Multi-omics data integration in the Cloud:
Analysis of Statistically Significant Associations Between Clinical and Molecular Features in Breast Cancer

Kawther Abdilleh[1,2], Boris Aguilar[1,3], J. Ross Thomson[4]

Institute for Systems Biology-Cancer Genomics Cloud (ISB-CGC)[1], General Dynamics Information Technology[2], Institute for Systems Biology,[3] Scientific Computing, Google Cloud[4]

Abstract

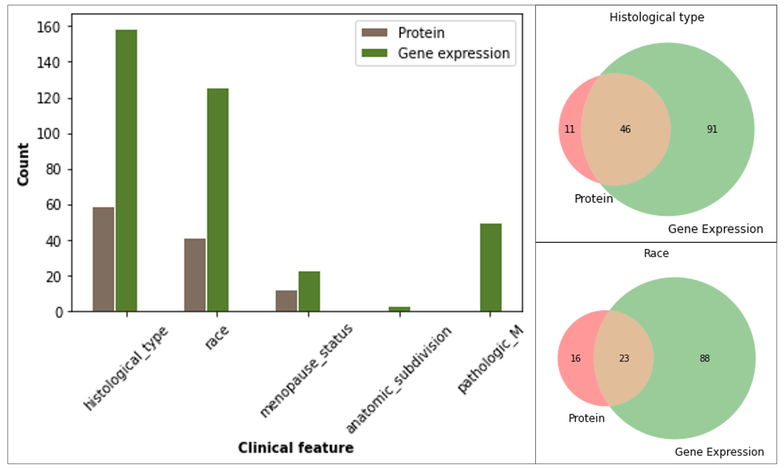
- Breast Cancers are among the most common forms of cancers impacting women with over 1 million diagnoses every year worldwide.
- They are complex cancers characterized by distinct clinical outcomes, morphological and molecular features.
- The availability of large-scale cancer data allows us to study these entities in concert to gain a more holistic picture of Breast Cancer

- This study aimed to:
  - Use a cloud-based multi-omics data analysis approach to identify novel associations between clinical outcomes and molecular features in Breast Cancer
  - Leverage available open-access National Cancer Institute (NCI) cancer data in ISB-CGC BigQuery tables to identify genes and proteins that are significantly associated with Breast cancer clinical features
  - Demonstrate how to compute and implement complex statistical methods in Google BigQuery directly on multi-omics cancer data on the Cloud

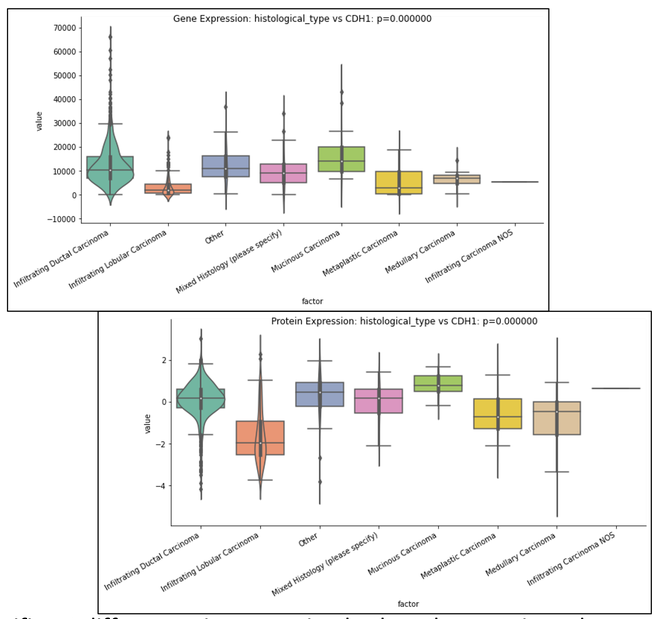
Methods

- Employed a novel multi-omics cloud-based approach to analyze statistical associations between available clinical, genomic and proteomic cancer data on the Google Cloud Platform through the ISB-CGC, one of the National Cancer Institute's (NCI) Cloud Resources.
- The Cancer Genome Atlas (TCGA) clinical, genomic, proteomic data stored in Google BigQuery tables hosted by ISB-CGC were analyzed using standard SQL.
- Google BigQuery tables, a managed service consisting of a columnar database backed by a massively parallel analytics engine. Data is stored in a highly distributed manner making it possible to split up SQL queries automatically, resulting in super-fast processing times.
- Statistical associations between TCGA categorical clinical features and protein/RNA expression were computed using the Kruskal Wallis test implemented as BigQuery user-defined functions.
- Only significant associations (p-value < 0.001) were used in the analysis.
- Data visualizations were implemented in cloud-based python notebooks that are publicly available.

Preliminary Results

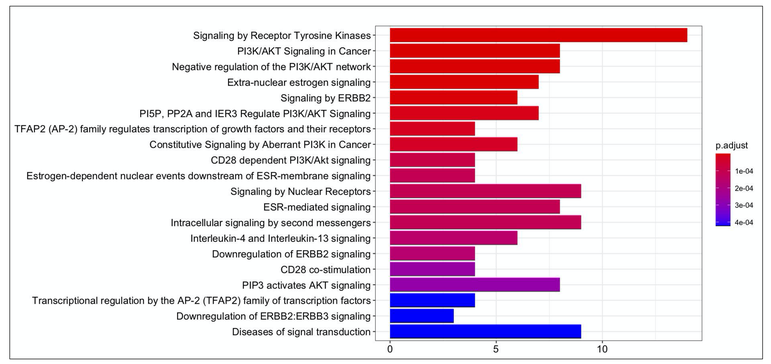


- Identified genes and proteins that are significantly associated with key clinical features of Breast Cancer.
- The clinical feature of ***histological type*** is of particular interest as it refers to the growth pattern of tumors.
- We find that this clinical feature had the most significant association with key proteins and genes. A total 46 genes and proteins that are significantly associated with the clinical feature of histological type. (Figure 1, 2a)



Significant differences in expression both at the protein and gene expression level are apparent in the ***CDH1*** gene between the two distinct Breast Cancer histological subtypes, infiltrating ductal carcinoma and infiltrating lobular carcinoma.

Preliminary Results



- The ReactomePA R package was used to identify significant pathway enrichment for genes and proteins associated with histological subtypes in Breast Cancer
- Genes and proteins involved in these pathways may be putative molecular markers to correctly classify and properly diagnose Breast Cancer

Summary

- All of the analyses presented here were conducted using open-access NCI cancer data stored in Google BigQuery tables on the ISB-CGC NCI Cloud Resource
- We demonstrate:
  - how powerful multi-omics cloud-based analytics can be used in identifying novel molecular relationships relevant for Breast Cancer
  - how to perform complex statistics using Google BigQuery to identify novel biological relationships
  - the potential big data analyses of this nature can have on the discovery of putative biomarkers in disease research
- **Notebooks** used in this work are available in: https://github.com/isb-cgc/Community-Notebooks/tree/master/FeaturedNotebooks

Acknowledgements: ISB-CGC has been funded in whole with Federal funds from the National Cancer Institute, National Institutes of Health, under contract No. HHSN261200800001E.