Award Number: 81XWH-11-1-0690

TITLE: NAU-TGen North Dangerous Pathogens DNA Forensics Center Upgrade

PRINCIPAL INVESTIGATOR: Jeffrey Foster, Ph.D.

CONTRACTING ORGANIZATION: Northern Arizona University , FLAGSTAFF, AZ 86011

REPORT DATE: April 2014

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE April 2014	**2. REPORT TYPE** Final	**3. DATES COVERED** 08Sep2011-7Apr2014

4. TITLE AND SUBTITLE
NAU-TGen North Dangerous Pathogens DNA Forensics Center Upgrade

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-11-1-0690

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Jeffrey Foster, James Schupp, Arun Rawat, Rebecca Colman,
Paul Keim
Keim
Betty Diamond email: jeff.foster@nau.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Northern Arizona University, Arizona Board of Regents
Flagstaff, AZ 86011

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Despite its importance for infectious disease diagnosis, the ability to rapidly identify the causative agent for infections remains an elusive. Whole genome sequencing for clinical diagnostics is emerging as the future of disease identification and management by providing a rapid and highly sensitive method of diagnosing and characterizing infection. We have created a metagenomic application for analysis of clinical samples titled MetaGenomics Explorer (MetaGeniE, previously called IRIMAS) that is undergoing optimization and statistical validation with a range of different datasets. MetaGeniE provides novel insights and analyses of clinical datasets, detecting and characterizing the major component of community that is usually the infectious agent of interest as well as identifying the other microbes in this community. Furthermore, we are able to validate these detections with laboratory culture, and available patient information. Finally, we are able to take these analyses a step beyond nearly all current approaches and can use both SNP genotyping and determine features such as antibiotic resistance to precisely characterize these pathogens.

15. SUBJECT TERMS
pathogen, genomics, next-generation sequencing, clinical diagnostics

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	97	**19b. TELEPHONE NUMBER** *(include area code)*

TABLE OF CONTENTS

Page

INTRODUCTION

Despite its importance for infectious disease diagnosis, the ability to rapidly identify the causative agent for infections remains an elusive goal. When a soldier enters a clinic with disease symptoms, the infectious agent is rarely known. This patient is then subjected to a battery of tests, often taking days to weeks for results, to narrow down the etiological agent. Whole genome sequencing for clinical diagnostics is emerging as the future of disease identification and management by providing a rapid and highly sensitive method of diagnosing and characterizing infection. The goal is to replace the battery of clinical tests with a single test—complete sequencing of the sample to identify pathogens likely responsible for disease. Next-Generation sequencing technologies have transformed our ability to rapidly generate sequence data. Samples from diseased individuals can now be taken, the microbial and host DNA in the sample sequenced, and the likely pathogens identified and characterized to streamline treatment. Despite this seemingly simple process, there are numerous obstacles to efficient and accurate identification of pathogens in clinical samples. Our research project did the following: 1) Rigorously tested DNA sequences using bioinformatics. Exhaustive testing of sequence data from experiments using patient samples is extremely costly and time consuming. By evaluating sequences from prior work or by using computer generated sequences (i.e. *in silico*), we could rigorously test the limits of this technology and statistically assess the outcomes in a manner that would not be possible through experimentation; 2) Determined specific genome targets for species identification. We also determined genome targets for select strains of clinical relevance and those containing known regions associated with drug resistance by compiling relevant data from internal resources and conducted an extensive literature search; 3) Determined genes present in all strains or pan-genome references of targets by identifying all genetic loci/targets present in all strains of the pathogen DNA and determined an appropriate statistical threshold for a match to the pan-genome; 4) Created mixtures of human and pathogen DNA by adding single pathogen DNA at varying quantities into human clinical samples and adding DNA samples of multiple pathogens into human clinical samples; 5) Sequenced DNA mixtures using Ion Torrent and Illumina platforms to determine the level of sensitivity and accuracy, using known mixtures of single and multiple pathogen experiments; and 6) Deep sequenced DNA from patient clinical samples using these sequencing techniques. These analyses provided a solid foundation for the further development of Next-Generation sequencing for clinical diagnostics and allowed us to develop a metagenomic pipeline for rapid and accurate identification of pathogens and other microbes in clinical samples.

BODY

Overview of Tasks

All tasks have been completed.

In silico bioinformatic testing

1. Create filters to remove reads of human DNA. *Completed.*
2. Run human genome reads against pathogen reference sequences. *Completed.*
3. Create mixtures of read data from prior human and pathogen sequencing projects. Mixtures will be created at a wide range of sequence quantities and genome coverage. *Completed.*

Determine genome targets for Species ID, subtyping characterization, drug resistance

4. Compiling relevant data from NAU and TGen, *Completed.*
5. Comprehensive literature search with an emphasis on subtyping (e.g. determining particular lineages, clonal complexes, etc.) and finding targets for drug resistance characterization, *Completed.*
6. Contact infectious disease experts, *Completed.*

Create pan-genome references of targets

7. Generate a reference pan-genome with sequences for all loci/targets. *Completed.*
8. Determine appropriate statistical threshold for a match to the pan-genome. *Completed*.

Create mixtures of human and pathogen DNA

9. Spike single pathogen DNA at varying quantities into human clinical samples*. Completed.*
10. Spike DNA samples of multiple pathogens into human clinical samples. *Completed.*

Sequence DNA mixtures on Ion Torrent and Illumina platforms

11. Determine level of sensitivity and accuracy using known mixtures of single and multiple pathogen experiments. *Completed*.

Sequence patient clinical samples

12. Compile all clinical samples then culture, extract and quantify DNA for all pathogens. *Completed*.
13. Run deep coverage sequencing of clinical samples. *Completed*.

Project Overview

The primary goal of our project was to develop a metagenomics pipeline for identification of pathogens in clinical samples. We have completed all tasks for this project, as outlined in the Overview of Tasks on the previous page. The completion of these tasks was essential for development and testing of this pipeline, MetaGenomic Explorer (MetaGeniE). But pipeline development and its evaluation are only part of the work we had planned. We received a no-cost extension to have sufficient time to complete the primary manuscript on the pipeline as well as apply our analyses to unique clinical cases where the pathogens causing disease were unknown or not confirmed using traditional methods. This Final Report summarizes all of the work we performed on the project and includes results from Annual Reports from Years 1 and 2 of the project all combined into a single document plus work that we completed during the six month no-cost extension period.

Overview of MetaGeniE

We initially provide an overview of the metagenomics pipeline we have designed since it is at the heart of our analyses. The overview is also provided to give context as to which analyses we conducted as part of evaluating pipeline performance that were specific to our Tasks from page 4.

Motivation

With the decreasing cost of next-generation sequencing, deep sequencing of metagenomes is an emerging field that provides unique opportunities to understand clinical samples. Among the primary translational goals of metagenomic sequencing of clinical samples is rapidly filtering out human reads to detect pathogens with high specificity and sensitivity. Metagenomes are inherently variable due to a number of distinct non-host populations in the samples, and their relative abundance, the size and architecture of genomes and factors such as tissue (human DNA versus pathogen DNA concentration) and sampling (localization) biases from patients. This variation in metagenomes manifests in sequencing datasets as low pathogen abundance, a high number of host reads, presence of close relatives and complex communities from multiple non-host community members. In addition to these challenges posed by the composition of metagenomes, high numbers of reads generated from high-throughput deep sequencing pose immense computational challenges. Accurate identification of pathogens is confounded by individual reads mapping to multiple different genomes due to gene similarity in different taxa present in community or close relative in reference database. Available global and local aligners also vary in sensitivity, specificity, speed of detection. The efficiency of detection in clinical samples can be estimated by the level of taxonomic resolution of the organisms. We have developed an efficient strategy that allows genetic characterization of pathogens with high accuracy, consistent with results from strain level SNP-genotyping and bacterial identification from laboratory culture and PCR assays.

Background

The metagenomic composition of microbial communities has typically been estimated with conserved gene amplicon sequencing (e.g., 16S rRNA for bacteria). More recently, whole genome sequencing (WGS) approaches have emerged as a powerful alternative that gives a relatively unbiased and global representation of all members of the microbial community (Tringe and Rubin 2005, Tringe et al. 2005, Eisen 2007, Simon and Daniel 2011). With the advances in next-generation sequencing technology (Mardis 2008, Shendure and Ji 2008) and decreasing cost, it is possible to identify in clinical samples such features as all of the pathogen sequences within a metagenome (Kostic et al. 2011, Kostic et al. 2012, Kuroda et al. 2012). The next step is to go beyond species identification. WGS allows for higher resolution characterization of genomes within a sample for sub-species or strain identification, or genetic attributes (e.g., presence of antibiotic resistance markers) that can provide important insights for accurate clinical management for these patients. For clinical diagnostics, genus or even species-level identification may not be sufficient for proper clinical treatment. For example, a patient suffering from *Staphylococcus aureus* TCH1516 (methicillin resistant strain) requires different treatment than a patient colonized by *S. aureus* Newman (methicillin sensitive strain).

Next-generation sequencing technologies are improving rapidly, decreasing in cost and generating 100s of millions of reads per run. Different metagenomic analysis pipelines incorporate available aligners (local/global) in computational infrastructure such as cloud computing or high performance computing (HPC) to provide speed and scalability. Each of the current pipelines have advantages and disadvantages, with the typical issue of computational speed and scalability of high number of reads against reference database for accurate detection. For instance, PathSeq utilizes Amazon cloud computing to expand computational scalability but has associated overheads (Dimon et al. 2013). PathSeq, PARSES and MEGAN (Huson et al. 2007) characterize unknown reads (>1 million) with BLAST, however with the high number of metagenome reads, BLAST is not optimal given the high computational time required (Schmieder and Edwards 2011, Liu et al. 2013). IMSA does not discuss the scalability to large reference database and huge metagenome read files (Dimon et al. 2013), uses BLAST and invokes some processes like BLAT without parallelization. DeconSeq only removes host sequences and therefore does not have direct applicability for complex communities (Schmieder and Edwards 2011).

Each metagenomic sequencing read, in theory, originates from a single genome. Assigning these high numbers of reads (especially 50-200 bp short reads) back to its genome of origin is problematic for multiple reasons including, a) the presence of overlapping/shared genomes from other organisms in the sample; b) the querying these reads against related genomes from publically available databases may result in a greater number of hits due to homology; and c) computational resources required to scan through huge reference databases. We have developed a pipeline designed for high performance computing (HPC) that manages sensitivity and specificity for detection of complex microbial samples and addresses all of the traditional limitations with many metagenomic analyses. The pipeline first builds all against all relations between read and reference database and then generate cumulative statistics from local and global alignment. The pipeline also incorporates features like comprehensive human read filtration and scalability to search large reference databases like microbial Refseq presently at ~19 GB.

Final Design, Development and Optimization of MetaGeniE

Computational Aspects

The goal of clinical metagenomics is often to detect a signal for the major cause of infection amidst a veritable sea of host and non-host sequences. No two metagenomes are identical and variation exists due to number of distinct populations in the community, the size and architecture of genomes and relative abundance of different microbes in the sampled community. Other factors that dictate variation in metagenomes are tissue (human DNA versus pathogen DNA concentration) and sampling (localization) biases. The variation in metagenomes makes it challenging to develop a “one size fits all” method, given the different needs of researchers. The characterization of community composition using microbial sequences can be approached at three taxonomic levels: genus, species and strain/genotype (Figure 1).

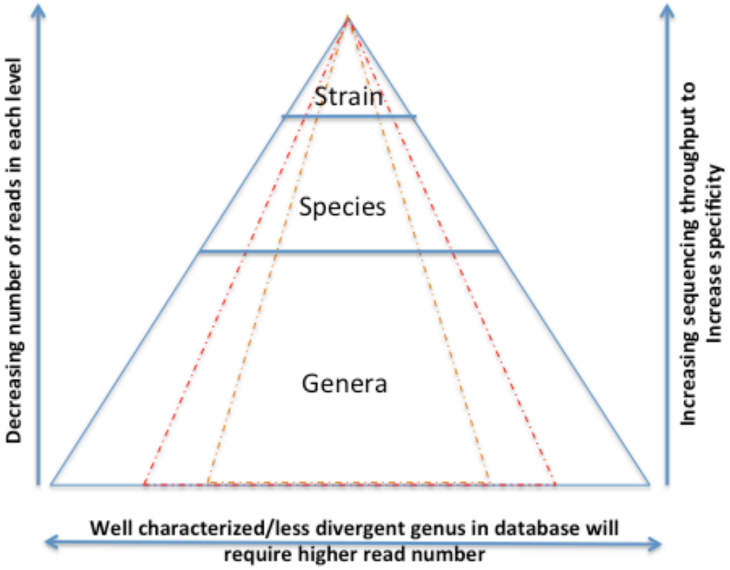


Figure 1. Hierarchical architecture of genomes and its relationship with sequencing throughput.

Usually the probability of sequencing strain or species-specific regions is less than genus-specific regions, given the hierarchical architecture of genome. Increasing sequencing throughput will allow better taxonomic resolution of any organism, especially the well-characterized or/and less divergent genus. For metagenome data, single reads may map to multiple organisms either due to conserved microbial genomic regions (e.g., genus specific genes) due to the presence of closely related organisms in the community being analyzed or in reference databases. Studies have shown metagenomic sequences share similar regions for even the simplest microbial communities. Assigning each read to all mapped genomes might be effective strategy as metagenome community can be considered to have no *a priori* knowledge about organism/community. The genus specific reads will map to higher number of organisms followed by reads specific to species and sub-species/strains. The organism with the highest shared (genus specific) regions, as well as unique regions, which generally belong to a species, and strain specific genes, will result in a higher % genome mapped. The taxonomic rank and the resolution for detection is proportional to sequencing throughput, richness of pathogen(s) in metagenome sampling and the presence of genomic data from the community (target pathogens) or its close relative in the reference database. We benchmarked the sensitivity and specificity of the detection step of the pipeline by evaluating simulated read libraries through identification of correct pathogen and its %read recalled and genome coverage detected, correct % of host reads filtered and false detection of host and non-host reads as discussed in Section Human Filtration and Pathogen Detection.

Implementation of MetaGeniE

The MetaGenomic Explorer (MetaGeniE) is a distributed and scalable microbial discovery pipeline that consists of two modules: Read-Reduct and Patho-Detect. MetaGeniE is fully functional metagenomic pipeline that can be installed in Unix operating system and documentation for installation is available. The pipeline provides configuration file that allows user to set parameters and paths easily for required applications. The Metagenie is supported to run on cluster computing with user defined processors, memory usage and modules to run. Directions to set the options are available both in the documentation as well as command option provided.

Design of Read Reduction Module of MetaGeniE

The Read-Reduct module sequentially filters and reduces the low quality, redundant and human reads. In this first module, to reduce overall computational processing time and memory, faster analyses such as quality filtering and BWT alignment are performed initially using BWA (Li and Durbin 2009, 2010) or BOWTIE2 (Langmead 2010). Higher CPU and memory intensive features such as data compression and hash-based sensitive alignment using STAMPY (Lunter and Goodson 2011) are then utilized to further reduce the overall number of reads. Finally the remaining reads are aligned against the repeat database (Repbase, http://www.girinst.org/repbase/) and mapped repeat reads are removed. Benchmarking was performed from seven human datasets downloaded from SRA and one from simulated dataset and discussed in the first section of our tasks, In silico bioinformatic testing.

Design of Pathogen Detection Module of MetaGeniE

The second module of the pipeline, Patho-Detect, aligns the remaining reads after human filtration against known bacterial, fungal and viral sequences with BWT alignment followed with local aligner BLAT (Figure 2). MetaGeniE aggregates global and local (BLAT) alignments and generates cumulative report from both alignments for sensitive detection. Detection by read numbers is often not an accurate predictor due to repeat elements, close relatives in the metagenome and PCR amplification biases. Instead, MetaGeniE detects microbial presence by genomic reconstruction using percentage of the genome mapped and therefore largely overcomes these biases. Incorporating a larger NCBI RefSeq DB rather than using just a few selected complete genomes helps in resolution of pathogens to subspecies/strain level. RefSeq DB has doubled from 8.7 GB in Release 54 to 19 GB in Release 60 for bacteria. MetaGeniE can now handle multiple partitions of the reference database and so is scalable to large databases for better memory management. NCBI indexes and hashes allow faster extraction of millions of reads. Custom parsers allocate all reads to all mapped genomes instead of SAMTools that map main hits only and therefore MetaGeniE limits underestimating pathogen detection in the metagenome.

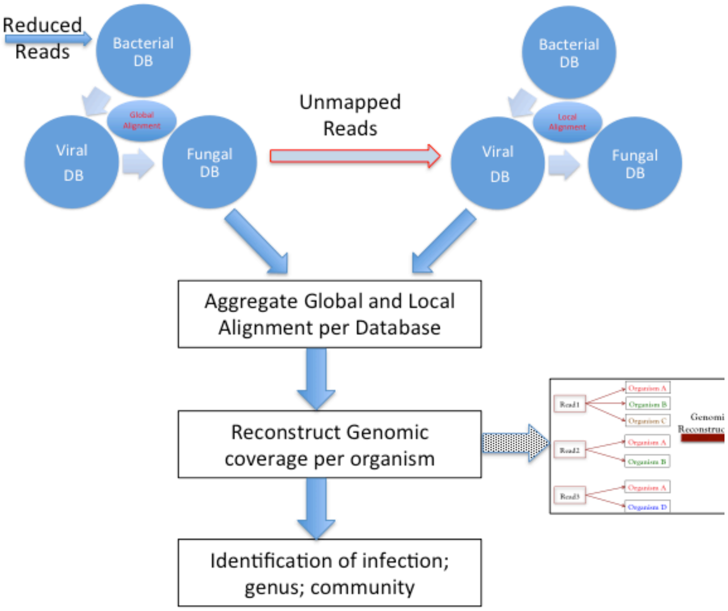


Figure 2. The pathogen detection module of the pipeline.

Features

Scalability

Incorporating a larger NCBI RefSeq database rather than using just a few selected complete genomes helps in closest resolution of identification to subspecies/strain level. RefSeq bacterial database has doubled from 8.7G in Release 54 to 19G in Release 60 for bacteria and increasing in future. This results in increasing demand for computational memory to scale huge reference databases.

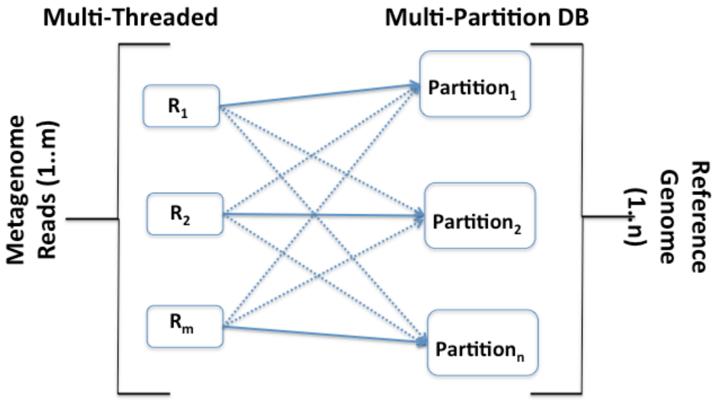


Figure 3. Multithreaded input sequence file query the multiple partition reference database to address the scalability.

To address the issue of scalability with large reference databases, we designed the pipeline to handle multiple partitions of reference database for better memory management (Figure 3). Each smaller database partitions (~1GB) can be queried by multithreaded input files iteratively therefore reducing overall memory footprint. This querying of each input file fragment generates higher number of mapped-unmapped relationships against the partitioned database results per iteration that increases the computational time. To address this issue, the pipeline utilizes custom hash functions and indexing tool formatdb and fastacmd (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/) to allow faster extraction of millions of reads as an input to next reference database search.

Normalized (%) Genome Coverage

Pathogen detection by the total number of reads that hit/align to respective genome is not always an accurate predictor of presence of an organism due to repeat elements, close relatives in metagenome and PCR amplification. To overcome these biases, the pipeline detects microbial presence by genomic reconstruction i.e. %genome mapped for each organism from initial mixture of non-host DNA of metagenome. The pipeline first converts the local and global alignment output to common BED format. Genome coverage of each mapped organism is then calculated from the global and local alignments with BEDTOOLS (Quinlan and Hall, 2010). The total genome reconstructed for each mapped organism is the sum of genome coverage from global and local alignments for any metagenome. The normalized genome coverage (%genome coverage) is calculated as follows:

$$\text{Normalized Genome Cov.} = \left(\frac{\text{Genome Cov. by local alignment} + \text{Genome Cov. by global alignment}}{Total\ Genome\ Size}\right) * 100$$

The normalized genome coverage allows comparison of different organism with different genome size. It is helpful in representing the abundance of various organisms in metagenome for community analysis.

Benchmark

Bacterial Datasets: Simulated reads of length 100 bases were generated from respective reference genome as discussed in each section. Average Illumina sequencing error of 0.4% is incorporated in all the simulated bacterial reads. To study divergence and its effect on detection, additional variability of 0.1%, 0.2%, 0.5% and 1% is incorporated in each simulated bacterial library. All these steps utilized the same parameters as follows, BWA (default) and BLAT (80% identity).

Pathogen Detection

Metagenome datasets derived from clinical samples might have challenges like low abundance of pathogens, the complexity of community might vary from single infection (with only one dominant infection), multiple infections from close relative and complex community (example cystic fibrosis patients). The sequencing reads to be aligned against reference genome might have high divergence resulting from sequencing error or/and SNP mutations. We created simulated libraries based on these complexities of community to estimate the efficiency of pathogen detection.

Simple Community

Metagenome sequences are often processed as a single genome alignment to a reference genome(s). In a single genome alignment, reads aligning to multiple loci to a reference genome are randomly assigned to a locus and SAMTools only parses these “main” hits. To compare this against our strategy to discern an organism among its close relatives with the all-against-all strategy between reads and reference genome, we utilized *S. aureus* TCH1516 to understand single infection, as Staphylococcus is well-characterized genus with high number of strains. This will allow us to test the specificity to detect correct organism not only among members of its own clade ST8-MRSA-IVa/USA300 but large number of well-characterized genome of *Staphylococcus* in reference genome. Typically the genus specific regions of *Staphylococcus* will be assigned to all the members of this taxonomic rank. The reads that will further contribute to unique regions, which belong to its species (aureus), and strain specific genes (clonal complex 5), will result in highest % genome coverage of correct organism. The single alignment coverage is performed with BWA – SAMTools – BEDTools against multiple genome RefSeq dataset, Metagenome alignment is performed with human read reduction and pathogen detection of pipeline against RefSeq dataset. The results are compared with actual coverage detected with BWA – SAMTools – BEDTools against single genome of *S. aureus* TCH1516 genome (Table 1).

Table 1. The comparison of percent genome coverage with increasing read number for the single genome alignment with metagenome alignment. * Genome coverage of top hit reported even if the organism detected is incorrect.

	100	**1K**	**10K**	**100K**	**250K**
Single Alignment Coverage	0.014*	0.059	0.62*	5.116	11.307
MetaGenome Coverage	0.352	3.557	30.126	97.004	99.983
Actual Coverage	0.348	3.427	29.494	96.955	99.982

We were also able to detect *S. aureus* TCH1516 in all the test sets as top hit even with very low number of reads while single genome alignment is not able to report correct detection. We found that the single alignment underestimate the genome coverage compared with the results from our pipeline (Table 1). The coverage detected by our approach was nearly equal to the actual coverage detected (actual coverage is slightly lower than the coverage detected by our pipeline as actual coverage is calculated from only BWA alignment while the pipeline calculate coverage from both BWA and BLAT alignment).

We also compared the effect on pathogen detection on number of reads, % genome coverage and read recall % with different features available for human filtration of the pipeline (Figure 4). Read recall % are the simulated reads that align back correctly to its reference genome after human filtration. As the read number increases, as expected genome coverage% also increases; approaching 99.9% at nearly 250K reads thus having coverage across the entire genome. 250K reads can be defined as approximate reads to reconstruct the entire genome of *S. aureus* TCH1516 from the metagenome.

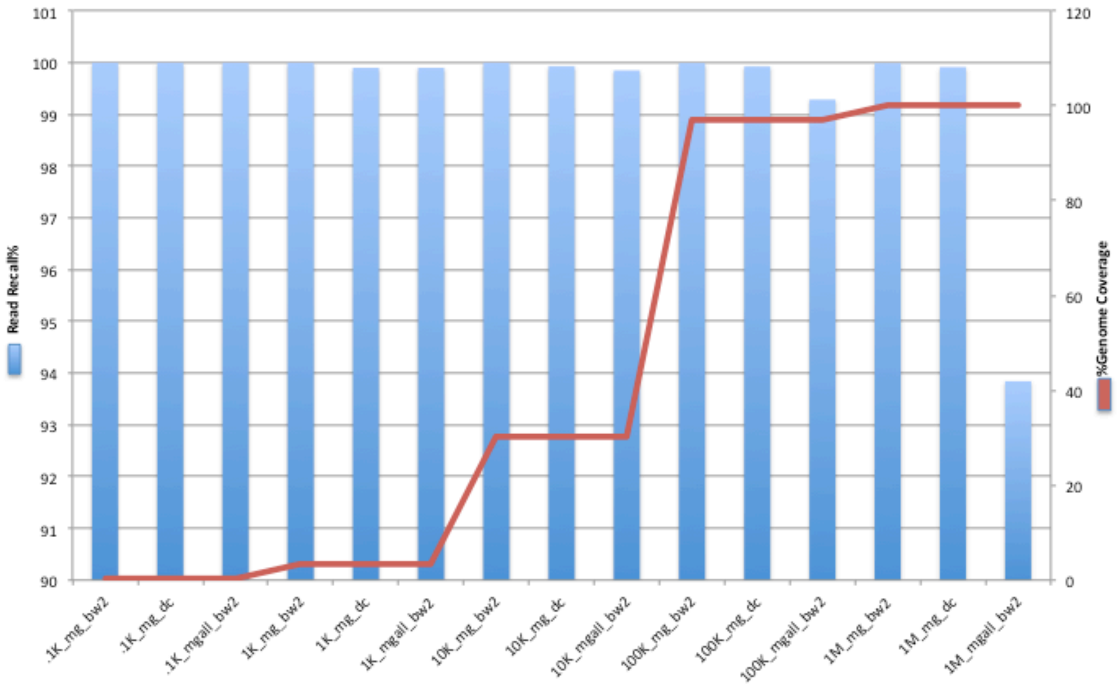


Figure 4. Effect of human filtration on percent genome coverage and read recall percentage of pathogen detection. The legends of the figure are prefixed with the number of reads (.1K=100; 1K=1000; 10K=10000; 100K=100000; 1M=1000000) followed by mg_bw2 for only fast alignment feature of human read reduction; mg_dc for all features of human read reduction except data compression; mgall_bw2 for all features of human read reduction module)

As more number of reads is sequenced (simulated), higher number of duplicate reads is also expected. Turning on data compression feature of human filtration of pipeline (*_mgall_bw2) to remove duplicates reduces the read recall% but has no effect on genome coverage% and detection of correct organism. The duplicate reads therefore do not add additional information and to manage computation scalability, removal of these duplicates is helpful. We also see that using all the human filtration steps of the pipeline (*_mgall-bw2) as compared to using just fast alignment (*_mg_bw2) or not utilizing data compression (*_mg_dc) does not lead to underestimation of %genome coverage for correct pathogen detection.

Complex Community

The ability to detect community in complex clinical samples such as cystic fibrosis might be helpful in generating insight for proper clinical recourse. The shared regions are expected in even the simplest microbial communities. The simulated library will allow us to evaluate the impact on detection due to presence of multiple organisms in community that have different genome size. We designed a complex community of five bacteria that is detected in one of the cystic fibrosis clinical sample. Simulated reads is generated from reference genome of each of the five organisms and four libraries with different read number (100, 1000, 10K, 100K per organism) is created. In metagenomes, many organisms might not have any complete or incomplete entries in the reference genome database. To test the specificity of detection with unknown organism, *V. dispar* ATCC 17748 that is not present in bacterial reference genome (RefSeq Build 60) is added in this complex community. Querying a large reference database usually results in detection of multiple organisms within same genus due to overlapping homology. Therefore for detection of organism, we selected the highest mapped genome (top-hit) within the same genus. The correct detection is confirmed for all the organisms except for *V. dispar* ATCC 17748. The pipeline allows detection of correct organisms even in complex community.

Different genera in complex community might share genomic regions. The robustness of detection can be measured by loss in sensitivity (genome coverage) of any organism in complex community versus single infection. We compared the %genome coverage of *E. coli* APEC O1 as single infection and in complex community. We found no difference and hence no loss in % genome coverage is reported and the *E. coli* APEC O1 in simple and complex community overlap completely (Figure 5). All against all relationship between the reads and reference database therefore allows us to detect any organism without any loss in sensitivity that might have shared genomic regions.

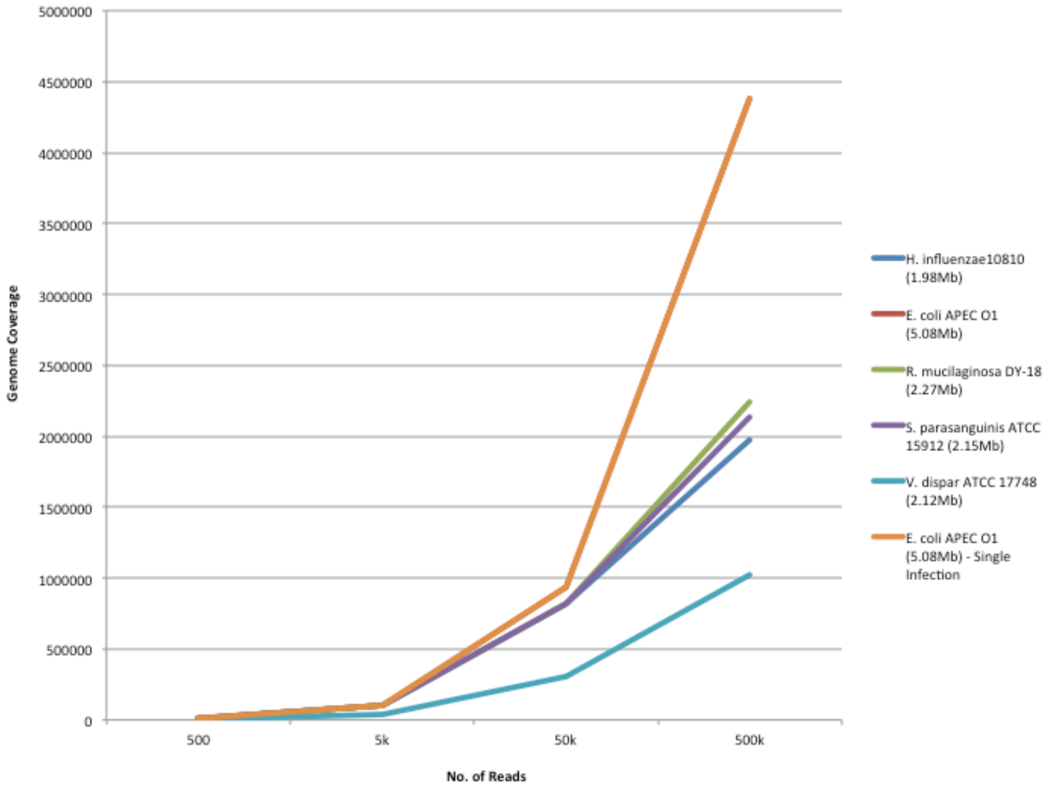


Figure 5. Detection of genomes in complex community. Relationship between genome size and genome coverage with increasing sequencing reads. Effect of detection on *E. coli* APEC O1 in simple and complex community.

E. coli APEC O1 has highest genome hit among this community, the increase in read number have higher impact on the increase in genome coverage (Figure 5). For simulated reads of *V. dispar* ATCC 17748 that is true negative, *V. parvula* DSM 2008 chromosome is detected as top hit with lower %genome coverage compared to other hits. We can infer that true calls may not always be possible, given the limited (albeit growing) nature of genomic databases and the resolution of detection might decrease.

Co-infections

We were able to accurately detect top hit per genus in complex community for known genomes in complex community as discussed above. However certain clinical samples might have pathogens from same species example S. *aureus* Newman and S. *aureus* TCH516. The *S. aureus* TCH1516 and S. *aureus* Newman belong to different clonal complex (CC8 & CC5) and resistant (MRSA) and sensitive to methicillin (MSSA) respectively. To test the specificity in these types of clinical samples, we created co-infection libraries consisting of simulated reads from S. *aureus* Newman and S. *aureus* TCH1516 genomes. The presence of *S. aureus* Newman in co-infection library (true positive) is compared with detection in single infection library (false positive) that only has simulated reads from *S. aureus* TCH1516 genome (Figure 6-A).

Any genome coverage% detected for *S. aureus* Newman in single infection library can be considered as false detection. The %genome coverage of *S. aureus* Newman (false call) is reported slightly less than its true presence in multiple-infection library, due to contribution of homologous reads from *S. aureus* TCH1516. However by top hit per genus, *S. aureus* Newman rank behind few other closely related genomes of *S. aureus* TCH1516 (CC5) in single infection library but is detected as top hit in co-infection library as shown in (Table 2).

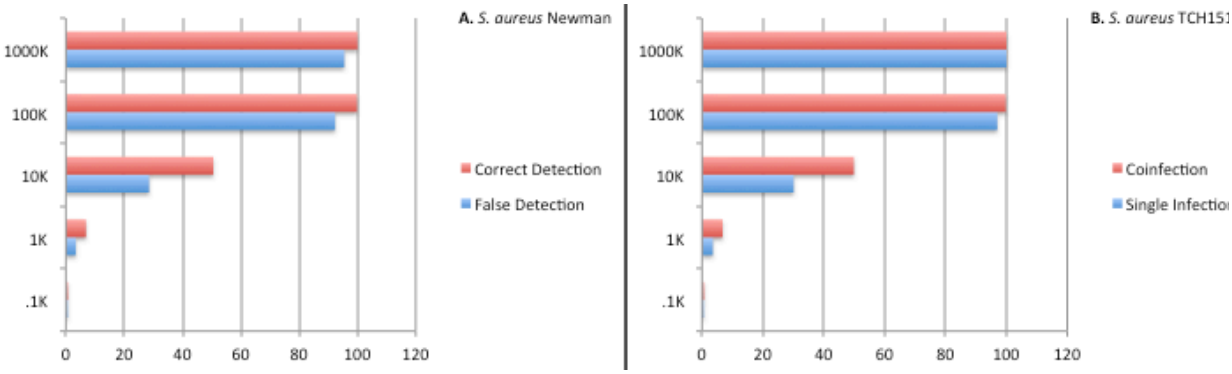


Figure 6. Comparison of detection of close relative in co-infection versus single infection. **A.** Comparison of %genome coverage of true detection in co-infection versus false detection of *S. aureus* Newman. **B.** Comparison of %genome coverage of *S. aureus* TCH1516 in co-infection versus simple infection.

A co-infection library consist of reads from both S. *aureus* TCH1516 and *S. aureus* Newman. Due to the all reads mapped against all reference strategy, the shared homology between these two organisms results in higher % genome coverage of *S. aureus* TCH1516 in co-infection library then single-infection library (Figure 6-B). The *S. aureus* TCH1516 is detected as top hit per genus in single infection library and as one of top hits with *S. aureus* Newman in co-infection library as represented in Table 2. We can infer that though detection as top hit per genus is correct in co-infection metagenome, however detection of co-infection is difficult and might require additional validation.

Table 2. The rank of *S. aureus* Newman and *S. aureus* TCH1516 in single infection and co-infection library. FP: False Positive; TP: True Positive.

	Ranking of Sa. Newman		**Ranking of Sa. TCH1516**	
Reads	**FP**	**TP**	**Single Inf (TP)**	**Co-Inf (TP)**
.1K	5	1	1	2
1K	4	1	1	3
10K	5	1	1	3
100K	5	1	1	2

Diversity

The metagenome reads might have artificial divergence due to sequencing error or variation incorporated due to selective mutation. The ability to assign these divergent reads back to its genome can allow the sensitivity for detection. Local alignment algorithms are considered to be more sensitive and accurate than global alignment algorithms. However, with the high number of metagenome reads, local aligner like BLAST is not optimal given the high computational time cost. Existing global aligners such as BWA, STAMPY, BOWTIE, BFAST, MAQ, NOVOALIGN, MOSAIK and SOAP2 are preferred over local alignments given high volumes of metagenome sequences. However utilizing only global aligner might result in loss of detection of divergent reads. To incorporate these divergent reads for sensitive detection, we utilized BLAT that is ~500 times faster than preexisting tools with comparable sensitivity.

We designed the simulated reads from *S. aureus* TCH1516 genome with increasing divergence (See Benchmark above). To evaluate the sensitivity, reads that global aligner is unable to map but are aligned by local aligner (BLAT) are categorized as divergent reads. With increasing divergence, higher number of reads is not aligned by global aligner (Figure 7).

The pipeline is able to incorporate these divergent reads through local alignment without decrease in the genome coverage for detection (Figure 7). In all the 25 simulated test case (0%, 0.1%, 0.2%, 0.5% and 1% divergence for 100, 1K, 10K, 100K, 250K reads), *S. aureus* TCH1516 is detected correctly in all except one, at 1% divergence with 100 reads. The limitation of detection for correct identification can therefore be seen at highest divergence with low number of reads.

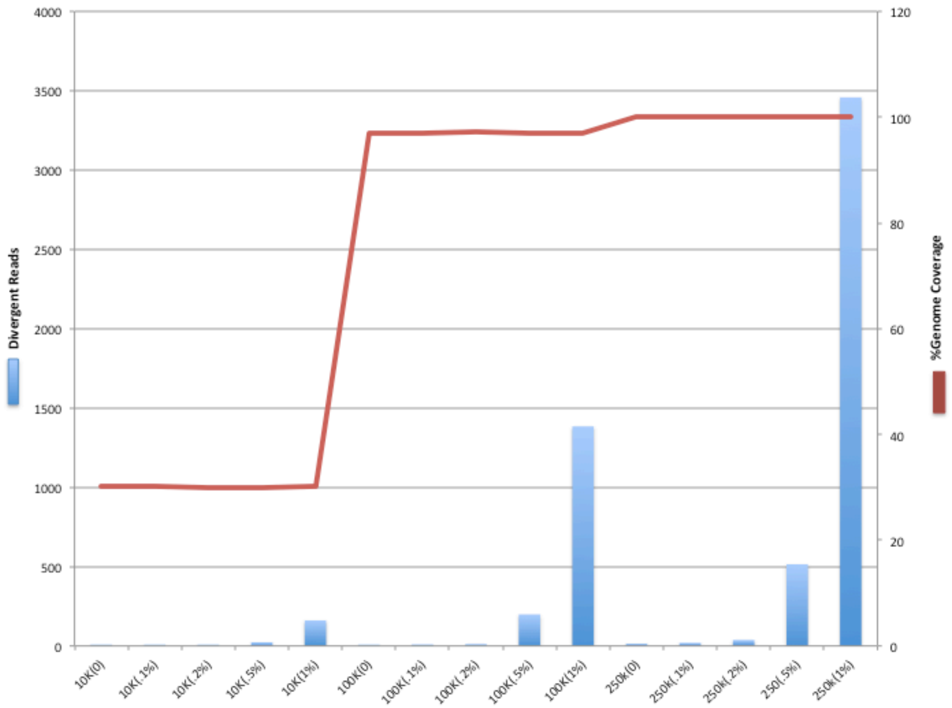


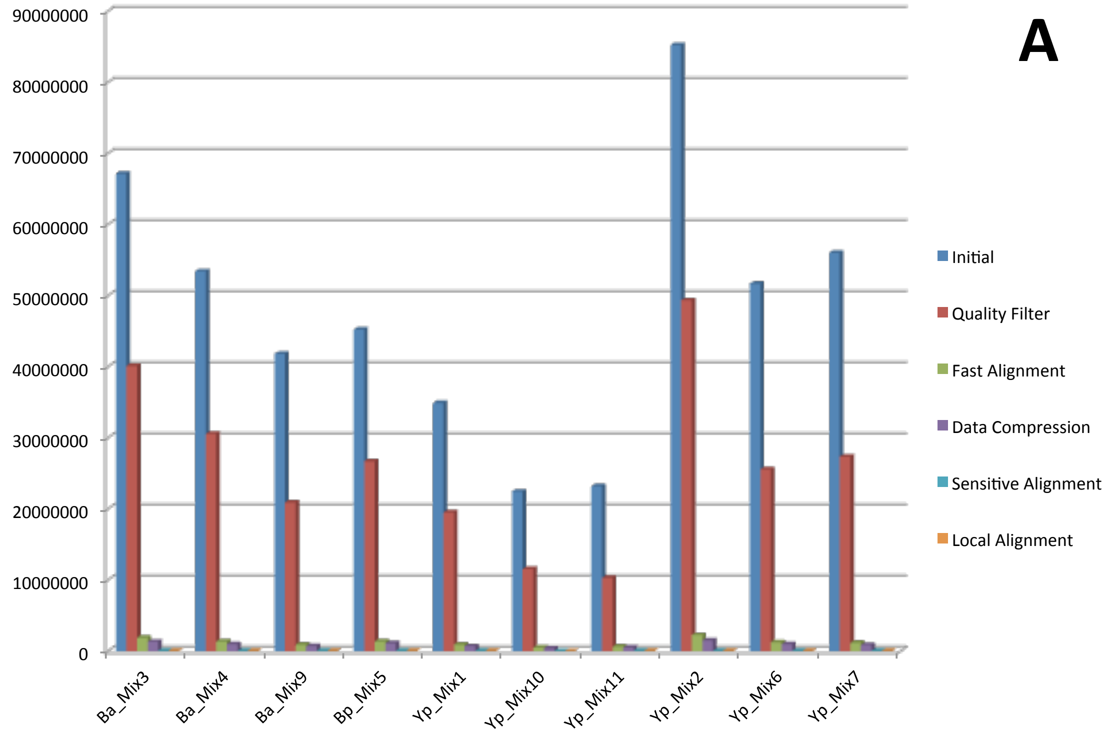
Figure 7. Relationship between percent genome coverage and read recall percentage with incremental divergence.

