**Methods and Brief Summary for Construction of *in silico* metagenomes for decontamination software evaluation”**

Report created by Shannon Joslin | April 15, 2018

**Background:**

Metagenomic samples may have host DNA concentration levels ranging from 20-80%1. In humans, host contamination of metagenomic datasets provide issues of security and inaccurate analysis. Human reads mistakenly included in publicly accessible metagenomics data may compromise the privacy of the sample’s host. Additionally, analysis of data including host reads will slow down assembly steps, cause misassembly and provide erroneous conclusions in downstream analysis. Therefore, it is important to remove host DNA from all metagenomic samples. The Human Microbiome Project (HMP), an NIH funded initiative providing standardized data resources, suggests using the software Deconseq2 and BMTagger3 as potential tools for decontaminating metagenomic sequencing reads4.

Deconseq software uses an alignment based search method to eliminate human reads from datasets. It was built for rapid identification and removal of sequence contamination in reads greater than a mean length of 150nt. Their paper, published in 2011, gives a detailed analysis comparing the accuracy of various aligner tools but does not do a direct analysis of against other decontamination software.

Unlike Deconseq’s alignment based evaluation of contaminants, BMTagger does not align all reads to the human genome. Rather BMTagger compares 18mers found in each sequencing read with 18mers found in the human genome. Any unclassifiable read from the afore mentioned heuristic discrimination is then fed into a short read alignment too, *srprism,* for alignment based classification. This allows BMTagger to be both fast and sensitive, however no software statistics were given in the Rotmistrovsky et al. paper3.

Here, I present simulated metagenomic profiles of infant gut microbiomes to evaluate the sensitivity, specificity, accuracy and runtime of two different decontamination tools for integration into the Mills’ lab metagenomics pipeline. Metagenomic profiles are simulated based of a proposed infant human gut sampling containing *Homo sapiens*(*H. sapiens*), *Bifidobacterium longum infantis* (*B. infantis*) and nine other microbial species commonly found in infant microbiomes. In infant guts *B. infantis* has been in both high and nominal abundances5,6. Varying *B. infantis* levels in the infant gut are of interest to the Mills lab and were preferentially simulated due to the importance of *B. infantis* in aiding healthy breakdown of milk carbohydrates7.

**Methods:**

***in silico* reference composition**

In order to evaluate the software, *in silico* metagenomic data were created using the next generation sequencing simulator software MetaSim8. All genomic sequences to be used in MetaSim were retrieved as fasta files from the NCBI Taxonomy database. Four different human gut metagenomes were simulated based on high and low contaminant DNA (*H. sapiens*) and high and low levels of *B. infantis.* Additionally, nine other microbial species commonly found in infant guts were included to best replicate the microbial conditions of infant fecal samples9. High abundance human contaminated metagenomic profiles contained 80% human reads overall and low abundance profile contained 20% human reads overall. High and low abundance *B. infantis* profiles were calculated based on the subtraction of human reads from total reads with 80 or 20% of the remaining reads being allocated to *B. infantis*, respectively. The outstanding read counts were equally distributed between non-microbial species (Table 1).

**Table 1.** Abundance profiles of *in silico* infant gut metagenome profiles for A. high *H. sapiens* contaminant, high *B. infantis* levels; B. high *H. sapiens* contaminant, low *B. infantis* levels; C. low *H. sapiens* contaminant, high *B. infantis* levels; D. low *H. sapiens* contaminant, low *B. infantis* levels. Each profile shows the percent abundance of nine different bacteria after subtraction of *H.sapiens* and *B.infantis*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | A. | B. | C. | D. |
|  |  | high *H.sapiens* high *B.infantis* | high *H.sapiens* low *B.infantis* | low *H.sapiens* high *B.infantis* | low *H.sapiens* low *B.infantis* |
| *Bifidobacterium* | *longum* | 17.00% | 4.00% | 68.00% | 16.00% |
| *Streptococcus* | *thermophilus* | 0.33% | 1.78% | 1.33% | 7.11% |
| *Enterococcus* | *raffinosus* | 0.33% | 1.78% | 1.33% | 7.11% |
| *Ruminococcus* | *gnavus* | 0.33% | 1.78% | 1.33% | 7.11% |
| *Bacteroides* | *vulgatus* | 0.33% | 1.78% | 1.33% | 7.11% |
| *Escherichia* | *coli* | 0.33% | 1.78% | 1.33% | 7.11% |
| *Lactobacillas* | *rhamnosus* | 0.33% | 1.78% | 1.33% | 7.11% |
| *Veillonella* | *dispar* | 0.33% | 1.78% | 1.33% | 7.11% |
| *Clostridium* | *bolteae* | 0.33% | 1.78% | 1.33% | 7.11% |
| *Eubacterium* | *eligens* | 0.33% | 1.78% | 1.33% | 7.11% |
| *Homo* | *sapiens* | 80.00% | 80.00% | 20.00% | 20.00% |

