Human DNA Decontamination Software for WGS Comparison and Optimization Using *In Silico* and *In Vivo* Datasets

Samir Akre[1], Shannon E.K. Joslin [4], Diana H. Taft[1,2], and David A. Mills[1,2,3].

UCDAVIS
UNIVERSITY OF CALIFORNIA

[1]Department of Food Science and Technology, [2]Foods for Health Institute, [3]Department of Viticulture and Enology, [4]Department of Molecular and Cellular Biology,
University of California, Davis, CA 95616.

Abstract

Whole metagenomic sequencing (WGS) is used to study human associated microbiomes. WGS sequencing files may contain human derived reads. This poses an ethical issue because personally identifiable information may be present in datasets. Therefore, the Human Microbiome Project recommends two potential tools: Deconseq and BMTagger to remove human reads. Both BMTagger and Deconseq have successfully identified a set of human-origin sequences in infant WGS datasets. Comparisons of the accuracy of BMTagger and Deconseq in removing human reads using an *in silico* database built with MetaSim suggest that both have a specificity and that BMTagger has better sensitivity. Overlap in reads identified as human is high between both tools. Identity checking of a random subset of reads checked by BMTagger and Deconseq is ongoing. We plan to use the BMTagger to remove human reads and the retain microbial reads in future WGS based studies of the human microbiome.

Materials & Methods

- The *in silico* database was created using metaSim[2] to mimic an infant microbiome with high or low levels of human contamination (labelled as highH or lowH) and high or low levels of *Bifidobacterium longum infantis* (highB or lowB)
- Deconseq was run using 95%, 90%, and 85% coverage and identity[3]
- BMTagger was run with default conditions[1]
- Infant samples were from a cohort of Bangladeshi infants sequenced on the Illumina HiSeq 2500 with 150 PE reads and a fragment size of ~250 bp

