Review of: "Scalable microbial strain inference in metagenomic data using StrainFacts"

Vladimir Ulyantsev

Potential competing interests: The author(s) declared that no potential competing interests exist.

The article addresses the crucial problem of strain deconvolution in metagenomic datasets. While the overall taxonomic profile of the dataset can be studied using reference databases, the major community functions are most likely driven by the unique strain composition. Authors present a novel algorithm StrainFacts for fast and accurate strain inference across large metagenomic datasets.

The key feature of the method is the support of continuous values in range [0; 1] for genotype variables. Unlike the common methods, which support only discrete values 0 (for reference allele) and 1 (for alternative allele), the designed solution allows the efficient gradient-based optimization of the underlying model. Nevertheless, the values are pushed towards zero or one during training steps and resulting values are discretized (to 0 or 1) for downstream analyses

The proposed method has been validated on simulated and real metagenomic data and compared to an existing tool StrainFinder. The results show the superior time performance of StrainFacts with a compared inference accuracy on various datasets. The algorithm was able to process metagenotype data from more than 20000 metagenomes and perform a strain inference for four species of interest. As a result, several clusters composed only of novel non-reference strains have been found, providing a space for future research.

The presented algorithm seems to be extremely promising in large metagenomic studies. However, I have a few comments as follows:

1. The authors constrain metagenotypes to be biallelic: reference or alternative. While there is a rationale behind this choice, it will be interesting to apply the model supporting all four nucleotide bases. That will provide the possibility not only to assess strain diversity, but to reconstruct the exact nucleotide sequences of inferred strains for various metagenomes. These sequences will be of utmost importance for biological interpretation of newly-detected strains.
2. In subsection “Strain Inference” hyperparameters values used for analysis are presented. It could be useful to discuss how those values were selected and how robust they are for various experimental designs (i.e. Do I need to tune those values for my dataset?).
3. The “Deconvolution of metagenotype data” subsection seems to have some flaws in formulas. Matrix form for alternative allele frequency should be written as $P=\Pi\Gamma$ to match the dimensions. In the same paragraph, there is a missing ‘s’ index for π in the summation formula.

Overall the article is of a high quality and presents a powerful algorithm for studying strain variations in the metagenomic datasets.