni* present, but 16 of those reads can be immediately eliminated because they also align to other genomes.

---

This process of virus and bacteria detection is very useful for drug studies. The presence of bacteria or virus can be quantified before and after treatment and can be further analyzed by looking at the sequences. With NextGENe's paired-end linking method, short paired Illumina and SOLiD reads can be combined to create 10 to 100 million reads of about 200 bp. This will allow researchers to accurately detect more pathogens, even those present at low levels in the sample.

NextGENe also includes software applications for a variety of other application types including expression studies like Digital Gene Expression, transcriptome analysis, microRNA studies and SAGE, as well as de novo assembly, SNP and indel detection, and ChIP-Seq.

## References

1. Quan, P. et al. Detection of respiratory viruses and subtype identification of influenza A viruses by GreeneChipResp oligonucleotide microarray. *J. Clin. Microbiol* 45, 2359-2364(2007).
2. Fox, J.D. Nucleic acid amplification tests for detection of respiratory viruses. *Journal of Clinical Virology* 40, S15-S23(2007).
3. Nakamura, S. et al. Metagenomic diagnosis of bacterial infections. *Emerging Infect. Dis* 14, 1784-1786(2008).
4. Nakamura, S. et al. Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. *PLoS ONE* 4, e4219(2009).

Trademarks are property of their respective owners.