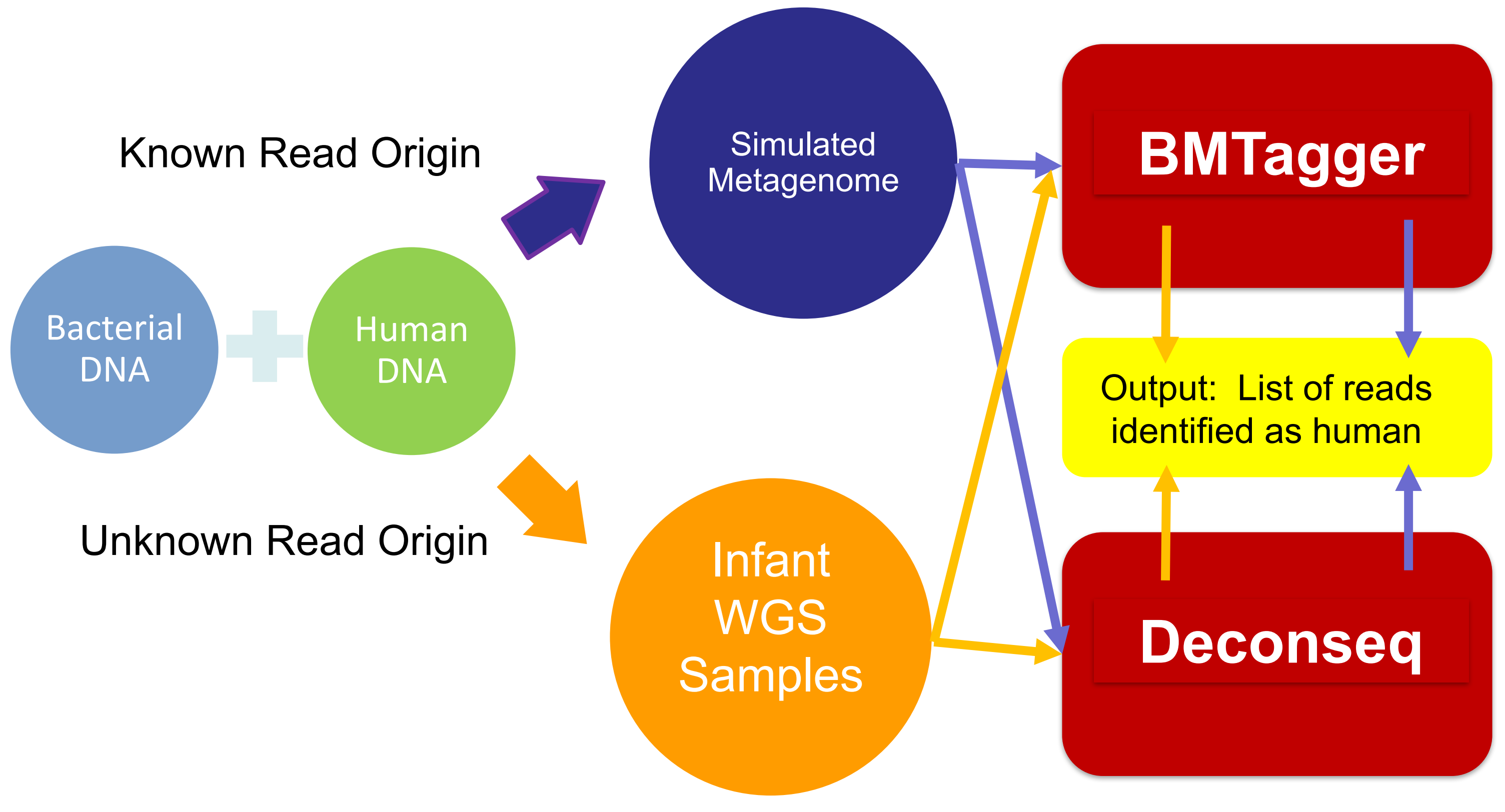


Figure 1: Comparing Deconseq and BMTagger
Both BMTagger and Deconseq are run on simulated data and actual infant data. For the *in silico* simulated data, results are compared to the known origin of reads. For the real data, results from BMTagger and Deconseq are compared to each other.

Results

Figure 2: Sensitivity and Specificity of BMTagger vs Deconseq
The sensitivity (true human DNA contaminant identification rate) and specificity (true bacterial DNA identification rate) of BMTagger and Deconseq (assessed at 85%, 90%, and 95% coverage and identity). Specificity not shown because it is 100% for all data.

