CLOUD-SCALE GENE EXPRESSION QUANTIFICATION OF THOUSANDS OF RNA-SEQ SAMPLES

SEAN DAVIS

Genetics Branch, Center for Cancer Research, National Cancer Institute

Large-scale collection of genomic data is now common in cancer research and many such datasets are now available as public resources. Transferring these data resources over a network to local compute facilities is becoming more problematic as datasets grow. In addition, executing computational workflows over these massive datasets can effectively utilize massively scaleable infrastructures to enhance efficiency and timeliness of results. The National Cancer Institute (NCI) has recently established the NCI Genomic Data Commons as a repository for valuable cancer genomics datasets (Grossman et al. 2016). In addition, the NCI has funded three pilot projects to test the feasibility of processing genomic data using commercial cloud computing infrastructure and platforms. We have applied a state-of-the-art transcript and gene quantification method, Salmon (Patro et al. 2017), to 11048 samples from the TCGA and CCLE projects at an average cost of less than $0.15 per sample. A total of 66.6 TiB of data were processed to produce bias-corrected gene and transcript expression values. Cloud-based infrastructure was built using the Google Genomics Pipeline API (Google 2017). The computational workflows were driven by the cRomwell R package (Davis 2017) and metadata for the TCGA and CCLE data collected and managed using the GenomicDataCommons package (Morgan and Davis 2017). Approximately 20,000 compute cores (the equivalent of 40% of Biowulf) and 5TB of total memory were simultaneously employed on the Google Compute Platform. Resulting files were processed further using Apache Spark, another distributed compute techology. To facilitate data mining and downstream analysis, all processed data were loaded into Google BigQuery massively scalable data warehouse. In summary, we have performed large-scale gene expression quantification as a proof-of-concept, scalable, next-generation computational analysis that combines several distributed computing technologies with reproducible and reusable computational research approaches that has yielded a data product that is immediately useful for cancer data science applications over the entire TCGA and CCLE RNA-seq collections.