Analysis Overview

Modifications to MetaGeniE including full automation, better workflow design, and fixes in scalability issues allowed us to rapidly and efficiently analyze 96 deep sequencing clinical samples from nine different projects (see clinical sequencing section below). This adds to analysis of over 5 billion sequence reads requiring 3 terabytes storage of sequence data. Nearly all of these datasets had a runtime of less than 24 hours and detected pathogens with high sensitivity and specificity as discussed in the analysis section. We think that further decreases in the computational runtime can be achieved by designing MetaGeniE with OpenMPI, a task we plan on initiating within the next year but is not required to be completed by the end of this current project.

Read Reduction

The first module of MetGeniE is Read-Reduction, a five-step process that removes low quality, redundant and human reads from sequence data (Figure 8A). The reads are iteratively mapped and only the unmapped reads remaining after each alignment step are processed as input to next step. The first step removal of low quality reads led to removal of ~40%-51% for 10 samples and ~31%-56% for the other 9 samples.



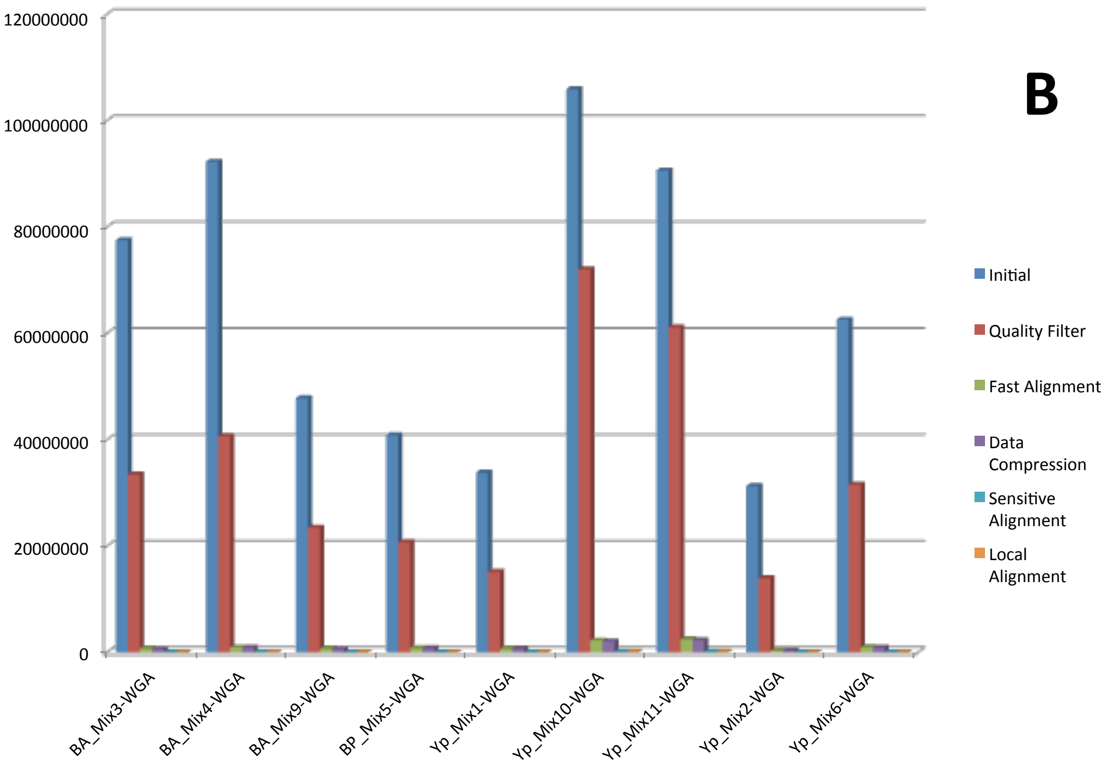
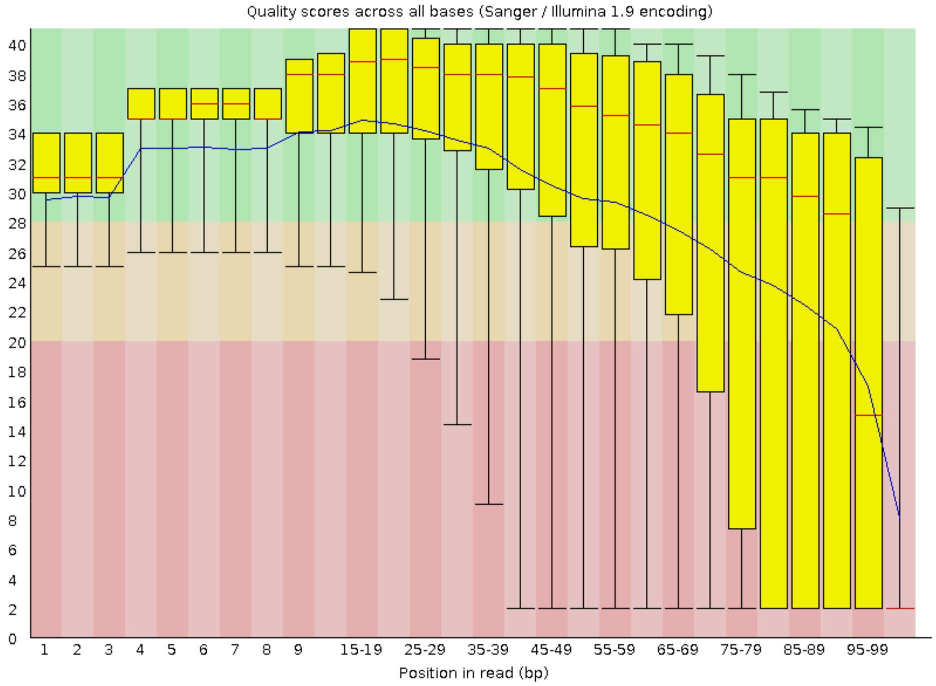


Figure 8. Iterative Reduction of the metagenome reads with Read-Reduct module of MetGeniE A. No whole genome amplification, B. Whole genome amplification reads

These low quality reads were observed to be much higher than other metagenomic samples that have been analyzed thus far.

The filtered (remaining) reads with high quality were aligned against human reference genome with BWA (Figure 8B). We detail a subset of the total samples analyzed; ~95% for 10 bacterial samples and ~97% for the 9 WGA samples aligned to human reference genome and were not considered for further analysis (i.e. removed from the analysis). The third data compression step resulted in removal of ~16%-32% of 10 samples and ~2.5%-9% of the 9 WGA samples. The filtered reads are aligned against human reference genome with STAMPY, an aligner that is considered sensitive. The more sensitive alignment resulted in reduction of ~93%-98% of these remaining reads for all 19 samples. Finally the remaining reads were aligned against human repeat database with help of local aligner BLAT and ~1%-4% of the reads were removed. The total number of remaining reads left after five-stage filtering ranged from 12,200 reads to 48,862 reads (SD ± 10,934) for 10 bacterial samples. For the WGA samples, remaining reads ranged from 12,254 reads to a maximum of 107,099 reads (SD ± 30,841). These remaining (filtered) reads were utilized for pathogen detection.

To analyze the anomaly of high number of low quality reads, we performed quality summarization with FASTQC (Figure 9). The box-plot representation of the quality is generated separately for the reads that were discarded and the reads that passed the quality filter were used for further detection. These low quality reads were observed to be much higher than other metagenomic samples that have been analyzed so far. To analyze this anomaly, we performed quality summarization with FASTQC. The box-plot representation of the quality is generated separately for the reads that were discarded and the reads that passed the quality filter were used for further detection.



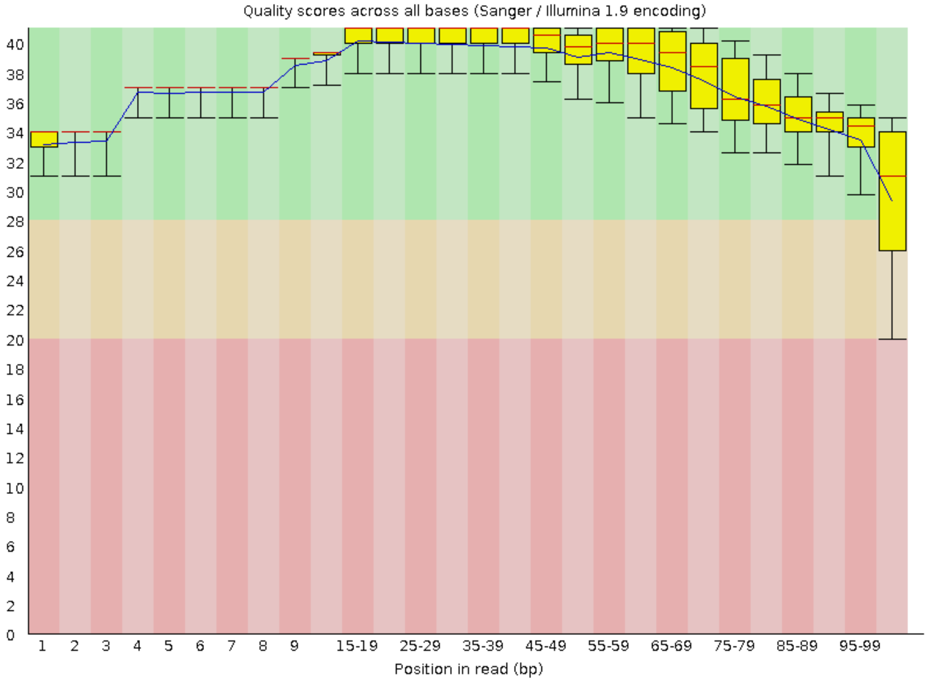


Figure 9. Quality summary per position. A. Discarded reads in yellow. B. Remaining reads that were further utilized for detection. The discarded reads have low quality per position (<15 threshold) as compared to the filtered reads that have high quality (median quality >32).

Detection of Bacteria

MetGeniE detects the presence of microbes by performing fast local and global alignments with BWA and BLAT respectively against the bacterial Refseq database (Figure 10). Results were that ~1.2%-64.7% and ~1.3%-7.2% aligned with BWA and BLAT, respectively, against the Refseq database for the 10 samples. Results for the 9 WGA samples were that ~0.4%-42% and 0.5%-4% aligned with BWA and BLAT, respectively, against the Refseq database. The total number of reads that were unknown, i.e. undetected by MetGeniE, is less than ~0.05% for all 19 samples.

MetaGeniE detects the bacterial presence by genome reconstruction by aggregating aligned reads (from Patho-Detect) of each individual genome. In each case, MetGeniE accurately identified each sample and its mixture to the correct species, except for *B. pseudomallei* being identified as *B. mallei* in one instance. This is understandable since *B. mallei* emerged as a clonal lineage from within *B. pseudomallei* so they share much of the genome composition with each other. Thus we have results indicating that at first pass, MetGeniE does quite well in species assignment in pure samples as well as mixed samples. MetGeniE detected one or more strains from different species in the mixture with different % genome. We observed that as the % genome gets <2%, the resolution for detection decreases and different strains (from same species) were difficult to discern. Different samples were mixed with different concentration of DNA for different species that is subjected to blind test. The accuracy of this approach is still being evaluated.

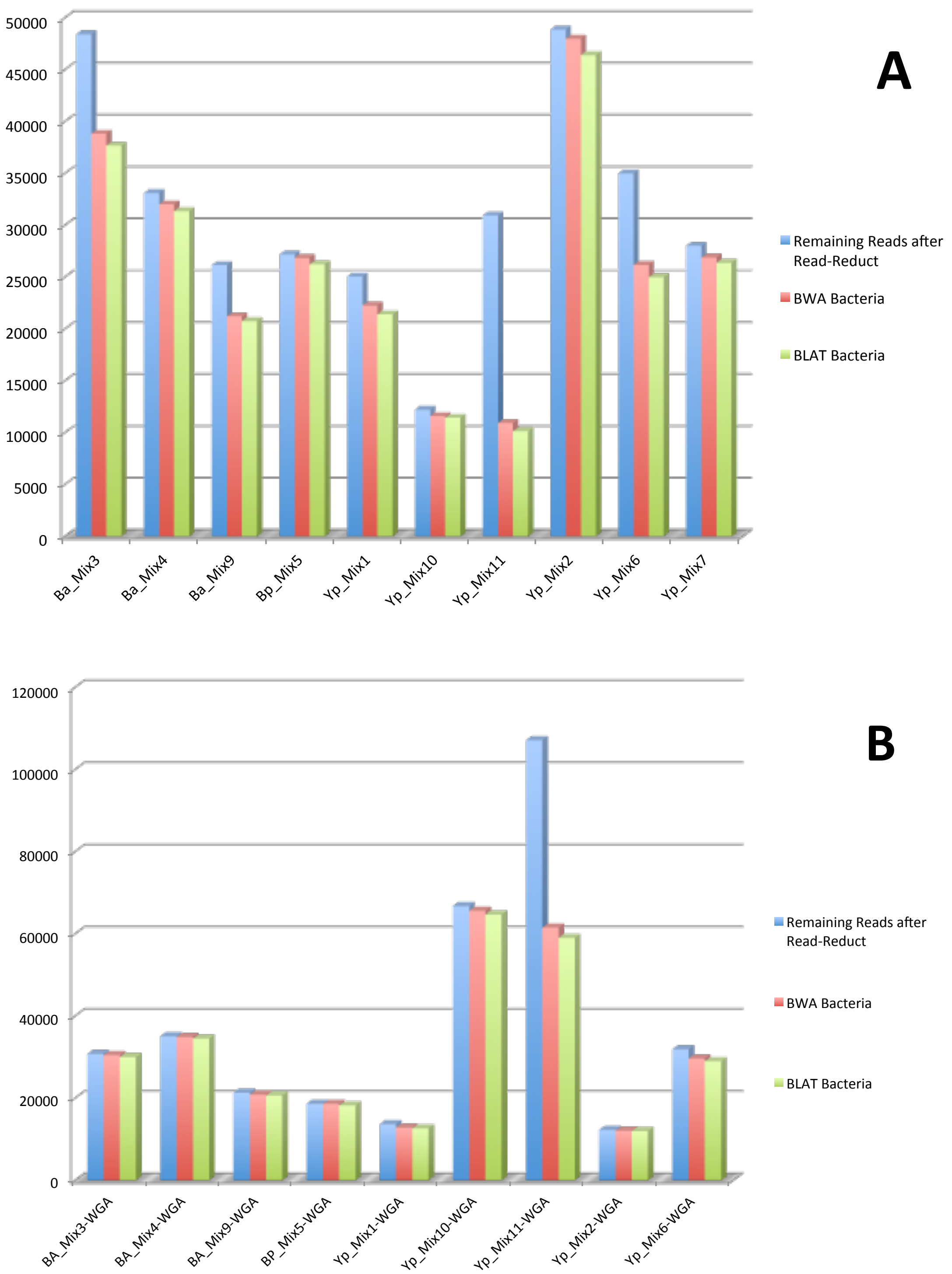


Figure 10. The bacterial identification with Patho-Detect module of MetGeniE. Figures are A. No whole genome amplification, B. Whole genome amplification reads

Details of Tasks and Task-specific Results

In silico bioinformatic testing

1. **Create filters to remove reads of human DNA**
2. **Run human genome reads against pathogen reference sequences**

Design of Human Read Reduction Module of MetaGeniE

As detailed above, MetaGeniE is designed as distributed and scalable software to analyze millions of reads and query large reference databases and consists of two modules: Read-Reduct and Patho-Detect. The Read-Reduct module sequentially filters and reduces the low quality, redundant and human reads (Figure 1). The low quality reads are filtered with the help of PRINSEQ. Human read filtration can be performed with the short read aligners, these are classified into Burrows-Wheeler Transform (BWT) and hash-based mappers. The BWT mappers such as BWA, SOAP2 and Bowtie are fast but considered less sensitive, while the hash-based aligners are slow but more accurate such as MAQ, ELAND, Novoalign and STAMPY. To reduce overall computational processing time and memory, faster analysis BWT aligner BWA or BOWTIE2 are utilized initially. Higher CPU and memory intensive features like data compression and hash-based sensitive alignment STAMPY are then utilized to further reduce the overall number of reads.

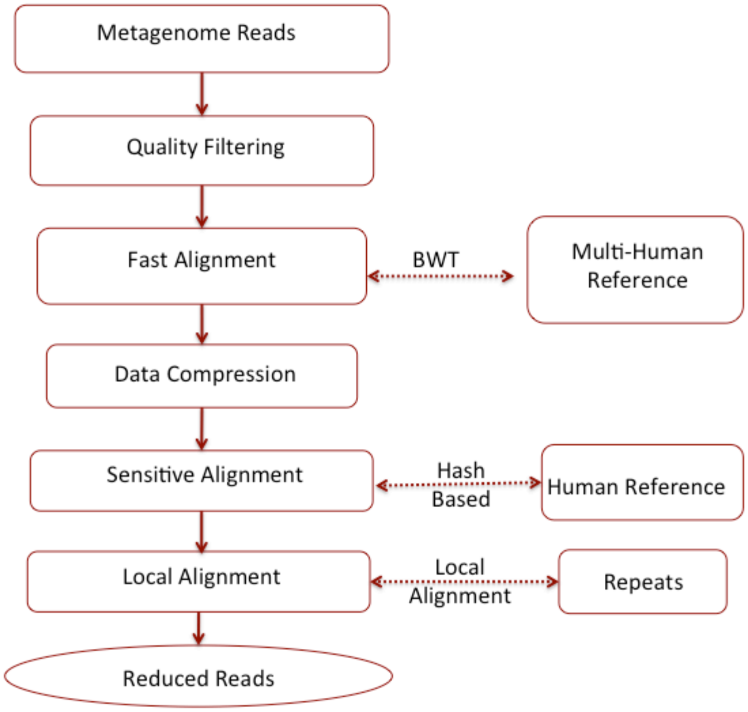


Figure 1. The workflow of Human read reduction module the pipeline

Benchmark

Human filtration is five step process and different steps (Table 1) are utilized to compare the sensitivity/specificity of detection of simulated reads, execution speed and memory usage.

Table 1. Description of different steps of human filtration of pipeline utilized to compare sensitivity/specificity of detection and performance of runtime and computational resources of the simulated reads.

	Quality Filter	**Fast Alignment**	**Data Compression**	**Sensitive Alignment**	**Repeat DB**
mg_bw2	yes	bowtie2	-	-	-
mg_bwa	yes	bwa	-	-	-
mg_dc	yes	bwa	-	stampy	yes
mgall_bw2	yes	bowtie2	Yes	stampy	yes
mgall_bwa	yes	bwa	Yes	stampy	yes

To test the sensitivity of the efficiency of human filtration with Read-Reduct module, we utilized only human reads to test the sensitivity to remove human reads (Table 2). Seven whole genome sequencing of human datasets were downloaded from SRA (http://www.ncbi.nlm.nih.gov/sra/). We simulated 30 million reads from human reference genome build 37.2 (ftp://ftp.ncbi.nih.gov/genomes/H_sapiens) with GRINDER version 0.5.3. These simulated human reads were 100 bases and were incorporated with total 0.5% variability, 0.01% as expected human SNP frequency and 0.4% as the average sequencing error for Illumina reads.

Table 2. Description of different human datasets for benchmarking human filtration.

Name	**Read Numbers**	**Source**
ERR191896	53.03 million	SRA
ERR218094	49.50 million	SRA
ERR237515	2.54 million	SRA
SRR032752	35.29 million	SRA
SRR033605	23.53 million	SRA
SRR054743	40.63 million	SRA
SRR054753	39.76 million	SRA
Simult_Hg19	30 million	Simulated datasets

All these steps utilized the same parameters as follows, BWA (default), BOWTIE2 (default with very sensitive mode), STAMPY (default), PHRED quality score > 15, minimum length >50, low complexity (dust) and BLAT (80% identity).

Results

The efficiency of human read filtration can be measured by total number of human reads removed from clinical samples. To test the effect on human read filtration with different parameters, these eight databases were filtered against human reference genome (Hg19) with MetaGeniE. The mg_bwa and mg_bw2 uses only BWA and Bowtie2 aligner only while mgall_bw2 and mgall_bwa uses all five steps of pipeline including fast alignment with BOWTIE2 and BWA respectively. We found that removing human reads with single aligner (mg_bwa/mg_bw2) might not be efficient in removing human reads. Utilizing all the features of the pipeline (mgall_bw2 and mgall_bwa) allowed higher filtration of human reads (Figure 2).

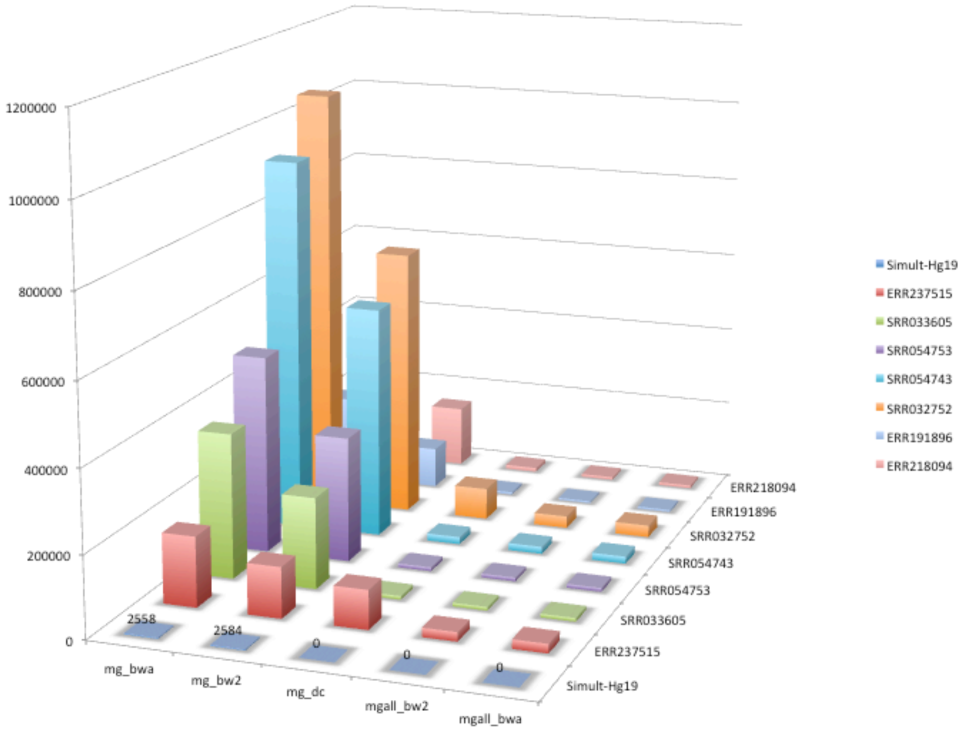


Figure 2. Total numbers of reads remaining after human read reduction with different filtration parameters.

The runtime of single step (mg_bwa/mg_bw2) is usually faster than running all steps of human filtration (mgall_bwa/mgall_bw2) (Figure 3). Keeping all constraints same, we found that BWA aligner runs faster than BOWTIE2 while BOWTIE is more sensitive than BWA as it aligns higher number of human reads (Figure 2). However the total number of reads removed by mgall_bw2 and mgall_bwa is nearly equal irrespective of whether BWA or BOWTIE2 aligner is used.

Remaining human reads that were not filtered were aligned against the NCBI Refseq bacterial database. These human reads mapped to the bacterial database and as expected were higher for single step alignment (*_mg_bw2) that have higher number of human reads than comprehensive reduction with pipeline (*_mgall_bw2) (Figure 4). The unfiltered human reads not only map incorrectly to microbial datasets, but also contribute to overall runtime during pathogen detection. Removal of human reads with high specificity is advantageous for sensitive clinical interpretation.

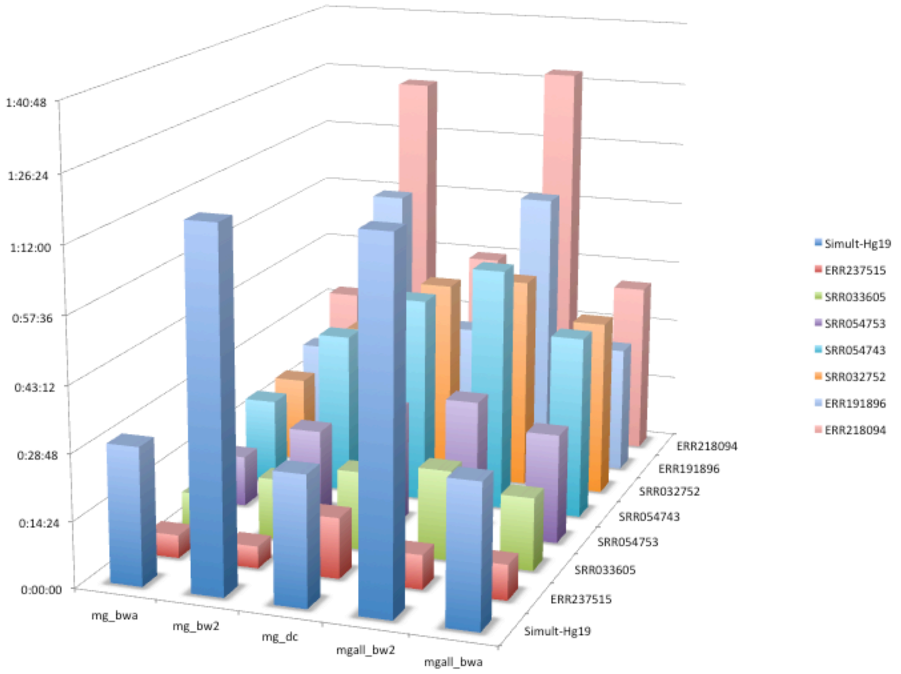


Figure 3. Runtime for human read filtration with different aligner and filtration parameters (in minutes).

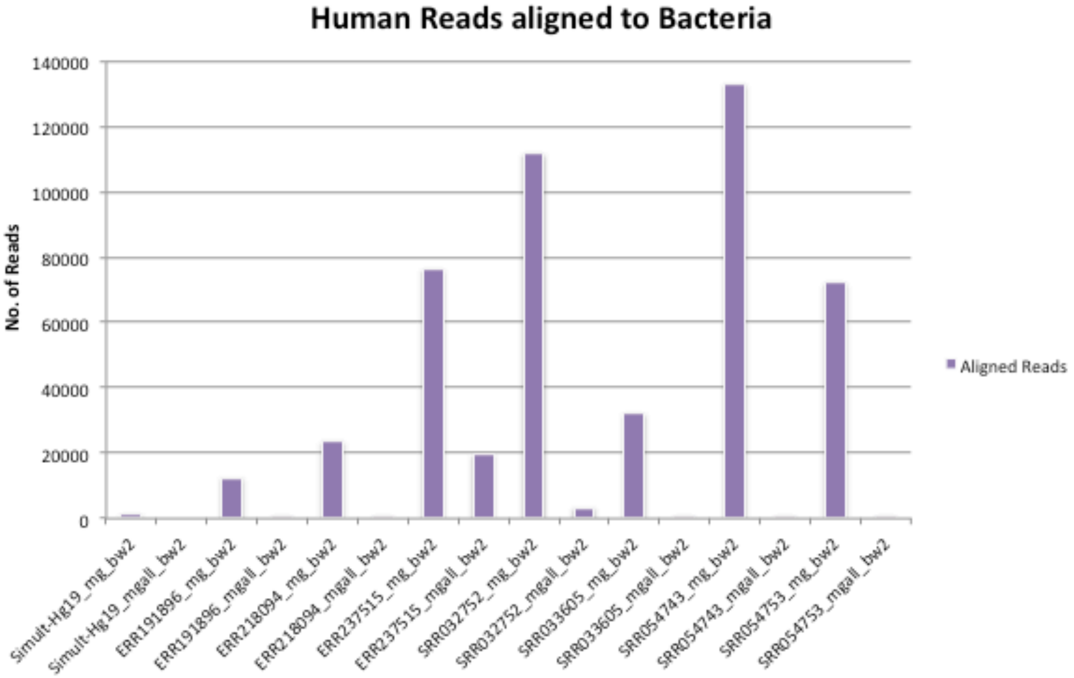


Figure 4. Total numbers of reads aligning to bacterial database after human read filtration. The legends are prefixed with sample name (Simult-Hg19, ERR191896, ERR218094, ERR237515, SRR032752, SRR033605, SRR054743, SRR054753) followed by mg_bw2 and mgall_bw2 that represent utilizing single aligner and multiple features of the pipeline.

Human DNA subtraction

In addition to *in silico* removal of human DNA reads, we evaluated wet bench approaches to remove human DNA prior to library preparation as well. Our initial focus hybridization or enzymatic reduction of human DNA within a sample. As this will increase the number of subsequent sequence reads originating from non-human sources within the sample, this should increase the sensitivity of detecting and characterizing the microbial content of the sample. A number of commercially available kits that promise to selectively enrich bacterial DNA from blood cultures are available. Most rely on differential lysis of the blood cells with subsequent degradation of the human DNA while the bacterial cells remain intact, then the bacterial cells are lysed. One such kit from Molzyme® has been tested here at TGen North on past projects and was found to remove significant amounts of bacterial DNA as well as the human DNA, and so was deemed unsatisfactory. We first attempted the development of a lysis independent human DNA subtraction method that can be applied directly to DNA solutions. Starting with a method under development by researchers at Sandia National Labs for a microfluidics application, subtraction probes are prepared by biotin labeling fragmented DNA, in this case human, which are then hybridized to a complex target sample. Human sequences in theory will hybridize to the biotin-labeled probes, and then be removed from the rest of the sample via streptavidin-labeled magnetic beads. Extensive development and evaluation indicated only minimal removal human DNA from mixed samples. As such, we decided to start working with a commercially available kit. Metagenomic DNA extracted from (sample origins) of four *Francisella tularensis* clinical cases was quantified using the Nanodrop 1000 spectrophotometer. Based on these quantifications, 1-2 μg of total DNA was enriched for microbial DNA using the NEBNext Microbiome DNA Enrichment Kit (New England Biolabs, Ipswich, Massachusetts, USA) according to the supplied protocol. This protocol captures DNA with CpG methylation, which is prominent in eukaryotic DNA and rare in prokaryotic DNA, and removes it from the metagenomic sample, effectively enriching microbial DNA.

Preliminary evaluation with quantitative PCR analysis of this subtraction protocol has resulted in up to 99% removal of human DNA alu qPCR assay target fragments. Upon post subtraction sequencing, we still see human reads present in the sequencing libraries, but their abundance is significantly reduced, providing for much more sensitive pathogen target discovery. This approach has broad applications beyond just clinical sequencing and includes such needs as removing host DNA from cell culturing work where microbial DNA needs to be separated from such vertebrate cell lines.

While we started this work early in the first year of the project, optimal results were not achieved until the last quarter of our work where we were able to evaluate our methods using DNA from four *Francisella tularensis* clinical cases (FT1, FT2, FT6, FT13).

Sequencing Library Preparation

Both native and subtracted DNA samples were prepared for multiplexed Illumina sequencing using the KAPA Illumina series library preparation kit (Kapa Biosystems, Wilmington, Massachusetts, USA) with adapter and index primer oligos as described in Kozarewa and Turner (2011). Libraries were size-selected using the dual-SPRI scheme described by Lundin et al. (2010) to achieve an average insert size of 700 bp. Sequencing was performed on an Illumina HiSeq 2000.

Human Subtraction Protocol Analysis

For the four *F. tularensis* samples, human DNA was subtracted using the CpG methylation protocol. We compared these four subtracted samples with the normal samples (Figure 5). We found that the subtracted human sequencing reduces the total percentage of human reads as against the normal sequencing from ~1% to ~6%.

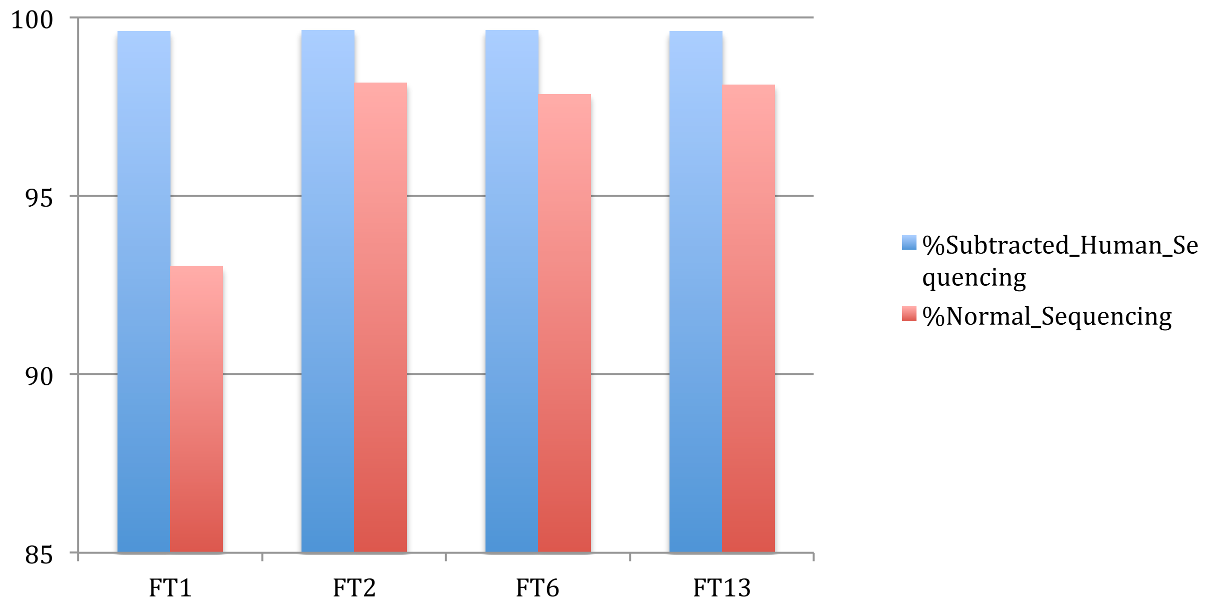


Figure 5. Comparison of subtracted human protocol sequencing versus normal sequencing.

We also found that the total number of reads aligning to bacterial reads significantly increased by factor of ~5 to ~90 (Figure 6). Comparing the community between the subtracted human (host) sequencing and normal sequencing, microbial detection was quite different. For normal sequencing, we were able to detect *Francisella tularensis* as major component of infection, for these patients suffering from tularensis besides other organisms like *Ralstonia picketti*. However comparing the distribution of population with the human subtracted sequencing, top hits were *Herbaspirillum seropedicae* SmR1 for Sample 1 and 2 and *Cupriavidus metallidurans* CH34 for Sample FT6 and FT13 (Table 3). *Francisella tularensis* is also detected but the genome coverage % was higher for two samples and lower for other two samples.

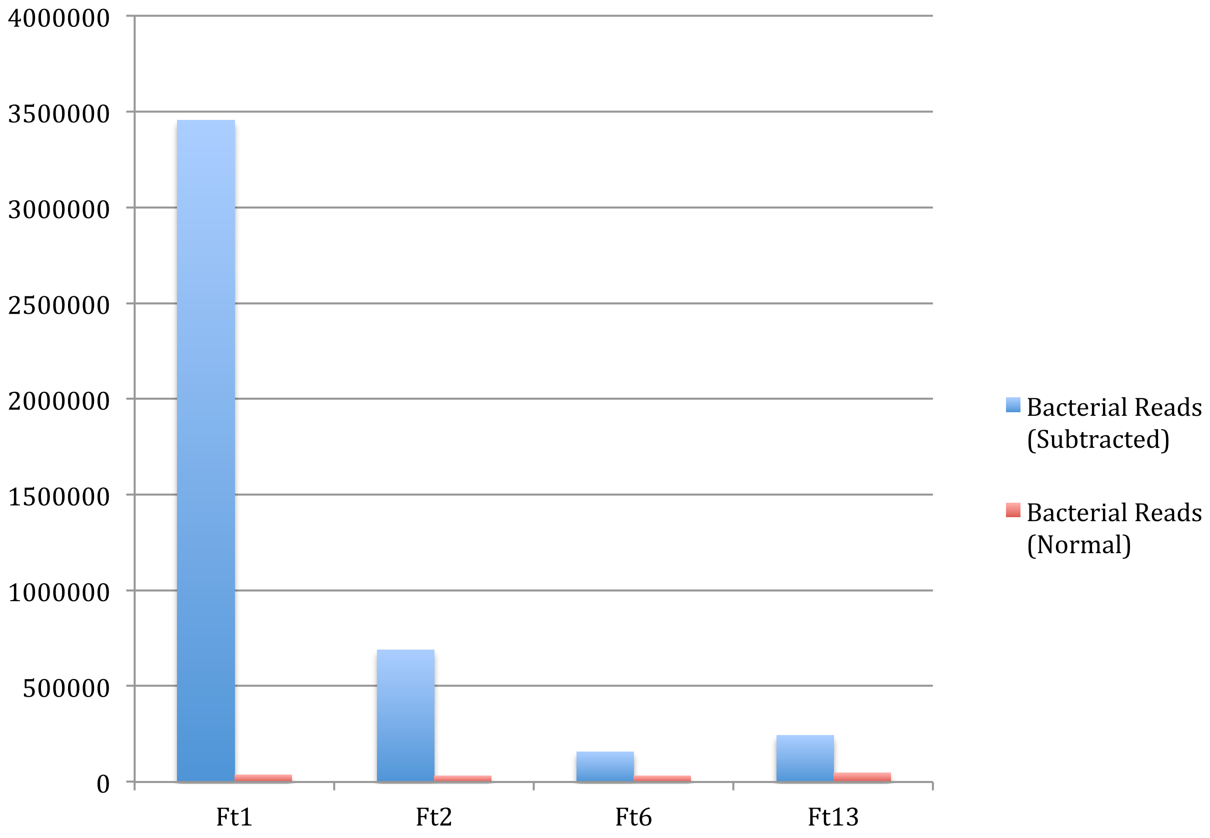


Figure 6. Comparison of bacterial read alignment of subtracted human protocol sequencing versus normal sequencing.