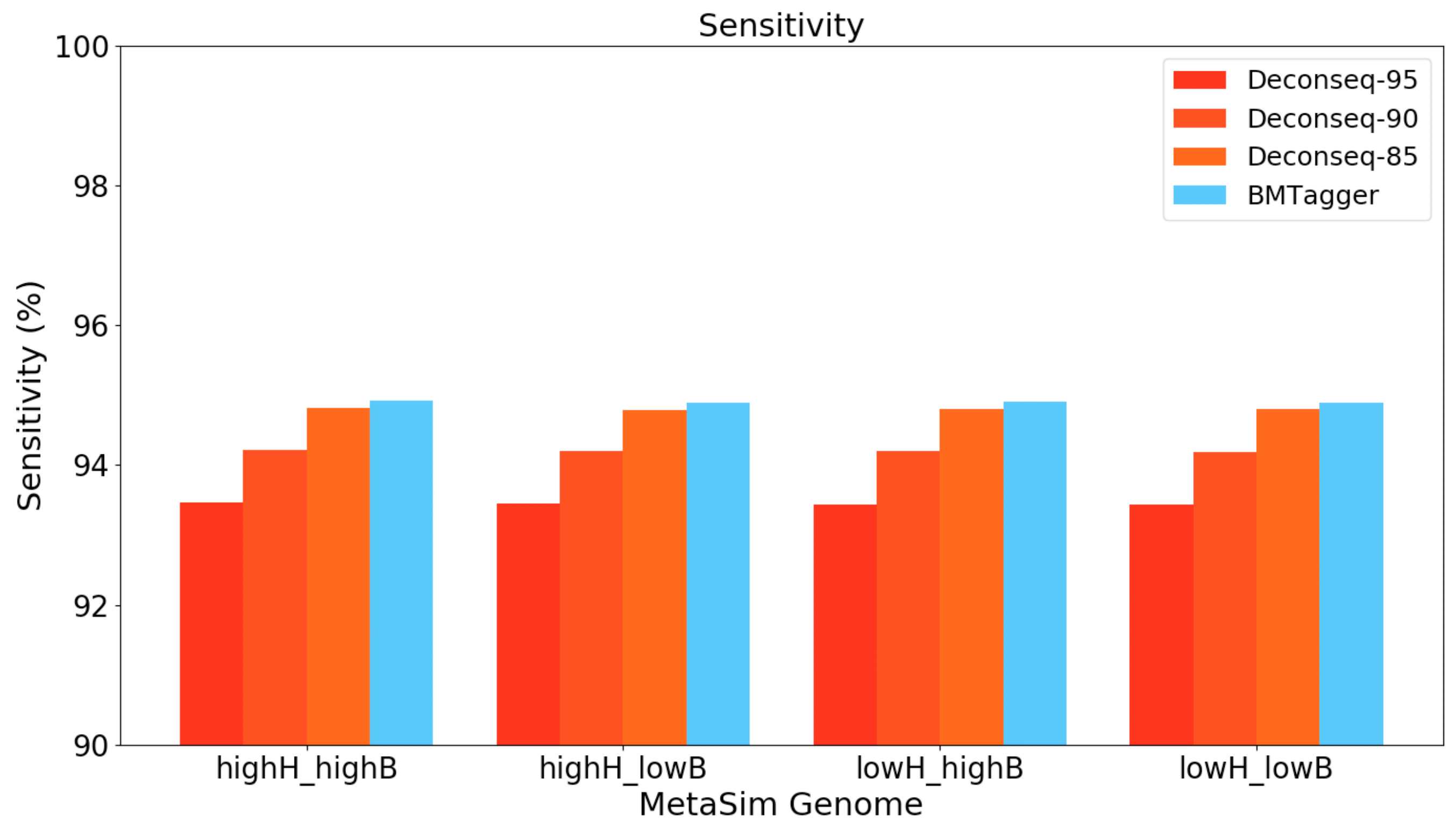
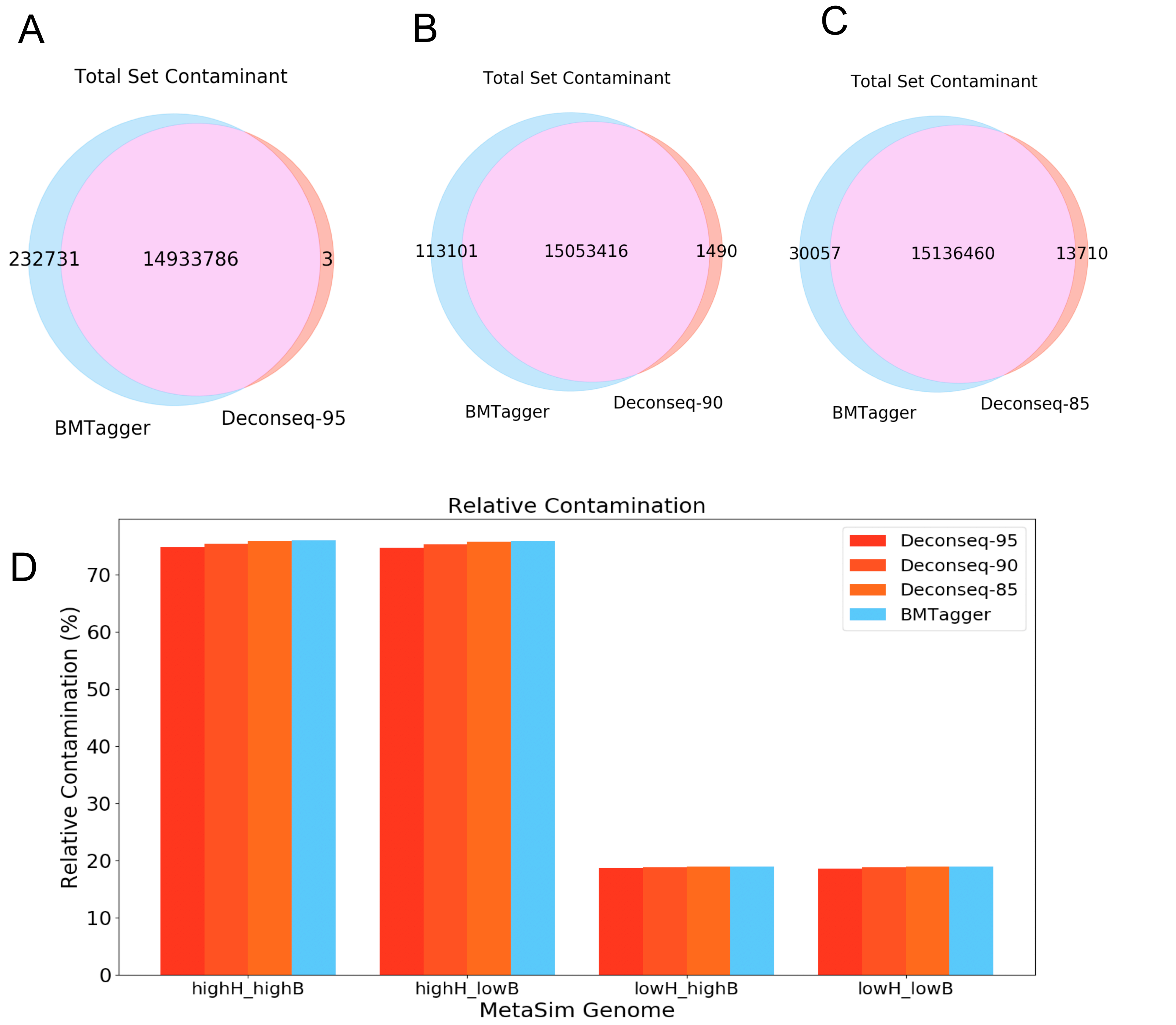
