Alignment-free quantification of genes and operons across metagenomes by assembly graph deconvolution

Byron J. Smith & Katherine S. Pollard

Quantifying the sequencing depth of genes, operons, or species in shotgun metagenomes is a key step in the analysis of complex microbial communities. Unfortunately, incomplete reference databases and complex relationships between sequences (recombination, duplication, deletion, rearrangement, etc.) lead to widespread issues for alignment-based quantification using short reads[^1]. Graph genomes and the alignment of metagenomic reads to these is a promising solution, but is held back by its computational complexity and incomplete reference databases. On the other side of the spectrum, alignment-free approaches to quantification, which harness ubiquitous and unique marker kmers, are computationally efficient, but limited by the availability and validation of such kmers. The challenge is especially acute at the scale of bacterial operons, where structural variation can have important consequences for function, but cannot be disambiguated using only short reads or kmers[^2]. While long reads solve this problem in theory, such data is still prohibitively expensive for quantifying rare microbes and their genes across large numbers of samples.

Here we explore an alternative approach that fuses sequence assembly and quantification in order to simultaneously overcome both database incompleteness and the ambiguity of read alignment. Concisely, our proposed approach statistically deconvolves sequences in a condensed de Bruijn graph based on kmer counts across multiple samples. We unify the problems of assembly and depth estimation and formulate them as a joint, sparse coding problem applied iteratively on an assembly graph. We describe novel algorithms for denoising depth estimates based on kmers, identifying paths through the assembly graph that represent real sequences, and estimating the depth of these.

We present an implementation of this method in software, analysis of its accuracy and computational complexity, and results from applying it to real metagenomes from a synthetic community derived from isolates with high-quality genomes. Ongoing work is also considering refinements to the method that improve the accuracy of assembly and quantification, heuristics for improved scalability, and extensions that incorporate additional sources of information. Our findings suggest that graph deconvolution is a promising approach for both assembly and quantification using short reads that overcomes inherent limitations of alignment-based approaches.

[^1]: Zhao et al. 2023. Cell Systems. 10.1016/j.cels.2022.12.007

[^2]: Meleshko et al. 2019. Genome Research 10.1101/gr.243477.118