Table 3. Top metagenomic hits for microbial discovery based on genome coverage for four *Francisella tularensis* clinical cases (FT1, FT2, FT6, FT13).

Subtracted Human Sequencing (FT1)

Organism	Genome Coverage	Genome Coverage (%)	Genome Size
Herbaspirillum_#_seropedicae_!_SmR1_{}_|gi|300309346|ref|NC_014323.1|	2870146	52.053	5513887
Collimonas_#_fungivorans_!_Ter331_{}_|gi|340785197|ref|NC_015856.1|	386377	7.449	5186898
Janthinobacterium_#_sp._!_CG3_{_JANGC3DRAFT1.1_}|gi|484156181|ref|NZ_KB467824.1|	267953	4.829	5549265
Delftia_#_acidovorans_!_SPH-1_{}_|gi|160895450|ref|NC_010002.1|	220753	3.262	6767514
Ralstonia_#_eutropha_!_H16_{}_1_|gi|113866031|ref|NC_008313.1|	220577	5.444	4052032
Achromobacter_#_xylosoxidans_!_A8_{}_|gi|311103224|ref|NC_014640.1|	220474	3.144	7013095
Cupriavidus_#_necator_!_N-1_{}_1|gi|339324158|ref|NC_015726.1|	212406	5.484	3872936
Ralstonia_#_solanacearum_!_CMR15_{}_|gi|523408232|ref|NC_017559.1|	206756	5.75	3596030
Delftia_#_sp._!_Cs1-4_{}_|gi|333911667|ref|NC_015563.1|	204803	3.063	6685842
Cupriavidus_#_taiwanensis_!_LMG_{_19424_}1|gi|188590795|ref|NC_010528.1|	201059	5.884	3416911
Francisella_#_tularensis_!_subsp._{_holarctica_}_LVS|gi|89255449|ref|NC_007880.1|	200591	10.58	1895994
Francisella_#_tularensis_!_subsp._{_holarctica_}_FTNF002-00|gi|156501369|ref|NC_009749.1|	200154	10.585	1890909
Francisella_#_tularensis_!_subsp._{_holarctica_}_FSC200|gi|422937995|ref|NC_019551.1|	199713	10.544	1894157
Ralstonia_#_solanacearum_!_Po82_{}_|gi|386331671|ref|NC_017574.1|	197334	5.669	3481091
Francisella_#_tularensis_!_subsp._{_holarctica_}_OSU18|gi|115313981|ref|NC_008369.1|	196393	10.36	1895727
Cupriavidus_#_metallidurans_!_CH34_{}_|gi|94308945|ref|NC_007973.1|	193310	4.921	3928089
Ralstonia_#_solanacearum_!_GMI1000_{}_|gi|17544719|ref|NC_003295.1|	192265	5.173	3716413
Ralstonia_#_solanacearum_!_IPO1609_{_|gi|207741818|ref|NW_002196569.1|	188862	5.599	3372855
Francisella_#_tularensis_!_subsp._{_holarctica_}_F92|gi|423049750|ref|NC_019537.1|	188158	9.972	1886888
Ralstonia_#_solanacearum_!_CFBP2957_{}_|gi|300702374|ref|NC_014307.1|	184200	5.39	3417386

Normal Human Sequencing (FT1)

Organism	Genome Coverage	Genome Coverage (%)	Genome Size
Francisella_#_tularensis_!_subsp._{_holarctica_}_LVS|gi|89255449|ref|NC_007880.1|	42495	2.241	1895994
Francisella_#_tularensis_!_subsp._{_holarctica_}_FTNF002-00|gi|156501369|ref|NC_009749.1|	42254	2.235	1890909
Francisella_#_tularensis_!_subsp._{_holarctica_}_FSC200|gi|422937995|ref|NC_019551.1|	42188	2.227	1894157
Francisella_#_tularensis_!_subsp._{_holarctica_}_OSU18|gi|115313981|ref|NC_008369.1|	42162	2.224	1895727
Francisella_#_tularensis_!_subsp._{_holarctica_}_F92|gi|423049750|ref|NC_019537.1|	40979	2.172	1886888
Francisella_#_tularensis_!_TIGB03_{}_|gi|379716390|ref|NC_016933.1|	38223	1.942	1968651
Francisella_#_tularensis_!_subsp._{_tularensis_}_NE061598|gi|385793751|ref|NC_017453.1|	35758	1.889	1892681
Francisella_#_tularensis_!_subsp._{_tularensis_}_SCHU_S4|gi|255961454|ref|NC_006570.2|	35758	1.889	1892775
Francisella_#_tularensis_!_subsp._{_tularensis_}_FSC198|gi|110669657|ref|NC_008245.1|	35707	1.887	1892616
Francisella_#_tularensis_!_subsp._{_tularensis_}_TI0902|gi|379725073|ref|NC_016937.1|	35707	1.887	1892744
Francisella_#_tularensis_!_subsp._{_tularensis_}_WY96-3418|gi|134301169|ref|NC_009257.1|	35445	1.867	1898476
Francisella_#_tularensis_!_subsp._{_mediasiatica_}_FSC147|gi|187930913|ref|NC_010677.1|	29870	1.577	1893886
Francisella_#_novicida_!_U112_{}_|gi|118496615|ref|NC_008601.1|	16928	0.886	1910031
Francisella_#_cf._!_novicida_{_Fx1_}|gi|385791932|ref|NC_017450.1|	15216	0.795	1913619
Francisella_#_tularensis_!_subsp._{_holarctica_}_2571.19|gi|254367826|ref|NZ_DS229056.1|	12257	2.205	555807
Streptomyces_#_griseoflavus_!_Tu4000_{_genomic_}1.1|gi|224581108|ref|NZ_GG657758.1|	8778	0.109	8047042
Francisella_#_tularensis_!_subsp._{_tularensis_}_FSC0331.4|gi|254370255|ref|NZ_DS264119.1|	8131	1.032	788151
Francisella_#_tularensis_!_subsp._{_novicida_}_FTE_FTE1|gi|224580220|ref|NZ_DS989818.1|	7110	0.837	849109
Francisella_#_tularensis_!_subsp._{_holarctica_}_FSC0221.8|gi|254369615|ref|NZ_DS264140.1|	7070	2.139	330502
Francisella_#_novicida_!_FTG_{_FTG1_}|gi|224580234|ref|NZ_DS995363.1|	6286	0.893	703773

Subtracted Human Sequencing (FT2)

Organism	Genome Coverage	Genome Coverage (%)	Genome Size
Herbaspirillum_#_seropedicae_!_SmR1_{}_|gi|300309346|ref|NC_014323.1|	3039047	55.116	5513887
Collimonas_#_fungivorans_!_Ter331_{}_|gi|340785197|ref|NC_015856.1|	455636	8.784	5186898
Janthinobacterium_#_sp._!_CG3_{_JANGC3DRAFT1.1_}|gi|484156181|ref|NZ_KB467824.1|	316372	5.701	5549265
Ralstonia_#_eutropha_!_H16_{}_1_|gi|113866031|ref|NC_008313.1|	261011	6.441	4052032
Achromobacter_#_xylosoxidans_!_A8_{}_|gi|311103224|ref|NC_014640.1|	257003	3.665	7013095
Delftia_#_acidovorans_!_SPH-1_{}_|gi|160895450|ref|NC_010002.1|	252615	3.733	6767514
Cupriavidus_#_necator_!_N-1_{}_1|gi|339324158|ref|NC_015726.1|	247458	6.389	3872936
Francisella_#_tularensis_!_subsp._{_holarctica_}_LVS|gi|89255449|ref|NC_007880.1|	245618	12.955	1895994
Francisella_#_tularensis_!_subsp._{_holarctica_}_FSC200|gi|422937995|ref|NC_019551.1|	245565	12.964	1894157
Ralstonia_#_solanacearum_!_CMR15_{}_|gi|523408232|ref|NC_017559.1|	243872	6.782	3596030
Francisella_#_tularensis_!_subsp._{_holarctica_}_FTNF002-00|gi|156501369|ref|NC_009749.1|	243843	12.896	1890909
Francisella_#_tularensis_!_subsp._{_holarctica_}_OSU18|gi|115313981|ref|NC_008369.1|	242455	12.79	1895727
Cupriavidus_#_taiwanensis_!_LMG_{_19424_}1|gi|188590795|ref|NC_010528.1|	240282	7.032	3416911
Ralstonia_#_solanacearum_!_Po82_{}_|gi|386331671|ref|NC_017574.1|	236551	6.795	3481091
Francisella_#_tularensis_!_subsp._{_holarctica_}_F92|gi|423049750|ref|NC_019537.1|	233405	12.37	1886888
Delftia_#_sp._!_Cs1-4_{}_|gi|333911667|ref|NC_015563.1|	232974	3.485	6685842
Ralstonia_#_solanacearum_!_GMI1000_{}_|gi|17544719|ref|NC_003295.1|	228347	6.144	3716413
Francisella_#_tularensis_!_TIGB03_{}_|gi|379716390|ref|NC_016933.1|	224368	11.397	1968651
Ralstonia_#_solanacearum_!_IPO1609_{_|gi|207741818|ref|NW_002196569.1|	222563	6.599	3372855
Ralstonia_#_solanacearum_!_CFBP2957_{}_|gi|300702374|ref|NC_014307.1|	215864	6.317	3417386

Normal Human Sequencing (FT2)

Organism	Genome Coverage	Genome Coverage (%)	Genome Size
Francisella_#_tularensis_!_subsp._{_holarctica_}_FSC200|gi|422937995|ref|NC_019551.1|	72600	3.833	1894157
Francisella_#_tularensis_!_subsp._{_holarctica_}_FTNF002-00|gi|156501369|ref|NC_009749.1|	72361	3.827	1890909
Francisella_#_tularensis_!_subsp._{_holarctica_}_LVS|gi|89255449|ref|NC_007880.1|	71914	3.793	1895994
Francisella_#_tularensis_!_subsp._{_holarctica_}_OSU18|gi|115313981|ref|NC_008369.1|	68360	3.606	1895727
Francisella_#_tularensis_!_subsp._{_holarctica_}_F92|gi|423049750|ref|NC_019537.1|	65721	3.483	1886888
Francisella_#_tularensis_!_TIGB03_{}_|gi|379716390|ref|NC_016933.1|	64022	3.252	1968651
Francisella_#_tularensis_!_subsp._{_tularensis_}_WY96-3418|gi|134301169|ref|NC_009257.1|	61612	3.245	1898476
Francisella_#_tularensis_!_subsp._{_tularensis_}_NE061598|gi|385793751|ref|NC_017453.1|	61426	3.245	1892681
Francisella_#_tularensis_!_subsp._{_tularensis_}_FSC198|gi|110669657|ref|NC_008245.1|	61174	3.232	1892616
Francisella_#_tularensis_!_subsp._{_tularensis_}_SCHU_S4|gi|255961454|ref|NC_006570.2|	61117	3.229	1892775
Francisella_#_tularensis_!_subsp._{_tularensis_}_TI0902|gi|379725073|ref|NC_016937.1|	60940	3.22	1892744
Francisella_#_tularensis_!_subsp._{_mediasiatica_}_FSC147|gi|187930913|ref|NC_010677.1|	51421	2.715	1893886
Francisella_#_novicida_!_U112_{}_|gi|118496615|ref|NC_008601.1|	29126	1.525	1910031
Francisella_#_cf._!_novicida_{_Fx1_}|gi|385791932|ref|NC_017450.1|	26557	1.388	1913619
Francisella_#_tularensis_!_subsp._{_holarctica_}_2571.19|gi|254367826|ref|NZ_DS229056.1|	22104	3.977	555807
Francisella_#_tularensis_!_subsp._{_tularensis_}_FSC0331.4|gi|254370255|ref|NZ_DS264119.1|	14356	1.821	788151
Francisella_#_tularensis_!_subsp._{_holarctica_}_FSC0221.5|gi|254368696|ref|NZ_DS264137.1|	13465	2.759	488102
Francisella_#_tularensis_!_subsp._{_novicida_}_FTE_FTE1|gi|224580220|ref|NZ_DS989818.1|	12673	1.493	849109
Francisella_#_novicida_!_FTG_{_FTG1_}|gi|224580234|ref|NZ_DS995363.1|	11472	1.63	703773
Francisella_#_tularensis_!_subsp._{_holarctica_}_FSC0221.6|gi|254369075|ref|NZ_DS264138.1|	10313	2.029	508365

Subtracted Human Sequencing (FT6)

Organism	Genome Coverage	Genome Coverage (%)	Genome Size
Cupriavidus_#_metallidurans_!_CH34_{}_|gi|94308945|ref|NC_007973.1|	2631037	66.98	3928089
Cupriavidus_#_metallidurans_!_CH34_{_megaplasmid_}|gi|291481467|ref|NC_007974.2|	1746897	67.707	2580084
Ralstonia_#_eutropha_!_H16_{}_1_|gi|113866031|ref|NC_008313.1|	194259	4.794	4052032
Cupriavidus_#_necator_!_N-1_{}_1|gi|339324158|ref|NC_015726.1|	192410	4.968	3872936
Cupriavidus_#_taiwanensis_!_LMG_{_19424_}1|gi|188590795|ref|NC_010528.1|	190485	5.575	3416911
Cupriavidus_#_metallidurans_!_CH34_{_plasmid_}_pMOL30|gi|56130627|ref|NC_006466.1|	184536	78.944	233755
Cupriavidus_#_metallidurans_!_CH34_{_plasmid_}_pMOL30|gi|291464753|ref|NC_007971.2|	184489	78.936	233720
Ralstonia_#_eutropha_!_JMP134_{}_1|gi|73539706|ref|NC_007347.1|	174988	4.597	3806533
Francisella_#_tularensis_!_subsp._{_holarctica_}_LVS|gi|89255449|ref|NC_007880.1|	166871	8.801	1895994
Francisella_#_tularensis_!_subsp._{_holarctica_}_FSC200|gi|422937995|ref|NC_019551.1|	166321	8.781	1894157
Francisella_#_tularensis_!_subsp._{_holarctica_}_FTNF002-00|gi|156501369|ref|NC_009749.1|	165408	8.748	1890909
Francisella_#_tularensis_!_subsp._{_holarctica_}_OSU18|gi|115313981|ref|NC_008369.1|	162954	8.596	1895727
Francisella_#_tularensis_!_subsp._{_holarctica_}_F92|gi|423049750|ref|NC_019537.1|	151481	8.028	1886888
Francisella_#_tularensis_!_TIGB03_{}_|gi|379716390|ref|NC_016933.1|	145410	7.386	1968651
Francisella_#_tularensis_!_subsp._{_tularensis_}_WY96-3418|gi|134301169|ref|NC_009257.1|	138407	7.29	1898476
Francisella_#_tularensis_!_subsp._{_tularensis_}_TI0902|gi|379725073|ref|NC_016937.1|	138093	7.296	1892744
Francisella_#_tularensis_!_subsp._{_tularensis_}_NE061598|gi|385793751|ref|NC_017453.1|	136778	7.227	1892681
Francisella_#_tularensis_!_subsp._{_tularensis_}_SCHU_S4|gi|255961454|ref|NC_006570.2|	136734	7.224	1892775
Francisella_#_tularensis_!_subsp._{_mediasiatica_}_FSC147|gi|187930913|ref|NC_010677.1|	134359	7.094	1893886
Francisella_#_tularensis_!_subsp._{_tularensis_}_FSC198|gi|110669657|ref|NC_008245.1|	96340	5.09	1892616

Normal Human Sequencing (FT6)

Organism	Genome Coverage	Genome Coverage (%)	Genome Size
Francisella_#_tularensis_!_subsp._{_holarctica_}_LVS|gi|89255449|ref|NC_007880.1|	296764	15.652	1895994
Francisella_#_tularensis_!_subsp._{_holarctica_}_FSC200|gi|422937995|ref|NC_019551.1|	295826	15.618	1894157
Francisella_#_tularensis_!_subsp._{_holarctica_}_FTNF002-00|gi|156501369|ref|NC_009749.1|	295286	15.616	1890909
Francisella_#_tularensis_!_subsp._{_holarctica_}_OSU18|gi|115313981|ref|NC_008369.1|	293857	15.501	1895727
Francisella_#_tularensis_!_subsp._{_holarctica_}_F92|gi|423049750|ref|NC_019537.1|	285104	15.11	1886888
Francisella_#_tularensis_!_TIGB03_{}_|gi|379716390|ref|NC_016933.1|	276026	14.021	1968651
Francisella_#_tularensis_!_subsp._{_tularensis_}_WY96-3418|gi|134301169|ref|NC_009257.1|	264917	13.954	1898476
Francisella_#_tularensis_!_subsp._{_mediasiatica_}_FSC147|gi|187930913|ref|NC_010677.1|	263190	13.897	1893886
Francisella_#_tularensis_!_subsp._{_tularensis_}_TI0902|gi|379725073|ref|NC_016937.1|	262822	13.886	1892744
Francisella_#_tularensis_!_subsp._{_tularensis_}_NE061598|gi|385793751|ref|NC_017453.1|	262607	13.875	1892681
Francisella_#_tularensis_!_subsp._{_tularensis_}_FSC198|gi|110669657|ref|NC_008245.1|	262370	13.863	1892616
Francisella_#_tularensis_!_subsp._{_tularensis_}_SCHU_S4|gi|255961454|ref|NC_006570.2|	262370	13.862	1892775
Francisella_#_novicida_!_U112_{}_|gi|118496615|ref|NC_008601.1|	195963	10.26	1910031
Francisella_#_cf._!_novicida_{_Fx1_}|gi|385791932|ref|NC_017450.1|	183855	9.608	1913619
Francisella_#_tularensis_!_subsp._{_holarctica_}_2571.19|gi|254367826|ref|NZ_DS229056.1|	86791	15.615	555807
Francisella_#_tularensis_!_subsp._{_tularensis_}_FSC0331.4|gi|254370255|ref|NZ_DS264119.1|	82045	10.41	788151
Francisella_#_tularensis_!_subsp._{_novicida_}_FTE_FTE1|gi|224580220|ref|NZ_DS989818.1|	80740	9.509	849109
Francisella_#_novicida_!_FTG_{_FTG1_}|gi|224580234|ref|NZ_DS995363.1|	67747	9.626	703773
Francisella_#_tularensis_!_subsp._{_holarctica_}_FSC0221.5|gi|254368696|ref|NZ_DS264137.1|	65767	13.474	488102
Francisella_#_cf._!_novicida_{_3523_}|gi|387823583|ref|NC_017449.1|	64634	3.323	1945310

Subtracted Human Sequencing (FT13)

Organism	Genome Coverage	Genome Coverage (%)	Genome Size
Cupriavidus_#_metallidurans_!_CH34_{}_|gi|94308945|ref|NC_007973.1|	3049075	77.622	3928089
Cupriavidus_#_metallidurans_!_CH34_{_megaplasmid_}|gi|291481467|ref|NC_007974.2|	2009785	77.896	2580084
Ralstonia_#_eutropha_!_H16_{}_1_|gi|113866031|ref|NC_008313.1|	271535	6.701	4052032
Cupriavidus_#_necator_!_N-1_{}_1|gi|339324158|ref|NC_015726.1|	260822	6.734	3872936
Cupriavidus_#_taiwanensis_!_LMG_{_19424_}1|gi|188590795|ref|NC_010528.1|	254368	7.444	3416911
Ralstonia_#_eutropha_!_JMP134_{}_1|gi|73539706|ref|NC_007347.1|	242231	6.364	3806533
Cupriavidus_#_metallidurans_!_CH34_{_plasmid_}_pMOL30|gi|56130627|ref|NC_006466.1|	197441	84.465	233755
Cupriavidus_#_metallidurans_!_CH34_{_plasmid_}_pMOL30|gi|291464753|ref|NC_007971.2|	197177	84.365	233720
Francisella_#_tularensis_!_subsp._{_holarctica_}_LVS|gi|89255449|ref|NC_007880.1|	132199	6.973	1895994
Francisella_#_tularensis_!_subsp._{_holarctica_}_FTNF002-00|gi|156501369|ref|NC_009749.1|	131543	6.957	1890909
Francisella_#_tularensis_!_subsp._{_holarctica_}_FSC200|gi|422937995|ref|NC_019551.1|	131258	6.93	1894157
Francisella_#_tularensis_!_subsp._{_holarctica_}_OSU18|gi|115313981|ref|NC_008369.1|	128893	6.799	1895727
Francisella_#_tularensis_!_subsp._{_holarctica_}_F92|gi|423049750|ref|NC_019537.1|	122915	6.514	1886888
Ralstonia_#_solanacearum_!_CMR15_{}_|gi|523408232|ref|NC_017559.1|	121624	3.382	3596030
Ralstonia_#_solanacearum_!_Po82_{}_|gi|386331671|ref|NC_017574.1|	121071	3.478	3481091
Francisella_#_tularensis_!_TIGB03_{}_|gi|379716390|ref|NC_016933.1|	113359	5.758	1968651
Ralstonia_#_solanacearum_!_PSI07_{}_|gi|300689714|ref|NC_014311.1|	113115	3.213	3520618
Ralstonia_#_solanacearum_!_CFBP2957_{}_|gi|300702374|ref|NC_014307.1|	113075	3.309	3417386
Ralstonia_#_solanacearum_!_GMI1000_{}_|gi|17544719|ref|NC_003295.1|	110070	2.962	3716413
Francisella_#_tularensis_!_subsp._{_tularensis_}_WY96-3418|gi|134301169|ref|NC_009257.1|	109645	5.775	1898476

Normal Human Sequencing (FT13)

Organism	Genome Coverage	Genome Coverage (%)	Genome Size
Francisella_#_tularensis_!_subsp._{_holarctica_}_FTNF002-00|gi|156501369|ref|NC_009749.1|	140625	7.437	1890909
Francisella_#_tularensis_!_subsp._{_holarctica_}_LVS|gi|89255449|ref|NC_007880.1|	140169	7.393	1895994
Francisella_#_tularensis_!_subsp._{_holarctica_}_FSC200|gi|422937995|ref|NC_019551.1|	139139	7.346	1894157
Francisella_#_tularensis_!_subsp._{_holarctica_}_OSU18|gi|115313981|ref|NC_008369.1|	138832	7.323	1895727
Francisella_#_tularensis_!_subsp._{_holarctica_}_F92|gi|423049750|ref|NC_019537.1|	132246	7.009	1886888
Francisella_#_tularensis_!_TIGB03_{}_|gi|379716390|ref|NC_016933.1|	122280	6.211	1968651
Francisella_#_tularensis_!_subsp._{_tularensis_}_WY96-3418|gi|134301169|ref|NC_009257.1|	117989	6.215	1898476
Francisella_#_tularensis_!_subsp._{_tularensis_}_TI0902|gi|379725073|ref|NC_016937.1|	115572	6.106	1892744
Francisella_#_tularensis_!_subsp._{_tularensis_}_NE061598|gi|385793751|ref|NC_017453.1|	115405	6.097	1892681
Francisella_#_tularensis_!_subsp._{_tularensis_}_FSC198|gi|110669657|ref|NC_008245.1|	115311	6.093	1892616
Francisella_#_tularensis_!_subsp._{_tularensis_}_SCHU_S4|gi|255961454|ref|NC_006570.2|	115311	6.092	1892775
Francisella_#_tularensis_!_subsp._{_mediasiatica_}_FSC147|gi|187930913|ref|NC_010677.1|	115096	6.077	1893886
Francisella_#_novicida_!_U112_{}_|gi|118496615|ref|NC_008601.1|	63192	3.308	1910031
Francisella_#_cf._!_novicida_{_Fx1_}|gi|385791932|ref|NC_017450.1|	55336	2.892	1913619
Francisella_#_tularensis_!_subsp._{_holarctica_}_2571.19|gi|254367826|ref|NZ_DS229056.1|	41831	7.526	555807
Francisella_#_tularensis_!_subsp._{_tularensis_}_FSC0331.4|gi|254370255|ref|NZ_DS264119.1|	29337	3.722	788151
Francisella_#_tularensis_!_subsp._{_novicida_}_FTE_FTE1|gi|224580220|ref|NZ_DS989818.1|	27185	3.202	849109
Francisella_#_tularensis_!_subsp._{_holarctica_}_FSC0221.5|gi|254368696|ref|NZ_DS264137.1|	23490	4.813	488102
Francisella_#_novicida_!_FTG_{_FTG1_}|gi|224580234|ref|NZ_DS995363.1|	23381	3.322	703773
Francisella_#_tularensis_!_subsp._{_holarctica_}_2571.13|gi|254367300|ref|NZ_DS229050.1|	21554	6.104	353090

3. **Create mixtures of read data from prior human and pathogen sequencing projects. Mixtures will be created at a wide range of sequence quantities and genome coverage.**

We used clinical samples from these types of patients due to the clinical data on infections already present plus the ability to assess different tissue types, including sputum, lymph nodes, and urine. We evaluated different sequence alignment applications and parameters to optimize detection of the pathogen reads in these complex clinical samples. Optimal aligners and parameters were then incorporated into the MetaGeniE application. An example of an alignment of a 1000 *B. pseudomallei* read mixture with one of the cystic fibrosis sequence data sets (50M reads) is shown in Figure 7. Reads aligning to the K96243 NCBI reference genome are shown. Note the dispersed alignment pattern across the genome. This would be the expected alignment pattern of an organism actually present in a sample at low levels, and is what is seen when viewing reads aligning to *S. thermophilus* and *R. mucilaginosa* in this particular sample.

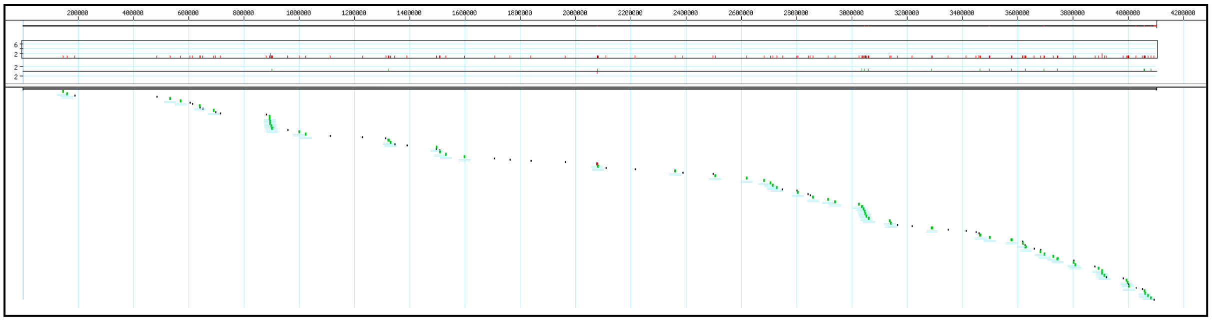


Figure 7. Lasergene Alignment of 50M cystic fibrosis throat swab Illumina PE 100bp sequence reads mixed with 1000 randomly sampled reads from a K96243 *B. pseudomallei* strain Illumina sequence data set to 2200 bacterial reference genomes, including the K96243 NCBI reference genome.

Figure 8 shows alignment of the same sequence read data set to the NCBI *Dyadobacter fermentans* genome sequence. Note the isolated locations in the genome where reads are clustered. These are primarily highly conserved genes such as tRNA's, rRNA genes and short sequence repeats. These aligned reads are sequences common to many different microbes and many are exchanged by lateral gene transfer so can be found in unrelated microbes. This type of alignment pattern would not be strong evidence of the presence of this organism in the sample. On chromosome one of two of the *B. mallei* sequences reads aligned to two 50 bp locations in the genomes, both of which were highly conserved tRNA genes. Tables 4-6 show the numbers of reads aligning to various *B. mallei* and *B. pseudomallei* genomes (closely related species, in fact *B. mallei* emerged as a clone from within *B. pseudomallei*) when 100 and 1000 *B. pseudomallei* K96243 reads are mixed with the clinical sample sequence reads. The low ratio of input reads aligning to the *B. pseudomallei* genomes likely indicates suboptimal alignment parameters.

Table 4. Results from early analyses where no *B. pseudomallei* reads were added to sample sequences.

# reads aligning	**NCBI ACC#**	**Organism**
107804	**NC_006449.1**	**StreptococcusthermophilusCNRZ1066chromosome,completegenome**
99381	**NC_013715.1**	**RothiamucilaginosaDY-18,completegenome**
97812	**NC_003454.1**	**Fusobacteriumnucleatumsubsp.nucleatumATCC25586chromosome,completegenome**
58276	**NC_013853.1**	**StreptococcusmitisB6,completegenome**
55579	**NC_013520.1**	**VeillonellaparvulaDSM2008,completegenome**
49071	**NC_014370.1**	**PrevotellamelaninogenicaATCC25845chromosomechromosomeI,completesequence**
35815	**NC_009785.1**	**Streptococcusgordoniistr.Challissubstr.CH1,completegenome**
30123	**NC_014371.1**	**PrevotellamelaninogenicaATCC25845chromosomechromosomeII,completesequence**
26592	**NC_013162.1**	**CapnocytophagaochraceaDSM7271,completegenome**
25684	**NC_013798.1**	**StreptococcusgallolyticusUCN34,completegenome**
24781	**NC_009009.1**	**StreptococcussanguinisSK36,completegenome**
19190	**NC_009796.1**	**Campylobacterconcisus13826plasmidpCCON16,completesequence**
18931	**NC_007146.2**	**Haemophilusinfluenzae86-028NP,completegenome**
17991	**NC_013192.1**	**LeptotrichiabuccalisDSM1135,completegenome**
15789	**NC_010120.1**	**Neisseriameningitidis053442,completegenome**
12410	**NC_014498.1**	**Streptococcuspneumoniae670-6Bchromosome,completegenome**
11941	**NC_002946.2**	**NeisseriagonorrhoeaeFA1090,completegenome**
11796	**NC_012471.1**	**Streptococcusequisubsp.equi4047,completegenome**
11751	**NC_009802.1**	**Campylobacterconcisus13826,completegenome**
9925	**NC_003383.1**	**ListeriainnocuaClip11262plasmidpLI100,completesequence**
9454	**NC_013895.1**	**Clostridialesgenomosp.BVAB3str.UPII9-5chromosome,completegenome**
7528	**NC_002951.2**	**Staphylococcusaureussubsp.aureusCOL,completegenome**
7473	**NC_004669.1**	**EnterococcusfaecalisV583plasmidpTEF1,completesequence**
5793	**NC_012778.1**	**EubacteriumeligensATCC27750,completegenome**
5452	**NC_013721.1**	**Gardnerellavaginalis409-05chromosome,completegenome**
5147	**NC_012923.1**	**StreptococcussuisBM407plasmidpBM407,completesequence**
5028	**NC_009089.1**	**Clostridiumdifficile630,completegenome**
[redacted]		
1	**NC_006348.1**	**BurkholderiamalleiATCC23344chromosome1,completesequence**
1	**NC_006349.2**	**BurkholderiamalleiATCC23344chromosome2,completesequence**
248*	**NC_008836.1**	**BurkholderiamalleiNCTC10229chromosomeI,completesequence**
393*	**NC_009080.1**	**BurkholderiamalleiNCTC10247chromosomeI,completegenome**
2	**NC_009076.1**	**Burkholderiapseudomallei1106achromosomeI,completegenome**
5	**NC_009078.1**	**Burkholderiapseudomallei1106achromosomeII,completegenome**
2	**NC_007434.1**	**Burkholderiapseudomallei1710bchromosomeI,completesequence**
2	**NC_007435.1**	**Burkholderiapseudomallei1710bchromosomeII,completesequence**
5	**NC_009074.1**	**Burkholderiapseudomallei668chromosomeI,completegenome**
2	**NC_012695.1**	**BurkholderiapseudomalleiMSHR346chromosomeI,completegenome**

*Vast majority of reads aligning to the B. mallei genomes are found in a single 50bp location in the genome

Table 5. Results from analysis where 100 *B. pseudomallei* reads were added to sequence reads from CF clinical sample reads.

# reads aligning	NCBI ACC#	Organism
3	NC_006348.1	BurkholderiamalleiATCC23344chromosome1,completesequence
5	NC_006349.2	BurkholderiamalleiATCC23344chromosome2,completesequence
251	NC_008836.1	BurkholderiamalleiNCTC10229chromosomeI,completesequence
1	NC_008835.1	BurkholderiamalleiNCTC10229chromosomeII,completesequence
417	NC_009080.1	BurkholderiamalleiNCTC10247chromosomeI,completegenome
3	NC_009079.1	BurkholderiamalleiNCTC10247chromosomeII,completegenome
1	NC_008785.1	BurkholderiamalleiSAVP1chromosomeI,completegenome
1	NC_008784.1	BurkholderiamalleiSAVP1chromosomeII,completegenome
8	NC_009076.1	Burkholderiapseudomallei1106achromosomeI,completegenome
7	NC_009078.1	Burkholderiapseudomallei1106achromosomeII,completegenome
11	NC_007434.1	Burkholderiapseudomallei1710bchromosomeI,completesequence
2	NC_007435.1	Burkholderiapseudomallei1710bchromosomeII,completesequence
4	NC_009074.1	Burkholderiapseudomallei668chromosomeI,completegenome
6	NC_009075.1	Burkholderiapseudomallei668chromosomeII,completegenome
14	NC_006350.1	BurkholderiapseudomalleiK96243chromosome1,completesequence
14	NC_006351.1	BurkholderiapseudomalleiK96243chromosome2,completesequence
3	NC_012695.1	BurkholderiapseudomalleiMSHR346chromosomeI,completegenome

Table 6. Results from analysis where 1000 *B. pseudomallei* reads were added to sequence reads from CF clinical sample reads.

# reads aligning	NCBI ACC#	Organism
74	NC_006348.1	Burkholderia mallei ATCC 23344 chromosome 1, complete sequence
33	NC_006349.2	Burkholderia mallei ATCC 23344 chromosome 2, complete sequence
79	NC_009076.1	Burkholderia pseudomallei 1106a chromosome I, complete genome
42	NC_009078.1	Burkholderia pseudomallei 1106a chromosome II, complete genome
91	NC_007434.1	Burkholderia pseudomallei 1710b chromosome I, complete sequence
49	NC_007435.1	Burkholderia pseudomallei 1710b chromosome II, complete sequence
64	NC_009074.1	Burkholderia pseudomallei 668 chromosome I, complete genome
52	NC_009075.1	Burkholderia pseudomallei 668 chromosome II, complete genome
177	NC_006350.1	Burkholderia pseudomallei K96243 chromosome 1, complete sequence
139	NC_006351.1	Burkholderia pseudomallei K96243 chromosome 2, complete sequence
59	NC_012695.1	Burkholderia pseudomallei MSHR346 chromosome I, complete genome

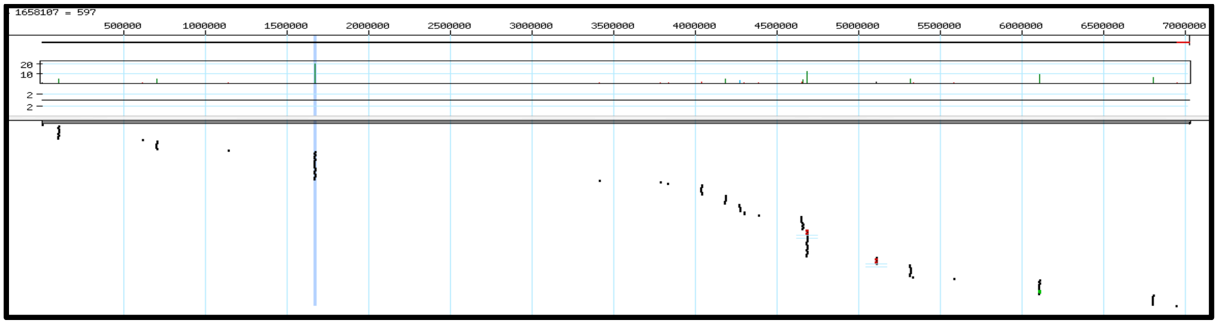


Figure 8. Lasergene Alignment of 50M cystic fibrosis throat swab Illumina PE 100bp sequence reads to 2200 bacterial reference genomes. Location within the genome of reads aligning to the NCBI *Dyadobacter fermentans* reference genome is shown.

Analysis of GAIIx and HiSeq Mixtures

We analyzed 14 samples from mixtures places on the GAIIx and 35 samples from mixtures placed on the HiSeq; 9 samples on the HiSeq were also run with whole genome amplified material. The initial number of reads for 14 GAIIx samples ranged from 6.38 million reads to 57.5 million reads (mean 25.0 million, SD ± 11.8 million). The initial number of reads for 35 HiSeq samples ranged from 70.8 million reads to 105.9 million reads (mean 53.1 million, SD ± 24.4 million). The other 9 whole genome amplified samples ranged from 31.3 million to 105.9 million (SD ± 27.9 million).

Table 7. Results of sequencing with mixture and insert size variation, read statistics, sequencing platform, and type of run for Illumina instruments.

Sample	**# total reads**	**platform**	**Run type**
DNA Mixture1_220bp*	34,872,808	HiSeq	100bpX2
DNA Mixture2_220bp	85,222,206	HiSeq	100bpX2
DNA Mixture3_220bp	67,146,924	HiSeq	100bpX2
DNA Mixture4_220bp	53,407,900	HiSeq	100bpX2
DNA Mixture5_220bp	45,209,146	HiSeq	100bpX2
DNA Mixture6_220bp	51,643,986	HiSeq	100bpX2
DNA Mixture7_220bp	55,998,636	HiSeq	100bpX2
DNA Mixture8_220bp	54,377,232	HiSeq	100bpX2
DNA Mixture9_220bp	41,804,982	HiSeq	100bpX2
DNA Mixture10_220bp	22,434,026	HiSeq	100bpX2
DNA Mixture11_220bp	23,179,078	HiSeq	100bpX2
DNA Mixture1-WGA_220bp**	33,791,610	HiSeq	100bpX2
DNA Mixture2-WGA_220bp	31,259,728	HiSeq	100bpX2
DNA Mixture3-WGA_220bp	77,460,036	HiSeq	100bpX2
DNA Mixture4-WGA_220bp	92,243,524	HiSeq	100bpX2

DNA Mixture5-WGA_220bp	40,840,362	HiSeq	100bpX2
DNA Mixture6-WGA_220bp	62,553,118	HiSeq	100bpX2
DNA Mixture7-WGA_220bp	51,846,948	HiSeq	100bpX2
DNA Mixture8-WGA_220bp	51,929,548	HiSeq	100bpX2
DNA Mixture9-WGA_220bp	47,806,890	HiSeq	100bpX2
DNA Mixture10-WGA_220bp	105,901,040	HiSeq	100bpX2
DNA Mixture11-WGA_220bp	90,600,348	HiSeq	100bpX2
YP-Blood repA-10e3_220bp[***]	30,454,020	HiSeq	100bpX2
YP-Blood repB-10e3_220bp	7,082,796	HiSeq	100bpX2
YP-Blood repC-10e3_220bp	69,589,268	HiSeq	100bpX2
YP-Blood repA-10e4_220bp	18,720,472	HiSeq	100bpX2
YP-Blood repB-10e4_220bp	46,880,850	HiSeq	100bpX2
YP-Blood repC-10e4_220bp	73,344,980	HiSeq	100bpX2
YP-Blood repA-10e7_220bp	77,696,268	HiSeq	100bpX2
YP-Blood repB-10e7_220bp	71,446,862	HiSeq	100bpX2
YP-Blood repC-10e7_220bp	20,101,380	HiSeq	100bpX2
YP-Blood repA-NTC_220bp	17,686,296	HiSeq	100bpX2
YP-Blood repB-NTC_220bp	64,977,842	HiSeq	100bpX2
YP-Blood repC-NTC_220bp	48,477,286	HiSeq	100bpX2
YP-Blood repA-NTC_600bp	27,659,534	GAIIx	100bpX2
YP-Blood repA-10e3_600bp	12,450,630	GAIIx	100bpX2
YP-Blood repA-10e4_600bp	22,147,300	GAIIx	100bpX2
YP-Blood repA-10e7_600bp	17,812,372	GAIIx	100bpX2
YP-Blood repA-NTC_220bp	28,401,622	GAIIx	100bpX2
YP-Blood repA-10e3_220bp	32,686,456	GAIIx	100bpX2
YP-Blood repA-10e4_220bp	30,348,982	GAIIx	100bpX2
YP-Blood repA-10e7_220bp	20,342,338	GAIIx	100bpX2
YP-Sputum TGN6576-NTC_220bp	88,936,920	HiSeq	100bpX2
YP-Sputum TGN6376-10e4_220bp	25,812,050	GAIIx	100bpX2
YP-Sputum TGN6376-10e7_220bp	57,500,138	GAIIx	100bpX2
YP-Sputum TGN6376_10e3_600bp	6,382,980	GAIIx	100bpX2
YP-Sputum TGN6376_10e3_220bp	17,661,352	GAIIx	100bpX2
YP-Sputum TGN6376_10e7_220bp	21,705,452	GAIIx	100bpX2
YP-Sputum TGN6376_10e7_600bp	29,781,930	GAIIx	100bpX2

[*] 220bp and 600bp refer to the targeted library fragment size.