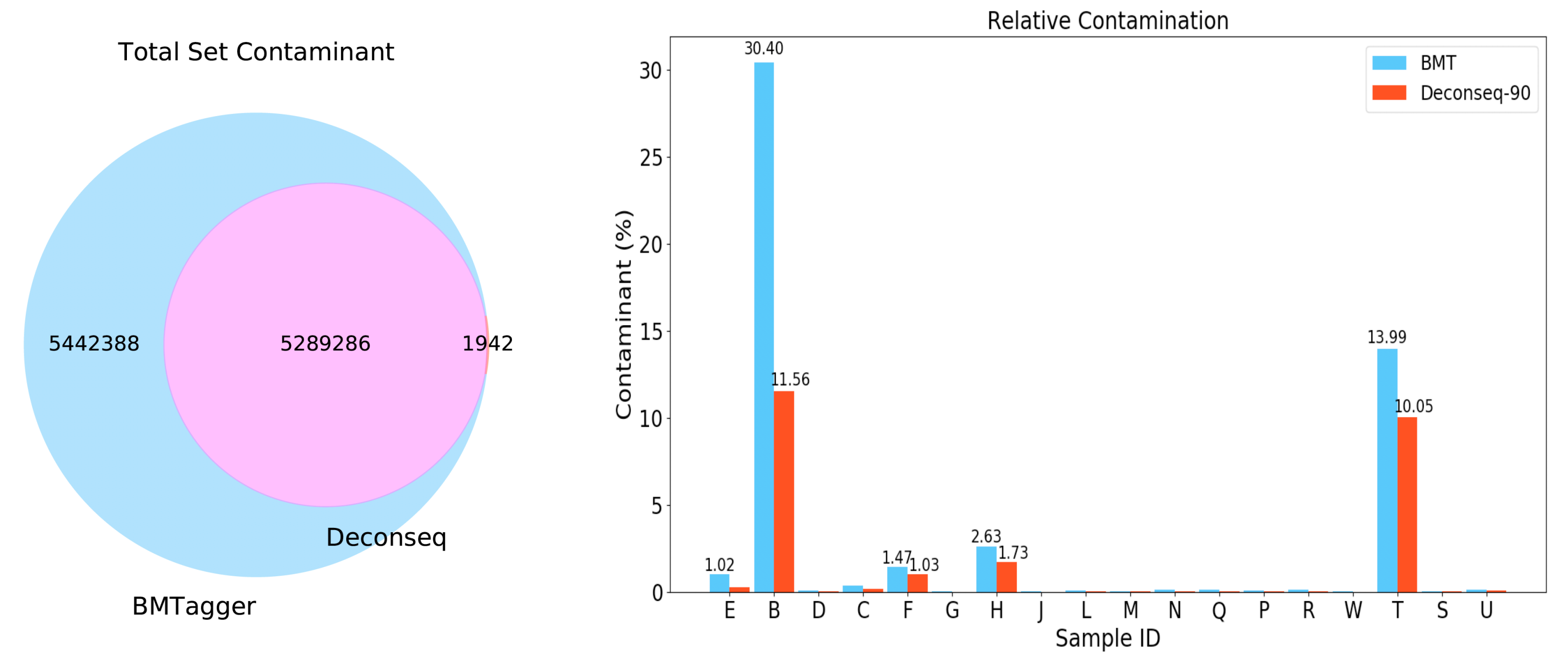


