May 20th, 12:30 PM

Using Next-Gen Sequencing to Estimate Strain Diversity and Frequency within Infections

Nicholas J. Hathaway
University of Massachusetts Medical School

Et al.

Using Next-Gen Sequencing to Estimate Strain Diversity and Frequency within Infections

Nicholas J. Hathaway (1) , Jeffrey A. Bailey (1,2)
University of Massachusetts Medical School, (1) Program in Bioinformatics and Integrative Biology and (2) Division of Transfusion Medicine
Contact: jeffrey.bailey@umassmed.edu

Abstract

Targeted deep sequencing has rapidly transformed our ability to investigate environmental and infectious microbial diversity. Our lab is focused on applying deep sequencing to diversity in malaria infections. A key challenge in all deep sequencing work is determining true sequence differences from errors. While several amplicon deep sequencing clustering tools exist these tools can be CPU intensive and/or lack the sensitivity to detect down to a single base pair difference between sequences, which is a necessity for examining intrapopulation differences in malaria. We have therefore created a novel clustering and statistical framework to overcome these limitations. Our clustering algorithm provides a rapid initial clusters using a step-wise heuristic process collapsing low base quality differences. These initial clusters are then subject to statistical simulations again incorporating quality to assign p-values and refine the clusters. Here, we used several control data sets of known mixtures of 16s sequence from bacterial, Plasmodium sequence, and Hepatitis-C sequence to benchmark our pipeline against other tools demonstrating equal or improved sensitivity and specificity while providing improved speed often by several orders of magnitude. Our method also offers additional benefits such as comparing PCR replicates thereby further reducing error, removing chimeras, and clustering parasites across individual patients for population-based analyses. Additionally, our methods are concrete allowing the user to target a given number of differences between clusters allowing biologic questions to be better framed. Thus, given our accuracy, speed and flexibility, our new program, SeekDeep, should be broadly applicable to deep sequencing applications from microbiomes to HIV diversity.