[**] WGA indicates that the sample DNA was subjected to Whole Genome Amplification prior to library preparation.

[***] The *Y. pestis*-human blood mixtures were prepared in triplicate at different bacterial loads, the sputum mixtures only one sample/bacterial load.

Determine genome targets for Species ID, subtyping characterization, drug resistance

4. Compiling relevant data from NAU and TGen

Compiled in-house on MGGen’s bio617 and bio653 servers. Our approach is listed in Tasks 5 and 7 and consists largely of data from phylogenetic trees for subtyping characterization and from on-line databases for species ID and drug resistance (listed below). For subtyping, we used a novel approach. Rreads from particular bacterial species can be collated into a single file and then analyzed in a phylogenetic framework using a program our labs at MGGen and TGen have developed called the whole genome focused array SNP typer (WG-FAST, https://github.com/jasonsahl/wgfast). The goal of WG-FAST is to phylogenetically genotype an unknown sample in the context of a well studied pathogen. This sample can be from either a metagenomics dataset, a metatranscriptomics dataset, or a single isolate sequencing dataset. WG-FAST works off of existing phylogenetic trees so is most informative for well characterized organisms where their evolutionary history is accurately described with a tree (i.e., largely clonal organisms).

5. Comprehensive literature search with an emphasis on subtyping (e.g. determining particular lineages, clonal complexes, etc.) and finding targets for drug resistance characterization

Completed for all target pathogens. After much initial work on our end, several on-line databases became available or were recommended to us so we switched to use these compiled databases. Genes coding for known antibiotic resistance mechanisms were downloaded in multi-FASTA format from the Antibiotic Resistance Genes Database (Liu and Pop 2009) and the Comprehensive Antibiotic Resistance Database (McArthur et al. 2013). Genes coding for known virulence factors were downloaded in multi-FASTA format from the Virulence Factors of Pathogenic Bacteria database (Chen et al. 2012). 16S (bacterial) and 18S (fungal) sequences for numerous known human pathogens were downloaded from GenBank. These were concatenated into a single multi-FASTA file bioinformatically and used as a consolidated reference for aligning sequence reads.

Pathogen strains targeted in this study
Acinetobacter baumannii
Aspergillus fumigatus
Bacillus anthracis
Brucella abortus
Brucella melitensis
Brucella suis
Burkholderia mallei
Burkholderia pseudomallei
Burkholderia thailandensis
Clostridium botulinum
Coxiella burnetii
Escherichia coli
Francisella tularensis
Haemophilus influenza
Klebsiella pneumonia
Mycobacterium avium
Pseudomonas aeruginosa
Rickettsia prowazekii
Rickettsia rickettsii
Staphylococcus aureus
Staphylococcus epidermidis
Yersinia pestis

6. Contact infectious disease experts

Given the focus on identifying genetic target sequences for agent identification and characterization, the extensive expertise in genetic signature and assay development for genetic subtyping of Select Agent pathogens and public health pathogens already in house at MGGEN and TGen North, respectively, obviated the need for contacting outside experts for the purposes of this project.

Create pan-genome references of targets

7. Generate a reference pan-genome with sequences for all loci/targets

Determining species identity by aligning reads against a reference genome for various targets is clear-cut for many species. This alignment process compares the read data from the sequencing run against the entire reference genome. This is relatively quick and straightforward when you know your target DNA and are only comparing to one reference genome. Due to improvements in the Refseq database and our ability to incorporate its genomes into our MetaGeniE analysis pipeline, we changed our analysis strategy of developing a reference pan-genome. While the reference pan-genome is intuitively simple, the Refseq database provides superior searching and matching capabilities. Full details of our new approach are given in our clinical sequencing results section as well as our description of the MetaGeniE pipeline.

8. Determine appropriate statistical threshold for a match to the pan-genome.

Due to the change in our approach and shift from a pan-genome approach to use of the RefSeq database this task in no longer entirely appropriate. However, the concept of what defines a “match” remains extremely valid. As discussed above, some measure of read alignment dispersal pattern, such as % reference genome covered, appears to be an important aspect of determining the presence/absence of a given organism within a sample. The challenge is that even with a particular cutoff or thresholds there are possibilities for detection and exact identification and characterization of a particular sample with only a few sequence reads. See Francisella tularensis section in Task 13.

Create mixtures of human and pathogen DNA

9. Spike single or multiple pathogens DNA at varying quantities into human DNA.

In order to evaluate the effect of exogenous DNA on the identification and classification of bacterial DNA we generated mixtures of varying proportions of bacterial DNA and human DNA. Results from this is the following section are fully detailed, only the experimental design and set up is given so they are relatively short. We generated the necessary bacterial DNA for this work, and acquired commercially available human DNA for the mixtures. Both the bacterial and the human DNA were quantified by Nanodrop and the following mixtures were made (Table 8). By using DNA mixtures we know what proportion of each mixture should be bacterial reads, allowing for the examination of bioinformatics bias in metagenomics analysis. We also examined the effect of whole genome amplification (WGA) on these mixtures. A WGA was performed on each of the mixtures. Sequencing libraries were created for both the neat DNA/DNA mixtures, and the WGAs of the mixtures.

Table 8. Experiment with 11 different mixtures of three different Select Agent bacteria spiked into human DNA. Percentage is amount of bacteria relative to human DNA. *Yersinia pestis* = Yp, *Bacillus anthracis* = Ba, *Burkholderia pseudomallei* = Burk.

	Percent of Mixture			Amount added (ng)		
Mixture ID	Yp	Ba	Burk	Yp	Ba	Burk
1	0.01%	--	--	0.2	0	0
2	0.00%	--	--	0.02	0	0
3	--	0.01%	--	0	0.2	0
4	--	0.00%	--	0	0.02	0
5	--	--	0.00%	0	0	0.02
6	0.01%	0.01%	--	0.2	0.2	0
7	0.00%	0.00%	--	0.02	0.02	0
8	0.01%	0.00%	--	0.2	0.02	0
9	0.00%	0.01%	--	0.02	0.2	0
10	0.00%	0.00%	0.00%	0.02	0.02	0.02
11	0.10%	--	0.10%	2	0	2

10. Spike cultures of multiple pathogens into human clinical samples.

In order to mimic a clinical sample more closely we prepared mixtures with *Y. pestis* cultures at different bacterial loads and either human sputum or human blood samples and extracted DNA (Table 9). For the sputum samples, the *Y. pestis* culture spikes were calculated based on OD and added to equal volumes of sputum for each sample. There were four different sputa used for these mixtures (Table 9). For the blood samples, there was one single blood sample used with the four concentrations of *Y. pestis*. Three concentrations (10^3, 10^4, 10^7) of *Y. pestis* were used along with a sample with no bacterial spike (NTC). Sequencing libraries were created for both the sputum and blood mixtures.

Table 9. Different bacterial loads spiked into the samples prior to extraction.

Sputum ID	*Y. pestis* cell loaded
TGN6576	0
TGN6576	10^3
TGN6576	10^4
TGN6576	10^7

TGN6376	0
TGN6376	10^3
TGN6376	10^4
TGN6376	10^7
TGN6380	0
TGN6380	10^3
TGN6380	10^4
TGN6380	10^7
TGN6388	0
TGN6388	10^3
TGN6388	10^4
TGN6388	10^7

Sequence DNA mixtures on Ion Torrent and Illumina platforms

11. Determine level of sensitivity and accuracy using known mixtures of single and multiple pathogen experiments.

To confirm the generated DNA/DNA mixtures, all samples were tested with qPCR assays specific to each bacterium (*Yersinia pestis*, *Bacillus anthracis*, *Burkholderia pseudomallei*). We used specific qPCR assays and not a general 16S qPCR, which would detect all bacteria in these samples, not just the targets. The goal of the qPCR assay was to verify the relative amount of DNA spiked into the sample with a method that is highly sensitive to the detection of trace amounts of target DNA present in a complex matrix of DNA, cells, and cell parts from the host and its natural microbial community. In all but two cases, both with 0.001% of bacteria that had been prepared using whole genome amplification (WGA), the target bacteria in the sample were detectable with qPCR. Also as expected, 1/10 dilutions of target DNA (e.g. as seen with mixtures 3 and 4) gave a ~3 Ct difference, consistent with typical qPCR dilutions. The eleven mixtures were run on at least one Illumina platform (GAIIx or HiSeq, 100-150 bp paired-end runs with~600 bp inserts) and on the Ion Torrent PGM instrument with 318 chips.

Table 10. Quantitative PCR assays run on DNA/DNA mixtures to independently test the amount of bacteria DNA spiked into human DNA samples. *Yersinia pestis* = Yp, *Bacillus anthracis* = Ba, *Burkholderia pseudomallei* = Burk. Amounts of bacterial DNA given in nanograms rather than percentages. Cycle threshold values (Ct) are given as a relative measure of bacterial DNA quantity in the sample, with samples run both neat (pure sample) and as a whole genome amplification (WGA) neat DNA. Bacteria not detected are in red and given as undetermined (und.).

Mixture	Yp amount (ng)	Ba amount (ng)	Burk amount (ng)	Yp Neat Ct	Yp WGA Ct	Ba Neat Ct	Ba WGA Ct	Burk Neat Ct	Burk WGA Ct
1	0.2	0	0	27.4	26.4	---	---	---	---

2	0.02	0	0	32.2	30.3	---	---	---	---
3	0	0.2	0	---	---	27.9	29.6	---	---
4	0	0.02	0	---	---	31.4	32	---	---
5	0	0	0.02	---	---	---	---	33.1	und.
6	0.2	0.2	0	28.2	25.8	27.4	29.6	---	---
7	0.02	0.02	0	31.5	29.5	31.1	31.5	---	---
8	0.2	0.02	0	28.2	25.7	30.7	32.9	---	---
9	0.02	0.2	0	31.1	30.3	27.7	29.2	---	---
10	0.02	0.02	0.02	31.1	30.4	31.7	32.3	32.5	und.
11	2	0	2	24.2	21.9	---	---	26.4	28.3

To confirm the generated blood and sputum mixtures, all samples were tested with qPCR assays specific to *Y. pestis*. Again, we used specific qPCR assays and not a general 16S qPCR, along with a plasmid dilution curve to allow for quantification of *Y. pestis*. The goal of the qPCR assay was to verify the relative amount of *Y. pestis* spiked into the sample with a method that is highly sensitive to the detection of trace amounts of target DNA present in a complex matrix of DNA, cells, and cell parts from the host and its natural microbial community. We ran a chromosome based assay to calculate the number of *Y. pestis* cells in each mixture (Table 11).

Table 11. Quantitative PCR assays run on blood and sputum mixtures to independently test the amount of *Y. pestis* spiked into human blood or sputum samples. The blood samples are denoted by A, B, and C. The sputum samples are denoted by TGN numbers. Cycle threshold values (Ct) are given as a relative measure of bacterial DNA quantity in the sample, the higher the Ct the lower the number of *Y. pestis* cells. The chromosome copy mean is calculated based off the plasmid quantification curve that was run with all the samples. If *Y. pestis* was not detected the Ct is given as undetermined (Und).

Sample Name	**Ct mean**	**Chromosome copy mean**
A 10^3	35.3	2
A 10^4	32.5	17
A 10^7	22.9	57,019
A NTC	Und	0
B 10^3	35.3	2
B 10^4	34.3	4
B 10^7	23.6	30,528
B NTC	Und	0
C 10^3	36.4	1
C 10^4	24.5	13,831
C 10^7	35.0	4
C NTC	Und	0
TG6376 10^3	20.8	316,414
TG6376 10^4	32.4	19
TG6376 10^7	33.6	8
TG6376 NTC	Und	0
TG6380 10^3	19.2	1,176,816
TG6380 10^4	29.9	153
TG6380 10^7	34.1	4
TG6380 NTC	Und	0
TG6388 10^3	20.9	288,999
TG6388 10^4	31.4	46
TG6388 10^7	36.1	1
TG6388 NTC	Und	0
TG6576 10^3	19.6	874,228
TG6576 10^4	29.7	182
TG6576 10^7	34.9	2
TG6576 NTC	Und	0

Detection of DNA in mixtures

Table 12. Detection of bacteria by identification of genome % for different mix for bacteria spiked in only. The top hit is in the Bacteria column and the percent of the reads is indicated in the % Genome column.

Samples	**Actual**	**Detected Genomically**	
	% mix bacterial/ human	Bacteria	% Genome
Yersinia-pestis_mix1	0.01%	Yp (antiqua)	7.8%
Yersinia-pestis_mix2	0.00%	Yp (angola/antiqua)	2.0%
Bacillus-anthracis_mix3		Ba (A0248/AmesAnc)	12.0%
Bacillus-anthracis_mix4		Ba (CDC684/AmesAnc)	1.0%
Burkholderia-pseudomallei_mix5		Bmallei(ATCC23344)	0.3%
Yersinia-pestis_mix6	0.01%	Ba (A0248/AmesAnc), Yp (antiqua)	8.3%,8.4%
Yersinia-pestis_mix7	0.00%	Yp (angola/antiqua)	1.3%
Yersinia-pestis_mix8	0.01%		
Bacillus-anthracis_mix9	0.00%	Ba (A0248/AmesAnc), Yp (antiqua)	7.0%, 1.0%
Yersinia-pestis_mix10	0.00%	Yp (antiqua), Ba (A0248/AmesAnc)	1.1%,0.4%
Yersinia-pestis_mix11	0.10%	Yp (antiqua), Bp(K9)	25.3%,8.6%

Table 13. Detection of bacteria by identification of genome % for different mix for whole genome amplified (WGA) samples only. The top hit is in the Bacteria column and its percent of the reads is indicated in the % Genome column.

Samples	**Actual mix**	**Detected mix**	
	% mix bacterial/ human	Bacteria	%Genome
Yersinia-pestis_mix1	NA	Yp (antiqua)	3.6%
Yersinia-pestis_mix2	NA	Yp (antiqua)	2%
Bacillus-anthracis_mix3	NA	Ba (A0248/AmesAnc)	0.2%
Bacillus-anthracis_mix4	NA	No significant hits	-
Burk.-pseudomallei_mix5	NA	No significant hits	-
Yersinia-pestis_mix6	NA	Yp (antiqua)	5.7%
Yersinia-pestis_mix7	NA		
Yersinia-pestis_mix8	NA		
Bacillus-anthracis_mix9	NA	Yp (antiqua), Ba (A0248/AmesAnc	2.7%,0.2%

Yersinia-pestis_mix10	NA	Yp (antiqua), Ba (CDC,A0248/AmesAnc)	3.2%,0.1%
Yersinia-pestis_mix11	NA	Yp (antiqua), Bmallei(ATCC23344	49%,1.1%

Detection for DNA Mixtures on HiSeq

Table 14 represents the 11 HiSeq mixtures that were mixed with various strains at different concentrations. Blind comparisons* were performed with the % genome coverage detected by MetaGeniE against the actual concentration per strain (*The actual organism and concentrations were released after *in silico* detection for non bias comparison).

Table 14. Comparison between the actual mixture (ng) and the detected mixture (% genome coverage). *Incorrect detection at species level (but detection of sample in correct genus) shown in red font.

Samples	Species	Actual Mixture	Detected Mixture
Ba_Mixture3	Ba	0.2ng	12%
Ba_Mixture4	Ba	0.02ng	1%
Ba_Mixture9	Ba;Yp	0.2 ng; 0.02ng	7%;1%
Bp_Mixture5	Bp	0.02ng	0.34%
Yp_Mixture1	Yp	0.2ng	7.80%
Yp_Mixture10	Ba;Yp;Bp	0.02ng;0.02ng;0.02ng	0.43%;1.1%;0.06%
Yp_Mixture11	Yp;Bp	2ng,2ng	25.3%;8.6%
Yp_Mixture2	Yp	0.02ng	2%
Yp_Mixture6	Yp;Ba	0.2ng;0.2ng	8.4%;8.3%
Yp_Mixture7	Yp;Ba	0.02ng;0.02ng	1.3%;83%
Yp_Mixture8	Yp;Ba	0.2ng;0.02ng	8.1%;0.9%

Table 15 shows a Lasergene NGEN™ analysis of 17.7M Illumina sequence reads derived from an artificial *Y. pestis*-human sputum mixture DNA extraction (10^7 CFU *Y. pestis* load). All columns shown are output by Lasergene except for the read dispersion ratio column. The table is split into two sections. The top section consists of all the reference genomes found to have reads aligned with dispersion ratios greater than 0.1, sorted by # reads aligned. All of the organisms in this section with greater than 100 reads aligned are likely present in this sample, and have been found in other human sputum samples we have analyzed for other research projects. The bottom section consists of the genomes with the nine highest # reads aligned with dispersion ratios less than 0.1. This value is calculated by dividing the actual length of the reference sequence covered by the theoretical maximum amount of sequence that would be covered if the sequence reads were completely dispersed around the reference genome (genome length-reference bases not covered)/(# Reads Aligned X 100bp read length). The larger this value (range 0-1), the greater the level of read dispersion around a given genome. This value, in conjunction with the % reference genome covered, can indicate whether a given organism was likely present within a sample. For example, *P. stutzeri* (green shading) had 51 reads align, resulting in only 0.06% of the reference genome covered but has a 0.44 dispersion ratio, indicating the reads are dispersed across the genome. Normally, this level of read dispersion would indicate that this organism was likely present in the sample. However, it is more likely that a number of the reads that mapped to the *P. aeruginosa* genome (29943 reads, dispersion ratio = 0.519) also mapped to this related species. In contrast, the *M. vannielii* genome had 494 reads align, resulting in 0.06% of the reference genome covered, but a dispersion ratio of only 0.019. This indicates these reads are not well dispersed and are likely all aligning to a very small portion of the genome that is likely highly conserved within bacteria. This organism was likely not present in the sample. This highlights the fact that closely related species can give false calls in metagenomic analyses due to genetic similarities and that care must be taken to identify the most likely microbe in a sample.

Table 15. Lasergene NGEN metagenomic analysis of *Y. pestis*-human sputum mixture. Preliminary metagenomic analysis of artificial *Y. pestis*-Sputum and blood mixtures. More detailed analyses were then done with MetaGeniE.

# Reads Aligned	NCBI referenc Accession	reference genome	genome length	reference bases not covered	% reference covered	read dispersion ratio
29943	NC_002516.2	Pseudomonas aeruginosa PAO1 chromosome	6264536	4709003	24.83%	0.519
2938	NC_003143.1	Yersinia pestis CO92 chromosome	4646554	4481240	3.56%	0.563
1312	NC_009708.1	Yersinia pseudotuberculosis IP 31758 chromosome	4705701	4631769	1.57%	0.564
225	NC_017671.1	Stenotrophomonas maltophilia D457	4758756	4745993	0.27%	0.567
172	NC_017958.1	Tistrella mobilis KA081020-065 plasmid pTM3	95201	85953	9.71%	0.538
114	NC_003131.1	Yersinia pestis CO92 plasmid pCD1	68166	62198	8.76%	0.524
99	NC_008782.1	Acidovorax sp. JS42 chromosome	3809387	3804328	0.13%	0.511
88	NC_003134.1	Yersinia pestis CO92 plasmid pMT1	90142	85375	5.29%	0.542
87	NC_007973.1	Cupriavidus metallidurans CH34 chromosome	2568261	2563258	0.19%	0.575
82	NC_003132.1	Yersinia pestis CO92 plasmid pPCP1	9202	5655	38.55%	0.433
77	NC_017731.1	Providencia stuartii MRSN 2154 chromosome	4304594	4300357	0.10%	0.550
62	NC_014640.1	Achromobacter xylosoxidans A8 chromosome	6920779	6917738	0.04%	0.490
58	NC_010554.1	Proteus mirabilis HI4320 chromosome	4029955	4026731	0.08%	0.556
55	NC_012590.1	Corynebacterium aurimucosum ATCC 700975	2732383	2729471	0.11%	0.529
51	**NC_009434.1**	**Pseudomonas stutzeri A1501 chromosome**	**3711059**	**3708817**	**0.06%**	**0.440**
42	NC_015436.1	Spirochaeta coccoides DSM 17374 chromosome	2209761	2209064	0.03%	0.166
34	NC_009349.1	Aeromonas salmonicida subsp. salmonicida A449 plasmid	96768	95124	1.70%	0.484
31	NC_007946.1	Escherichia coli UTI89 chromosome	2198538	2196887	0.08%	0.533
23	NC_007164.1	Corynebacterium jeikeium K411 chromosome	2084729	2083235	0.07%	0.650
23	NC_018107.1	Klebsiella oxytoca E718 plasmid pKOX_R1	340976	339771	0.35%	0.524
22	NC_015410.1	Pseudomonas mendocina NK-01 chromosome	5208340	5207059	0.02%	0.582
21	NC_008027.1	Pseudomonas entomophila L48 chromosome	5166302	5165277	0.02%	0.488
20	NC_007972.2	Cupriavidus metallidurans CH34 plasmid pMOL28	35742	34727	2.84%	0.508
20	NC_015556.1	Pseudomonas fulva 12-X chromosome	4531508	4530572	0.02%	0.468
1292	NC_015958.1	Thermoanaerobacter wiegelii Rt8.B1 chromosome	2266998	2266406	0.03%	0.005
1101	NC_009441.1	Flavobacterium johnsoniae UW101 chromosome	5718680	5718523	0.00%	0.001
973	NC_014738.1	Riemerella anatipestifer ATCC 11845 = DSM 15868 chromosome	1522151	1522074	0.01%	0.001
494	**NC_009634.1**	**Methanococcus vannielii SB chromosome**	**1691825**	**1690884**	**0.06%**	**0.019**
478	NC_007355.1	Methanosarcina barkeri str. Fusaro chromosome	4481158	4480637	0.01%	0.011

448	NC_009135.1	Methanococcus maripaludis C5 chromosome	1090695	1090596	0.01%	0.002
445	NC_014933.1	Bacteroides helcogenes P 36-108 chromosome	3845056	3844935	0.00%	0.003
405	NC_003901.1	Methanosarcina mazei Go1 chromosome	4058071	4057483	0.01%	0.015
378	NC_017941.1	Haloferax mediterranei ATCC 33500 chromosome	1245902	1245843	0.00%	0.002

Table 16 shows sequence analysis data comparing artificial mixtures of *Y. pestis* culture at different CFU loads with human blood. Two libraries with different insert sizes were prepared from each mixture. The 220 bp insert library is designed to provide overlapping reads, in order to test SNP analysis capability with low numbers of reads at 1X and 2X coverage at SNP loci. The use of overlapping reads to improve SNP reliability is conceptually simple but few people have been using this strategy. The data indicate that, at the indicated # of total reads, the 220 bp insert library, not surprisingly, is less sensitive at detecting *Y. pestis* in the mixtures, having lower reference genome coverage and dispersion ratio values. The 600 bp insert library promises to be more sensitive, as the reads will be more dispersed around the genome, resulting in greater reference genome coverage. Thus, our preliminary results suggest that overlapping reads drastically improve SNP calling data but come at a cost of less data. These samples will be resequenced on the HiSeq platform to obtain greater numbers of reads to ideally remove the impediment of lower coverage. Further analysis, including establishing optimal alignment parameters and BLAST analysis of mapped reads, will be needed to establish a robust threshold of detection.

Table 16. *Y. pestis*-human blood mixture sequence analysis: *Y. pestis* detection sensitivity. Preliminary bacterial CFU load sensitivity analysis.

Sample	**Y. pestis CFU load**	**Library inse size**	**Total # Illumina 100bp reads**	**# reads mapping to CO92 reference genome**	**% reference genome coverage**	**dispersion ratio**
Mock YP Blood mixture	0	220	28.4M	355	0.04%	0.046
Mock YP Blood mixture	0	600	27.7M	349	0.03%	0.03
Mock YP Blood mixture	1000	220	32.7M	397	0.02%	0.02
Mock YP Blood mixture	1000	600	12.5M	178	0.04%	0.11
Mock YP Blood mixture	10000	220	30.3M	415	0.16%	0.18
Mock YP Blood mixture	10000	600	22.1M	376	0.18%	0.224
Mock YP Blood mixture	10000000	220	20.3M	34446	37.30%	0.503
Mock YP Blood mixture	10000000	600	18.8M	31797	46.30%	0.678

Identification of microbial species in a sample however is relatively commonplace in research labs, particularly if the focus is only on 16S and ITS sequencing for species identification. The next step is to not just identify the organisms but to fully characterize them, including such traits as antibiotic resistance as well as to identify specific lineages of epidemiological importance. In order to address the capability of next-generation sequence technology for direct genotyping of organisms found within clinical samples, we are exploring the use of short insert sequencing libraries. The challenge here is to extract reliable SNP data from a limited # of reads that are dispersed around any given genome, resulting in very low coverage depth, as little as 1X, for any SNP call. The error rate of

Illumina sequencing can be as high as 0.5-1.0%, depending upon a number of factors, resulting in roughly 1 base call error in every 100 bp read. One way to address this is to only call SNPs at loci that have been previously validated to be variable using a set of high quality genomic sequences, or a Master SNP database. This approach requires previous whole genome sequence analyses be done with the target organisms. We have these databases in place for a number of Select Agent organisms as well as numerous pathogens important from a public health perspective and are continuously adding to this reference dataset. A second strategy would be to try to generate overlapping reads, so that any given SNP locus has a high probability of having at least one read on each strand. This may increase the reliability of a SNP call, as the probability of an error occurring in the same position in 2 overlapping reads is theoretically the product of a single error occurring in one read (e.g. 0.05 x 0.05 = 0.0025). In addition, having a read on each strand may be more reliable than 2 reads on the same strand. We are still exploring the use of short insert libraries to generate overlapping reads. A 220 bp fragment library contains on average 100 bp of genomic sequence between the adapters. With the use of a 100 bp paired-end read run, this will provide a read on each strand of the genomic sequence covered. We have generated both 220 and 600 bp fragment libraries for a number of the DNA mixtures as well as the mock *Y. pestis*-blood and sputum mixtures and compared the reliability of resulting SNP calls against a Master SNP database that includes genome sequence for the *Y. pestis* strain used in the mixtures.

We note another avenue of research we are exploring for analysis of clinical samples. Some clinical samples may not yield enough total DNA for reliable preparation of sequencing libraries. One mitigating methodology is Whole Genome Amplification (WGA), a generally accepted method for synthesizing microgram quantities of high molecular weight (10-20 Kbp) DNA from nanograms to picogram quantities of genomic DNA. However, this technique may result in a biased representation of genomes present in a sample. For example, it is also known to amplify high GC organisms very poorly. We are evaluated the impact of Whole Genome Amplification on the detection and characterization of a number of target organisms in the artificial DNA mixtures. While this technique has been used to prepare single cell sequencing libraries, avoiding WGA processing is preferable if possible. New developments in library preparation reagents are promising to generate libraries from less than 1ng of total DNA, further minimizing the need for using the WGA technique for any but the most extreme low level DNA sample types.

Mock mixture sample sequencing

We created various mixtures such as *B. pseudomallei* Illumina sequence read data and the cystic fibrosis (CF) samples, which generated sequence data in the ranges of 10-72000 reads. Initial sequencing of a subset of samples was done on the Illumina GAIIx platform to evaluate library qualities and estimate required coverage levels. The bulk of the more recent sequencing is being done on the Illumina HiSeq platform due to the ca. 5 fold increase in data yield, allowing for increased multiplexing of samples/lane. The DNA-DNA mixtures are being blinded to the data analyst for unbiased analysis.

MetaGeniE is able to recall correct number of organisms up to species level even at very low concentration (0.02 ng) except for *Burkholderia*. Considering the blind test analysis, no false positives and false negatives were detected, which is important for reliable detection in clinical datasets. Figure 9 shows the Pearson correlation between the actual mixture and the detected genome coverage %. The correlation between percent of the genome detected and concentration (in ng) for actual mixture for all samples excluding *B. pseudomallei* was found to be 0.9. This high correlation suggests that the percent genome mapped is directly proportional to the sampling effort and thus makes it an important step for metagenomic studies. In our experience, however, high GC% organisms such as *B. pseudomallei* is more difficult to detect especially at low concentration as compared to other organisms, due to lower PCR amplification efficiency during library prep and during the sequencing process. Including *B. pseudomallei* (but excluding the incorrect species detected) reduced the Pearson correlation to 0.73.

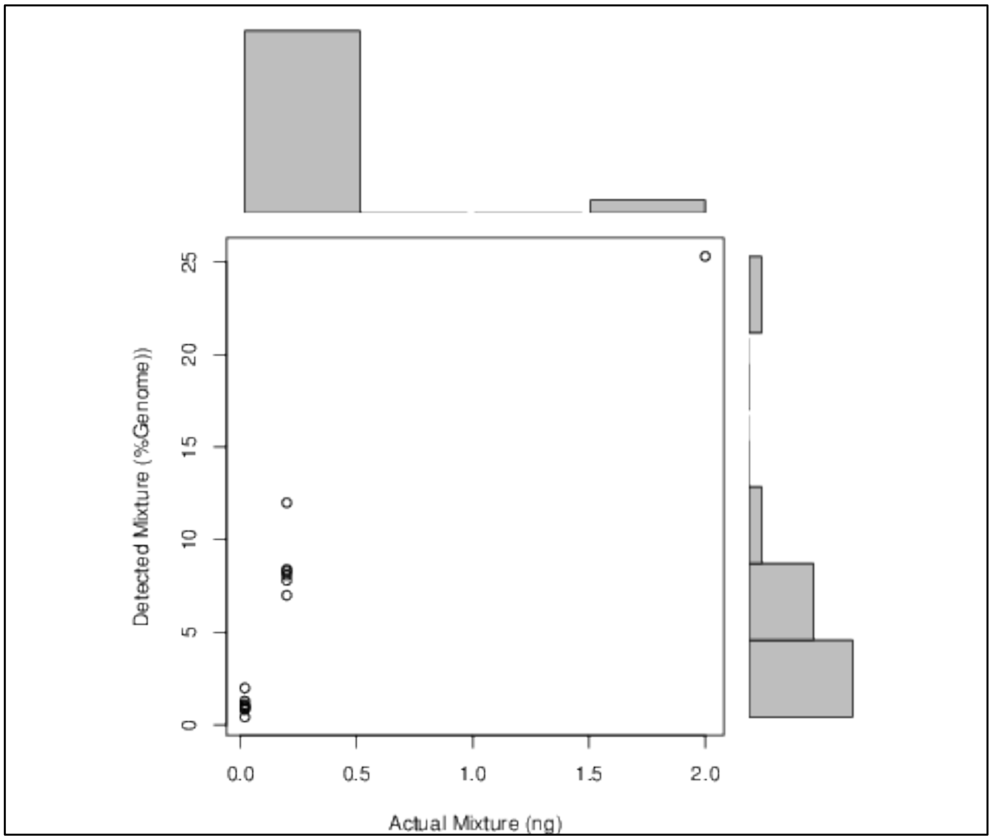


Figure 9. Pearson correlation between the actual mixture and detected genome coverage.

Detection for WGA Mixture HiSeq

Often samples may have pathogens at extremely low levels. We included a step that involved whole genome amplification (WGA) of the sample to assess if WGA increased sensitivity/detectability in metagenomic analyses. Table 17 shows the detection of bacteria in WGA mixtures. In comparison to Table 14, we see the WGA mixtures do not correlate with the actual mixture either in number of the correct strain(s) detected or/and with the actual concentration of the mixture. We therefore conclude that WGA shows biases in detection.

Table 17. Bacterial detection in WGA Mixture.

Samples	Bacteria	%Genome
Ba_Mixture3-WGA	Ba(A0248/AmesAnc)	0.24%
Ba_Mixture4-WGA	No significant hits	
Ba_Mixture9-WGA	Yp(antiqua), Ba(A0248/AmesAnc)	2.7%,.18%
Bp_Mixture5-WGA	No significant hits	
Yp_Mixture1-WGA	Yp(antiqua)	3.60%
Yp_Mixture10-WGA	Yp(antiqua), Ba(CDC,A0248/AmesAnc)	3.2%,.07%
Yp_Mixture11-WGA	Yp(antiqua), Bmallei(ATCC23344)	49%,1.1%
Yp_Mixture2-WGA	Yp(antiqua)	2%
Yp_Mixture6-WGA	Yp(antiqua)	5.70%
Yp_Mixture7-WGA	Yp(antiqua)	2.30%
Yp_Mixture8-WGA	Yp(antiqua)	2.10%

Detection of Mock Yp using HiSeq

Bacterial detection for sputum and blood samples for Mock Yp HiSeq is shown in Table 18.

Table 18. Bacterial detection in Mock Yp HiSeq samples.

Bacteria	%Genome	Tissue
P. aeruginosa LESB58, N mucosa C102, Strep mitis B6	67.9%, 86.5%, 41%	Sputum
Myco tuberculosis W-148, Bp 1026b	2.8%, 0.27%	Blood
Myco tuberculosis H37Ra, Bp 1026b	1.6%, 0.2%	Blood
Yp (antiqua), Myco tuberculosis W-148	77.3%, 4.3%	Blood
Myco tuberculosis W-148, Bp 1026b	2.3% , 21%	Blood
No significant hits		Blood
No significant hits		Blood
Yp (antiqua)	56%	Blood
No significant hits		Blood
No significant hits		Blood
No significant hits		Blood
Yp (antiqua)	21.80%	Blood
No significant hits		Blood

Sequence patient clinical samples

After development of MetaGeniE and validating its results using a variety of methods, we shifted our focus to actual clinical samples from different sources. In the following section, we detail our results for different datasets. However, we first discuss in more depth how MetaGeniE works.

Detection of Mock Yp on HiSeq and GAIIx

Blood Samples

The 12 mock YP HiSeq dataset belonging to blood tissue has three replicates NTC_RepA, NTC_RepB and NTC_RepC. The three NTC replicates are the baseline samples without any bacterial presence. These 3 replicates were then spiked with Yersinia *pestis* at three different concentrations 10e3, 10e4 and 10e7 (Figure 10). MetaGeniE did not detect any bacterial population in 2 NTC replicates (RepB-NTC and RepC-NTC) confirming the experimental design. However contamination in Replicate A (RepA-NTC) due to Mycobacterium *tuberculosis* is detected. This contamination is consistently detected in all the three subsamples of this replicate at different concentration (also discussed in next Section). This finding is significant in metagenomic sequencing as contamination can result in false detection.
The Yersinia *pestis* spiked at three concentrations for each replicate is detected only at concentration 10E7 and not in lower concentrations (10e3 and 10e4). This suggests that detection of an organism is limited by the concentration/sampling. We expect that the threshold might vary according to the genome (example GC rich pathogens) and the yield per sequencing lane.

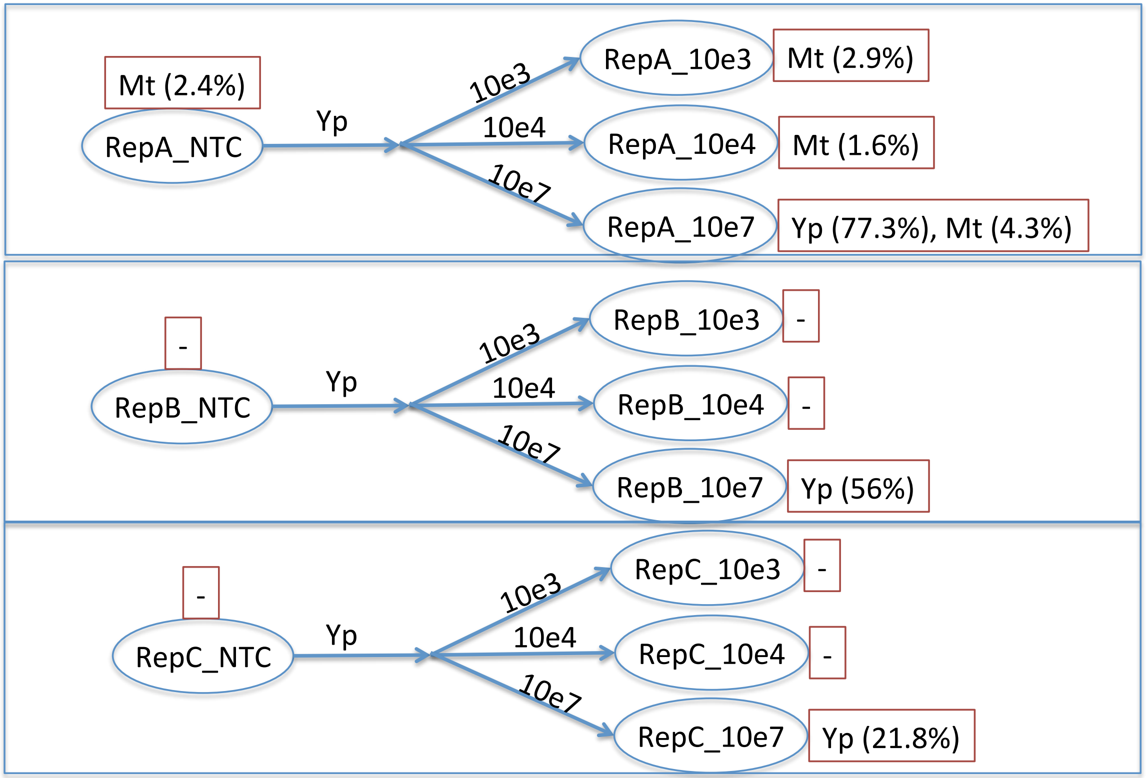


Figure 10. Three replicates for blood sample spiked with *Yersinia pestis* at 3 different concentrations with HiSeq. Yp: *Yersinia pestis;* Mt: Mycobacterium *tuberculosis*. *GAIIx sequencing for RepA similar to HiSeq and not shown here.

Contamination
Contamination in the blood sample YP-Blood repA-NTC_220bp in HiSeq due to *Mycobacterium tuberculosis* is detected as discussed above. This contamination is also consistently detected in three concentration subsamples. *Mycobacterium tuberculosis* and *Y. pestis* do not share any reads and thus no homology exist between these two organisms (Figure 11). BLAST analysis further confirmed the presence of *M. tuberculosis.*

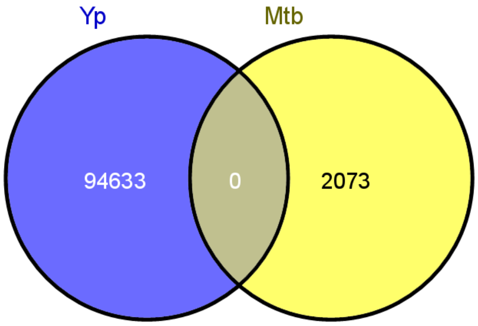


Figure 11. Total number of reads overlapping between *Mycobacterium tuberculosis* and *Yersinia pestis.*

Contamination can lead to incorrect detection of pathogen, as it is not possible to otherwise separate it from the true community without *a priori* information about it. We were able to confirm *M. tuberculosis* contamination as all the NTC samples were suppose to be baseline sample without any bacterial presence. We believe that contamination can usually assimilate during sample prep and more rigorous laboratory protocols are required to handle contamination.

