Cloud-based Interactive Analytics for Very Large Scale Genomic Data

Cuiping Pan1,2,3,6, Gregory McInnes3,6, Nicole DeFlauro*, Michael Snyder2,3, Jonathan Bingham*, Somalee Datta3,6, Philip Tsao1,5,6

1) Palo Alto Veterans Institute for Research; 2) Department of Genetics, Stanford University; 3) Stanford Center for Genomics and Personalized Medicine, Stanford University; 4) Google Inc, Mountain View; 5) Division of Cardiovascular Medicine, Stanford University; 6) VA Palo Alto Health Care System, CA. *contributed equally

Introduction

Large scale genomic sequencing studies are underway to decipher biological functions, human diseases, ecosystems and to meet agricultural needs. With the quantity and diversity these data harbor, a robust and scalable data handling and analysis solution is desired. Here we present interactive analytics for large volumes of genomic data, developed according to the standards of Global Alliance for Genomics and Health on public cloud infrastructure (e.g. Google Cloud Platform) and distributed computing database Dremel, to perform information compression, comprehensive quality controls, and biological information retrieval. We demonstrate that such computing paradigms can provide orders of magnitude faster turnaround for common analyses, transforming long-running batch jobs submitted via a Linux shell into questions that can be asked from a web browser in seconds.

Workflow for Genomic Data Analysis in GCP

1. Data Transformation
   - Genomics Storage
   - MapReduce/Cluster
   - Data Science Scripts
   - Interactive Exploration
2. Biological Queries
   - BigQuery
3. Quality Control
   - references, phenotypic data
   - annotation

We implemented a three-level quality control pipeline utilizing the GCP infrastructure over the DNA variation, at both the individual and population genomics levels. A majority of these QC steps were performed on BigQuery, while a few others were performed using Cloud Dataflow and Apache Spark.

Conclusion

We presented the computational paradigm combining cloud-based systems and distributed query engines to address the scalability and interactivity in large genomic data analytics.

1. Our database schema facilitated variant data representation in a compressed format yet retaining many critical variant calling parameters for each genome.
2. We developed a comprehensive cloud-based quality control pipeline considering both principles of individual and population genomics.
3. Solutions for querying statistical, biological and medical information were demonstrated.
4. Performance assessment indicated that our approach was interactive and scalable, suitable for analyzing very large volumes of genomics data.

More at GitHub: StanfordBioinformatics/mvg_aaa_codeLab; Contact: cuiping@stanford.edu
Preprint manuscript on bioRxiv: "Interactive Analytics for Very Large Scale Genomic Data."