Tell-TaxContigs: microbial metagenome assembly, taxonomy, and abundance estimation using UST TELL-Seq long-range read information

Bioinformatics

Colin Heberling (cheberling@universalsequencing.com), Universal Sequencing Technology, **Long Pham**, Universal Sequencing Technology, **Sree Krishna Chanumolu**, GeneFront, **Hasan Otu**, OTUFY, **Yu Xia**, Universal Sequencing Technology, **Peter Chang**, Universal Sequencing Technology, **Andrew Anfora**, Universal Sequencing Technology, **Ivan Garcia-Bassets**, Universal Sequencing Technology, **Tom Chen**, Universal Sequencing Technology Corp, **Yong Wang**, Universal Sequencing Technology

Deconvolving diversity of a metagenome is critical for understanding the role of a given microbial community in human health and disease, small molecule biosynthesis, and other complex ecosystems where more reductive analysis proves to be elusive. Metagenomic assembly is one common method for characterizing a metagenome, especially for identifying novel gene content or novel organisms. However, analyzing sequencing data from microbial mixtures with a high dynamic range of relative abundance of strains, close relatedness, and repetitive genomic content amongst members can vastly complicate the genome assembly process of individual microorganisms and strains. Furthermore, efficient assembly requires high fidelity sequencing reads to avoid ambiguities. We previously developed a method that captures long-range molecular origin information from kilobase-long genomic fragments by a process of DNA barcoding—transposase enzyme-linked long read sequencing (TELL-Seq) developed by Universal Sequencing Technologies (UST). TELL-Seq barcoded fragments can be sequenced with instruments that process short reads (i.e., high fidelity sequencing). Here, we show that integration of TELL-Seq data with a computational pipeline that combines de novo genome assembly (Tell-Link) with taxonomic classification and abundance estimation (Tell-TaxContigs) provides highly accurate metagenomic analyses. We show how the application of Tell-Link and Tell-TaxContigs on sequencing data generated from commercially available microbial mixture standards results in genome assemblies with contiguities larger than 1Mbp (N50), and highly accurate classification and relative abundance estimation for organisms at 0.18% or greater relative abundance, respectively. Therefore, Tell-Link, in combination with genome binning software (e.g. metabat2) provides highly contiguous and high fidelity genome assemblies of abundant organisms in a metagenomic sample. Tell-TaxContigs classifies contigs and unassembled reads with BLASTn and using a deep learning approach resolves ambiguities, rules out false positive classifications, and accurately estimates relative abundances of classified species. This approach has an average margin of error of lower than 1% in enumerating relative abundance for the microbial mixture standards tested.