Sputum samples
Twelve mock YP HiSeq and GAIIx dataset belonging to sputum tissue were sequenced (Table 19). These sputum samples belong to two different NTC samples (TGN6376 and TGN6576). These two NTC samples are spiked with *Yersinia pestis* at different concentrations. *Pseudomonas aeruginosa* was detected in TGN6376 NTC sputum samples while *Pseudomonas aeruginosa*, *Neisseria meningitis* and *Streptococcus pneumonia* were detected in TGN6576 NTC sputum samples.

Table 19. Bacterial presence in sputum samples. Pa: *Pseudomonas aeruginosa*; Yp: *Yersinia pestis*; Nm: *Neisseria meningitis*; Sp: *Streptococcus pneumonia.*

Samples	Platform	Bacteria	%Genome
YP-Sputum TGN6376-NTC_220bp	Hiseq	Pa	24.10%
YP-Sputum TGN6376_10e3_220bp*	GA IIx	Pa;Yp	10%;7.3%

YP-Sputum TGN6376-10e3_220bp	Hiseq	Pa, Yp	8.4%, 2.9%
YP-Sputum TGN6376_10e3_600bp*	GA IIx	Pa;Yp	3.9%;5.1%
YP-Sputum TGN6376-10e4_220bp*	GA IIx	Pa	7.21%
YP-Sputum TGN6376_10e7_220bp*	GA IIx	Pa	13.02%
YP-Sputum TGN6376-10e7_220bp*	GA IIx	Pa	17.35%
YP-Sputum TGN6376_10e7_600bp*	GA IIx	Pa	13.5%
YP-Sputum TGN6576-NTC_220bp	Hiseq	Pa, Nm, Sm	67.9%, 86.4%, 41.1%
YP-Sputum TGN6576-10e3_220bp	Hiseq	Pa, Yp, Nm, Sp	21.3%, 25%, 55.8%, 22
YP-Sputum TGN6576-10e4_220bp	Hiseq	Pa, Nm, Sp	43.9%, 81.8%, 37.2%
YP-Sputum TGN6576-10e7_220bp	Hiseq	Pa, Nm, Sp	20.4%, 50.8%, 22.6%

12. Compile all clinical samples then culture, extract and quantify DNA for all pathogens.

Table 20. Overview of total number of clinical sample per project.

Id	Sample Type	Number of Samples	Project	Goals
1	Mixtures HiSeq	11	DoD (Artificial)	Validation of clinical samples
2	WGA Mixtures HiSeq	11	DoD (Artificial)	Validation of clinical samples
3	Mock Yp HiSeq	18	DoD (Spiked)	Validation of clinical samples
4	Mock Yp GAIIx	15	DoD (Spiked)	Validation of clinical samples
5	Cystic Fibrosis	4	CF	Bacterial, Fungal & Viral community
6	African Samples	3	Unknown disease	Bacterial, Fungal & Viral community
7	Francisella tularensis	8	Tularemia	Bacterial community,

				Antibiotic resistance
8	Burkholderia Samples	4	Melioidosis	Bacterial community
9	Myeloma Cell lines + blood	22	TGEN Phoenix	Viral scan

13. Run deep coverage sequencing of clinical samples

Workflow of Clinical Samples
Due to the variations and limitations in metagenome analyses as discussed and importance of detection accuracy given clinical perspective, the analyses of clinical samples might require the Detection → Validation → Confirmation cycle (Figure 12).

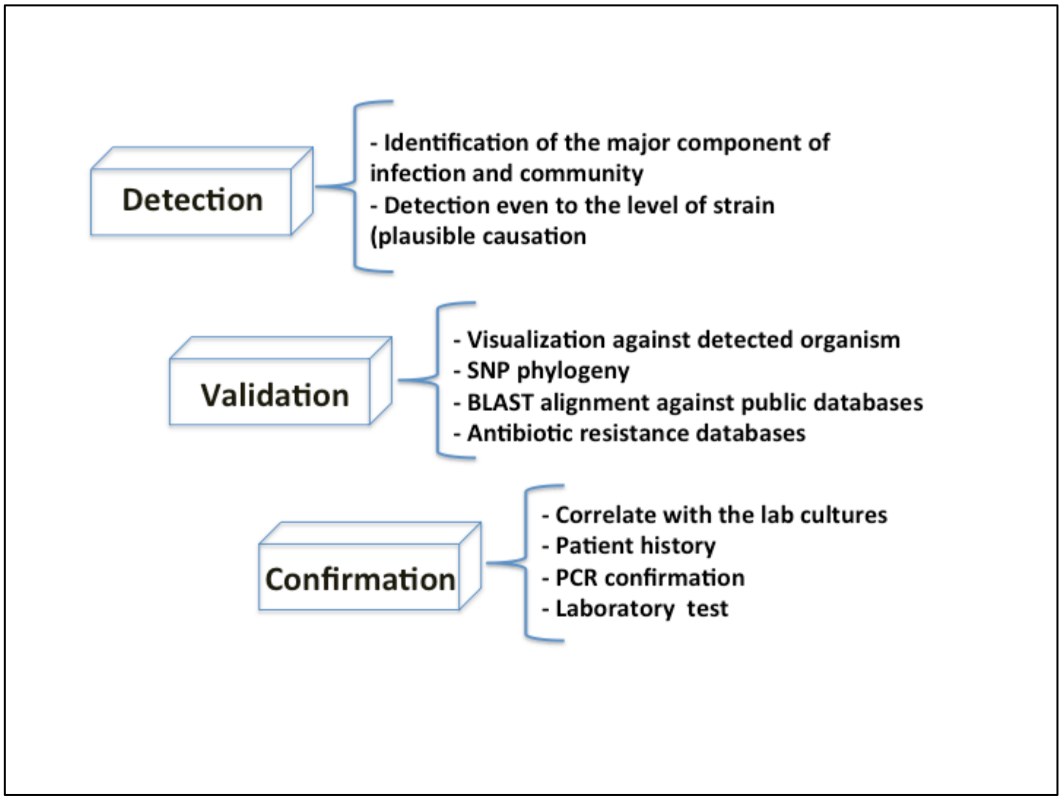


Figure 12. The workflow of the clinical sample analysis.

After detection of infection and community, the validation of clinical datasets can be done through analysis like SNP genotyping, BLAST analysis etc., depending on number of reads aligned to detected organism, type of organism. These inferences from clinical datasets can finally be confirmed with laboratory test/culture, PCR or patients clinical history. We performed Detection→ Validation → Confirmation workflow to evaluate overall performance of MetaGeniE for most of the clinical datasets presented in this Report.

We analyzed 14 clinical samples that are suspected or known to contain *F. tularensis*

DNA using our typical approach of analyzing directly extracted DNA from the tissue sample following standard extraction procedures. We also generated human DNA subtracted sequencing libraries from these for deep sequencing to evaluate the efficacy of wet bench host subtraction techniques (i.e., not bioinformatically). In addition, we evaluated a smaller insert size library preparation for increasing the quality of SNP calls from dispersed reads. A SNP call based upon a single Illumina read will be of very questionable quality, given the 0.5-1.0% error rate. Increasing the coverage of any given SNP locus to even 2X, with a read in each direction, can significantly increase the reliability of a SNP call, as the probability of an error occurring in the same location in two reads would be ca. 0.0001 (0.01 x 0.01). By decreasing the library insert size to the length of the read of the sequencing run, we achieved this with dispersed reads. This is based on a similar approach recently published by Schmitt et al. (2012). This approach does not rely on achieving an overall average coverage depth of >2X across the entire genome. An example of this is shown in Figure 13.

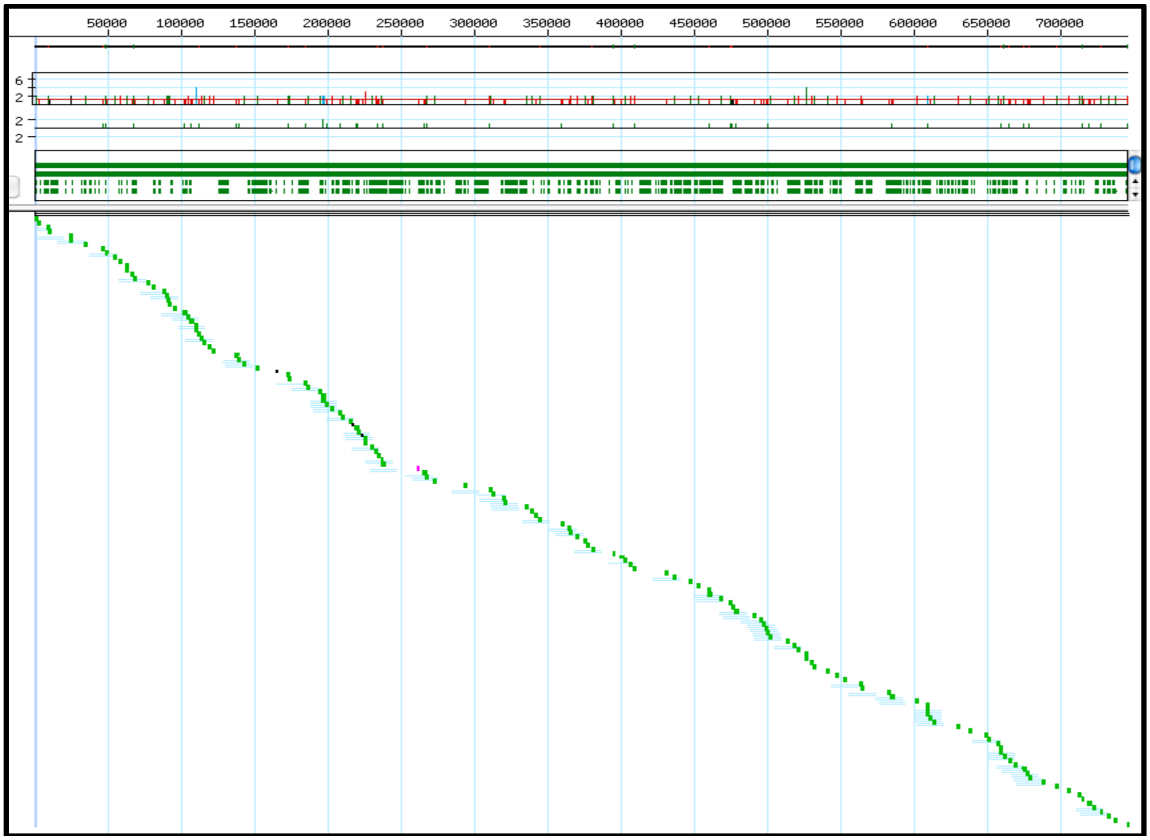


Figure 13. Alignment of ca. 40M Illumina 100bp paired reads against the NCBI *F. tularensis* OSU18 genome.

Panel A shows a pair of overlapping reads with potential SNP calls in only one of the paired reads (Figure 14). These are sequencing read errors that would not be detectable as such if only a single read were present in this location. Panel B shows a SNP call found in both reads, a potentially more reliable true SNP call. This approach is effective with actual sequencing libraries as well as supported using *in silico* modeling.

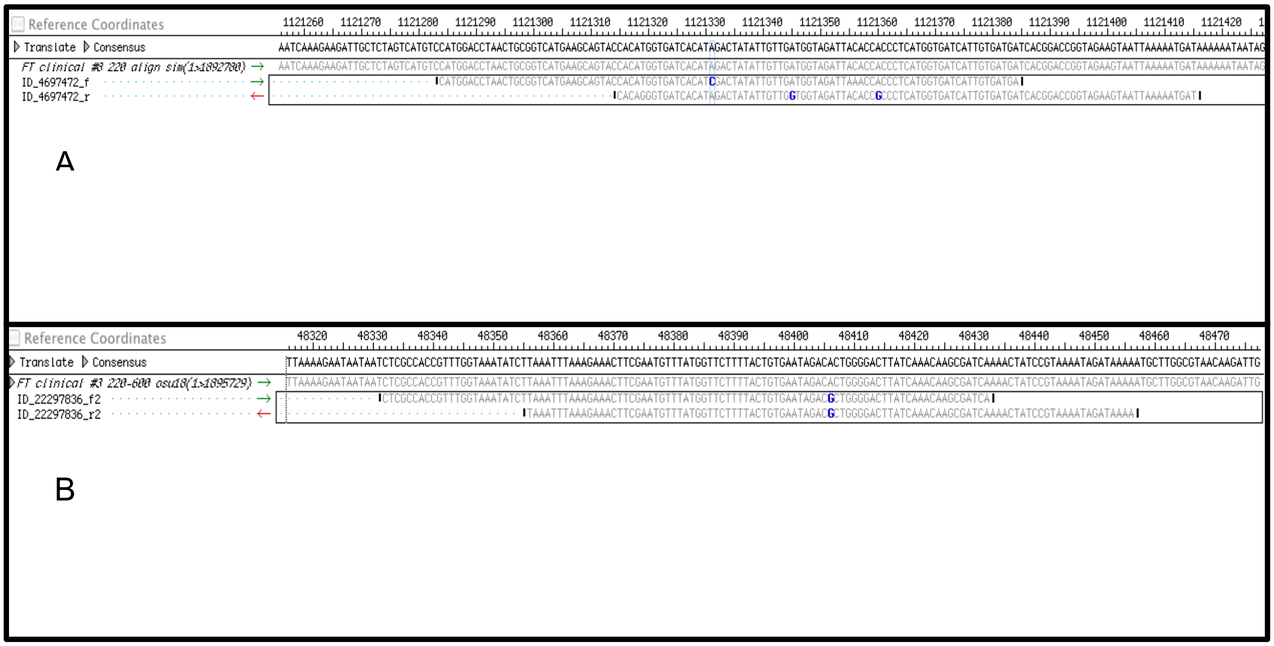


Figure 14. SNP locus found in *F. tularensis* clinical sample with overlapping paired reads. Panel A shows an overlapping read pair with SNP calls only on one of the two reads. Panel B shows a SNP call found in both reads in the sample position. SNPs are shown in blue.

Cystic Fibrosis Community

We applied MetaGeniE analysis to four clinical samples from cystic fibrosis patients, three throat and one nasopharyngeal swab sample (CF6612, CF6778, CF6780, CF6998). We presented some initial results for the cystic fibrosis dataset in previous Annual Report and report the results more fully here where we first present the complete breakdown of alignment statistics, identification and validation of four CF samples. Due small amount of DNA present in the sample, the DNA was subjected to Whole Genome Amplification prior to library preparation. We have optimized library preparation protocols to allow library preparation without WGA, as this process is known to be subject to biased amplification, especially with high GC DNA, typical of some organisms (Pinard et al. 2006). The libraries were sequenced on both the GAIIx and HiSeq Illumina NGS instruments. The sequencing of these four samples resulted in 47.01M, 36.96M, 57.91M and 53.95M reads respectively. After running the Read-Reduction module of MetaGeniE, the total number of reads were reduced to 4.86m, 15.71m, 20.01m and 32.20m, respectively. The Pathogen-Detection module resulted in total of 3.43m, 5.69m, 3.88m and 4.47m hits, respectively, against the bacterial pathogen genome database. We detected genome signatures unique to individual samples as well as genomic entities common among all four samples (Figure 15). Tables 21-24 show the pathogen reference genomes with the greatest portion of the genome covered by mapped reads (Coverage % column) for the four samples.

Table 21. Top organisms detected for sample 6612.

Organism	Coverage(%)	Coverage	Genome Size
Staphylococcus_#_aureus_!_subsp._{_aureus_}_USA300_TCH1516_chromosome	99.995	2872779	2872915
Staphylococcus_#_aureus_!_subsp._{_aureus_}_USA300_FPR3757_chromosome	99.988	2872432	2872769
Staphylococcus_#_aureus_!_subsp._{_aureus_}_JH9_chromosome	96.853	2815239	2906700
Staphylococcus_#_aureus_!_subsp._{_aureus_}_JH1_chromosome	96.859	2815213	2906507
Staphylococcus_#_aureus_!_subsp._{_aureus_}_str._Newman_chromosome	97.049	2793937	2878897
Staphylococcus_#_aureus_!_subsp._{_aureus_}_COL_chromosome	98.381	2763926	2809422
Staphylococcus_#_aureus_!_subsp._{_aureus_}_NCTC_8325_chromosome	97.79	2758996	2821361
Staphylococcus_#_aureus_!_subsp._{_aureus_}_Mu50_chromosome	95.792	2757396	2878529
Staphylococcus_#_aureus_!_subsp._{_aureus_}_Mu3_complete	95.669	2755433	2880168
Staphylococcus_#_aureus_!_subsp._{_aureus_}_N315_complete	97.881	2755176	2814816
Staphylococcus_#_aureus_!_subsp._{_aureus_}_ED98_complete	95.439	2695569	2824404
Staphylococcus_#_aureus_!_subsp._{_aureus_}_VC40_chromosome	99.847	2688446	2692570
Staphylococcus_#_aureus_!_subsp._{_aureus_}_MW2_complete	94.98	2678885	2820462
Staphylococcus_#_aureus_!_subsp._{_aureus_}_MSSA476_chromosome	94.581	2648072	2799802
Staphylococcus_#_aureus_!_subsp._{_aureus_}_TCH70_genomic_scaffold	94.82	2642227	2786578
Staphylococcus_#_aureus_!_subsp._{_aureus_}_ATCC_BAA-39_genomic_scaffold	90.478	2623968	2900112
Staphylococcus_#_aureus_!_subsp._{_aureus_}_MRSA252_chromosome	85.133	2471098	2902619
Staphylococcus_#_aureus_!_subsp._{_aureus_}_M013_chromosome	87.626	2443571	2788636
Staphylococcus_#_aureus_!_subsp._{_aureus_}_MN8_chromosome	82.709	2398214	2899588
Staphylococcus_#_aureus_!_RF122_{_complete_}	85.31	2339666	2742531

Table 22. Top organisms detected for sample 6780.

Organism	Coverage(%)	Coverage	Genome Size
Enterobacter_#_cloacae_!_subsp._{_cloacae_}_ATCC_13047	55.936	2972741	5314581
Enterobacter_#_hormaechei_!_ATCC_{_49162_}_genomic_scaffold	93.983	2216087	2357960
Enterobacter_#_cloacae_!_EcWSU1_{_chromosome	45.875	2171935	4734438
Klebsiella_#_oxytoca_!_KCTC_{_1686_}_chromosome	35.327	2110459	5974109
Streptococcus_#_salivarius_!_CCHSS3_{_complete	93.861	2081061	2217184
Enterobacter_#_hormaechei_!_ATCC_{_49162_}_genomic_scaffold	93.823	2036816	2170907
Streptococcus_#_parasanguinis_!_ATCC_{_15912_}_chromosome	92.922	2001226	2153652
Streptococcus_#_parasanguinis_!_ATCC_{_903_}_genomic_scaffold	91.657	1927010	2102412
Enterococcus_#_italicus_!_DSM_{_15952_}_genomic_scaffold	85.551	1920292	2244623
Streptococcus_#_australis_!_ATCC_{_700641_}_genomic_scaffold	80.417	1706071	2121533
Enterobacter_#_asburiae_!_LF7a_{_chromosome	33.979	1635357	4812833
Streptococcus_#_vestibularis_!_ATCC_{_49124_}_genomic_scaffold	81.505	1421522	1744094
Veillonella_#_dispar_!_ATCC_{_17748_}_genomic_scaffold	85.802	1334727	1555587
Streptococcus_#_sp._!_C150_{_genomic_}_scaffold	82.347	1329633	1614676
Veillonella_#_parvula_!_DSM_{_2008_}_chromosome	60.868	1297793	2132142
Klebsiella_#_pneumoniae_!_subsp._{_pneumoniae_}_HS11286	23.492	1253046	5333942
Klebsiella_#_pneumoniae_!_NTUH-K2044_{_chromosome	22.8	1196660	5248520
Klebsiella_#_pneumoniae_!_subsp._{_pneumoniae_}_MGH_78578	21.989	1168723	5315120
Escherichia_#_coli_!_S88_{_chromosome_}_complete	20.267	1019880	5032268
Escherichia_#_coli_!_APEC_{_O1_}_chromosome	19.914	1012049	5082025

Table 23. Top organisms detected for sample 6998.

Organism	Coverage(%)	Coverage	Genome Size
Escherichia_#_coli_!_APEC_{_O1_}_chromosome	93.321	4742614	5082025
Escherichia_#_coli_!_S88_{_chromosome	93.982	4729449	5032268
Escherichia_#_coli_!_UTI89_{_chromosome	92.683	4695079	5065741
Escherichia_#_coli_!_CFT073_{_chromosome	84.56	4423684	5231428
Escherichia_#_coli_!_536_{_complete	85.257	4210783	4938920
Escherichia_#_coli_!_ED1a_{_chromosome	78.196	4073662	5209548
Escherichia_#_coli_!_O55:H7_{_str._}_CB9615_chromosome	72.016	3879035	5386352
Escherichia_#_coli_!_KO11FL_{_chromosome	76.928	3784970	4920168
Escherichia_#_coli_!_ATCC_{_8739_}_chromosome	78.728	3736612	4746218
Escherichia_#_coli_!_O127:H6_{_str._}_E2348/69_chromosome	75.148	3731497	4965553
Shigella_#_sonnei_!_53G_{_complete	73.695	3676268	4988504
Shigella_#_flexneri_!_2a_{_str._}_301_chromosome	76.94	3544800	4607202
Escherichia_#_coli_!_IAI39_{_chromosome	63.947	3281793	5132068
Escherichia_#_coli_!_SMS-3-5_{_chromosome	61.794	3131984	5068389
Escherichia_#_coli_!_UMN026_{_chromosome	57.134	2972169	5202090
Escherichia_#_coli_!_55989_{_chromosome	57.146	2945816	5154862
Escherichia_#_coli_!_O26:H11_{_str._}_11368_chromosome	51.131	2913077	5697240
Escherichia_#_coli_!_O103:H2_{_str._}_12009_complete	52.85	2879966	5449314
Escherichia_#_coli_!_O157:H7_{_str._}_EC4115_chromosome	50.655	2822534	5572075
Escherichia_#_coli_!_O157:H7_{_str._}_TW14588_chromosome	50.484	2816387	5578816

Table 24. Top organisms detected for sample 6778.

Organism	Coverage(%)	Coverage	Genome Size
Staphylococcus_#_aureus_!_subsp._{_aureus_}_str._Newman_chromosome	73.3	2110219	2878897
Staphylococcus_#_aureus_!_subsp._{_aureus_}_VC40_chromosome	77.669	2091285	2692570
Staphylococcus_#_aureus_!_subsp._{_aureus_}_JH1_chromosome	71.782	2086341	2906507
Staphylococcus_#_aureus_!_subsp._{_aureus_}_JH9_chromosome	71.768	2086087	2906700
Staphylococcus_#_aureus_!_subsp._{_aureus_}_ED98	73.829	2085241	2824404
Staphylococcus_#_aureus_!_subsp._{_aureus_}_COL_chromosome	74.077	2081142	2809422
Staphylococcus_#_aureus_!_subsp._{_aureus_}_Mu50_chromosome	71.848	2068177	2878529
Staphylococcus_#_aureus_!_subsp._{_aureus_}_MSSA476_chromosome	73.728	2064228	2799802
Staphylococcus_#_aureus_!_subsp._{_aureus_}_N315	73.221	2061049	2814816
Streptococcus_#_salivarius_!_CCHSS3_{_complete_}_genome	92.536	2051701	2217184
Staphylococcus_#_aureus_!_subsp._{_aureus_}_NCTC_8325_chromosome	70.503	1989134	2821361
Staphylococcus_#_aureus_!_subsp._{_aureus_}_USA300_TCH1516_chromosome	69.149	1986597	2872915
Staphylococcus_#_aureus_!_subsp._{_aureus_}_USA300_FPR3757_chromosome	69.136	1986108	2872769
Staphylococcus_#_aureus_!_subsp._{_aureus_}_MW2_complete_genome	70.025	1975034	2820462
Staphylococcus_#_aureus_!_subsp._{_aureus_}_Mu3_complete_genome	68.502	1972985	2880168
Prevotella_#_pallens_!_ATCC_{_700821_}_genomic_scaffold	84.552	1957298	2314907
Staphylococcus_#_aureus_!_subsp._{_aureus_}_TCH70_genomic_scaffold	69.783	1944549	2786578
Staphylococcus_#_aureus_!_subsp._{_aureus_}_ATCC_BAA-39_genomic_scaffold	64.102	1859019	2900112
Streptococcus_#_parasanguinis_!_ATCC_{_15912_}_chromosome	86.23	1857095	2153652
Streptococcus_#_parasanguinis_!_ATCC_{_903_}_genomic_scaffold	85.684	1801440	2102412

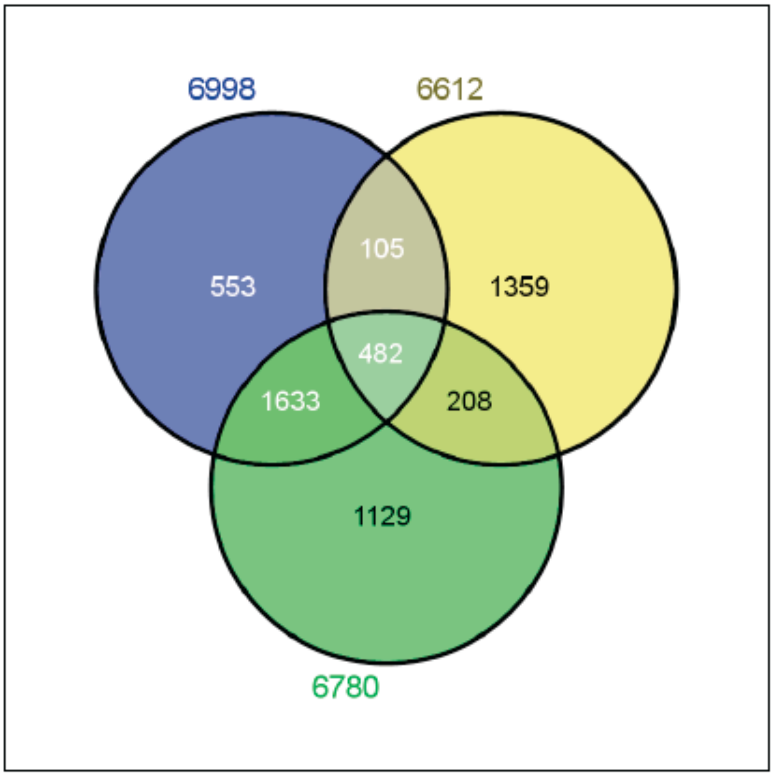


Figure 15. Overlap among the total number of organisms detected in the 3 throat samples. Venn diagram is from prior analyses but is presented to show that some organisms overlap among samples.

Alignment Statistics

We first removed low quality, redundant and human reads with the Read-Reduct module of MetaGeniE from the initial metagenomic reads (Figure 16). For these 4 samples, total ~33%-90% of the reads were filtered out. Different steps utilized by MetGeniE have varying effects of reduction/filtration on these metagenomes.

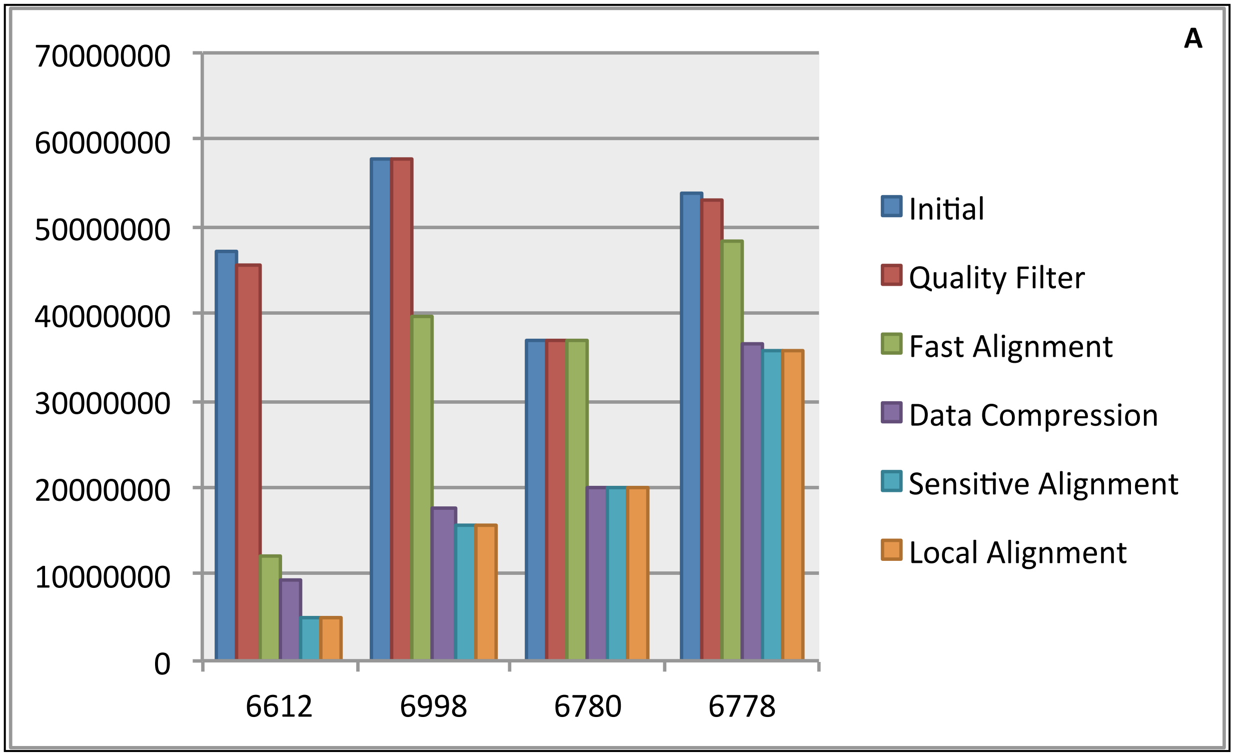


Figure 16. Iterative reduction of the metagenome reads for four clinical samples from cystic fibrosis patients. The bars represent the remaining reads after each processing step.

The remaining reads after running Read-Reduct module are input into the Patho-Detect module (Figure 17). The total number of reads mapped against public genome databases ranged from 24-68%.

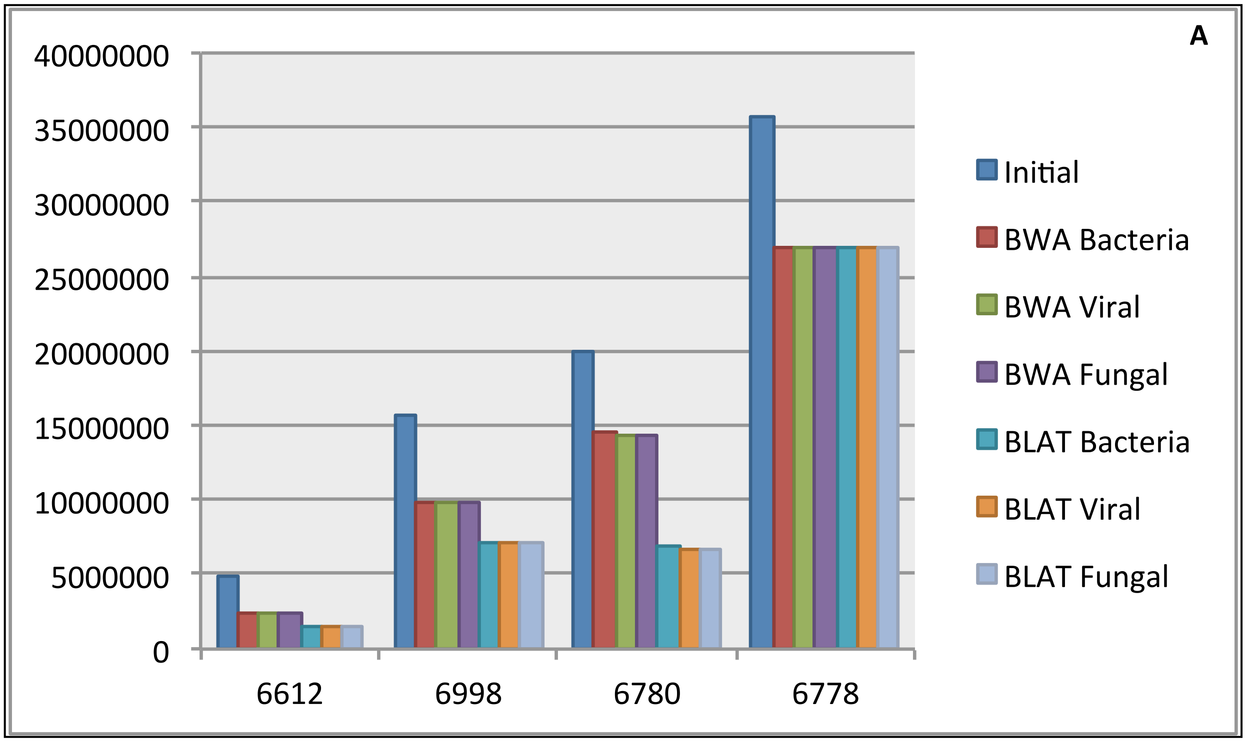


Figure 17. Pathogen detection of the metagenome reads for 4 clinical samples from cystic fibrosis patients. Each bar represents the remaining (unmapped) reads after aligning against reference database at each processing step.

Identification of Microbial Community through Genomic Reconstruction

The mapped reads in the Patho-Detect module are utilized for genomic reconstruction of the community. The genomic reconstruction (i.e., % genome mapped) of the top hits for CF samples ranged from ~55-99% (Table 25). The four CF samples represented different genomic signatures and thus different infections in each patient.

Table 25. Top five hits sorted by the genome coverage mapped per organism for four cystic fibrosis samples.

Sample	Organism	Genome Coverage %	Genome Coverage	Genome Size
6612	*Staphylococcus aureus* subsp. *aureus* USA300 TCH1516	99.995	2872779	2872915
	Streptococcus sanguinis SK36	75.607	1805813	2388435
	Enterococcus faecalis V583	50.015	1609507	3218031
	Rothia mucilaginosa DY-18	42.98	973323	2264603
	Granulicatella adiacens	88.068	669773	760519

	ATCC 49175 genomic scaffold			
6998	*Escherichia coli* APEC O1	93.321	4742614	5082025
	Shigella sonnei 53G	73.695	3676268	4988504
	Streptococcus parasanguinis ATCC 903	82.959	1744143	2102412
	Haemophilus influenzae 10810	78.98	1565017	1981535
	Veillonella dispar ATCC 17748 genomic scaffold	61.605	958312	1555587
6780	*Enterobacter cloacae* subsp. *cloacae* ATCC 13047	55.936	2972741	5314581
	Klebsiella oxytoca KCTC 1686	35.327	2110459	5974109
	Streptococcus salivarius CCHSS3	93.861	2081061	2217184
	Veillonella dispar ATCC 17748	85.802	1334727	1555587
	Escherichia coli S88	20.267	1019880	5032268
6778	*Staphylococcus aureus* subsp. *aureus* str. Newman	73.3	2110219	2878897
	Streptococcus salivarius CCHSS3	92.536	2051701	2217184
	Prevotella pallens ATCC 700821	84.552	1957298	2314907
	Fusobacterium nucleatum subsp. *polymorphum* ATCC 10953	71.999	1749354	2429698
	Haemophilus parainfluenzae T3T1	74.169	1547812	2086875

The transmission of pathogens to CF patients occurs through nosocomial, social and environmental routes. These transmissions as well as commensal microbiota represent the microbial community in CF patients. The community for each CF metagenome sample is represented by top hit per each genus detected by MetaGeniE (Figure 18). Organisms such as *Gemella*, *Granulicatella*, *Haemophilus*, *Neisseria* and *Streptococcus* are genera that are commonly found in the oral microbiome, including samples from CF patients and dominate the communities in these four samples.

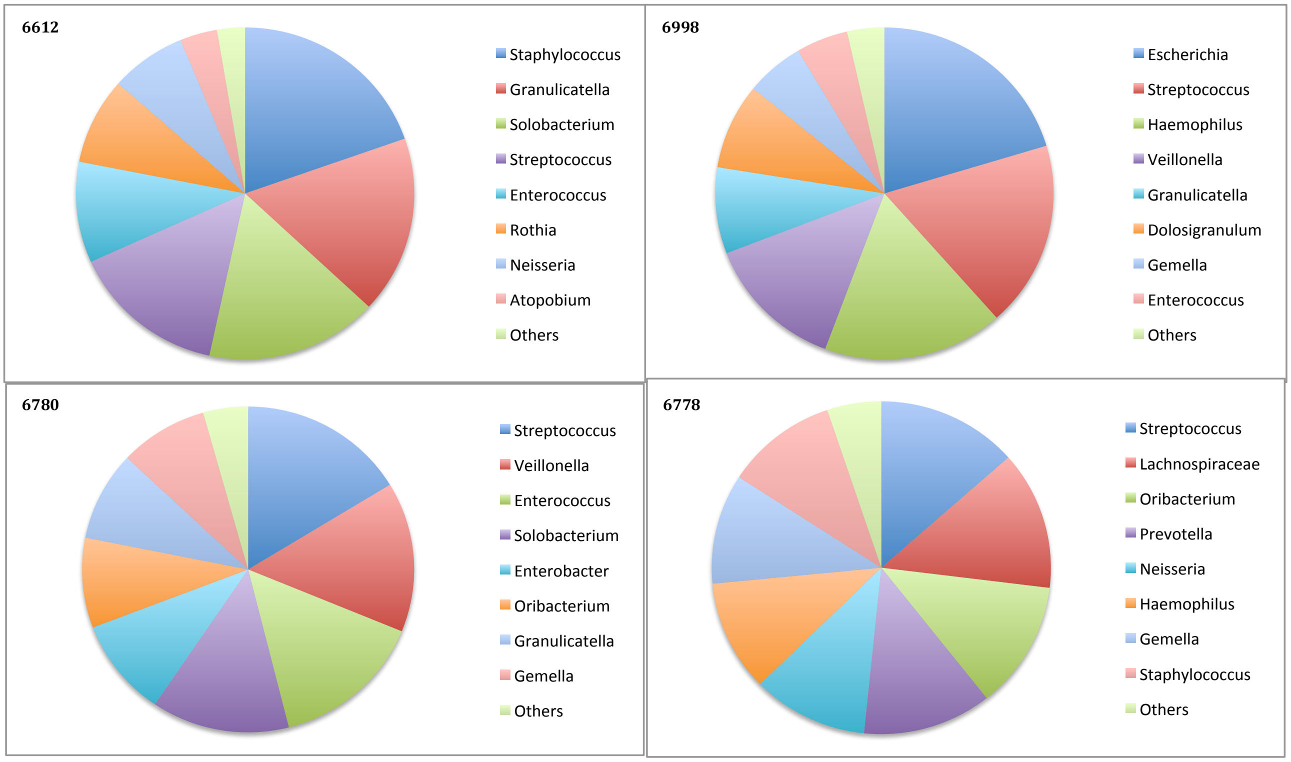


Figure 18. Community detected by deep sequencing Metagenome for four CF datasets.

Validation with Laboratory Culture

The clinical laboratory detected microbial infection across these four samples using culture-based methods (Table 27). The top hit (Table 24) and community (Figure 18) identified by MetaGeniE confirmed the lab culture results (Table 28). We were able to demonstrate that MetaGeniE can detect organisms at very low coverage, which is confirmed with culture results. The ability of MetaGeniE to correctly identify infections to the strain level, for example MRSA versus MSSA detection, demonstrates higher resolution than amplicon sequencing community analysis.

