STRAIN LEVEL CLASSIFICATION OF THE INFANT GUT MICROBIOME from shotgun metagenome data

E. Klaassens, N. Mekkes, R. Bongoni -- BaseClear, Leiden, The Netherlands

At birth there is a critical window of opportunity where exposure of microbial strains in the gut shape the infants microbiome and its function. Strain specificity is key in tracking this development. Therefor, large scale trials and clients working in infant nutrition (probiotics, human milk oligosaccharides, supplements) benefit from strain level identification of the infant gut microbiome. With the increase of available genomes and knowledge of microbiome taxonomic trees, the performance of a strain-level metagenomic classifier will only increase over the time to come.

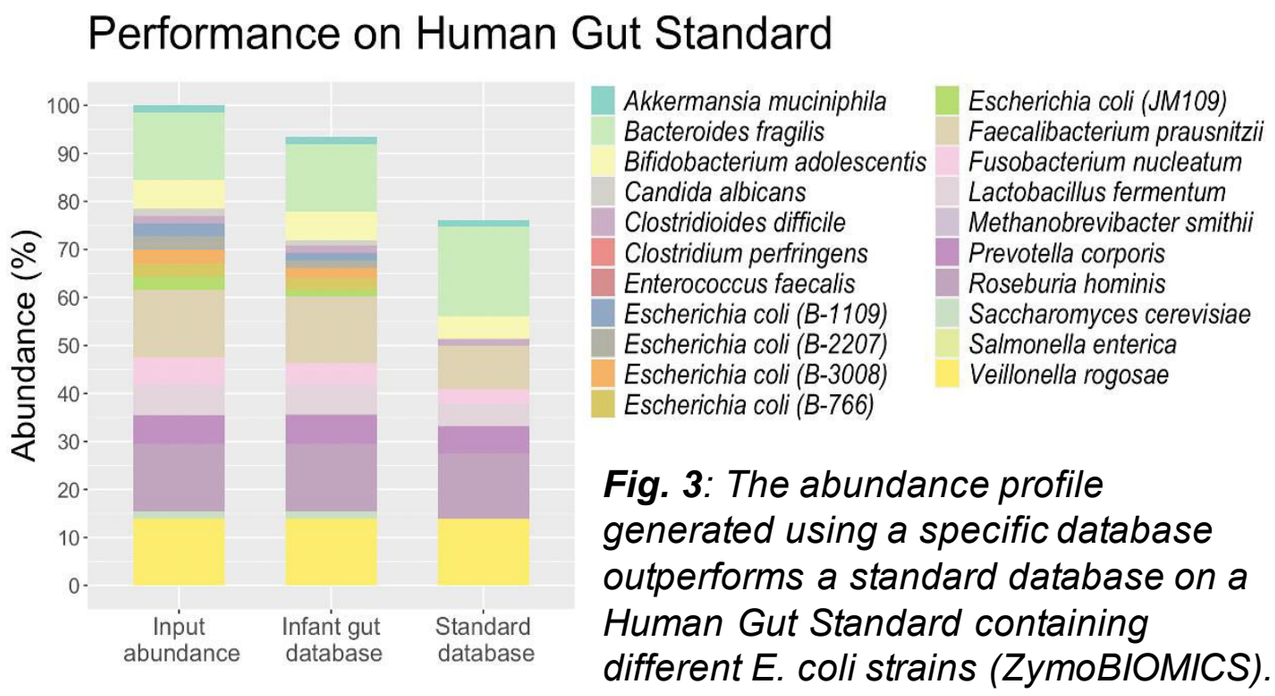

***Fig. 3**: The abundance profile generated using a specific database outperforms a standard database on a Human Gut Standard containing different E. coli strains (ZymoBIOMICS).*

