SAND2011-9420C

Pathogen detection in clinical samples by high-throughput sequencing

Owen Solberg*, Milind Misra†, Joseph S. Schoeniger*, Kelly P. Williams*

Sandia National Laboratories: Livermore CA* and Albuquerque NM†

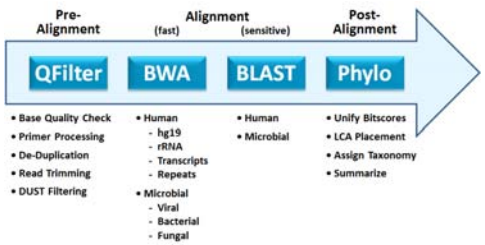
PROJECT BACKGROUND AND OVERVIEW

A main task in sequence-based detection of pathogens in clinical samples is to quickly identify and remove human sequences, a problem benefitting from the large phylogenetic separation between humans and their pathogens. We have developed a computational pipeline for identifying likely pathogens in high-throughput sequence data from clinical samples. The pre-alignment phase of our pipeline removes minimum-quality tails and rejects sequences with low average quality, removes artifacts involving primer sequences used in library preparation, and rejects low-complexity sequences. The alignment phase begins with Burrows-Wheeler (BW) alignments that provide quick identifications (first for human, then for bacterial, viral and fungal sequences) for the bulk of reads. With the smaller remaining readset, more sensitive BLAST-based alignment is applied to reach more divergent sequences. The post-alignment phase assigns taxonomy to the reads hit during alignment and summarizes abundances at multiple taxonomic levels. The summary often provides strong evidence for taxonomic identification of threat organisms.

We have generated sequence data from nasopharyngeal samples from patients with respiratory complaints, and applied our pipeline. Unlike more protected regions of the human body, the nasopharynx contains great diversity of organisms, and variability among patients. Some samples yield clear diagnoses (eg, large fractions of Streptococcus), whereas others yield near-equal distributions of multiple potential pathogens.

METHODS

Our metagenomics pipeline employs a custom pre-alignment filter, followed by a fast then sensitive phases of alignment using standard aligners, followed by a custom phylogenetic analysis tool. Our pipeline quickly removes host (human) sequences to focus on phylogeny of the microbiome.



The prealignment filter was written to deal with artifacts that we uncovered. It has three phases: Phase 1 is per-read: it looks at quality strings of each read trimming ends where each position has received the minimum quality score. Phase 2 is per unique sequence: it identifies and trims primer sequences (as detailed below). Any of the above trimming processes can trigger rejection, if length falls below a minimum. Phase 2 also has a filter for ambiguous base calls and low-complexity sequences (DUST). Phase is again per-read, calculating average quality score.

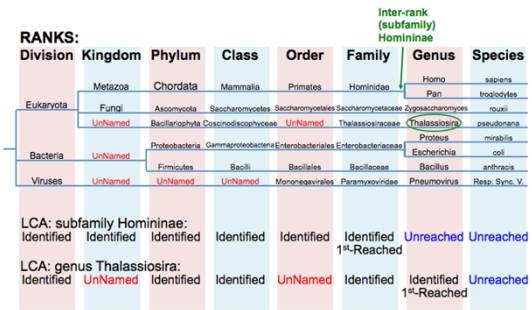
	Hits		Rejections	
	Process	% Reads	Process	% Reads
Minimum-Q 3' Tail	tail3	31.777	tail3	4.294
Minimum-Q 5' Tail	tail5	0.035	tail5	0.003
Primer	primer	4.390	primer	1.098
Any-Ambigous Call	anyN	1.022	anyN	1.022
Low-Complexity	loCmplx	37.599	loCmplx	27.873
	reject	39.627	qscore	5.337

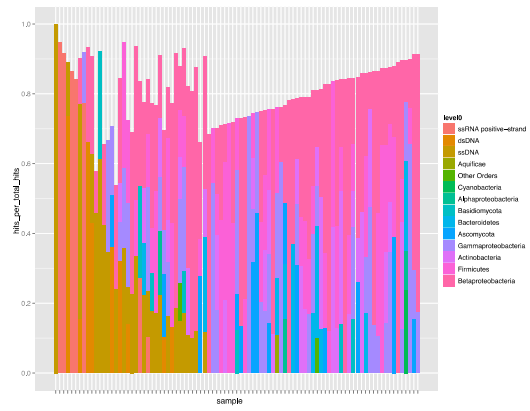
Careful primer searching was called for when we observed large amounts of tandem-array primer artifacts. Primer search uses 14-mers from each end of each primer in both orientations presuming that finding the ends will also treat the middle. Hits to a 14-mer or any of its 1-ntd mismatch variant trigger trimming, first finding the implied 3' end of the primer and trimming all sequence in the 5' direction. A second phase of primer searching looks for smaller 3' primer fragments at the extreme end of reads. Even when artifact levels are low, careful primer trimming improves yields of BW hits both proportionally and absolutely.

[placeholder for primer hit figure]

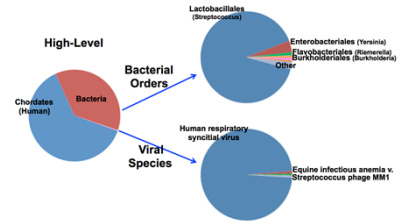
RESULTS

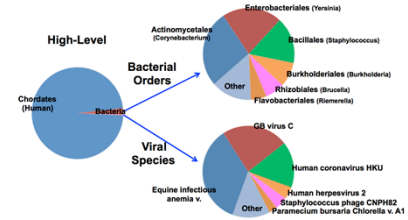
RapTOR's metagenomics pipeline quickly removes host (human) sequences to focus on phylogeny of the microbiome. In prealigment, reads may be trimmed or rejected based on quality, primer content, and complexity. Fast alignment finds close matches first to human then microbial references, reducing the load for slower but more sensitive searches. A 66% Lowest Common Ancestor algorithm assigns taxonomy.

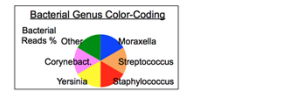


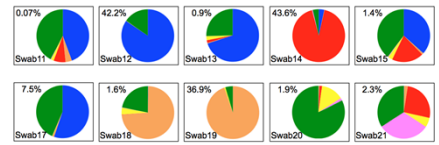


Diversity of nasopharyngeal samples. Patients with respiratory complaints were swabbed and sequences were processed through our pipeline. Often a compelling infectious agent appears, but there are also many cases where several potential pathogens appear with a balanced distribution. Future improvements to our pipeline will focus on pathogenicity genes per se and should help adjudicate tough calls.









Recently added functionality includes protein family profile analysis and assembly of viral genomes.

CONCLUSIONS AND FUTURE DIRECTIONS

We have completed a fully automated analytical pipeline for the detection of known and novel pathogens. The pipeline outputs detailed analysis of individual samples. Some samples have an obvious pathogen candidate, while others balance several pathogens.

We have observed that bacterial may swamp virus signal. Suppression of bacterial sequences may help uncover viral signal.

Future work will include:

- Simulations for confidence levels
- Focusing on pathogenicity genes
- Replace LCA algorithm with tree-traverser