Performing Phylogenomic Analysis on Unassembled Short-Read Genome Sequences

Christopher Bun[1] and Catherine Putonti[1,2,3,*]

1 Department of Computer Science, Loyola University Chicago, Chicago, IL 60611
2 Department of Biology, Loyola University Chicago, Chicago, IL 60660
3 Bioinformatics Program, Loyola University Chicago, Chicago, IL 60660
* Corresponding Author: Catherine Putonti cputonti@luc.edu

With the advent of the latest sequencing technologies, the number of eukaryotic organisms being sequenced has rapidly increased. Phylogenomic analysis of these sequencing projects is resource intensive when traditional alignment-based approaches are used as they necessitate genome assembly. Herein, we present an alignment-free method for comparing unassembled genome sequencing reads. In contrast to methods which compare genomes based upon the frequency of occurrence of subsequences, the method presented here is founded upon the frequency of appearance (presence/absence) of subsequences of a fixed length *k* (or *k*-mers). This method, referred to as the *Frappe* method, represents genomes as a frequency of appearance profile. Considering only the absence or presence of *k*-mers eliminates the need to filter repetitive elements as well as the need to have uniform coverage of the genomic sequence. Using this approach, analysis of 30+ publicly available mammalian genome data (assembled, WGS, and trace) was conducted. Different depths of coverage, sizes of *k*, and read lengths were considered in an effort to assess the method's ability to perform phylogenomic analysis on unassembled short-read data. Even with low coverage, *Frappe* generated phylogenies analogous to those derived by traditional phylogenetic methods.