Table 27. Bacterial infection detected by Laboratory Culture. MRSA: Methicillin resistant *Staphylococcus aureus;* ENCL: *Enterobacter cloacae;* PSAR: *Pseudomonas aeruginosa;* MSSA: Methicillin sensitive *S. aureus;* PSARM: *Pseudomonas aeruginosa* – mucoid; ECOL: *Escherichia coli;* SESP: *Serratia* sp.; ENSP: *Enterococcus* sp.; HAEM: *Haemophilus influenza;* YT: Yeast not *Cryptococcus.*

BarCode	**Source Code**	**Isolate 1**	**Isolate 2**	**Isolate 3**
6612	Throat	MRSA	ENCL	PSAR*
6998	Throat	ECOL	HAEM	YT
6780	Throat	ECOL	SESP	ENSP
6778	NP	MSSA	PSAR	PSARM*

Table 28. Bacterial infection detected by MetaGeniE confirmed with the culture media as above

Sample	Culture Report	Metagenome Detection
6612	MRSA	*S. aureus* subsp *aureus* USA300 TCH1516
	ENCL	*Enterobacter cloacae* subsp *cloacae* ATCC 13047
	PSAR	*
6998	ECOL	*Escherichia coli* APEC O1
	HAEM	*Haemophilus influenzae* 10810
	YT	Fungal infection detected
6780	ECOL	*Escherichia coli* S88
	SESP	*Serratia* sp AS9
	ENSP	*Enterococcus italicus* DSM 15952
6778	MSSA	*Staphylococcus aureus* subsp *aureus* str Newman
	PSAR	P *aeruginosa* PAO1
	PSARM	*

Validation with SNP genotyping

Single nucleotide polymorphism (SNP) genotyping is widely used in analysis of WGS to accurately identify and discriminate between strains of a species. Figures 19-21 represents the phylogenetic tree for the metagenome sequences mapping to top hit detected by MetaGeniE for sample 6612, 6998 and 6778. We found that top-hits detected by MetaGeniE are also confirmed through SNP genotyping. With respect to detection up to strain level, *Staphylococcus aureus* subsp. *aureus* USA300 TCH1516 (MRSA) and *Escherichia coli* APEC O1 detected in clinical sample 6612 and 6998 respectively is also confirmed with SNP analysis. *Staphylococcus aureus* subsp. *aureus* str. Newman (MSSA) detected by MetaGeniE and confirmed by culture report is not accurately detected by SNP genotyping due to low depth of this organism in metagenome sample.

Sample 6612:

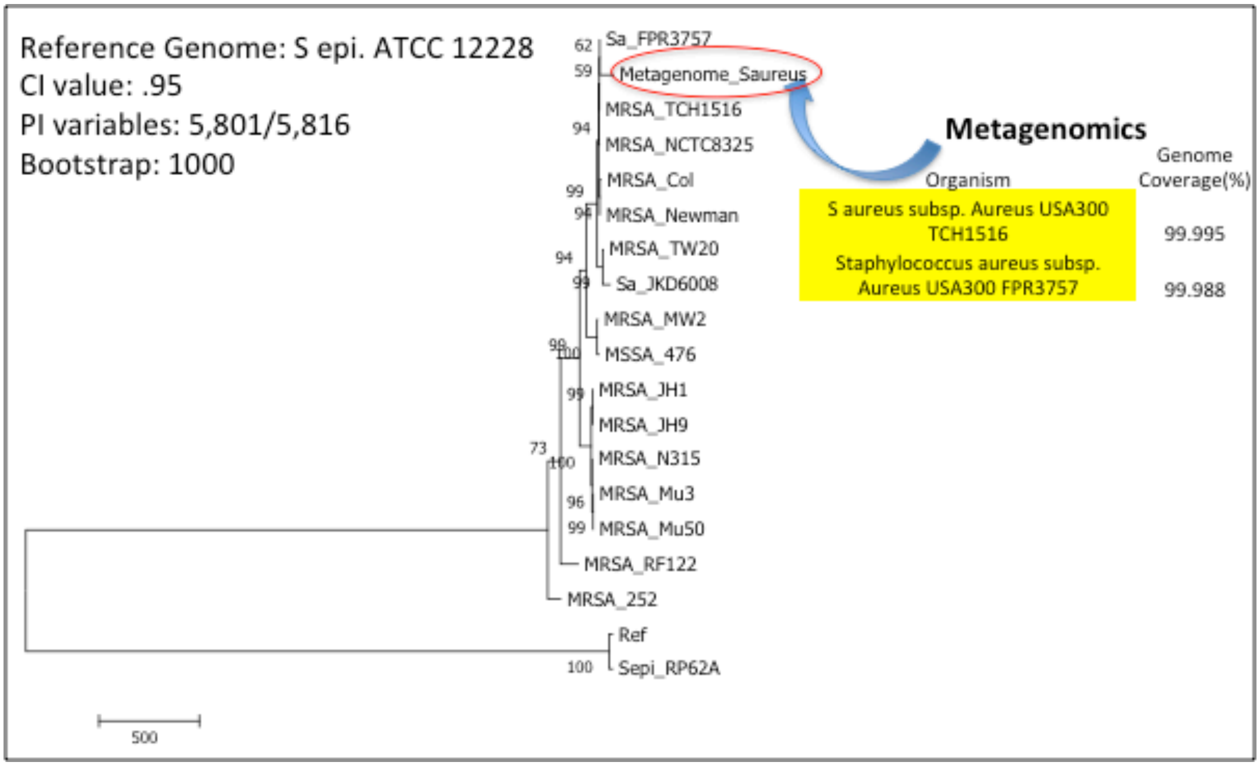
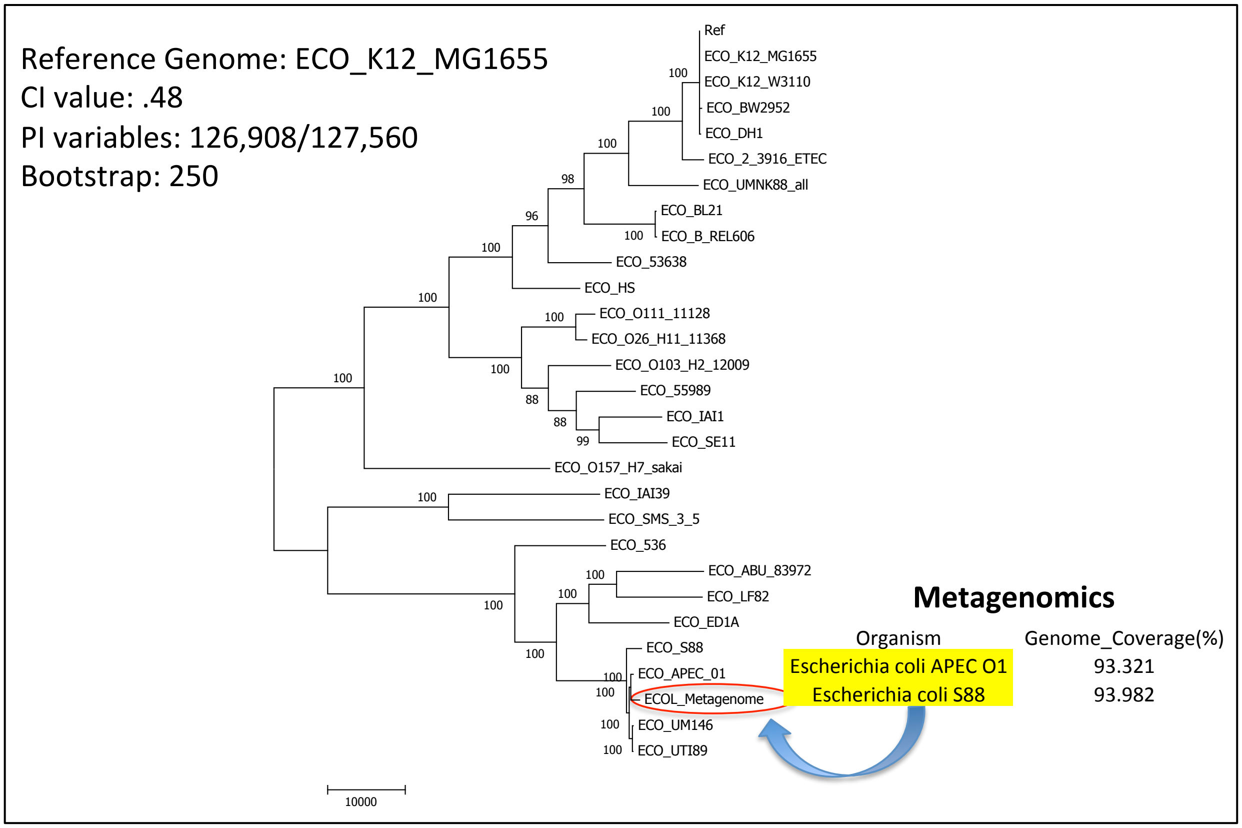


Figure 19. Phylogenetic tree representing the mapped reads of *E. coli* APEC O1 from clinical dataset and available genome in GenBank (CI – Consistency Index; PI – Parsimony Informative).

Sample 6998:



Reference Genome: ECO_K12_MG1655
CI value: .48
PI variables: 126,908/127,560
Bootstrap: 250

Metagenomics

Organism	Genome_Coverage(%)
Escherichia coli APEC O1	93.321
Escherichia coli S88	93.982

Figure 20. Phylogenetic tree representing the mapped reads of *E. coli* APEC O1 from clinical dataset and available genome in GenBank (CI – Consistency Index; PI – Parsimony Informative).

Sample 6778:

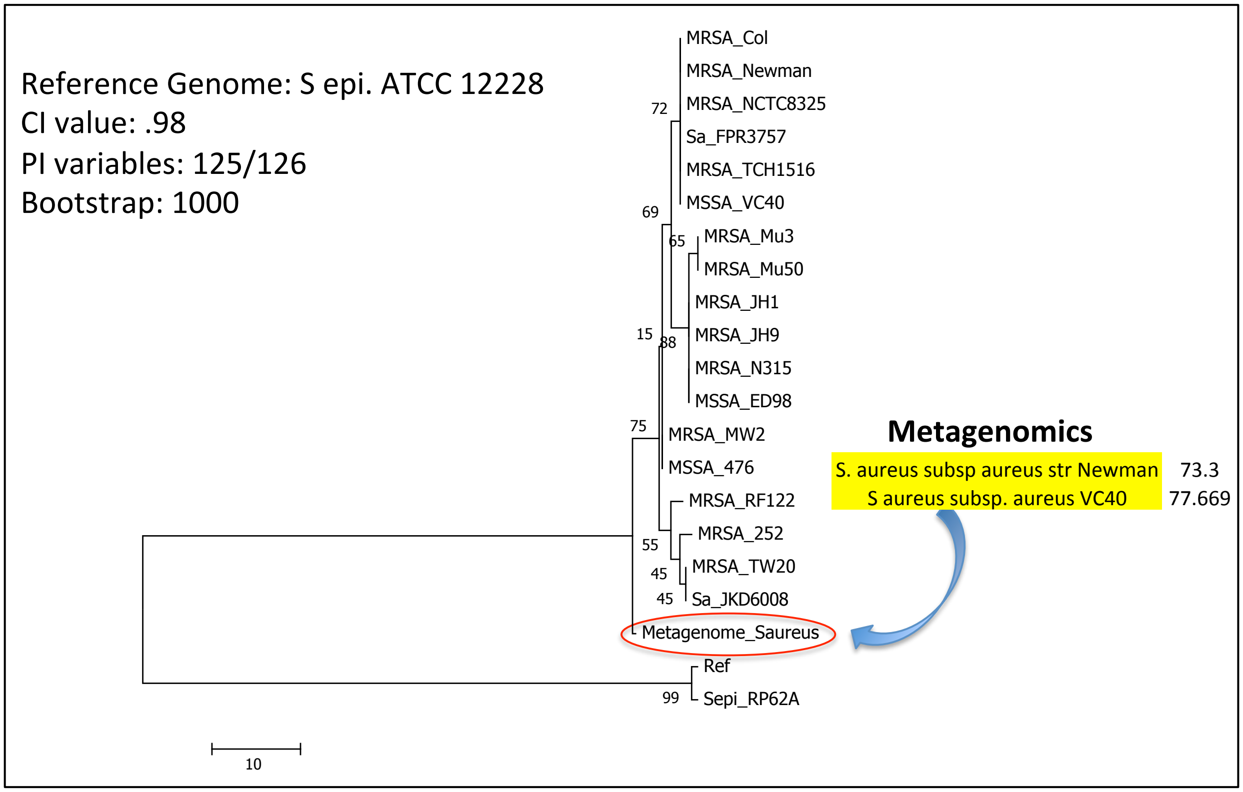
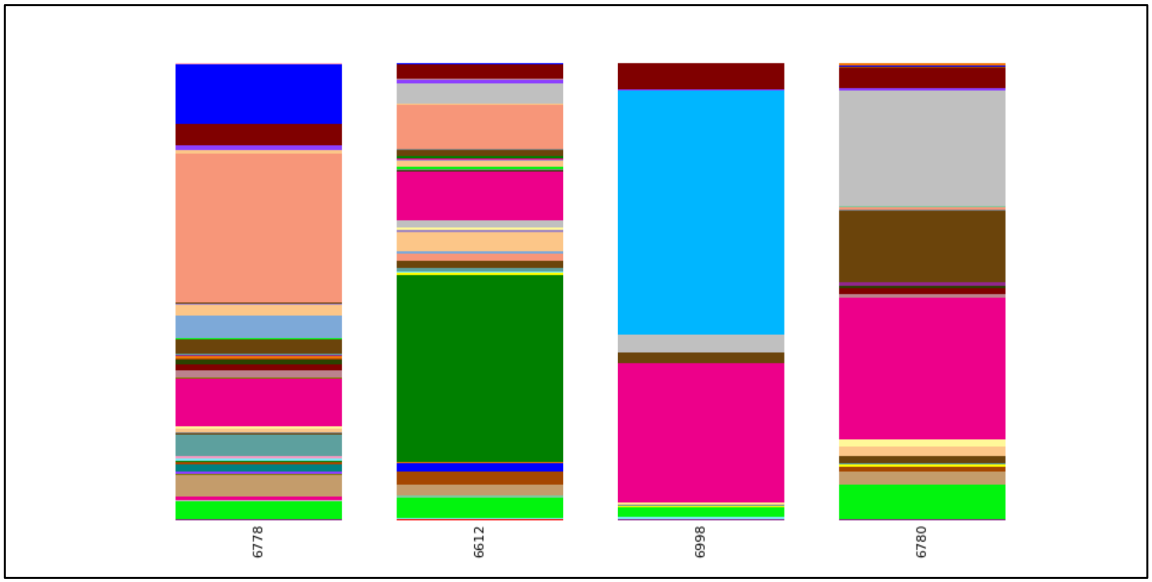


Figure 21. Phylogenetic tree representing the mapped reads of *S. aureus* Newman from clinical dataset and available genome in GenBank (CI – Consistency Index; PI – Parsimony Informative).

16S RNA versus Metagenome datasets

Comparison of community analysis between 16S RNA and deep sequencing metagenome datasets was performed with four cystic fibrosis samples.



6778	6612	6998	6780
🟦 Pseudomonadaceae	🟩 Staphylococcus	🟦 Escherichia ⬜ Enterobacteriaceae	⬜ Enterobacteriaceae 🟦 Pseudomonadaceae
Common is 3 or all samples 🟩 Rothia; 🟥 Haemophilus; 🟫 Veillonella; 🟪 Streptococcus; 🟧 Granulicatella; 🟧 Neisseria; 🟦 Gemella			

Figure 22. Community analysis with 16S rRNA analysis for 4 cystic fibrosis samples.

Figure 22 represent the community analysis performed for 4 cystic fibrosis dataset with QIIME on 16S rRNA sequencing. Figure 18 shows the top 9 organism (genus level) detected by MetaGeniE for deep sequencing. Based on these two figures, we see there is considerable overlap between the communities detected between these two technologies. Pathogens like Staphylococcus for 6612, *E coli* for 6998 and *Enterobacter* for 6780 is found in both technologies. There is also considerable overlap between species in the commensal community such as *Rothia* spp. and *Gemella* spp. among others. Many organisms detected by 16S rRNA and deep sequencing could not be compared due to higher phylogenetic level reported by 16S rRNA sequencing.
We observed that deep metagenome sequencing can lead to higher resolution, to species and even strain level. In Annual Report, we presented that MRSA and MSSA that is also confirmed with laboratory culture is detected in sample 6612 and 6778 respectively. This characterization to strain level is not possible with the 16S rRNA technology as demonstrated here. Other observation is that there seems to be no correlation between the abundance and genome coverage of bacteria found between these two technologies. Example for *Staphylococcus* detection for sample 6778, the abundance reported by QIIME is .2% while MetaGeniE reports the same organism as top hit at 73.3% genome coverage. *Pseudomonas* is detected by MetaGeniE in sample 6778 and 6780 at low genome coverage% (3.4% and 5.3%) while QIIME reports for sample 6612, 6778 and 6780.

We conclude that deep metagenome sequencing is more accurate in identifying pathogens in clinical samples at sensitivity levels meeting or exceeding traditional 16S rRNA analyses while also providing characterization of bacterial phenotypes not possible with 16S rRNA.

Francisella Samples

Eight of 14 clinical samples we sequenced from patients suffering from tularemia. These patients showed resistance to antibiotic treatment and thus were treated for longer duration.

Bacterial Interrogation

Out of 800 million total reads sequenced for these eight clinical samples, very few reads (~9,000) belonged to major component of infection *F. tularensis* (Figure 23). *Francisella tularensis* subsp. *holarctica* (with possibility of LVS/FTNF as possible strain) is detected as major component of infection in all the eight samples despite low genome coverage % (ranging from 2% to 15%).

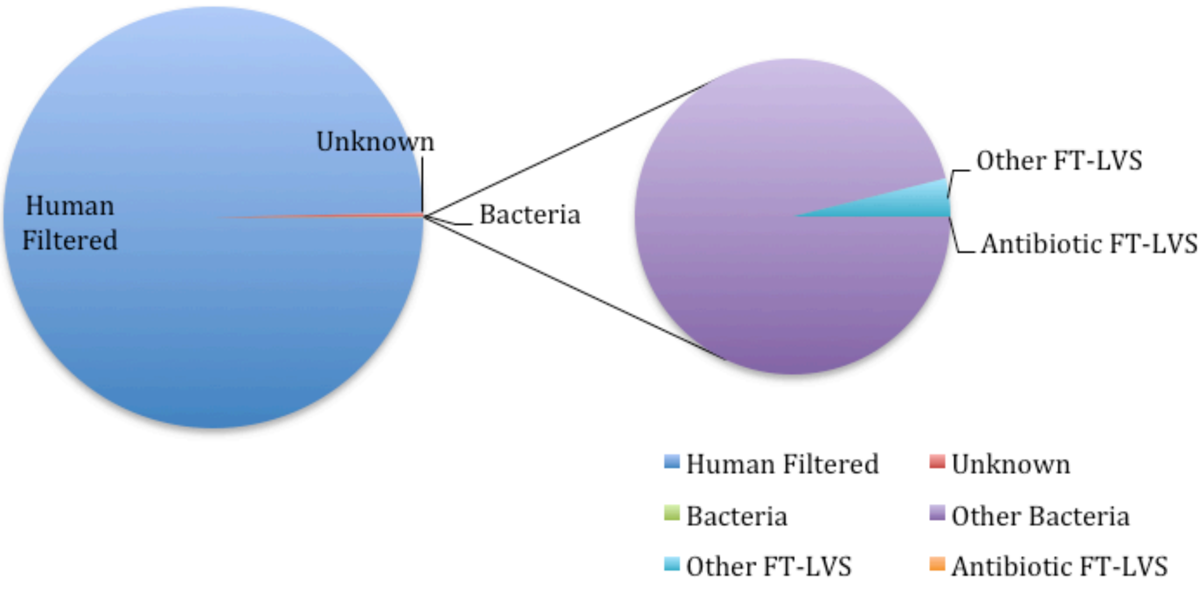


Figure 23. Overview of data analysis of *Francisella tularensis* samples.

Quantitative PCR also confirmed the presence of *Francisella tularensis* subsp. *holarctica* LVS. Further no conflict between manual canonical SNP detection despite 1x-3x coverage and assay data were found (Table 29).

We were able to detect few antibiotic resistance genes that are native to *F. tularensis* subsp. *holarctica* LVS. However as the overall genome coverage is low to report all antibiotic resistance genes, we are trying to build antibiotic resistance profile of close relative of *F. tularensis* LVS. Interestingly, coinfection with *Ralstonia picketti* is found in all the 8 samples at low coverage. *Ralstonia picketti* also natively consist of few antibiotic resistance genes and is usually a nosocomial pathogen affecting immune compromised patients (Table 30).

Table 29. Bioinformatic data and assay data.

clade	LVS SNP position	Genome SNP state (D/A)[a]	SNPs found	Depth	Ft1	Ft2	Ft3	Ft4	Ft5	Ft6	Ft7	Ft8	Ft9	Ft10	Ft11	Ft12	Ft13	F
SNP subclade					B.20/21/33	B.28/29	B.28/29	B.7/8	B.20/21/33	B.20/21/33	B.20/21/33???	B.20/21/33	B.20/21/33	B.20/21/33	B.20/21/33	B.20/21/33	B.20/21/33	B.20/21/33
B.20/21/33	14156	T/C	C	2													C	
B.20/21/33	37416	A/G	NA	0														
B.20/21/33	295469	A/G	NA	0														
B.20/21/33	359903	G/A	NA	0														
B.20/21/33	*761441	T/C	T	2						T								
B.20/21/33	1053647	A/G	NA	0														
B.20/21/33	1128028	A/C	NA	0														
B.20/21/33	1890907	C/T	NA	0														
B.Br.002	20235	A/C[b]	A	4						T								
B.Br.003	1540897	A/G[b]	NA	0	A			A										
B.Br.004	1347181	T/A	NA	0	T			T										
B.Br.005	85267	T/C[b]	T	3	T			T		T		T						
B.Br.006	1471852	A/G	NA	0	G			A										
B.Br.007	1265378	C/T	T	2	T			C			T		T					
B.Br.008	912987	T/G	NA	0	G			G			G							
B.Br.009	1219069	G/A	NA	0	A			A			A							
B.Br.010	426876	G/A[b]	NA	0	A			A			(-)							
B.Br.011	672014	A/G	G	1	G			G			G		G		G			
B.Br.012	1607966	T/A[b]	NA	0														
B.Br.013	1213436	G/A[b]	NA	0	G	G	G	A	G	G	G	G	G	G	G	G	G	G
B.Br.014	1083162	C/A[b]	A	5	A	A	A	(-)	A	A	(-)	A	A	A	A	A	A	A
B.Br.020	1102781	C/T[b]	NA	0	C	T	T	T	C	C	C	C	C	C	C	C	C	C
B.Br.021	927939	T/C[b]	NA	0	C	C	C	C	C	C	C	C	C	C	C	C	C	C
B.Br.022	1044580	A/G	NA	0	G	G	G	G	G	G	G	G	G	G	G	G	G	G
B.Br.023	174752	A/C[b]	C	1			C											
B.Br.025	299703	A/G[b]	NA	0														
B.Br.026	611155	A/C	2C/2A	4	A	C	C	C	A	A	A	A	A	A	A	A	A	A
B.Br.027	382784	T/G[b]	NA	0		T	T											
B.Br.028	1832065	T/G	NA	0		T	T											
B.Br.029	769531	A/G	NA	0		G	G											
B.Br.030	407121	T/G	G	3		G	G										G	
B.Br.031	519200	A/G	NA	0		G	G											
B.Br.032	155001	A/G[b]	NA	0		G	G											
B.Br.033	1,716,449	T/C	NA	0	C	C	C	C	C	C	C	C	C	C	C	C	C	C
B.Br.034	1,302,760	A/G	G	2			G											
B.Br.035	164,258	A/C	C	6						C								
B.Br.036	554,148	A/C	C	2					C									
B.Br.037	215,319	T/C	C	2						C								
B.Br.038	1,669,446	C/T	NA	0														

Color code	Sequenced	Bioinformatic data	Assay data
⬛	No	Not tested	Not tested
▩	(+)	(-)	Not tested
🟥	No	Not tested	(-)
⬜	No	Not tested	(+)
🟦	(+)	(+)	Not tested
🟨	(+)	(-)	(+)
🟩	(+)	(+)	(+)

Table 30. The major component of infection *F. tularensis* and coinfection *R. picketti* detected in eight samples.

Patient number	City of turkey	Fever	FT Reads	Picketti Reads	Previous antibiotics	Response of therapy	Hospitali-zation	Treatment of tularemia
Patient 2	cankırı	yes	483	167	cam	no	yes	cipro,amikasin
Patient 1	corum	yes	261	89	gentamisin	no	yes	genta,doksisik,tetrasik,cipro
Patient 3	yozgat	yes	1835	10	penis,sefaz,genta,tetrasiklin	no	yes	cipro,rifampisin
Patient 5	corum	yes	137	158	sam,bit	no	yes	cipro, genta, tetrasiklin
Patient 6	corum	yes	3291	371	cam,gentamisin	no	no	gentamisin, doksisiklin
Patient 8 (adult)	bala/ankara	yes	1041	3	cam	no	no	
Patient 9	ankara	yes	837	38	cam	no	yes	genta,doksi, siproflok,amik,ornid,vanko
Patient 13	ankara	yes	953	106	cam, sam, klindamisin	no	yes	ganta, doksisilin, cipro

Additional validation shows that *F. tularensis* and *R. picketti* do not share any homology with each other as no reads are shared between these two organisms (Figure 24).

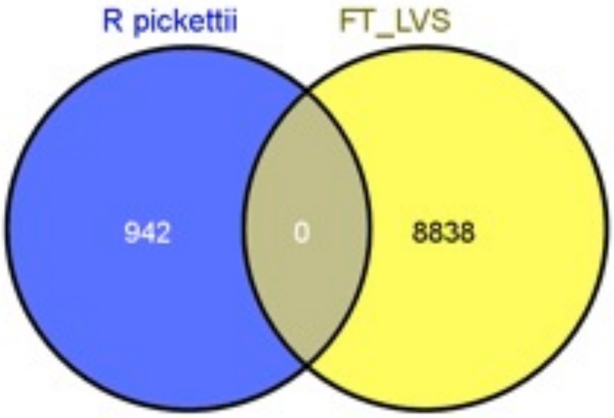


Figure 24. The overlap between the major component of infection with *F. tularensis* and coinfection with *R. picketti.*

Extracting the reads mapping to *R. picketti* and then performing blast on these sequences shows that these sequences very specifically map only to the *R. pickettii* confirming their presence. Aligning these sequences against antibiotic resistance genes resulted in few of the reads hit the resistance database. Few of the reads mapped to multidrug efflux system subunit MdtA. This efflux is part of a tripartite efflux system composed of MdtA, MdtB and MdtC, which confers resistance against novobiocin and deoxycholate. These finding are indicator of presence of co-infection that might be linked with increased antibiotic resistance.

Viral Interrogation

The viral interrogation is performed for the unmapped *Francisella* read against latest viral database (build 61). The bacterial interrogation was already submitted with the 2013 Annual Report.

Table 31. Detection of Viral infection in 3 samples.

Sample Id	*Pathogen*
FT3	Human parvovirus B19; Hepatitis B virus
FT5	Human parvovirus B19;
FT6	Human parvovirus B19;

We were able to detect two viral infections in 3 patients at very low-level (Table 31). For sample 3, only 12 reads mapped to Hepatitis B virus out of nearly 100 million sequencing read. After manual inspection of these low number of reads, we confirmed the presence of this viral infection. These 12 reads that mapped to Hepatitis B virus is visualized against Hepatitis B genome in Figure 25.

Figure 25. Visualization of 12 reads mapped against Hepatitis B reference genome

Personal communication with the doctor about detection of Hepatitis B in Patient 3 that is usually asymptomatic led to further investigation. After clinical testing, it was confirmed that patient 3 is infected with Hepatitis B.

Burkholderia Samples

For melioidosis the numbers of reads removed ranged from 76%-99% (Figure 26).

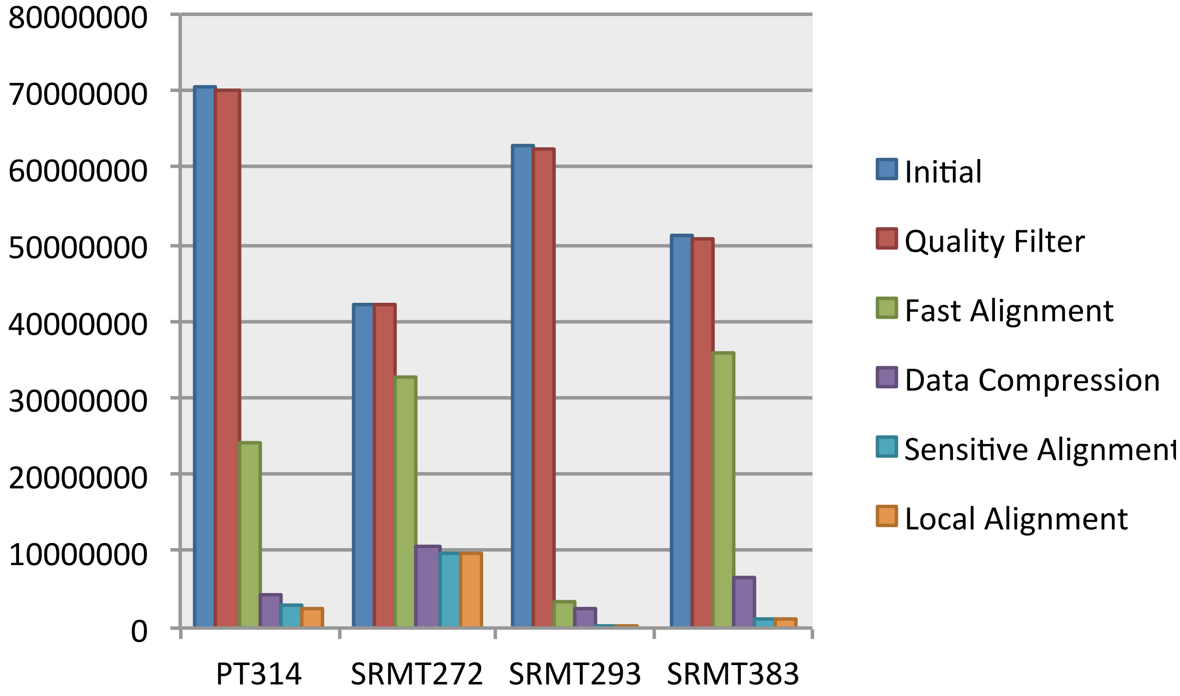


Figure 26. Iterative reduction of the metagenome reads from 4 clinical samples for melioidosis patients. The bars represent the remaining reads after each processing step.

The total number of reads mapped against public genome databases was 23-47% for melioidosis samples (Figure 27). From the four samples (2 urine and 2 sputum) collected from melioidosis patients, *Burkholderia pseudomallei* is identified as a dominant bacterium in 2 urine samples. Besides *Burkholderia pseudomallei,* species commonly found oral bacterial communities such as *Prevotella* sp., *Streptococcus* sp., and *Rothia* sp., were also detected in 2 sputum samples.

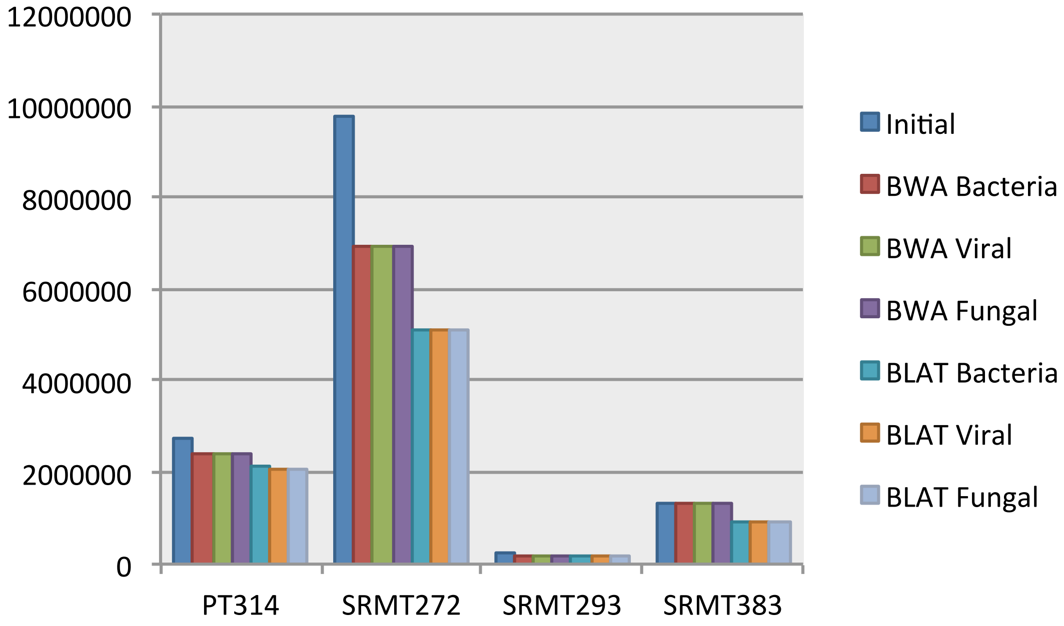


Figure 27. Pathogen detection of the metagenome reads for melioidosis patients. Each bar represents the remaining (unmapped) reads after aligning against reference database at each processing step.

Myeloma Samples

For the myeloma samples, the numbers of reads removed were higher, with removal of 98%-99% reads (Figure 28). The myeloma cell line samples have high human DNA content and therefore resulted in higher amounts of read reduction than the other sample type.

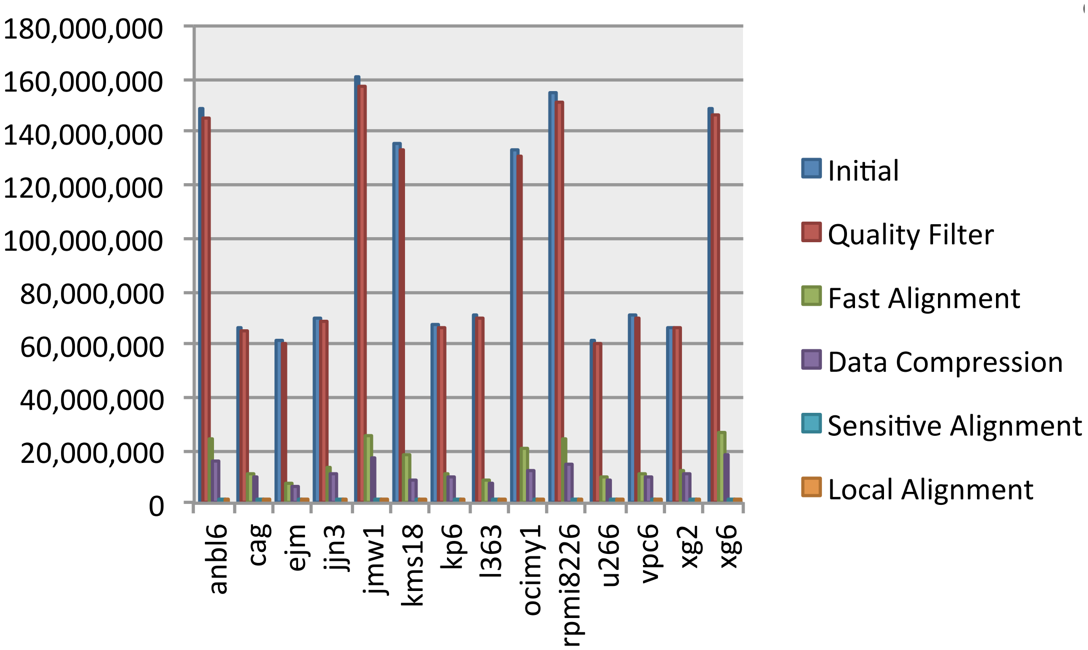


Figure 28. Iterative reduction of the metagenome reads from 14 clinical samples from myeloma patients. The bars represent the remaining reads after each processing step.

For myeloma, the number of mapped reads against viral database was low (<1%) for 6 out of 14 samples and ranged from 0.5%-29% (Figure 29).

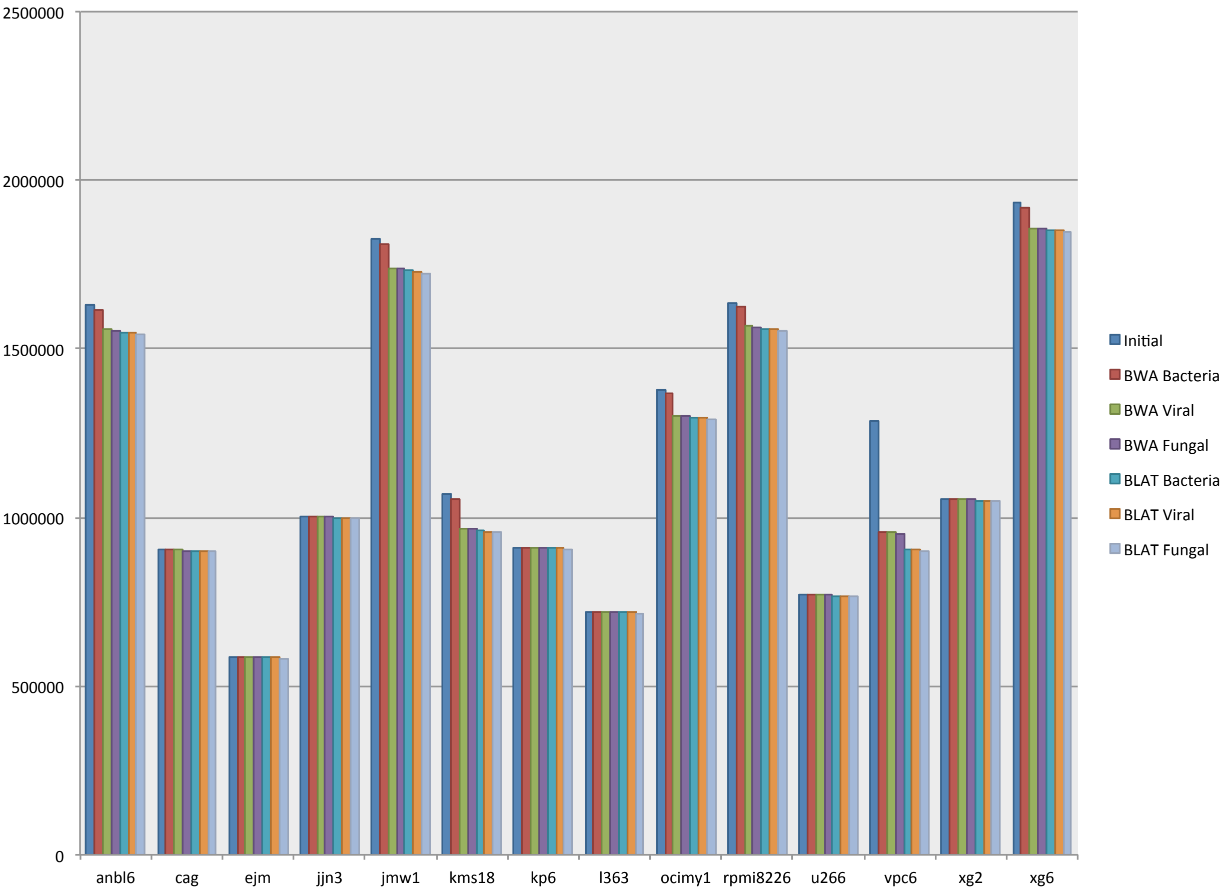


Figure 29. Pathogen detection of the metagenome reads from 14 clinical samples from myeloma patients. Each bar represents the remaining (unmapped) reads after aligning against reference database at each processing step.