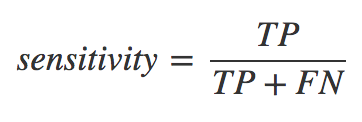
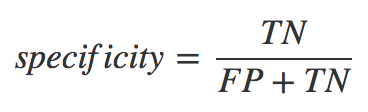
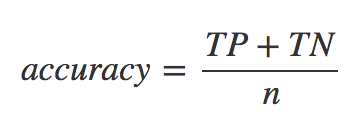
***in silico* metagenome construction**

Eight million single ended (SE) reads with a length of 225nt were simulated for each of the four different microbiome profiles. A read per nucleotide Illumina HiSeq error profile was created for all reads according to Minoche et al.10 (Supplemental 1).

**Evaluation of decontamination tools**

The simulated metagenomes were decontaminated with Deconseq and BMTagger software (Supplemental 2). Deconseq ran with 90 percent coverage and 90 percent identity. Both tools used human genome build 37. Each software was assessed for performance based on sensitivity, specificity, accuracy and runtime.

The calculations of sensitivity (the true positive rate) and specificity (true negative rate) were used to evaluate classification performance of each software at identifying human reads. Accuracy was used to assess the total number of true classifications.

The number of microbial reads classified as human were considered false positives (FP), human reads classified as microbial were considered false negatives (FN), human reads classified as human were considered true positives (TP) and microbial reads classified as microbial were true negatives (TN). The number of total reads was equal to *n*.

**Results:**

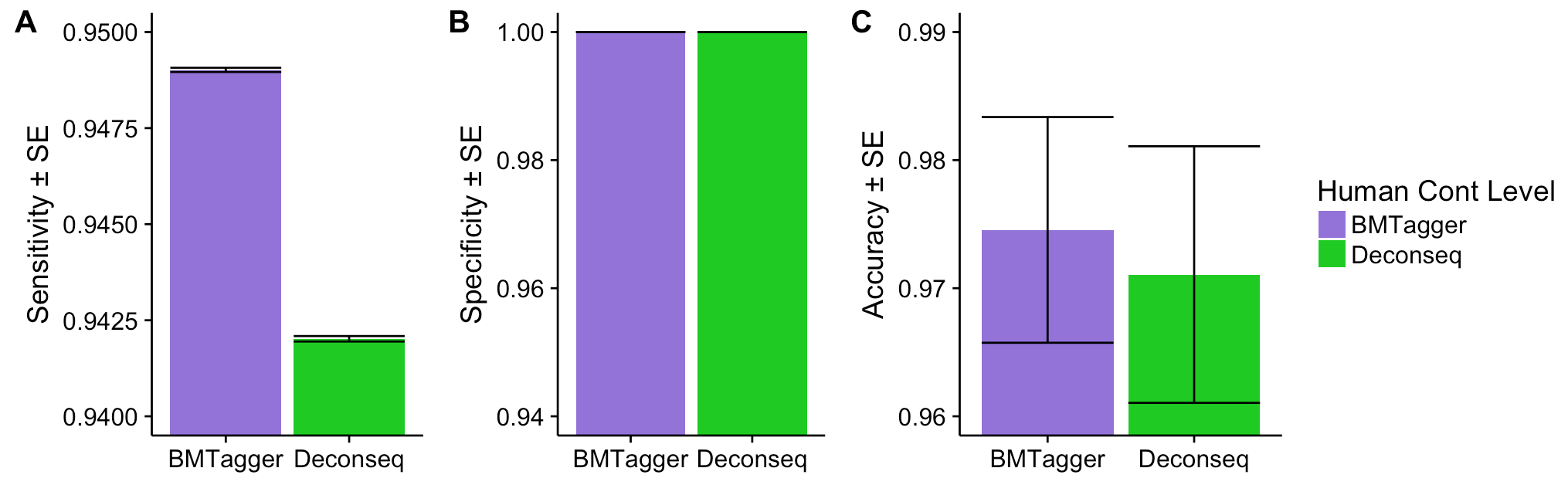
Overall, BMTagger was slightly more accurate and significantly more sensitive than Deconseq. The two software were equally specific. BMTagger significantly outperformed Deconseq’s standalone command line interface with respect to temporal efficiency each with runtimes of 1:00:21 and 69:46:35 (hrs:min:sec), respectively (Figure 1, Table 2). In both Deconseq and BMTagger lower human contamination levels resulted in significantly increased accuracy. Interestingly, lower human contamination also showed a decrease in sensitivity for software, though not significantly so (Figure 2, Table 2).

