High concordance between short and long read sequencing for genomics-based species identification and antimicrobial resistance

Jason D. Wittenbach, Paul Knysf, Grace Cox, Ian C. Herriott, Nicole Billings, Tim Farrell, Linnea Sahlgren, Cabell Maddux, Miriam H. Huntley

Day Zero Diagnostics, Boston, MA 02135

Introduction

Traditional laboratory techniques for the diagnosis of bacterial infections, consisting of species identification (ID) and antimicrobial susceptibility testing (AST), require time-intensive culturing and phenotyping steps which can take days, delaying appropriate therapy during a critical time in patient care. The availability of high quality and low-cost rapid DNA sequencing -- as provided by recent advances in nanopore sequencing -- has the potential to transform infectious disease diagnosis with the use of whole genome sequencing (WGS). In previous work, we built a bioinformatics pipeline for species ID (Keynome ID) and a machine learning system for genomic AST prediction (Keynome g-AST) for performing these tasks from WGS inputs, when paired with our sample preparation technology this process provides pathogen ID and AST diagnosis from whole blood samples in hours instead of days.

Here we assess the differences in performance of these algorithms when Illumina short-read versus ONT long-read WGS data is used as input. Bacterial isolate strains across multiple species were selected based on phenotypic and genotypic diversity and genomic DNA was sequenced on both platforms. We report a high degree of concordance for ID (99.4%) and AST (97.7%) between the two sequencing platforms, demonstrating the suitability of ONT sequencing to support such applications.

However, it should be noted that on the g-AST task, for all but one of the small number of discordant predictions, the prediction from Illumina sequencing was correct and that from ONT was incorrect when compared to ground truth, suggesting further improvements in long-read accuracy would still be beneficial. We are currently assessing whether Guppy v5 or other rapid error correction methods could bridge this remaining gap.

Results: species ID

We analyzed 168 strains across 50 species, and found that paired species ID predictions from long- and short-read sequencing were 99.4% concordant (167/168 samples). The lone error came from a sample that was predicted to be Enterobacter cloacae with Illumina sequencing but Enterobacter hormaechei with ONT sequencing. These two species both belong to the "Enterobacter cloacae complex" or closely related organisms, making the predictions concordant at the complex level, not too strict at the species level.

Results: g-AST

We analyzed g-AST predictions from 35 models for species-drug combinations across 9 species and 15 unique antimicrobial agents, making predictions on 10 strains per species for all drugs where a high-performing model was available. This resulted in a total of 350 unique predictions, which were 97.7% (342/350 predictions) concordant between long- and short-read sequencing data.

Discordant predictions did show some tendency to cluster by species-drug combination, indicating that some of the discordance might be related to specific models failing to generalize. Though, when compared to ground truth, 7 of the 8 discordant predictions showed correct Illumina predictions and incorrect ONT predictions.