No significant viral community was detected for these 14 human cell line samples. The eight blood samples showed viral community that seems to be mostly related to dietary habits of the patients. The myeloma samples were spiked with PhiX174 at about 1% in all the samples. MetaGeniE is able to reconstruct 99% of this “control” genome from these samples showing high sensitivity for its detection.

African Skin Snip Samples

Identification of unknown community is especially important in clinical samples for rapid identification in the case of a sudden disease outbreak. Three clinical samples collected from skin snips of children suffering from nodding disease from Africa were collected. We were able to detect *Wolbachia* endosymbiont of *Onchocerca ochengi*, commonly found bacteria in Africa along with some lesser-known pathogens that might be causative agent for this disease (Table 32).

Table 32. The bacterial community detected in three skin snip samples (Only one representative is selected per genus).

P145	137	140
Comamonas testosteroni CNB-2 chromosome	Pantoea vagans C9-1 chromosome	Wolbachia endosymbiont of Onchocerca ochengi
Delftia acidovorans SPH-1 chromosome	Erwinia billingiae Eb661 chromosome	Leptospira borgpetersenii serovar Hardjo-bovis
Dechlorosoma suillum PS chromosome	Klebsiella pneumoniae subsp. pneumoniae	
Sphingobium yanoikuyae ATCC 51230	Serratia proteamaculans 568 chromosome	
Ochrobactrum anthropi ATCC 49188	Cronobacter turicensis z3032 chromosome	
Acinetobacter sp. NIPH 809	Salmonella enterica subsp. arizonae	
Brevundimonas diminuta ATCC 11568	Enterobacter asburiae LF7a chromosome	
Citrobacter sp. 30 2 genomic	Escherichia coli IAI1 chromosome	
Ochrobactrum anthropi ATCC 49188	Edwardsiella tarda EIB202 chromosome	
Sinorhizobium fredii USDA chromosome	Rahnella sp. Y9602 chromosome	
Bordetella pertussis CS chromosome	Brenneria sp. EniD312 chromosome	

Tuberculosis Patients

Clinical samples from tuberculosis patients were analyzed with MetaGeniE (Figure 30).

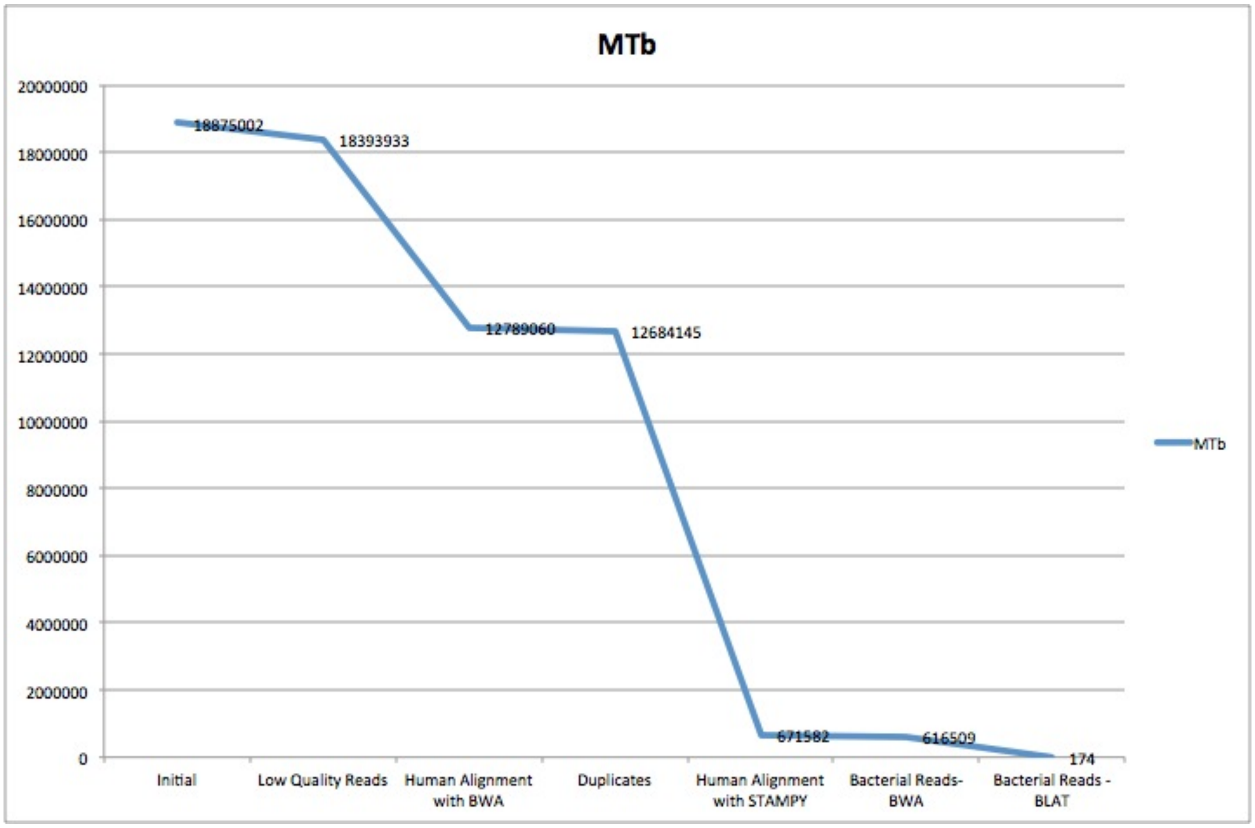


Figure 30. The breakdown of the total number of reads filtered against human database and aligned against bacterial datasets.

We were not only able to detect multi-drug resistant strains of tuberculosis for these samples but also the presence of *Acinetobacter baumannii* (Table 33).

Organism	Coverage(%)	Coverage	Genome Size
Mycobacterium_#_tuberculosis_!_CCDC5079_{_complete	26.351	1163227	4414325
Mycobacterium_#_tuberculosis_!_X122_{_scaffold	26.323	1159439	4404714
Mycobacterium_#_tuberculosis_!_HN878_{_scaffold	26.23	1155365	4404672
Acinetobacter_#_baumannii_!_1656-2_{_chromosome	11.397	449127	3940614
Acinetobacter_#_baumannii_!_TYTH-1_{_chromosome	11.222	444102	3957368
Acinetobacter_#_baumannii_!_TCDC-AB0715_{_chromosome	10.684	442145	4138388

Table 33. The top three hits per genus detected in tuberculosis samples.

To further confirm the presence of *A. baumannii*, we tried to look at the overlapping regions (reads) shared between *M. tuberculosis* and *A. baumannii* (Figure 31). Any significant overlap between these two organisms can be attributed to homology and usually presence of organism with higher genome coverage. However we found overlap between only the strains of *Mycobacterium* (as expected) and none with *Acinetobacter*, we conclude that both *M. tuberculosis* and *Acinetobacter baumannii* are present.

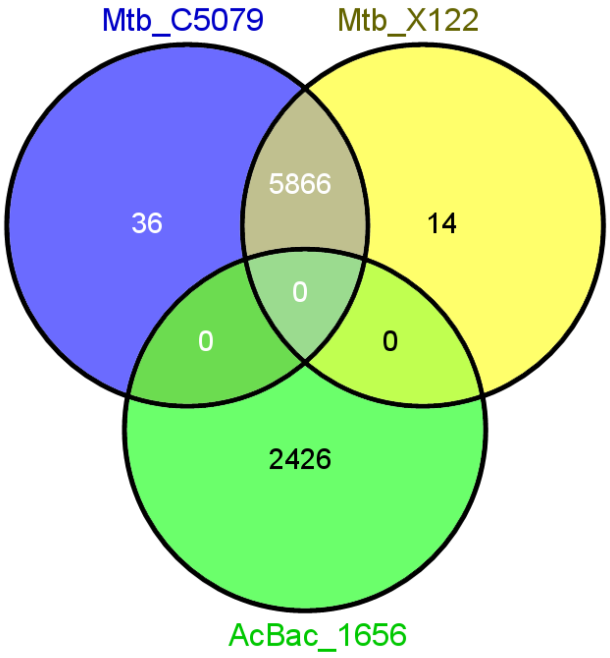


Figure 31. The overlap between two strains of *Mycobacterium tuberculosis* and *Acinetobacter baumannii*.

KEY RESEARCH ACCOMPLISHMENTS

- MetaGeniE Completion. We have created a metagenomic application for analysis of clinical samples previously referred to as IRIMAS.
- Submission of MetaGeniE for publication
- Development of the Read-Reduct module that sequentially filters and reduces the low quality, redundant and human reads
- Efficient read mapping of pathogen sequences using Patho-Detect module
- Accurate and sensitive detection of pathogens in clinical samples
- Development of SNP based framework and workflow for pathogen specific reads for detailed characterization of samples
- Evaluated human DNA subtraction methodology
- Overlapping paired-end read methodology for dispersed read SNP calling

REPORTABLE OUTCOMES

Presentations

Birdsell, D., Y. Özsürekci, M. Çelik, E. Karadağ-Öncel, A. Johansson, A.J. Vogler, A. Rawat, P. Keim, M. Ceyhan, and D.M Wagner. Whole-genome genetic analysis of *F. tularensis*-positive clinical samples from Turkey. Medical Biodefense Conference. Bundeswehr Medical Academy. Munich, Germany. October 2013.

Ozsurekci, Y., D. Birdsell, A.E. Aycan, V. Gurbuz, A. Rawat, J. Schupp, A.J. Vogler, P. Keim, M. Ceyhan, and D.M. Wagner. Whole genome genetic analysis of *F. tularensis*-positive clinical samples from Turkey. Annual meeting of the Interscience Conference on Antimicrobial Agents and Chemotherapy, Denver, Colorado, September, 2013.

Özsürekci, Y., D. Birdsell, M. Çelik, E. Karadağ-Öncel, A.J. Vogler, P. Keim, M. Ceyhan, and D.M. Wagner. Phylogenetic analysis of *Francisella tularensis* subspecies *holarctica* isolates from Turkey. International meeting on *Francisella tularensis* and Tularemia, Ürgüp, Turkey, June 2013.

Wagner, D.M. Signatures, membership, and whole genome SNP typing: examples using *Yersinia pestis*. 11th International Symposium on Protection against Chemical and Biological Warfare Agents, Stockholm, Sweden, June 2013.

Foster, J.T. Invited Seminar. Genomic approaches to disease epidemiology. University of New Hampshire. May 2013

Foster, J.T. Invited Seminar. Genomic approaches to disease epidemiology in livestock and wildlife. University of Vermont. March 2013

Keim, P. Genomics for Public Health Applications. CDC invitation lecture. Atlanta GA.

September 2012.

Foster, J.T. et al. Testing Limits of Next Generation Sequencing for Diagnostics. ASM2012—American Society for Microbiology Annual Meeting, San Francisco, CA, June 2012. Invited Talk at ASM2012.

Rawat, A. et al. Characterization of Microbial Infection in Cystic Fibrosis Clinical Samples using Deep Metagenomic Sequencing. ASM2012—American Society for Microbiology Annual Meeting, San Francisco, CA, June 2012.

Publications

Please note that while we have numerous manuscripts in preparation or resubmission after review at a journal, we do not currently have any publications as a result of this work.

CONCLUSION

Among the primary translational goals of metagenomic sequencing of clinical samples is accurate and comprehensive identification of pathogens within a sample. Identification of these microbes with high fidelity can help in generating insights for personalized medicine and rapid and accurate clinical management for patients suffering from infections. However, metagenomic samples have high variation in terms of human reads, bacterial community that is oftentimes limited by sampling, community complexity as seen in CF samples, and GC rich pathogens like *Burkholderia*.

MetaGeniE provides novel insights and analyses of clinical datasets, detecting and characterizing the major component of community that is usually the infectious agent of interest as well as identifying the other microbes in this community. Furthermore, we are able to validate these detections with laboratory culture, and available patient information. Finally, we are able to take these analyses a step beyond nearly all current approaches and can use both SNP genotyping and determine features such as antibiotic resistance to precisely characterize these pathogens.

Our study is able to provide insights into clinical samples where the pathogen causing infection is known (*F. tularensis*, *Burkholderia*), while also detecting additional co-infections (e.g., *Ralstonia* in *F. tularensis* samples), as well as identifying samples in unknown communities (artificial mixtures, CF, skin snip samples). Considering the blind test analysis of using known mixtures on the HiSeq instrument, no false positives were detected, which is quite important for reliable detection in clinical datasets. With samples with high genome coverage percentages, we are able to strain type detected organisms in our samples and confirm their placement in a SNP based phylogeny. As the genome coverage percentage decreases, the ability to strain type an organism is reduced but the ability to use multiple potential loci across the genome makes this approach superior to targeted approaches such as those using 16S rRNA. For low genome coverage (~ <15%), mapping canonical SNPs manually can be applied as seen in the *F. tularensis* samples. Our novel approach for metagenomic analysis has been accurate in identifying pathogens in clinical samples at sensitivity levels meeting or exceeding traditional 16S rRNA analyses, while also providing characterization of bacterial phenotypes not possible with 16S rRNA. Additional analysis besides SNP genotyping such BLAST analysis, antibiotic resistance detection, lab culture and annotation can benefit the study of various metagenomic datasets.

SO WHAT

Rapid and accurate identification of pathogens responsible for infectious disease is one of the Holy Grails of medicine. Our research provides a vital next step in this work where all of the genetic pieces of different organisms in a sample can be taken, a veritable soup of host and microbial DNA, and then the specific sequences of interest can be found and identified. These basic results are not new. Identifying various organisms in complex backgrounds such as amidst host DNA or in environmental samples has been done for nearly ten years. What is new however with our approach is the ability to know far beyond simple species identification and be able to fully characterize the genomes of the organisms detected at only trace levels. For example, we can determine which antibiotics a bacterium is resistant to, which has immediate relevance for clinical treatment. We also may know which lineage a microbe belongs to, a fact critically important for disease epidemiology especially if one lineage has low pathogenicity and the other is highly pathogenic. Moreover, our approach looks for a whole range of different pathogens. We were looking specifically for bacterial pathogens in most instances but found infections with viral agents as well. Thus, the ability to detect coinfections is possible as well. We expect our approach will be a part of the future of medical diagnostics.

REFERENCES

Chen, L., Z. Xiong, L. Sun, J. Yang, and Q. Jin. 2012. VFDB 2012 update: toward the genetic diversity and molecular evolution of bacterial virulence factors. Nucleic Acids Research **40**:D641-645.

Dimon, M. T., H. M. Wood, P. H. Rabbitts, and S. T. Arron. 2013. IMSA: Integrated Metagenomic Sequence Analysis for identification of exogenous reads in a host genomic background. PLoS ONE **8**:e64546.

Eisen, J. A. 2007. Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. PLoS Biology **5**:e82.

Huson, D. H., A. F. Auch, J. Qi, and S. C. Schuster. 2007. MEGAN analysis of metagenomic data. Genome Research **17**:377-386.

Kostic, A. D., D. Gevers, C. S. Pedamallu, M. Michaud, F. Duke, A. M. Earl, A. I. Ojesina, J. Jung, A. J. Bass, J. Tabernero, J. Baselga, C. Liu, R. A. Shivdasani, S. Ogino, B. W. Birren, C. Huttenhower, W. S. Garrett, and M. Meyerson. 2012. Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. Genome Research **22**:292-298.

Kostic, A. D., A. I. Ojesina, C. S. Pedamallu, J. Jung, R. G. W. Verhaak, G. Getz, and M. Meyerson. 2011. PathSeq: software to identify or discover microbes by deep sequencing of human tissue. Nat Biotechnol **29**:393-396.

Kozarewa, I., and D. Turner. 2011. 96-Plex Molecular Barcoding for the Illumina Genome Analyzer. Pages 279-298 *in* Y. M. Kwon and S. C. Ricke, editors. High-Throughput Next Generation Sequencing. Humana Press.

Kuroda, M., T. Sekizuka, F. Shinya, F. Takeuchi, T. Kanno, T. Sata, and S. Asano. 2012. Detection of a possible bioterrorism agent, *Francisella* sp., in a clinical specimen by use of next-generation direct DNA sequencing. J Clin Microbiol **50**:1810-1812.

Langmead, B. 2010. Aligning short sequencing reads with Bowtie. Curr Protoc Bioinformatics, Chapter 11, Unit 11 17.

Li, H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics **25**:1754-1760.

Li, H., and R. Durbin. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics **26**:589-595.

Liu, B., and M. Pop. 2009. ARDB--Antibiotic Resistance Genes Database. Nucleic Acids Research **37**:443-447.

Liu, J., H. Wang, H. Yang, Y. Zhang, J. Wang, F. Zhao, and J. Qi. 2013. Composition-based classification of short metagenomic sequences elucidates the landscapes of taxonomic and functional enrichment of microorganisms. Nucleic Acids Research **41**:e3.

Lundin, S., H. Stranneheim, E. Pettersson, D. Klevebring, and J. Lundeberg. 2010. Increased throughput by parallelization of library preparation for massive sequencing. PLoS ONE **5**:e10029.

Lunter, G., and M. Goodson. 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Research **21**:936-939.

Mardis, E. R. 2008. Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet **9**:387-402.

McArthur, A. G., N. Waglechner, F. Nizam, A. Yan, M. A. Azad, A. J. Baylay, K. Bhullar, M. J. Canova, G. De Pascale, L. Ejim, L. Kalan, A. M. King, K. Koteva, M. Morar, M. R. Mulvey, J. S. O'Brien, A. C. Pawlowski, L. J. Piddock, P. Spanogiannopoulos, A. D. Sutherland, I. Tang, P. L. Taylor, M. Thaker, W. Wang, M. Yan, T. Yu, and G. D. Wright. 2013. The comprehensive antibiotic resistance database. Antimicrob Agents Chemother **57**:3348-3357.

Pinard, R., A. de Winter, G. Sarkis, M. Gerstein, K. Tartaro, R. Plant, M. Egholm, J. Rothberg, and J. Leamon. 2006. Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. BMC Genomics **7**:216.

Schmieder, R., and R. Edwards. 2011. Quality control and preprocessing of metagenomic datasets. Bioinformatics **27**:863-864.

Schmitt, M. W., S. R. Kennedy, J. J. Salk, E. J. Fox, J. B. Hiatt, and L. A. Loeb. 2012. Detection of ultra-rare mutations by next-generation sequencing. Proceedings of the National Academy of Sciences USA **109**:14508–14513.

Shendure, J., and H. Ji. 2008. Next-generation DNA sequencing. Nat Biotechnol **26**:1135-1145.

Simon, C., and R. Daniel. 2011. Metagenomic analyses: past and future trends. Appl Environ Microbiol **77**:1153-1161.

Tringe, S. G., and E. M. Rubin. 2005. Metagenomics: DNA sequencing of environmental samples. Nat Rev Genet **6**:805-814.

Tringe, S. G., C. von Mering, A. Kobayashi, A. A. Salamov, K. Chen, H. W. Chang, M. Podar, J. M. Short, E. J. Mathur, J. C. Detter, P. Bork, P. Hugenholtz, and E. M. Rubin. 2005. Comparative metagenomics of microbial communities. Science **308**:554-557.

Appendix 1. Quarterly Report (POP 08 December 2013 - 07 March 2014). Submission of this report as an appendix to the Final Report approved by Ms. Danielle Nolitt.

Summary of work conducted in this quarter

Our research in the last quarter was multifaceted, with work on enriching for bacterial DNA during the extraction process rather than simply trying to do it bioinformatically, reducing the amount of host (human) DNA during the extraction process, and finalizing our bioinformatic work on pathogen detection. Some of these results were also presented in the final report for completeness. We appreciate the extra time afforded to us by the no-cost extension because it allowed us to complete some extremely interesting data sets.

Microbial DNA Enrichment

Metagenomic DNA extracted from (sample origins) of four *Francisella tularensis* clinical cases was quantified using the Nanodrop 1000 spectrophotometer. Based on these quantifications, 1-2 μg of total DNA was enriched for microbial DNA using the NEBNext Microbiome DNA Enrichment Kit (New England Biolabs, Ipswich, Massachusetts, USA) according to the supplied protocol. This protocol captures DNA with CpG methylation, which is prominent in eukaryotic DNA and rare in prokaryotic DNA, and removes it from the metagenomic sample, effectively enriching microbial DNA.

Sequencing Library Preparation

Both native and subtracted DNA samples were prepared for multiplexed Illumina sequencing using the KAPA Illumina series library preparation kit (Kapa Biosystems, Wilmington, Massachusetts, USA) with adapter and index primer oligos as described in Kozarewa and Turner (2011). Libraries were size-selected using the dual-SPRI scheme described by Lundin et al. (2010) to achieve an average insert size of 700 bp. Sequencing was performed on an Illumina HiSeq 2000.

Human Subtraction Protocol Analysis

For the four *F. tularensis* samples, human reads were subtracted using methylation protocol. We compared these four subtracted samples with the normal samples (Figure 1). We found that the subtracted human sequencing reduces the total percentage of human reads as against the normal sequencing from ~1% to ~6%.

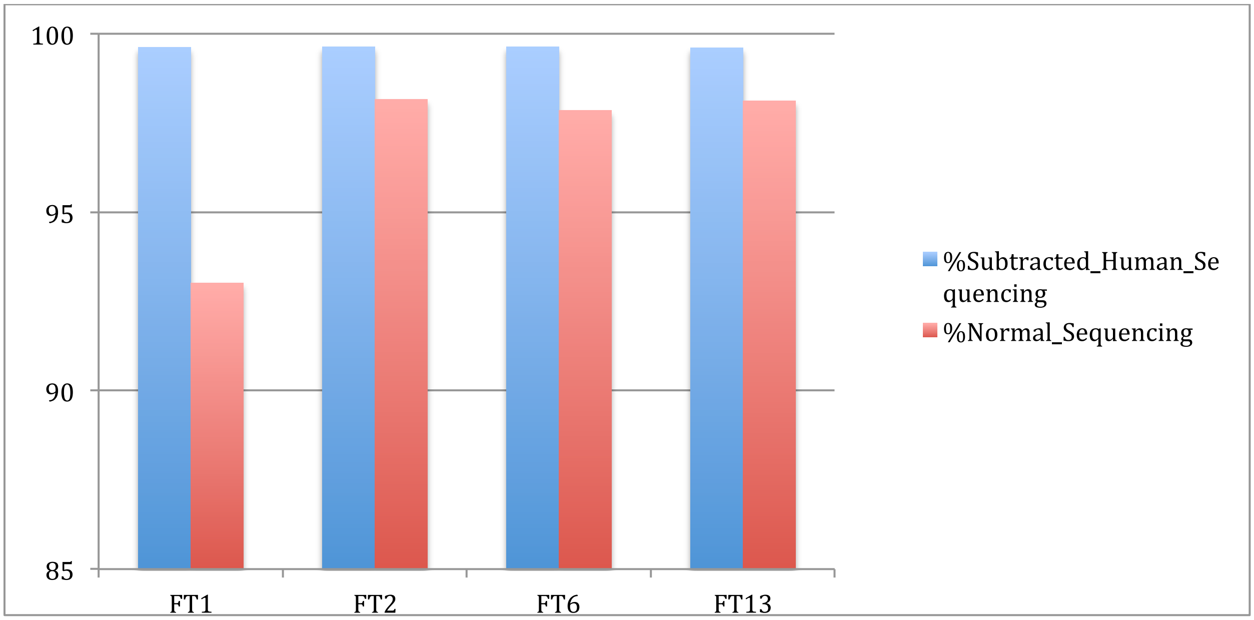


Figure 1. Comparison of subtracted human protocol sequencing versus normal sequencing

We also found that the total number of reads aligning to bacterial reads significantly increased by factor of ~5 to ~90 (Figure 2).

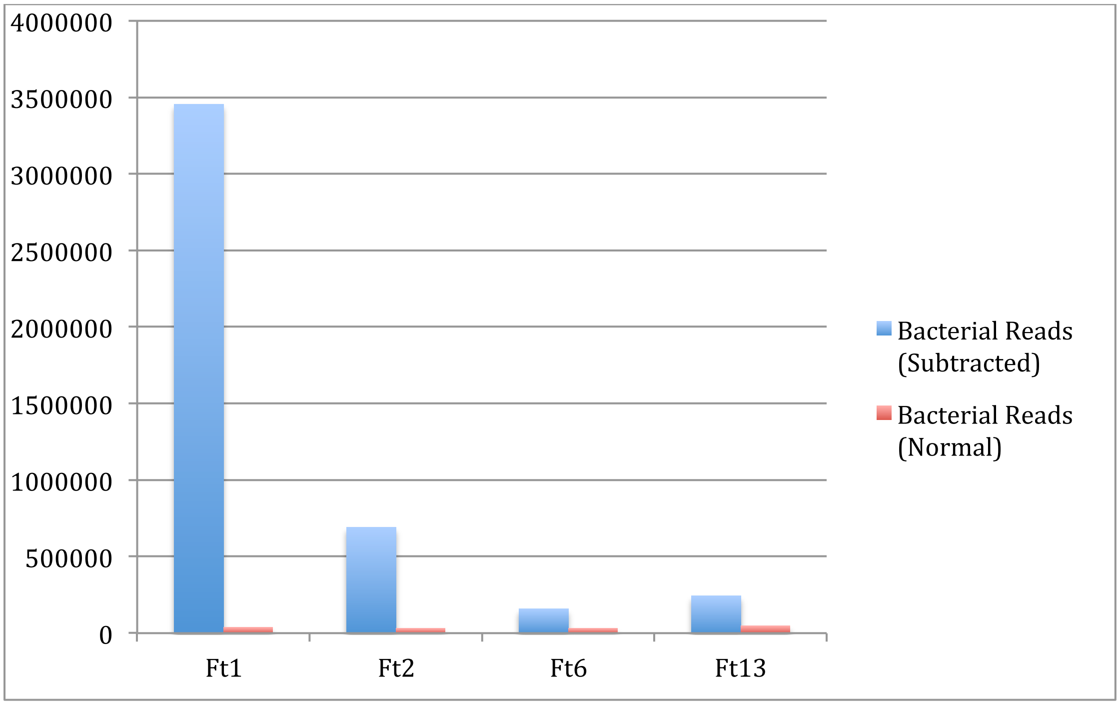


Figure 2. Comparison of bacterial read alignment of subtracted human protocol sequencing versus normal sequencing

Comparing the community between the subtracted human (host) sequencing and normal sequencing, microbial detection was quite different. For normal sequencing, we were able to detect *Francisella tularensis* as major component of infection, for these patients suffering from tularensis besides other organisms like *Ralstonia picketti*. However comparing the distribution of population with the human subtracted sequencing, top hits were *Herbaspirillum seropedicae* SmR1 for Sample 1 and 2 and *Cupriavidus metallidurans* CH34 for Sample FT6 and FT13 (See). *Francisella tularensis* is also detected but the genome coverage % was higher for two samples and lower for other two samples.

Subtracted Human Sequencing (FT1)

Organism	Genome Coverage	Genome Coverage (%)	Genome Size
Herbaspirillum_#_seropedicae_!_SmR1_{}_|gi|300309346|ref|NC_014323.1|	2870146	52.053	5513887
Collimonas_#_fungivorans_!_Ter331_{}_|gi|340785197|ref|NC_015856.1|	386377	7.449	5186898
Janthinobacterium_#_sp._!_CG3_{_JANGC3DRAFT1.1_}|gi|484156181|ref|NZ_KB467824.1|	267953	4.829	5549265
Delftia_#_acidovorans_!_SPH-1_{}_|gi|160895450|ref|NC_010002.1|	220753	3.262	6767514
Ralstonia_#_eutropha_!_H16_{}_1_|gi|113866031|ref|NC_008313.1|	220577	5.444	4052032
Achromobacter_#_xylosoxidans_!_A8_{}_|gi|311103224|ref|NC_014640.1|	220474	3.144	7013095
Cupriavidus_#_necator_!_N-1_{}_1|gi|339324158|ref|NC_015726.1|	212406	5.484	3872936
Ralstonia_#_solanacearum_!_CMR15_{}_|gi|523408232|ref|NC_017559.1|	206756	5.75	3596030
Delftia_#_sp._!_Cs1-4_{}_|gi|333911667|ref|NC_015563.1|	204803	3.063	6685842
Cupriavidus_#_taiwanensis_!_LMG_{_19424_}1|gi|188590795|ref|NC_010528.1|	201059	5.884	3416911
Francisella_#_tularensis_!_subsp._{_holarctica_}_LVS|gi|89255449|ref|NC_007880.1|	200591	10.58	1895994
Francisella_#_tularensis_!_subsp._{_holarctica_}_FTNF002-00|gi|156501369|ref|NC_009749.1|	200154	10.585	1890909
Francisella_#_tularensis_!_subsp._{_holarctica_}_FSC200|gi|422937995|ref|NC_019551.1|	199713	10.544	1894157
Ralstonia_#_solanacearum_!_Po82_{}_|gi|386331671|ref|NC_017574.1|	197334	5.669	3481091
Francisella_#_tularensis_!_subsp._{_holarctica_}_OSU18|gi|115313981|ref|NC_008369.1|	196393	10.36	1895727
Cupriavidus_#_metallidurans_!_CH34_{}_|gi|94308945|ref|NC_007973.1|	193310	4.921	3928089
Ralstonia_#_solanacearum_!_GMI1000_{}_|gi|17544719|ref|NC_003295.1|	192265	5.173	3716413
Ralstonia_#_solanacearum_!_IPO1609_{_|gi|207741818|ref|NW_002196569.1|	188862	5.599	3372855
Francisella_#_tularensis_!_subsp._{_holarctica_}_F92|gi|423049750|ref|NC_019537.1|	188158	9.972	1886888
Ralstonia_#_solanacearum_!_CFBP2957_{}_|gi|300702374|ref|NC_014307.1|	184200	5.39	3417386

Normal Human Sequencing (FT1)

Organism	Genome Coverage	Genome Coverage (%)	Genome Size
Francisella_#_tularensis_!_subsp._{_holarctica_}_LVS|gi|89255449|ref|NC_007880.1|	42495	2.241	1895994
Francisella_#_tularensis_!_subsp._{_holarctica_}_FTNF002-00|gi|156501369|ref|NC_009749.1|	42254	2.235	1890909
Francisella_#_tularensis_!_subsp._{_holarctica_}_FSC200|gi|422937995|ref|NC_019551.1|	42188	2.227	1894157
Francisella_#_tularensis_!_subsp._{_holarctica_}_OSU18|gi|115313981|ref|NC_008369.1|	42162	2.224	1895727
Francisella_#_tularensis_!_subsp._{_holarctica_}_F92|gi|423049750|ref|NC_019537.1|	40979	2.172	1886888
Francisella_#_tularensis_!_TIGB03_{}_|gi|379716390|ref|NC_016933.1|	38223	1.942	1968651
Francisella_#_tularensis_!_subsp._{_tularensis_}_NE061598|gi|385793751|ref|NC_017453.1|	35758	1.889	1892681
Francisella_#_tularensis_!_subsp._{_tularensis_}_SCHU_S4|gi|255961454|ref|NC_006570.2|	35758	1.889	1892775
Francisella_#_tularensis_!_subsp._{_tularensis_}_FSC198|gi|110669657|ref|NC_008245.1|	35707	1.887	1892616
Francisella_#_tularensis_!_subsp._{_tularensis_}_TI0902|gi|379725073|ref|NC_016937.1|	35707	1.887	1892744
Francisella_#_tularensis_!_subsp._{_tularensis_}_WY96-3418|gi|134301169|ref|NC_009257.1|	35445	1.867	1898476
Francisella_#_tularensis_!_subsp._{_mediasiatica_}_FSC147|gi|187930913|ref|NC_010677.1|	29870	1.577	1893886
Francisella_#_novicida_!_U112_{}_|gi|118496615|ref|NC_008601.1|	16928	0.886	1910031
Francisella_#_cf._!_novicida_{_Fx1_}|gi|385791932|ref|NC_017450.1|	15216	0.795	1913619
Francisella_#_tularensis_!_subsp._{_holarctica_}_2571.19|gi|254367826|ref|NZ_DS229056.1|	12257	2.205	555807
Streptomyces_#_griseoflavus_!_Tu4000_{_genomic_}1.1|gi|224581108|ref|NZ_GG657758.1|	8778	0.109	8047042
Francisella_#_tularensis_!_subsp._{_tularensis_}_FSC0331.4|gi|254370255|ref|NZ_DS264119.1|	8131	1.032	788151
Francisella_#_tularensis_!_subsp._{_novicida_}_FTE_FTE1|gi|224580220|ref|NZ_DS989818.1|	7110	0.837	849109
Francisella_#_tularensis_!_subsp._{_holarctica_}_FSC0221.8|gi|254369615|ref|NZ_DS264140.1|	7070	2.139	330502
Francisella_#_novicida_!_FTG_{_FTG1_}|gi|224580234|ref|NZ_DS995363.1|	6286	0.893	703773

Subtracted Human Sequencing (FT2)

Organism	Genome Coverage	Genome Coverage (%)	Genome Size
Herbaspirillum_#_seropedicae_!_SmR1_{}_|gi|300309346|ref|NC_014323.1|	3039047	55.116	5513887
Collimonas_#_fungivorans_!_Ter331_{}_|gi|340785197|ref|NC_015856.1|	455636	8.784	5186898
Janthinobacterium_#_sp._!_CG3_{_JANGC3DRAFT1.1_}|gi|484156181|ref|NZ_KB467824.1|	316372	5.701	5549265
Ralstonia_#_eutropha_!_H16_{}_1_|gi|113866031|ref|NC_008313.1|	261011	6.441	4052032
Achromobacter_#_xylosoxidans_!_A8_{}_|gi|311103224|ref|NC_014640.1|	257003	3.665	7013095
Delftia_#_acidovorans_!_SPH-1_{}_|gi|160895450|ref|NC_010002.1|	252615	3.733	6767514
Cupriavidus_#_necator_!_N-1_{}_1|gi|339324158|ref|NC_015726.1|	247458	6.389	3872936
Francisella_#_tularensis_!_subsp._{_holarctica_}_LVS|gi|89255449|ref|NC_007880.1|	245618	12.955	1895994
Francisella_#_tularensis_!_subsp._{_holarctica_}_FSC200|gi|422937995|ref|NC_019551.1|	245565	12.964	1894157
Ralstonia_#_solanacearum_!_CMR15_{}_|gi|523408232|ref|NC_017559.1|	243872	6.782	3596030
Francisella_#_tularensis_!_subsp._{_holarctica_}_FTNF002-00|gi|156501369|ref|NC_009749.1|	243843	12.896	1890909
Francisella_#_tularensis_!_subsp._{_holarctica_}_OSU18|gi|115313981|ref|NC_008369.1|	242455	12.79	1895727
Cupriavidus_#_taiwanensis_!_LMG_{_19424_}1|gi|188590795|ref|NC_010528.1|	240282	7.032	3416911
Ralstonia_#_solanacearum_!_Po82_{}_|gi|386331671|ref|NC_017574.1|	236551	6.795	3481091
Francisella_#_tularensis_!_subsp._{_holarctica_}_F92|gi|423049750|ref|NC_019537.1|	233405	12.37	1886888
Delftia_#_sp._!_Cs1-4_{}_|gi|333911667|ref|NC_015563.1|	232974	3.485	6685842
Ralstonia_#_solanacearum_!_GMI1000_{}_|gi|17544719|ref|NC_003295.1|	228347	6.144	3716413
Francisella_#_tularensis_!_TIGB03_{}_|gi|379716390|ref|NC_016933.1|	224368	11.397	1968651
Ralstonia_#_solanacearum_!_IPO1609_{_|gi|207741818|ref|NW_002196569.1|	222563	6.599	3372855
Ralstonia_#_solanacearum_!_CFBP2957_{}_|gi|300702374|ref|NC_014307.1|	215864	6.317	3417386

Normal Human Sequencing (FT2)

Organism	Genome Coverage	Genome Coverage (%)	Genome Size
Francisella_#_tularensis_!_subsp._{_holarctica_}_FSC200|gi|422937995|ref|NC_019551.1|	72600	3.833	1894157
Francisella_#_tularensis_!_subsp._{_holarctica_}_FTNF002-00|gi|156501369|ref|NC_009749.1|	72361	3.827	1890909
Francisella_#_tularensis_!_subsp._{_holarctica_}_LVS|gi|89255449|ref|NC_007880.1|	71914	3.793	1895994
Francisella_#_tularensis_!_subsp._{_holarctica_}_OSU18|gi|115313981|ref|NC_008369.1|	68360	3.606	1895727
Francisella_#_tularensis_!_subsp._{_holarctica_}_F92|gi|423049750|ref|NC_019537.1|	65721	3.483	1886888
Francisella_#_tularensis_!_TIGB03_{}_|gi|379716390|ref|NC_016933.1|	64022	3.252	1968651
Francisella_#_tularensis_!_subsp._{_tularensis_}_WY96-3418|gi|134301169|ref|NC_009257.1|	61612	3.245	1898476
Francisella_#_tularensis_!_subsp._{_tularensis_}_NE061598|gi|385793751|ref|NC_017453.1|	61426	3.245	1892681
Francisella_#_tularensis_!_subsp._{_tularensis_}_FSC198|gi|110669657|ref|NC_008245.1|	61174	3.232	1892616
Francisella_#_tularensis_!_subsp._{_tularensis_}_SCHU_S4|gi|255961454|ref|NC_006570.2|	61117	3.229	1892775
Francisella_#_tularensis_!_subsp._{_tularensis_}_TI0902|gi|379725073|ref|NC_016937.1|	60940	3.22	1892744
Francisella_#_tularensis_!_subsp._{_mediasiatica_}_FSC147|gi|187930913|ref|NC_010677.1|	51421	2.715	1893886
Francisella_#_novicida_!_U112_{}_|gi|118496615|ref|NC_008601.1|	29126	1.525	1910031
Francisella_#_cf._!_novicida_{_Fx1_}|gi|385791932|ref|NC_017450.1|	26557	1.388	1913619
Francisella_#_tularensis_!_subsp._{_holarctica_}_2571.19|gi|254367826|ref|NZ_DS229056.1|	22104	3.977	555807
Francisella_#_tularensis_!_subsp._{_tularensis_}_FSC0331.4|gi|254370255|ref|NZ_DS264119.1|	14356	1.821	788151
Francisella_#_tularensis_!_subsp._{_holarctica_}_FSC0221.5|gi|254368696|ref|NZ_DS264137.1|	13465	2.759	488102
Francisella_#_tularensis_!_subsp._{_novicida_}_FTE_FTE1|gi|224580220|ref|NZ_DS989818.1|	12673	1.493	849109
Francisella_#_novicida_!_FTG_{_FTG1_}|gi|224580234|ref|NZ_DS995363.1|	11472	1.63	703773
Francisella_#_tularensis_!_subsp._{_holarctica_}_FSC0221.6|gi|254369075|ref|NZ_DS264138.1|	10313	2.029	508365

Subtracted Human Sequencing (FT6)