Figure 3: *In silico* BMTagger and Deconseq Findings
The Venn diagrams (A, B, and C) show the reads labeled as human when comparing BMTagger and Deconseq with each coverage percentage for Deconseq. Relative contamination (D) was calculated by finding percent of total file identified as human DNA for both tools. BMTagger overall finds more contaminants in a given file.



Results, Continued

Figure 4: *In vivo* BMTagger and Deconseq Findings.
The Venn diagram shows the difference between the number of reads identified as human using BMTagger and the 90% coverage and identity Deconseq settings identified using the *in silico* dataset. The barchart shows the range of human contamination present in the *in vivo* samples, estimated using both BMTagger and the optimal Deconseq settings.



Conclusions

- Deconseq recommends using a coverage and identity setting of using 90% or 95% if uncertain of the error rate
- A major limitation is our use of the *in silico* database
  - Bacteria in database are by necessity well sequenced
  - Unknown how less characterized bacteria will influence BMTagger or Deconseq operation
- BMTagger requires less computational resources and has a better sensitivity
  - We will use this tool in future studies of the infant microbiome
- Future work
  - Confirm identity of subset of reads using BLAST
  - Compare these tools for work with non-infant fecal microbiomes or the microbiome at other body sites

References

[1] K. Rotmistrovsky, R. Agarwala (2011) BMTagger: Best Match Tagger for removing human reads from metagenomics datasets
[2] Richter DC, Ott F, Auch AF, Schmid R, Huson DH (2008) MetaSim—A Sequencing Simulator for Genomics and Metagenomics. *PLOS ONE* 3(10): e3373.https://doi.org/10.1371/journal.pone.0003373
[3] Schmieder R, Edwards R (2011) Fast Identification and Removal of Sequence Contamination from Genomic and Metagenomic Datasets. *PLOS ONE* 6(3): e17288.https://doi.org/10.1371/journal.pone.0017288

Acknowledgements

This work was supported by National Institutes of Health awards R01AT007079, R01AT008759, the Peter J. Shields Endowed Chair in Dairy Food Science and Mengnui Dairy.