**Figures and Tables:**

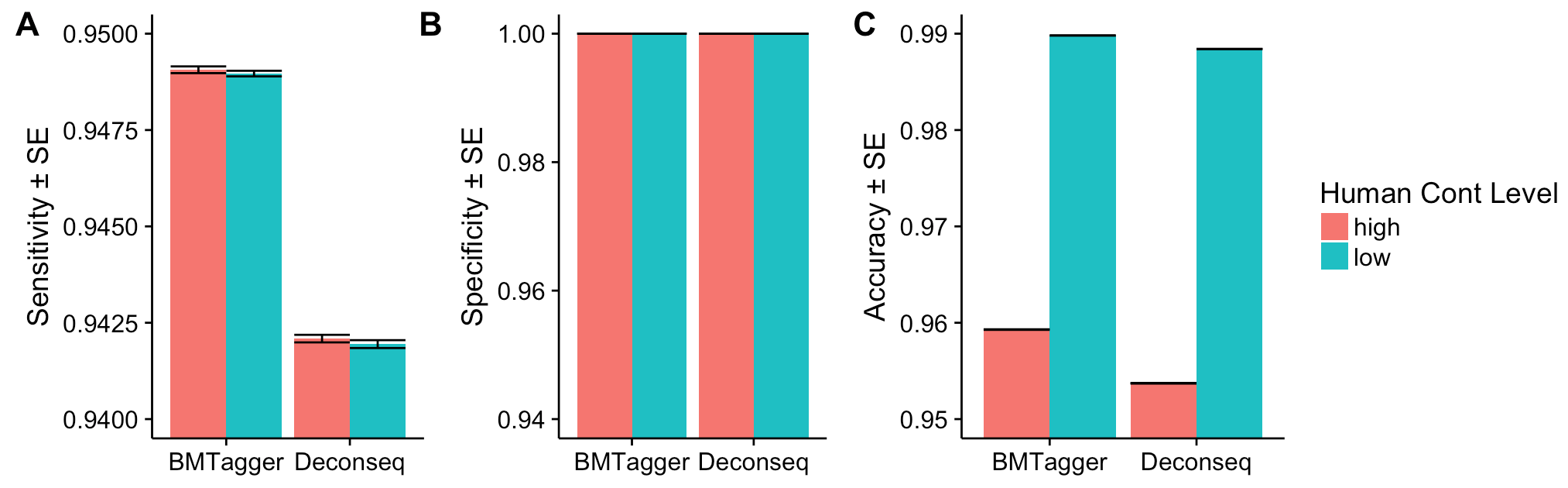
**Table 2.** The output for sensitivity, specificity, accuracy and runtime of the human decontamination software Deconseq and BMTagger. An overall assessment as well as different assessment of human contaminants is listed for each of the afore mentioned specifications. The run time is the total time taken to run all four datasets using the maximum memory of a supercomputing node.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Sensitivity | Specificity | Accuracy | Runtime (hrs:min:sec) | n |
| Deconseq Overall | 0.94201 ± 7.10e-05 | 1.00 ± 0.00 | 0.97106 ± 0.01002 | 69:46:35 | 4 |
| Deconseq High Human | 0.94209 ± 9.80e-05 | 1.00 ± 0.00 | 0.95371 ± 3.02e-05 | n/a | 2 |
| Deconseq Low Human | 0.94195 ± 1.02e-04 | 1.00 ± 0.00 | 0.98841 ± 1.17e-05 | n/a | 2 |
| BMTagger Overall | 0.94901 ± 5.45e-05 | 1.00 ± 0.00 | 0.97455 ± 0.00881 | 1:00:21 | 4 |
| BMTagger High Human | 0.94906 ± 8.76e-05 | 1.00 ± 0.00 | 0.95929 ± 2.77e-05 | n/a | 2 |
| BMTagger Low Human | 0.94896 ± 7.23e-05 | 1.00 ± 0.00 | 0.98981 ± 6.75e-05 | n/a | 2 |

