Large-scale cloud-based inference of differential breast cancer-related network gene hubs between patient cohorts

Kawther Abdilleh[1,3], Boris Aguilar[2,3], Ronald C. Taylor[4], George Acquaah-Mensah[5]

[1]General Dynamics Information Technology; [2]Institute for Systems Biology, [3]ISB-CGC- Cancer Gateway in the Cloud

[4]National Cancer Institute, Developmental Therapeutics Program, Rockville MD; [5]Massachusetts College of Pharmacy and Health Sciences, SOP-Worcester

Background

Breast invasive carcinoma (BrCA) remains a leading cause of mortality. Prognosis is worse for Black/African-American stage II BrCA patients 50 years old or younger (B/AA50) than for White patients of similar age and stage (W50).

Our analysis used a novel cloud-based approach, performing cohort analysis on multi-omic data sets from The Cancer Genome Atlas (TCGA) in the Google Cloud environment provided by ISB-CGC, an NCI-funded Cloud Resource. We combined miRNA expression, gene expression and somatic mutation data from TCGA breast cancer (TCGA BrCA) across multiple Google BigQuery database tables using SQL queries.

Using these correlations, we inferred notable network hubs expressed at higher levels in B/AA50 than in W50, using microRNA and gene expression correlations. Such hubs include microRNAs hsa-mir-93, hsa-mir-92a-2, hsa-mir25, hsa-mir200c, hsa-mir-519a-2 and hsa-mir-1304.

Rationale: To characterize the transcriptional regulatory differences between these two patient cohorts using multi-omics data in the cloud

Methods

- TCGA data stored in Google BigQuery tables hosted by the ISB-CGC - Cancer Gateway in the Cloud were analyzed using standard SQL.
- All data analyses compared two TCGA-BrCA cohorts (Black/African American women <=50 years of age and White women <=50 years of age).
- TCGA BigQuery tables containing information for the following data types were used for the analyses presented here: gene and microRNA expression and somatic mutation data.
- Network analysis was conducted which uses significant correlations between RNA-seq and miRseq data , which were generated using user-defined functions written in SQL.
- BigQuery tables were accessed and queried in a cloud-based R notebook using the bigRquery R package.
- Differentially expressed genes and miRNA were characterized and determined using the siggenes R package.
- Functional overrepresentation analysis was conducted using the Reactome package in R.
- Somatic mutation differences and patterns were determined using the R maftools package.

Results

Figure 1: We looked for the significant associations between all possible pairs of genes and unique miRNA identifiers available in the RNAseq and microRNA TCGA BrCA data. In particular, we used BigQuery to compute Pearson correlation between all possible pairs. The total number of computed correlations was approximately 100 million. Master regulator transcription factor HES4 is in significant ($p < 0.0001$) correlation with miR-663b, miR-548ag-1, etc. Brown miR nodes expressed more in B/AA50 than in W50.

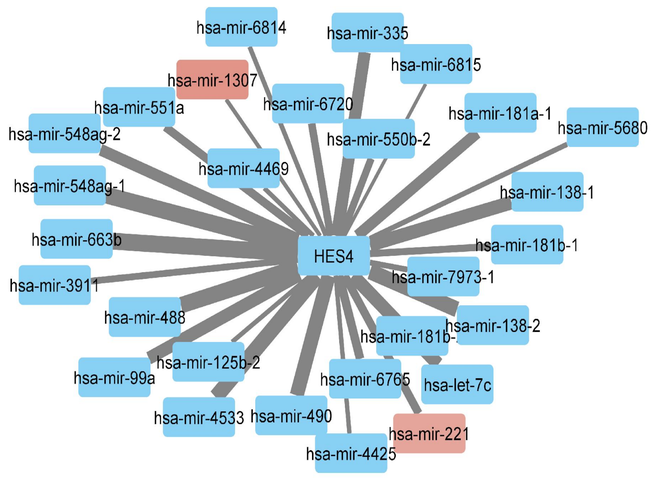
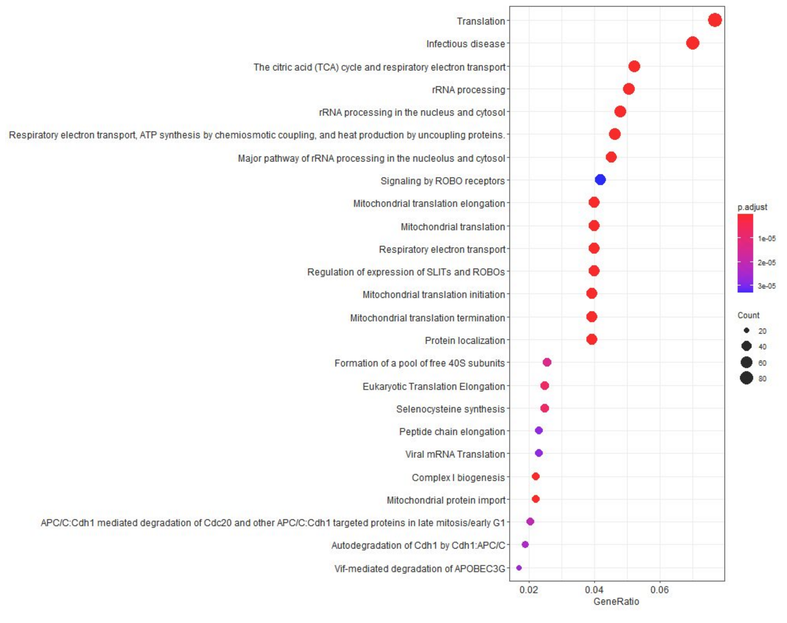
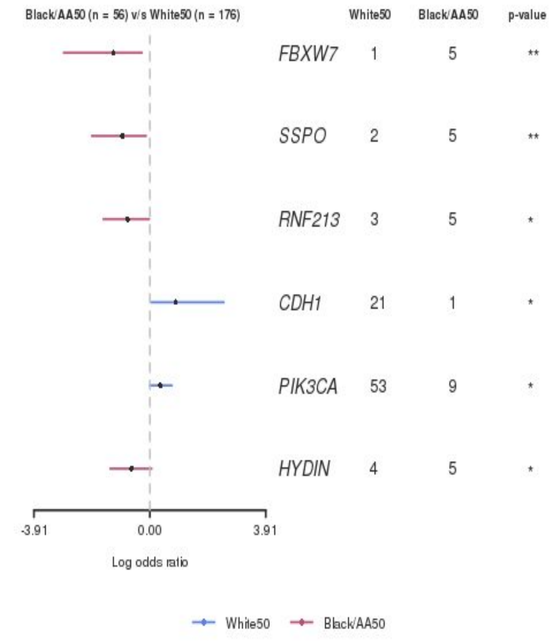


Figure 2. Processes that are over-represented among genes with elevated expression in B/AA50 relative to W50.



Results

Figure 3. Mutations found at different frequencies between B/AA50 and W50. CDH1 and PIK3CA mutations occur more frequently in W50; the other genes are more often mutated in B/AA50.



Summary

Cloud-based interrogations of a variety of BrCA and related data have tremendous potential. In this instance, we uncovered the following:

- Expressions of miR-663b, miR-548ag, and other cancer-regulating miRs are statistically tied to that of master regulator HES4.
- Among genes expressed more in B/AA50 than W50, pathways associated with proliferation and the response to cellular stress are over-represented.
- Mutations of tumor suppressor CDH1 and PIK3CA occur more frequently and co-occur in W50.

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