Cloud-Scale Gene Expression Quantification Of Thousands Of Rna-Seq Samples

Sean Davis, Center for Cancer Research, National Cancer Institute, NIH

Abstract

Transferring large-scale genomic data compendia–now intrinsic to cancer research–to local compute facilities has become infeasible. In addition, executing computational workflows over these massive datasets can effectively utilize massively scaleable infrastructures to enhance efficiency and timeliness of results. Here we apply 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 Cloud Platform. The computational workflows were driven by the cRomwell R package (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 raw files were processed further using Apache Spark. 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.

Samples by Project

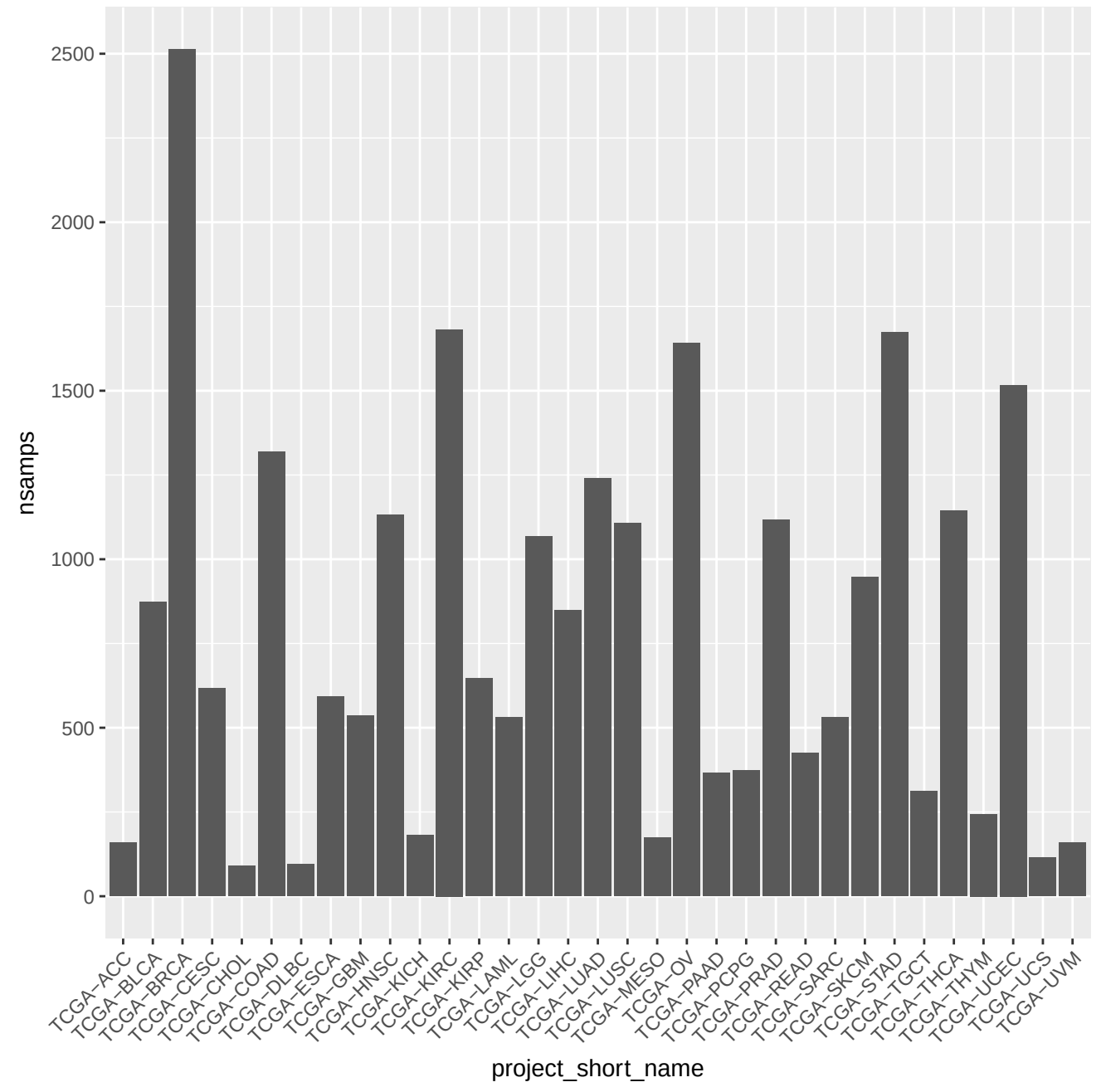


Figure 1: abc

Infrastructure and Implementation

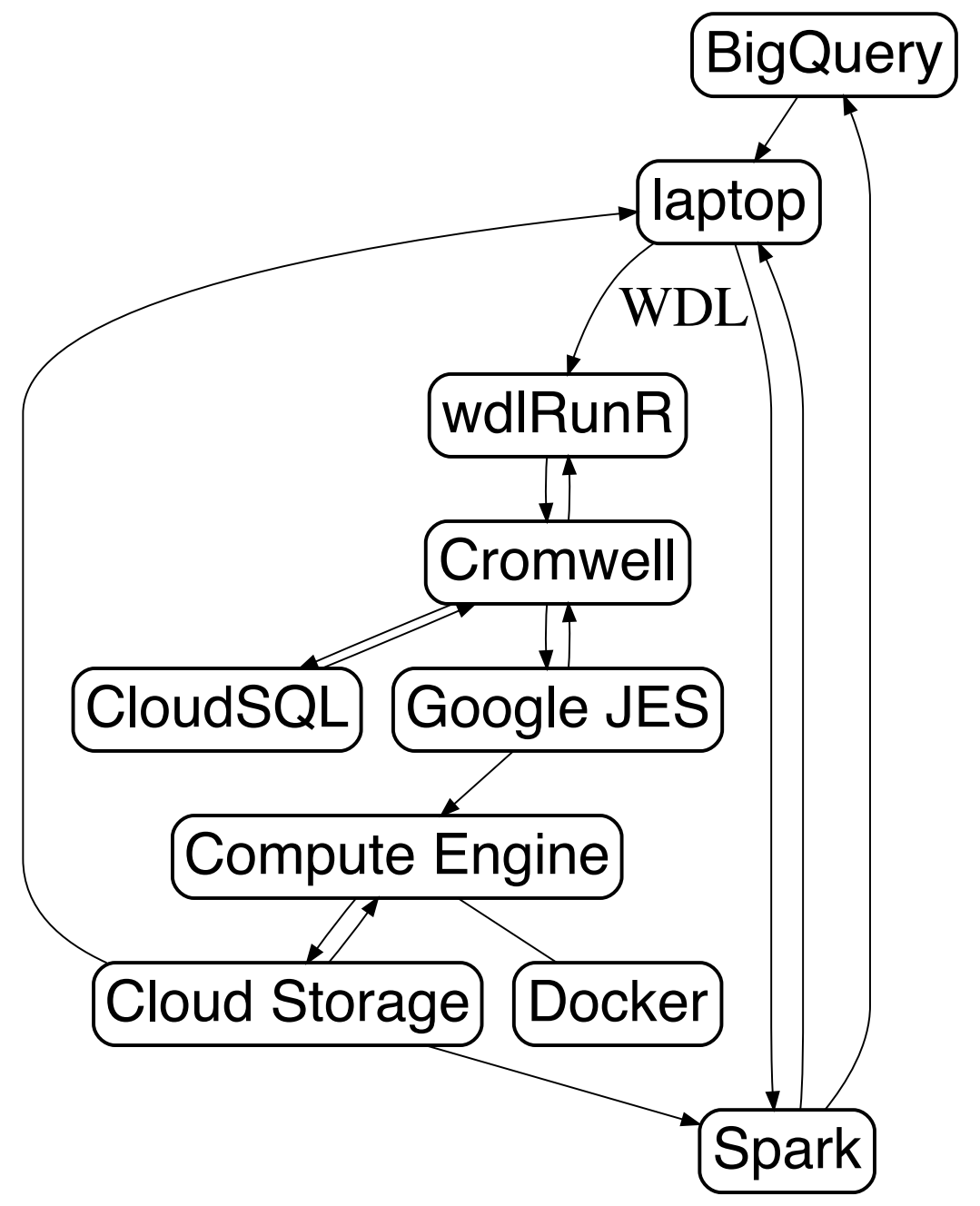


Figure 2: Data and information flow through the data ecosystem.

