Challenges in the bioinformatics analysis of next-generation, high-throughput sequencing data and its applications

Yongming A. Sun, Francisco M. De La Vega, Applied Biosystems

Next generation, rapid, low-cost genome sequencing promises to address a broad range of genetic analysis applications. This is achieved by carrying out massively parallel reactions. Despite the fact the read-length is short (25-50 bp), the overall throughput is enormous, each run producing up to several hundreds of million reads and billions of base-pairs of sequence data. The massive volumes and differences in the data produced by these technologies pose new challenges for data acquisition, management, and analysis. Software and analysis approaches are under continuous development to keep pace with the rapid evolution of next generation sequencing technologies. We have developed integrated sets of tools for processing, QC, filtering, aligning, visualizing and interpreting short read data from the SOLiD™ System. This system allows sequencing single or paired 25-50bp reads of 10^8 -10^9 templates on a single array containing beads with clonally amplified templates by a unique di-base probe ligation chemistry. This talk will describe our approach for dealing with the major challenges in the bioinformatics analysis of this data, including real-time monitoring, quality control, alignments to reference, and SNP and structural variation finding. We also discuss results of the analysis for biological applications including resequencing, small RNA discovery, and gene expression, with the SOLiD System.