­



**Figure 1.** Comparison of A. sensitivity B. specificity and C. accuracy for Deonseq and BMTagger on all four simulated metagenomes. BMTagger is significantly more sensitive than Deconseq, both software are equally specific and BMTagger marginally outperforms Deconseq in regard to accuracy but not specifically so.



**Figure 2.** Comparison Deconseq and BMTagger performance on high and low human contaminants for A. sensitivity B. specificity C. accuracy. No significant difference is found for either sensitivity or specificity, however both BMTagger and Deconseq become more accurate with lower human contamination level.

Other Files:

All files can be found on the Farm cluster with the path /group/kmkalangrp/SEKJ\_work/

**Discussion:**

The efficacy of removing human contaminants from metagenomic data by BMTagger and Deconseq revealed BMTagger to be more sensitive and time-efficient than Deconseq software. Both tools were completely specific–– identifying no bacterial reads as human–– but showed no significant difference in overall accuracy. Additionally, both tools showed lower accuracy in highly contaminated datasets; suggesting it may be worthwhile to use two different programs to remove human reads on highly contaminated metagenomics data. This may be carried out with BMTagger first, followed by Deconseq, due to the slow run time of the latter. Further testing is necessary to see if the two-step method would remove significantly more human reads. Adjustment of percent coverage and percent identity may lead to more accurate or more sensitive results by the Deconseq software. However, it is uncertain how or if Deconseq may become more time efficient. These data suggest BMTagger to be a superior tool for implementation into metagenomic pipelines.

**References:**

1. Ames, S. K. *et al.* Using populations of human and microbial genomes for organism detection in metagenomes.

2. Schmieder, R. & Edwards, R. Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets. **6,** (2011).

3. Rotmistrovsky, K. & Agarwala, R. BMTagger : Best Match Tagger for removing human reads from metagenomics datasets BMTagger screening. 2–7

4. Nih, T. & Working, H. M. P. The NIH Human Microbiome Project. 2317–2323 (1977). doi:10.1101/gr.096651.109.

5. Favier, C. F., Vaughan, E. E., Vos, W. M. De & Akkermans, A. D. L. Molecular Monitoring of Succession of Bacterial Communities in Human Neonates Molecular Monitoring of Succession of Bacterial Communities in Human Neonates. *Appl. Environ. Microbiol.* **68,** 219–226 (2002).

6. Hopkins, M. J., Macfarlane, G. T., Furrie, E., Fite, A. & Macfarlane, S. Characterisation of intestinal bacteria in infant stools using real-time PCR and northern hybridisation analyses. *FEMS Microbiol. Ecol.* **54,** 77–85 (2005).

7. Underwood, M. a, German, J. B., Lebrilla, C. B. & Mills, D. a. Bifidobacterium longum subspecies infantis: champion colonizer of the infant gut. *Pediatr. Res.* **77,** 229–35 (2015).

8. Richter, D. C., Ott, F., Auch, A. F., Schmid, R. & Huson, D. H. MetaSim — A Sequencing Simulator for Genomics and Metagenomics. **3,** (2008).

9. Turroni, F. *et al.* Diversity of bifidobacteria within the infant gut microbiota. *PLoS One* **7,** 20–24 (2012).

10. Minoche, A. E., Dohm, J. C. & Himmelbauer, H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biol.* **12,** R112 (2011).