<table>
<tr><th>Organism</th><th>Genome Coverage</th><th>Genome Coverage (%)</th><th>Genome Size</th></tr>
<tr><td>Cupriavidus_#_metallidurans_!_CH34_{}_|gi|94308945|ref|NC_007973.1|</td><td>2631037</td><td>66.98</td><td>3928089</td></tr>
<tr><td>Cupriavidus_#_metallidurans_!_CH34_{_megaplasmid_}|gi|291481467|ref|NC_007974.2|</td><td>1746897</td><td>67.707</td><td>2580084</td></tr>
<tr><td>Ralstonia_#_eutropha_!_H16_{}_1_|gi|113866031|ref|NC_008313.1|</td><td>194259</td><td>4.794</td><td>4052032</td></tr>
<tr><td>Cupriavidus_#_necator_!_N-1_{}_1|gi|339324158|ref|NC_015726.1|</td><td>192410</td><td>4.968</td><td>3872936</td></tr>
<tr><td>Cupriavidus_#_taiwanensis_!_LMG_{_19424_}1|gi|188590795|ref|NC_010528.1|</td><td>190485</td><td>5.575</td><td>3416911</td></tr>
<tr><td>Cupriavidus_#_metallidurans_!_CH34_{_plasmid_}_pMOL30|gi|56130627|ref|NC_006466.1|</td><td>184536</td><td>78.944</td><td>233755</td></tr>
<tr><td>Cupriavidus_#_metallidurans_!_CH34_{_plasmid_}_pMOL30|gi|291464753|ref|NC_007971.2|</td><td>184489</td><td>78.936</td><td>233720</td></tr>
<tr><td>Ralstonia_#_eutropha_!_JMP134_{}_1|gi|73539706|ref|NC_007347.1|</td><td>174988</td><td>4.597</td><td>3806533</td></tr>
<tr><td>Francisella_#_tularensis_!_subsp._{_holarctica_}_LVS|gi|89255449|ref|NC_007880.1|</td><td>166871</td><td>8.801</td><td>1895994</td></tr>
<tr><td>Francisella_#_tularensis_!_subsp._{_holarctica_}_FSC200|gi|422937995|ref|NC_019551.1|</td><td>166321</td><td>8.781</td><td>1894157</td></tr>
<tr><td>Francisella_#_tularensis_!_subsp._{_holarctica_}_FTNF002-00|gi|156501369|ref|NC_009749.1|</td><td>165408</td><td>8.748</td><td>1890909</td></tr>
<tr><td>Francisella_#_tularensis_!_subsp._{_holarctica_}_OSU18|gi|115313981|ref|NC_008369.1|</td><td>162954</td><td>8.596</td><td>1895727</td></tr>
<tr><td>Francisella_#_tularensis_!_subsp._{_holarctica_}_F92|gi|423049750|ref|NC_019537.1|</td><td>151481</td><td>8.028</td><td>1886888</td></tr>
<tr><td>Francisella_#_tularensis_!_TIGB03_{}_|gi|379716390|ref|NC_016933.1|</td><td>145410</td><td>7.386</td><td>1968651</td></tr>
<tr><td>Francisella_#_tularensis_!_subsp._{_tularensis_}_WY96-3418|gi|134301169|ref|NC_009257.1|</td><td>138407</td><td>7.29</td><td>1898476</td></tr>
<tr><td>Francisella_#_tularensis_!_subsp._{_tularensis_}_TI0902|gi|379725073|ref|NC_016937.1|</td><td>138093</td><td>7.296</td><td>1892744</td></tr>
<tr><td>Francisella_#_tularensis_!_subsp._{_tularensis_}_NE061598|gi|385793751|ref|NC_017453.1|</td><td>136778</td><td>7.227</td><td>1892681</td></tr>
<tr><td>Francisella_#_tularensis_!_subsp._{_tularensis_}_SCHU_S4|gi|255961454|ref|NC_006570.2|</td><td>136734</td><td>7.224</td><td>1892775</td></tr>
<tr><td>Francisella_#_tularensis_!_subsp._{_mediasiatica_}_FSC147|gi|187930913|ref|NC_010677.1|</td><td>134359</td><td>7.094</td><td>1893886</td></tr>
<tr><td>Francisella_#_tularensis_!_subsp._{_tularensis_}_FSC198|gi|110669657|ref|NC_008245.1|</td><td>96340</td><td>5.09</td><td>1892616</td></tr>
</table>

Normal Human Sequencing (FT6)

<table>
<tr><th>Organism</th><th>Genome Coverage</th><th>Genome Coverage (%)</th><th>Genome Size</th></tr>
<tr><td>Francisella_#_tularensis_!_subsp._{_holarctica_}_LVS|gi|89255449|ref|NC_007880.1|</td><td>296764</td><td>15.652</td><td>1895994</td></tr>
<tr><td>Francisella_#_tularensis_!_subsp._{_holarctica_}_FSC200|gi|422937995|ref|NC_019551.1|</td><td>295826</td><td>15.618</td><td>1894157</td></tr>
<tr><td>Francisella_#_tularensis_!_subsp._{_holarctica_}_FTNF002-00|gi|156501369|ref|NC_009749.1|</td><td>295286</td><td>15.616</td><td>1890909</td></tr>
<tr><td>Francisella_#_tularensis_!_subsp._{_holarctica_}_OSU18|gi|115313981|ref|NC_008369.1|</td><td>293857</td><td>15.501</td><td>1895727</td></tr>
<tr><td>Francisella_#_tularensis_!_subsp._{_holarctica_}_F92|gi|423049750|ref|NC_019537.1|</td><td>285104</td><td>15.11</td><td>1886888</td></tr>
<tr><td>Francisella_#_tularensis_!_TIGB03_{}_|gi|379716390|ref|NC_016933.1|</td><td>276026</td><td>14.021</td><td>1968651</td></tr>
<tr><td>Francisella_#_tularensis_!_subsp._{_tularensis_}_WY96-3418|gi|134301169|ref|NC_009257.1|</td><td>264917</td><td>13.954</td><td>1898476</td></tr>
<tr><td>Francisella_#_tularensis_!_subsp._{_mediasiatica_}_FSC147|gi|187930913|ref|NC_010677.1|</td><td>263190</td><td>13.897</td><td>1893886</td></tr>
<tr><td>Francisella_#_tularensis_!_subsp._{_tularensis_}_TI0902|gi|379725073|ref|NC_016937.1|</td><td>262822</td><td>13.886</td><td>1892744</td></tr>
<tr><td>Francisella_#_tularensis_!_subsp._{_tularensis_}_NE061598|gi|385793751|ref|NC_017453.1|</td><td>262607</td><td>13.875</td><td>1892681</td></tr>
<tr><td>Francisella_#_tularensis_!_subsp._{_tularensis_}_FSC198|gi|110669657|ref|NC_008245.1|</td><td>262370</td><td>13.863</td><td>1892616</td></tr>
<tr><td>Francisella_#_tularensis_!_subsp._{_tularensis_}_SCHU_S4|gi|255961454|ref|NC_006570.2|</td><td>262370</td><td>13.862</td><td>1892775</td></tr>
<tr><td>Francisella_#_novicida_!_U112_{}_|gi|118496615|ref|NC_008601.1|</td><td>195963</td><td>10.26</td><td>1910031</td></tr>
<tr><td>Francisella_#_cf._!_novicida_{_Fx1_}|gi|385791932|ref|NC_017450.1|</td><td>183855</td><td>9.608</td><td>1913619</td></tr>
<tr><td>Francisella_#_tularensis_!_subsp._{_holarctica_}_2571.19|gi|254367826|ref|NZ_DS229056.1|</td><td>86791</td><td>15.615</td><td>555807</td></tr>
<tr><td>Francisella_#_tularensis_!_subsp._{_tularensis_}_FSC0331.4|gi|254370255|ref|NZ_DS264119.1|</td><td>82045</td><td>10.41</td><td>788151</td></tr>
<tr><td>Francisella_#_tularensis_!_subsp._{_novicida_}_FTE_FTE1|gi|224580220|ref|NZ_DS989818.1|</td><td>80740</td><td>9.509</td><td>849109</td></tr>
<tr><td>Francisella_#_novicida_!_FTG_{_FTG1_}|gi|224580234|ref|NZ_DS995363.1|</td><td>67747</td><td>9.626</td><td>703773</td></tr>
<tr><td>Francisella_#_tularensis_!_subsp._{_holarctica_}_FSC0221.5|gi|254368696|ref|NZ_DS264137.1|</td><td>65767</td><td>13.474</td><td>488102</td></tr>
<tr><td>Francisella_#_cf._!_novicida_{_3523_}|gi|387823583|ref|NC_017449.1|</td><td>64634</td><td>3.323</td><td>1945310</td></tr>
</table>

Subtracted Human Sequencing (FT13)

Organism	Genome Coverage	Genome Coverage (%)	Genome Size
Cupriavidus_#_metallidurans_!_CH34_{}_|gi|94308945|ref|NC_007973.1|	3049075	77.622	3928089
Cupriavidus_#_metallidurans_!_CH34_{_megaplasmid_}|gi|291481467|ref|NC_007974.2|	2009785	77.896	2580084
Ralstonia_#_eutropha_!_H16_{}_1_|gi|113866031|ref|NC_008313.1|	271535	6.701	4052032
Cupriavidus_#_necator_!_N-1_{}_1|gi|339324158|ref|NC_015726.1|	260822	6.734	3872936
Cupriavidus_#_taiwanensis_!_LMG_{_19424_}1|gi|188590795|ref|NC_010528.1|	254368	7.444	3416911
Ralstonia_#_eutropha_!_JMP134_{}_1|gi|73539706|ref|NC_007347.1|	242231	6.364	3806533
Cupriavidus_#_metallidurans_!_CH34_{_plasmid_}_pMOL30|gi|56130627|ref|NC_006466.1|	197441	84.465	233755
Cupriavidus_#_metallidurans_!_CH34_{_plasmid_}_pMOL30|gi|291464753|ref|NC_007971.2|	197177	84.365	233720
Francisella_#_tularensis_!_subsp._{_holarctica_}_LVS|gi|89255449|ref|NC_007880.1|	132199	6.973	1895994
Francisella_#_tularensis_!_subsp._{_holarctica_}_FTNF002-00|gi|156501369|ref|NC_009749.1|	131543	6.957	1890909
Francisella_#_tularensis_!_subsp._{_holarctica_}_FSC200|gi|422937995|ref|NC_019551.1|	131258	6.93	1894157
Francisella_#_tularensis_!_subsp._{_holarctica_}_OSU18|gi|115313981|ref|NC_008369.1|	128893	6.799	1895727
Francisella_#_tularensis_!_subsp._{_holarctica_}_F92|gi|423049750|ref|NC_019537.1|	122915	6.514	1886888
Ralstonia_#_solanacearum_!_CMR15_{}_|gi|523408232|ref|NC_017559.1|	121624	3.382	3596030
Ralstonia_#_solanacearum_!_Po82_{}_|gi|386331671|ref|NC_017574.1|	121071	3.478	3481091
Francisella_#_tularensis_!_TIGB03_{}_|gi|379716390|ref|NC_016933.1|	113359	5.758	1968651
Ralstonia_#_solanacearum_!_PSI07_{}_|gi|300689714|ref|NC_014311.1|	113115	3.213	3520618
Ralstonia_#_solanacearum_!_CFBP2957_{}_|gi|300702374|ref|NC_014307.1|	113075	3.309	3417386
Ralstonia_#_solanacearum_!_GMI1000_{}_|gi|17544719|ref|NC_003295.1|	110070	2.962	3716413
Francisella_#_tularensis_!_subsp._{_tularensis_}_WY96-3418|gi|134301169|ref|NC_009257.1|	109645	5.775	1898476

Normal Human Sequencing (FT13)

Organism	Genome Coverage	Genome Coverage (%)	Genome Size
Francisella_#_tularensis_!_subsp._{_holarctica_}_FTNF002-00|gi|156501369|ref|NC_009749.1|	140625	7.437	1890909
Francisella_#_tularensis_!_subsp._{_holarctica_}_LVS|gi|89255449|ref|NC_007880.1|	140169	7.393	1895994
Francisella_#_tularensis_!_subsp._{_holarctica_}_FSC200|gi|422937995|ref|NC_019551.1|	139139	7.346	1894157
Francisella_#_tularensis_!_subsp._{_holarctica_}_OSU18|gi|115313981|ref|NC_008369.1|	138832	7.323	1895727
Francisella_#_tularensis_!_subsp._{_holarctica_}_F92|gi|423049750|ref|NC_019537.1|	132246	7.009	1886888
Francisella_#_tularensis_!_TIGB03_{}_|gi|379716390|ref|NC_016933.1|	122280	6.211	1968651
Francisella_#_tularensis_!_subsp._{_tularensis_}_WY96-3418|gi|134301169|ref|NC_009257.1|	117989	6.215	1898476
Francisella_#_tularensis_!_subsp._{_tularensis_}_TI0902|gi|379725073|ref|NC_016937.1|	115572	6.106	1892744
Francisella_#_tularensis_!_subsp._{_tularensis_}_NE061598|gi|385793751|ref|NC_017453.1|	115405	6.097	1892681
Francisella_#_tularensis_!_subsp._{_tularensis_}_FSC198|gi|110669657|ref|NC_008245.1|	115311	6.093	1892616
Francisella_#_tularensis_!_subsp._{_tularensis_}_SCHU_S4|gi|255961454|ref|NC_006570.2|	115311	6.092	1892775
Francisella_#_tularensis_!_subsp._{_mediasiatica_}_FSC147|gi|187930913|ref|NC_010677.1|	115096	6.077	1893886
Francisella_#_novicida_!_U112_{}_|gi|118496615|ref|NC_008601.1|	63192	3.308	1910031
Francisella_#_cf._!_novicida_{_Fx1_}|gi|385791932|ref|NC_017450.1|	55336	2.892	1913619
Francisella_#_tularensis_!_subsp._{_holarctica_}_2571.19|gi|254367826|ref|NZ_DS229056.1|	41831	7.526	555807
Francisella_#_tularensis_!_subsp._{_tularensis_}_FSC0331.4|gi|254370255|ref|NZ_DS264119.1|	29337	3.722	788151
Francisella_#_tularensis_!_subsp._{_novicida_}_FTE_FTE1|gi|224580220|ref|NZ_DS989818.1|	27185	3.202	849109
Francisella_#_tularensis_!_subsp._{_holarctica_}_FSC0221.5|gi|254368696|ref|NZ_DS264137.1|	23490	4.813	488102
Francisella_#_novicida_!_FTG_{_FTG1_}|gi|224580234|ref|NZ_DS995363.1|	23381	3.322	703773
Francisella_#_tularensis_!_subsp._{_holarctica_}_2571.13|gi|254367300|ref|NZ_DS229050.1|	21554	6.104	353090

Design of Pathogen Detection Module of MetaGeniE

The second module of the pipeline, Patho-Detect, aligns the remaining reads after human filtration against known bacterial, fungal and viral sequences with BWT alignment followed with local aligner BLAT (Figure 1). MetaGeniE aggregates global and local (BLAT) alignments and generates a cumulative report from both alignments for sensitive detection. Detection by read numbers is often not an accurate predictor due to repeat elements, close relatives in the metagenome and PCR amplification biases. Instead, MetaGeniE detects microbial presence by genomic reconstruction using percentage of the genome mapped and therefore largely overcomes these biases. Incorporating a larger NCBI RefSeq DB, which includes sequences from multiple strains of important pathogenic species, rather than using just a few selected complete genomes allows for finer scale resolution of clinical sample community memebers to the subspecies/strain level contained within the RefSeq DB. RefSeq DB has doubled from 8.7 GB in Release 54 to 19 GB in Release 60 for bacteria. MetaGeniE can now handle multiple partitions of the reference database and so is scalable to large databases for better memory management. NCBI indexes and hashes allow faster extraction of millions of reads. Custom parsers allocate all reads to all mapped genomes instead of SAMTools that map main hits only and therefore MetaGeniE provides more sensitive pathogen detection in the metagenome.

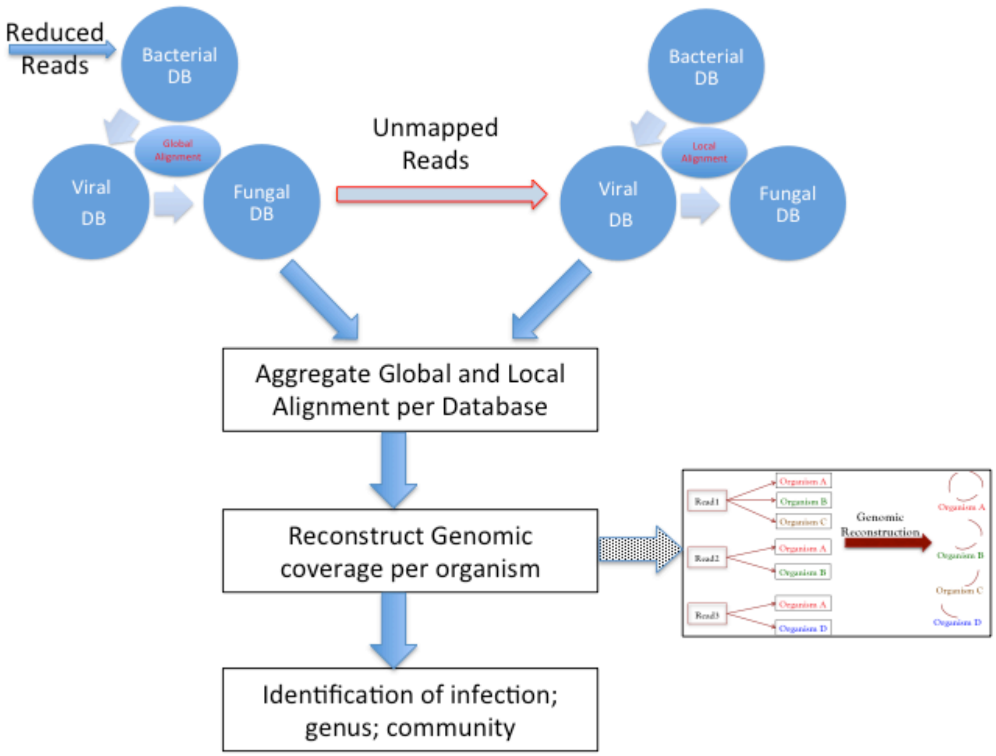


Figure 1. The Pathogen detection module of the pipeline.

Features

Scalability

Incorporating a larger NCBI RefSeq database rather than using just a few selected complete genomes helps in closest resolution of identification to subspecies/strain level. However, this results in increasing demand for computational memory to scale for growing reference databases.

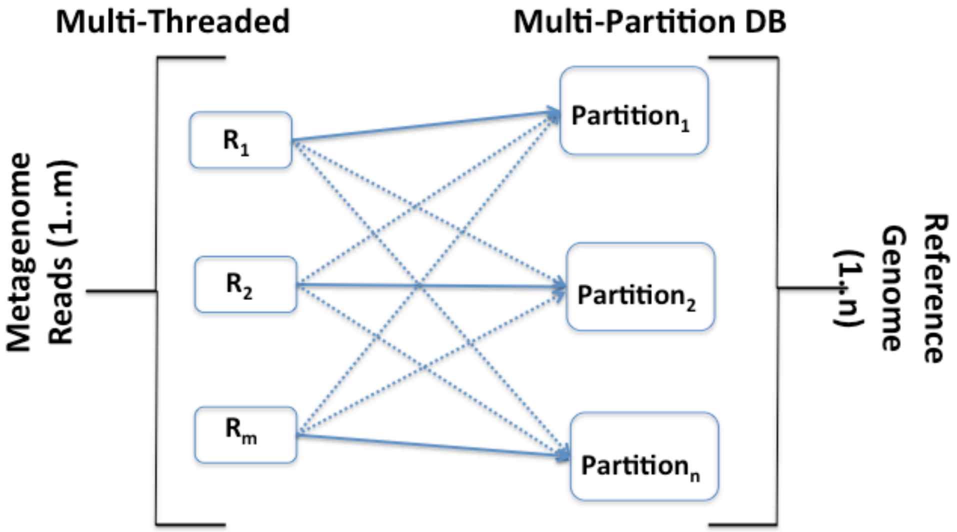


Figure 2. Multithreaded input sequence file query the multiple partition reference database to address the scalability.

To address the issue of scalability with large reference databases, we designed the pipeline to handle multiple partitions of reference database for better memory management (Figure 2). Each smaller database partitions (~1 GB) can be queried by multithreaded input files iteratively therefore reducing overall memory footprint. This querying of each input file fragment generates higher number of mapped-unmapped relationships against the partitioned database results per iteration that increases the computational time. To address this issue, the pipeline utilizes custom hash functions and indexing tool formatdb and fastacmd (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/) to allow faster extraction of millions of reads as an input to next reference database search.

Normalized (%) Genome Coverage

Pathogen detection by the total number of reads that hit/align to respective genome is not always an accurate predictor of presence of an organism due to repeat elements, close relatives in metagenome and PCR amplification. To overcome these biases, the pipeline detects microbial presence by genomic reconstruction i.e. %genome mapped for each organism from initial mixture of non-host DNA of metagenome. The pipeline first converts the local and global alignment output to common BED format. Genome coverage of each mapped organism is then calculated from the global and local alignments with BEDTOOLS (Quinlan and Hall, 2010). The total genome reconstructed for each mapped organism is the sum of genome coverage from global and local alignments for any metagenome. The normalized genome coverage (%genome coverage) is calculated as follows:

$$\text{Normalized Genome Cov.} = \left(\frac{\text{Genome Cov. by local alignment} + \text{Genome Cov. by global alignment}}{\textit{Total Genome Size}}\right) * 100$$

The normalized genome coverage allows comparison of different organism with different genome size. It is helpful in representing the abundance of various organisms in metagenome for community analysis.

Benchmark

Bacterial Datasets: Simulated reads of length 100 bases were generated from respective reference genomes as discussed in each section. Average Illumina sequencing error of 0.4% is incorporated in all the simulated bacterial reads. To study divergence and its effect on detection, additional variability of 0.1%, 0.2%, 0.5% and 1% is incorporated in each simulated bacterial library. All these steps utilized the same parameters as follows, BWA (default) and BLAT (80% identity).

Pathogen Detection

Metagenome datasets derived from clinical samples present numerous challenges, such as 1) low abundance of pathogens; 2) the complexity of community might vary from a single infecting agent (with only one dominant infection) to multiple infectious agents, some being closely related, within a complex community (example cystic fibrosis patients); 3) available reference genomes may be quite distinct from novel pathogens in the clinical community, minimizing the number of reads that will map. We created simulated libraries based on these complexities of community to estimate the efficiency of pathogen detection.

Simple Community

Metagenome sequences are often processed as a single genome alignment to a reference genome(s). In a single genome alignment, reads aligning to multiple loci to a reference genome are randomly assigned to a locus and SAMTools only parses these “main” hits. To compare this against our strategy to discern an organism among its close relatives with the all-against-all strategy between reads and reference genome, we utilized *S. aureus* TCH1516 to understand single infection, as Staphylococcus is well-characterized genus with high number of strains. This will allow us to test the specificity to detect correct organism not only among members of its own clade ST8-MRSA-IVa/USA300 but large number of well-characterized genome of *Staphylococcus* in reference genome. Typically the genus specific regions of *Staphylococcus* will be assigned to all the members of this taxonomic rank. The reads that will further contribute to unique regions, which belong to its species (aureus), and strain specific genes (clonal complex 5), will result in highest % genome coverage of correct organism. The single alignment coverage is performed with BWA – SAMTools – BEDTools against multiple genome RefSeq dataset, Metagenome alignment is performed with human read reduction and pathogen detection of pipeline against RefSeq dataset. The results are compared with actual coverage detected with BWA – SAMTools – BEDTools against single genome of *S. aureus* TCH1516 genome (Table 1).

Table 1. The comparison of percent genome coverage with increasing read number for the single genome alignment with metagenome alignment. * Genome coverage of top hit reported even if the organism detected is incorrect

	100	**1K**	**10K**	**100K**	**250K**
Single Alignment Coverage	0.014*	0.059	0.62*	5.116	11.307
MetaGenome Coverage	0.352	3.557	30.126	97.004	99.983
Actual Coverage	0.348	3.427	29.494	96.955	99.982

We were also able to detect *S. aureus* TCH1516 in all the test sets as top hit even with very low number of reads while single genome alignment is not able to report correct detection. We found that the single alignment underestimate the genome coverage compared with the results from our pipeline (Table 1). The coverage detected by our approach was nearly equal to the actual coverage detected (actual coverage is slightly lower than the coverage detected by our pipeline as actual coverage is calculated from only BWA alignment while the pipeline calculates coverage from both BWA and BLAT alignment).

We also compared the effect on pathogen detection on number of reads, % genome coverage and read recall % with different features available for human filtration of the pipeline (Figure 3). Read recall % are the simulated reads that align back correctly to its reference genome after human filtration. As the read number increases, as expected genome coverage% also increases; approaching 99.9% at nearly 250K reads thus having coverage across the entire genome. 250K reads can be defined as approximate reads to reconstruct the entire genome of *S. aureus* TCH1516 from the metagenome.

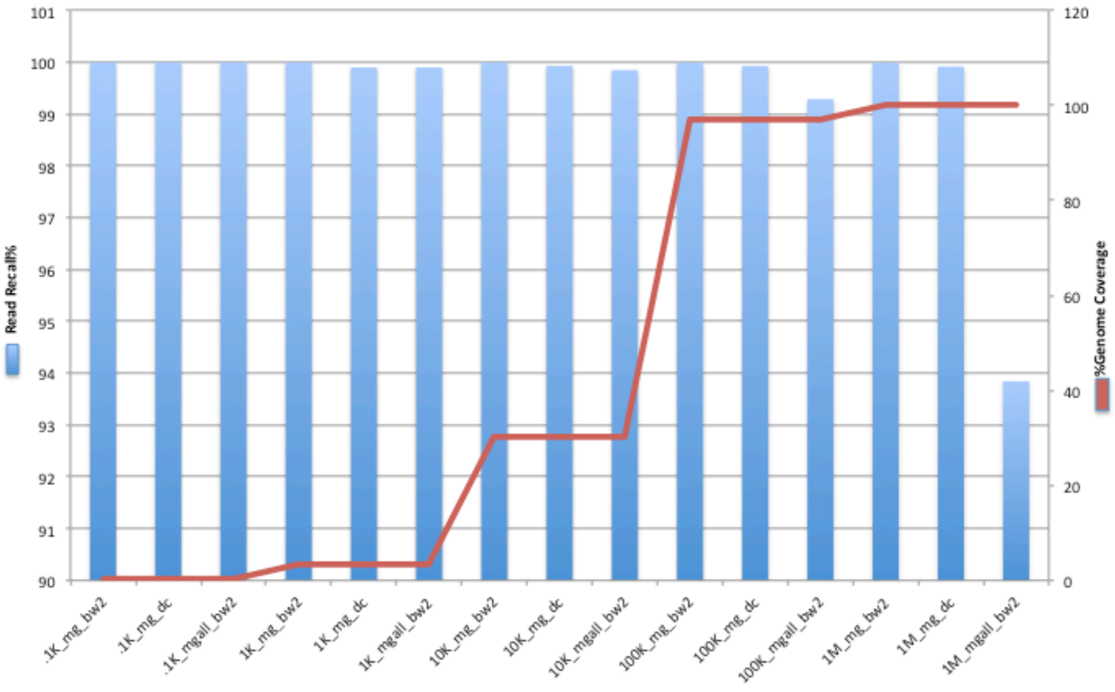


Figure 6. Effect of human filtration on percent genome coverage and read recall percentage of pathogen detection. The legends of the figure are prefixed with the

number of reads (.1K=100; 1K=1000; 10K=10000; 100K=100000; 1M=1000000) followed by mg_bw2 for only fast alignment feature of human read reduction; mg_dc for all features of human read reduction except data compression; mgall_bw2 for all features of human read reduction module)

As more number of reads is sequenced (simulated), higher number of duplicate reads is also expected. Turning on data compression feature of human filtration of pipeline (*_mgall_bw2) to remove duplicates reduces the read recall% but has no effect on genome coverage% and detection of correct organism. The duplicate reads therefore do not add additional information and to manage computation scalability, removal of these duplicates is helpful. We also see that using all the human filtration steps of the pipeline (*_mgall-bw2) as compared to using just fast alignment (*_mg_bw2) or not utilizing data compression (*_mg_dc) does not lead to underestimation of %genome coverage for correct pathogen detection.

Complex Community

The ability to detect community in complex clinical samples such as cystic fibrosis might be helpful in generating insight for proper clinical recourse. The shared regions are expected in even the simplest microbial communities. The simulated library will allow us to evaluate the impact on detection due to presence of multiple organisms in community that have different genome size. We designed a complex community of five bacteria that is detected in one of the cystic fibrosis clinical sample. Simulated reads is generated from reference genome of each of the five organisms and four libraries with different read number (100, 1000, 10K, 100K per organism) is created. In metagenomes, many organisms may not be represented in the reference genome database. To test the specificity of detection with unknown organism, *V. dispar* ATCC 17748 that is not present in bacterial reference genome (RefSeq Build 60) is added in this complex community. Querying a large reference database usually results in detection of multiple organisms within same genus due to overlapping homology. Therefore for detection of organism, we selected the highest mapped genome (top-hit) within the same genus. The correct detection is confirmed for all the organisms except for *V. dispar* ATCC 17748. The pipeline allows detection of correct organisms even in complex community.

Different genera in complex community might share genomic regions. The robustness of detection can be measured by loss in sensitivity (genome coverage) of any organism in complex community versus single infection. We compared the %genome coverage of *E. coli* APEC O1 as single infection and in complex community. We found no difference and hence no loss in % genome coverage is reported and the *E. coli* APEC O1 in simple and complex community overlap completely (Figure 4). All against all relationship between the reads and reference database therefore allows us to detect any organism without any loss in sensitivity that might have shared genomic regions.

E. coli APEC O1 has highest genome hit among this community, the increase in read number have higher impact on the increase in genome coverage (Figure 4). For simulated reads of *V. dispar* ATCC 17748 that is true negative, *V. parvula* DSM 2008 chromosome is detected as top hit with lower %genome coverage compared to other hits. We can infer that true calls may not always be possible, given the limited (albeit growing) nature of genomic databases and the resolution of detection might decrease.

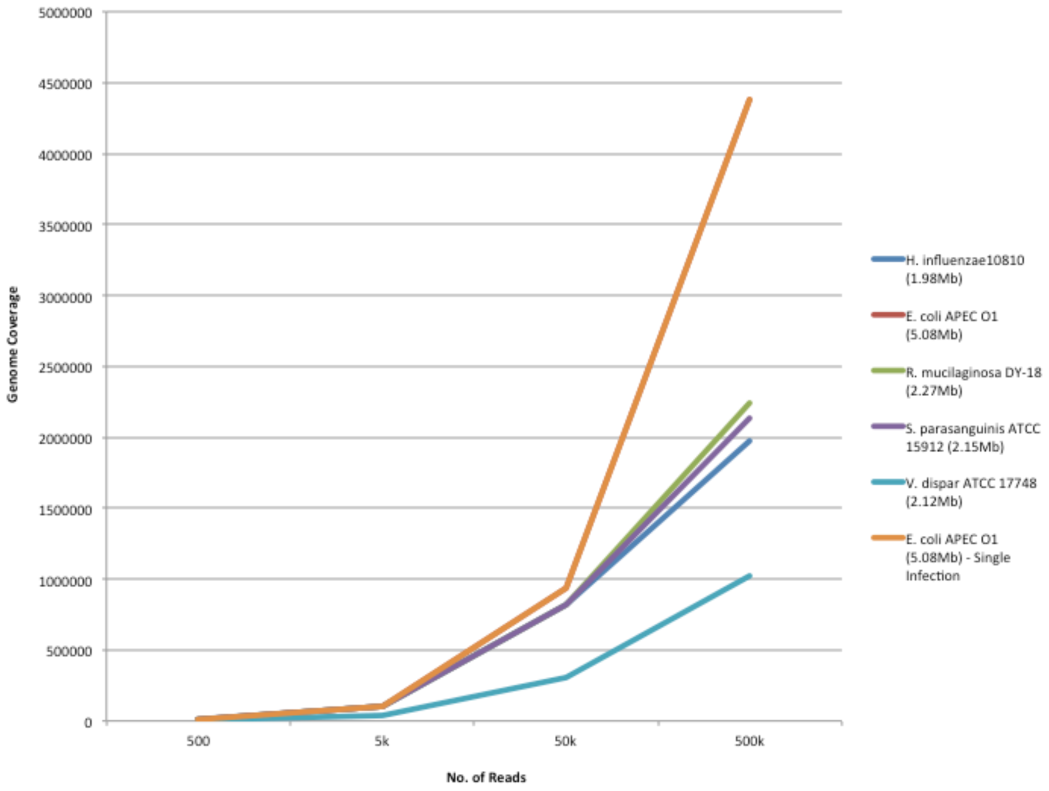


Figure 7. Detection of genomes in complex community. Relationship between genome size and genome coverage with increasing sequencing reads. Effect of detection on *E. coli* APEC O1 in simple and complex community.

Co-infections

We were able to accurately detect top hit per genus in complex community for known genomes in complex community as discussed above. However certain clinical samples might have pathogens from same species example S. *aureus* Newman and S. *aureus* TCH516. The *S. aureus* TCH1516 and S. *aureus* Newman belong to different clonal complex (CC8 & CC5) and resistant (MRSA) and sensitive to methicillin (MSSA) respectively. To test the specificity in these types of clinical samples, we created co-infection libraries consisting of simulated reads from S. *aureus* Newman and S. *aureus* TCH1516 genomes.

The presence of *S. aureus* Newman in co-infection library (true positive) is compared with detection in single infection library (false positive) that only has simulated reads from *S. aureus* TCH1516 genome (Figure 5-A).

Any genome coverage% detected for *S. aureus* Newman in single infection library can be considered as false detection. The %genome coverage of *S. aureus* Newman (false call) is reported slightly less than its true presence in multiple-infection library, due to contribution of homologous reads from *S. aureus* TCH1516. However by top hit per genus, *S. aureus* Newman rank behind few other closely related genomes of *S. aureus* TCH1516 (CC5) in single infection library but is detected as top hit in co-infection library as shown in (Table 2).

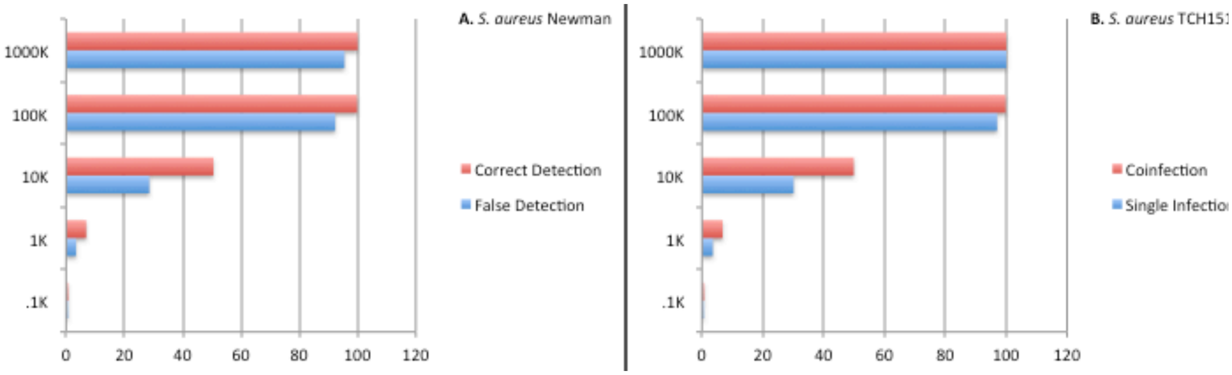


Figure 8. Comparison of detection of close relative in co-infection versus single infection. **A.** Comparison of %genome coverage of true detection in co-infection versus false detection of *S. aureus* Newman. **B.** Comparison of %genome coverage of *S. aureus* TCH1516 in co-infection versus simple infection.

A co-infection library consist of reads from both *S. aureus* TCH1516 and *S. aureus* Newman. Due to the all reads mapped against all reference strategy, the shared homology between these two organisms results in higher % genome coverage of *S. aureus* TCH1516 in co-infection library then single-infection library (Figure 5-B). The *S. aureus* TCH1516 is detected as top hit per genus in single infection library and as one of top hits with *S. aureus* Newman in co-infection library as represented in Table 2. We can infer that though detection as top hit per genus is correct in co-infection metagenome, however detection of co-infection is difficult and might require additional validation.

Table 2. The rank of *S. aureus* Newman and *S. aureus* TCH1516 in single infection and co-infection library. FP: False Positive; TP: True Positive.

	Ranking of Sa. Newman		**Ranking of Sa. TCH1516**	
Reads	**FP**	**TP**	**Single Inf (TP)**	**Co-Inf (TP)**
.1K	5	1	1	2
1K	4	1	1	3
10K	5	1	1	3
100K	5	1	1	2

Diversity

The metagenome reads might have artificial divergence due to sequencing error or variation incorporated due to selective mutation. The ability to assign these divergent reads back to its genome can allow the sensitivity for detection. Local alignment algorithms are considered to be more sensitive and accurate than global alignment algorithms. However, with the high number of metagenome reads, local aligner like BLAST is not optimal given the high computational time cost. Existing global aligners such as BWA, STAMPY, BOWTIE, BFAST, MAQ, NOVOALIGN, MOSAIK and SOAP2 are preferred over local alignments given high volumes of metagenome sequences. However utilizing only global aligner might result in loss of detection of divergent reads. To incorporate these divergent reads for sensitive detection, we utilized BLAT that is ~500 times faster than preexisting tools with comparable sensitivity.

We designed the simulated reads from *S. aureus* TCH1516 genome with increasing divergence (See Benchmark above). To evaluate the sensitivity, reads that global aligner is unable to map but are aligned by local aligner (BLAT) are categorized as divergent reads. With increasing divergence, higher number of reads is not aligned by global aligner (Figure 6).

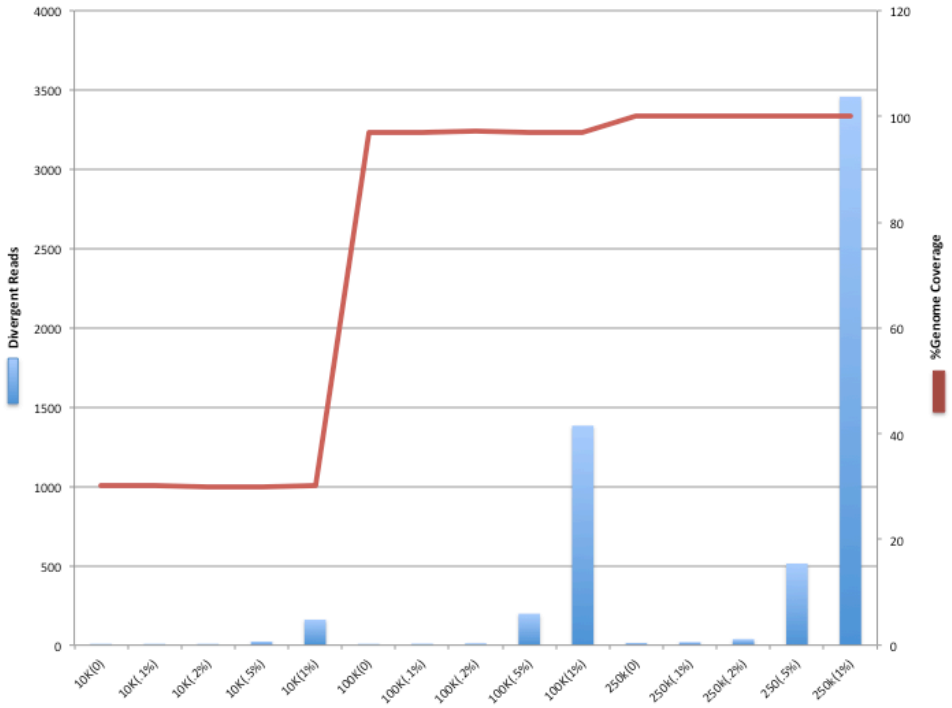


Figure 9. Relationship between percent genome coverage and read recall percentage with incremental divergence.

The pipeline is able to incorporate these divergent reads through local alignment without decrease in the genome coverage for detection (Figure 6). In all the 25 simulated test case (0%, 0.1%, 0.2%, 0.5% and 1% divergence for 100, 1K, 10K, 100K, 250K reads), *S*. *aureus* TCH1516 is detected correctly in all except one, at 1% divergence with 100 reads. The limitation of detection for correct identification can therefore be seen at highest divergence with low number of reads.