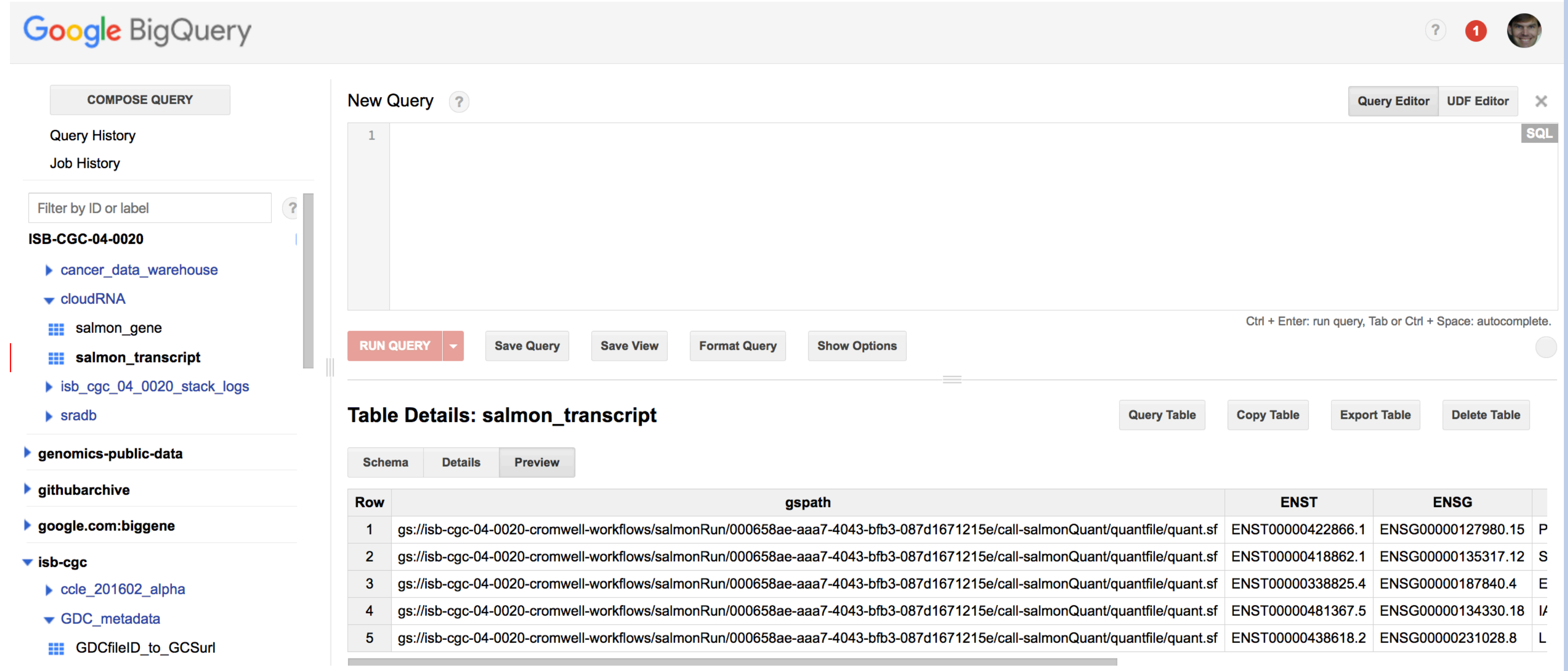
Technologies utilized

- Bigquery: massively scaleable cloud-based database as a service
- wdlRunR: R package for orchestrating cloud-scale workflows from R
- Cromwell: Server developed by the Broad Institute for managing execution of WDL workflows
- CloudSQL: Cloud-based SQL database
- Google JES: Google Job Execution Service (JES), also known as Google Genomics API
- Docker: Common container technology used to encapsulate workflow steps
- Spark: A distributed computing kernel (system) for massive-scale data analysis

Conclusion

This work represents a simple proof-of-concept of using a Data Lake to form the basis for a simple Data Commons. Technologies available as services from commercial cloud providers suffice to allow complex, highly-scaleable workflows, data engineering, and can be used to enhance data science. When combined with lightweight open source software, cloiud-scale computing can likely be leveraged by traditional bioinformaticians. Significant challenges to uptake including costs, security models, training, and policy barriers remain.

Data Warehouse of Results



Ctrl + Enter: run query, Tab or Ctrl + Space: autocomplete.

Table Details: salmon_transcript

Row	gspath	ENST	ENSG	
1	gs://isb-cgc-04-0020-cromwell-workflows/salmonRun/000658ae-aaa7-4043-bfb3-087d1671215e/call-salmonQuant/quantfile/quant.sf	ENST00000422866.1	ENSG00000127980.15	P
2	gs://isb-cgc-04-0020-cromwell-workflows/salmonRun/000658ae-aaa7-4043-bfb3-087d1671215e/call-salmonQuant/quantfile/quant.sf	ENST00000418862.1	ENSG00000135317.12	S
3	gs://isb-cgc-04-0020-cromwell-workflows/salmonRun/000658ae-aaa7-4043-bfb3-087d1671215e/call-salmonQuant/quantfile/quant.sf	ENST00000338825.4	ENSG00000187840.4	E
4	gs://isb-cgc-04-0020-cromwell-workflows/salmonRun/000658ae-aaa7-4043-bfb3-087d1671215e/call-salmonQuant/quantfile/quant.sf	ENST00000481367.5	ENSG00000134330.18	I/
5	gs://isb-cgc-04-0020-cromwell-workflows/salmonRun/000658ae-aaa7-4043-bfb3-087d1671215e/call-salmonQuant/quantfile/quant.sf	ENST00000438618.2	ENSG00000231028.8	L

Key Concepts and Definitions

Object Storage Data are stored as "objects" rather than "files". However, more important is that such "object storage" is implemented in commercial clouds as highly parallel and scales to arbitrary size.

Data Lake A virtual or physical collection of data objects, typically stored in a flat collection in their native formats (as opposed to loaded to a database).

Data Commons A shared data ecosystem consisting of *data*, *computing*, and *software and tools*.

Containers A technology that allows lightweight (fast, small) encapsulation of entire computational environments, ideal for "wrapping" tools and workflows portably.

Application Programming Interface (API) A set of protocols and tools for allowing software to interact with other software, often impelmented using web technologies.

Key Online References

- https://spark.apache.org
- https://cloud.google.com
- https://bigquery.cloud.google.com
- https://github.com/seandavi/wdlRunR
- https://github.com/broadinstitute/wdl

More information at `https://github.com/seandavi/wdlRunR` and `https://seandavi.github.io`