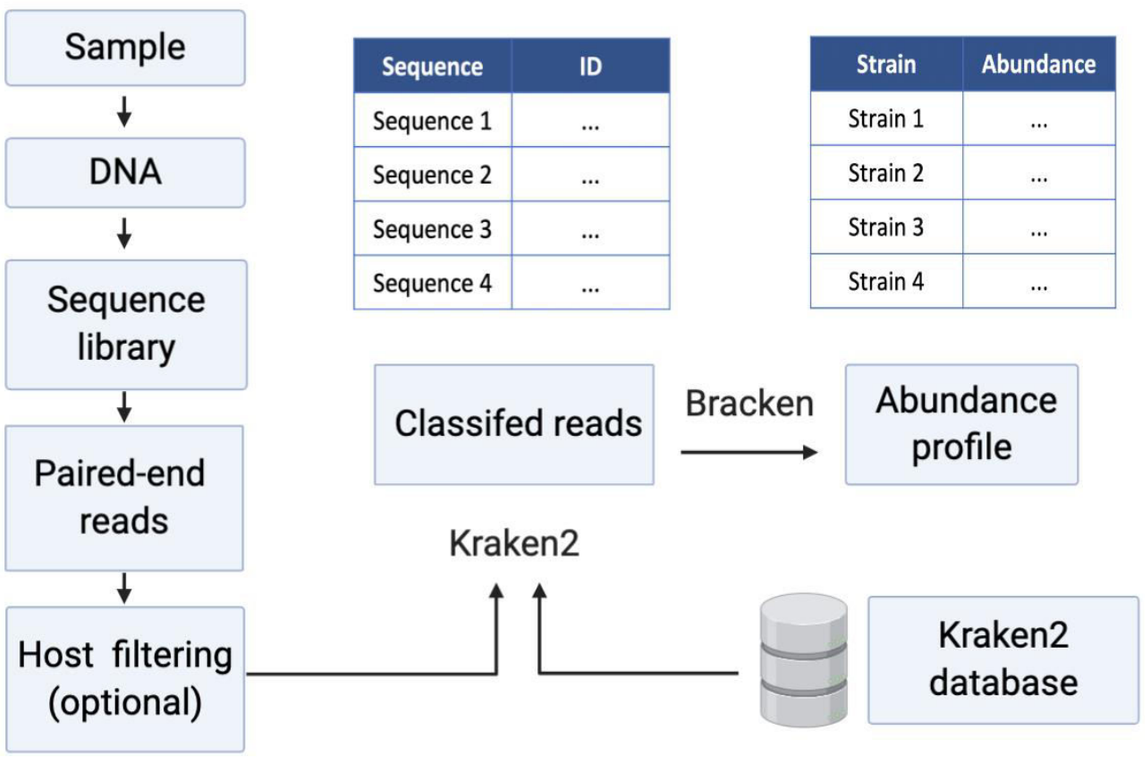

***Fig. 1**: Overview of the infant gut microbiome specific metagenomics pipeline*

Here, we present an integrated computational shotgun metagenomics pipeline (Fig. 1) for strain-level classification and abundance profiling of microbes specific for the infant gut microbiome. The existing tools Kraken2 (1) and Bracken (2) are combined with an infant-specific database. This infant-specific database was constructed based on the latest human gut genome catalogue (3), enriched with ~50.000 strains belonging to bacteria commonly found in the developing infant gut.

The performance of the pipeline was validated on different sets of simulated reads generated with CAMISIM (4), and tested for repeatability using real-life communities. Here we show that this pipeline is capable of correctly distinguishing between a set of simulated highly similar *Bifidobacterium* strain (Fig. 2). In addition, we show that our pipeline can create an accurate abundance profile on Illumina reads subsampled to a depth of 250Mbp of a Human Gut Standard (ZymoBIOMICS), with strain-level differentiation (Fig. 3). These results show how database optimization leads to faster, more accurate and more meaningful strain identification when testing on gut metagenome datasets, especially compared to more generic databases.

Conclusions

- **The infant specific database and pipeline is capable of classifying the gut microbiome down to strain-level**
- **This specific database offers an improvement over the current standard, which goes down to species level only**
- **This pipeline has many uses for research into the infant and pregnancy microbiome from shotgun metagenome data.**

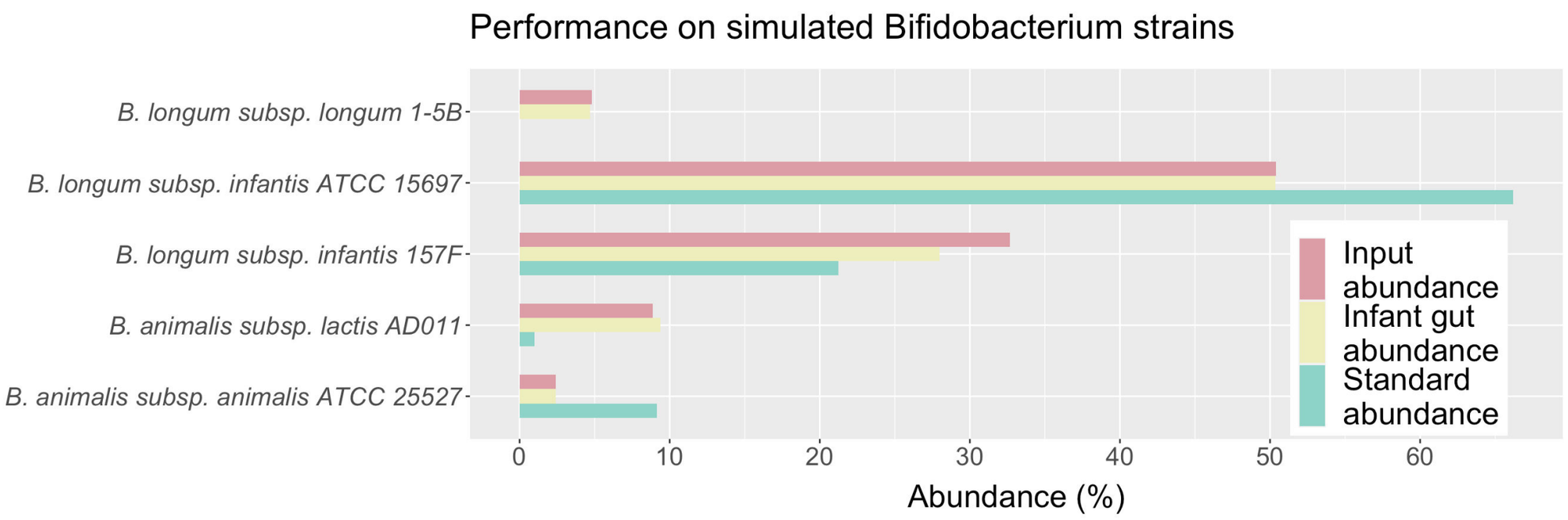

***Fig. 2:** Comparison in abundance profiles outputted using a standard database versus a specific database on highly similar, Bifidobacterium strains (CAMISIM). The standard database (blue) fails to create an accurate abundance profile.*

1) Wood, D et al *Genome biology* 20.1 (2019)
2) Lu, J et al *Computer Science* 3 (2017)
3) Almeida, A et al. *Nature Biotechnology* (2020)
4) Fritz, A et al. *Microbiome* 7.1 (2019)