Multi-omics data analysis in the cloud: Inference of differential breast cancer-related network hubs between TCGA 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] Institute for Systems Biology- Cancer Genomics 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).

Goal: To characterize the differences between these two cohorts using multi-omics data, including transcriptional regulatory relationships.

Methods

- TCGA data stored in Google BigQuery tables that are hosted by The Institute for Systems Biology - Cancer Genomics Cloud (ISB-CGC) were analyzed using standard SQL.
- All data analyses compared two two 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, DNA methylation, 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.
- ISB-CGC TCGA BigQuery tables were accessed and queried in cloud-based R implementation 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

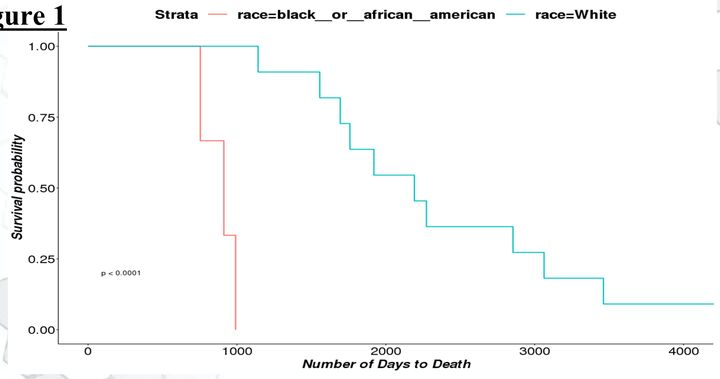

Figure 2

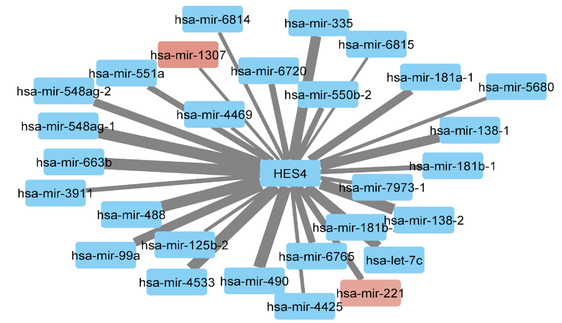

Figure 3

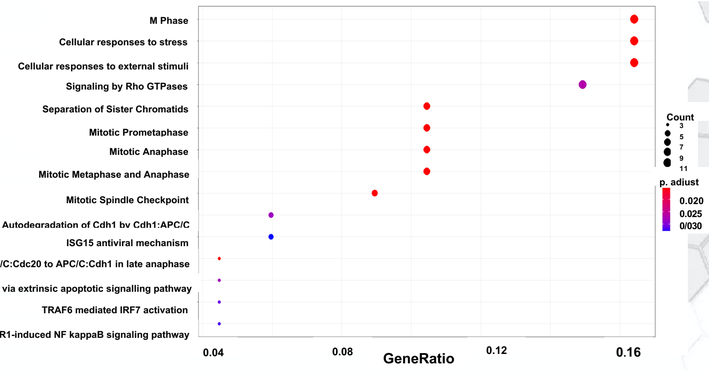

Figure 4

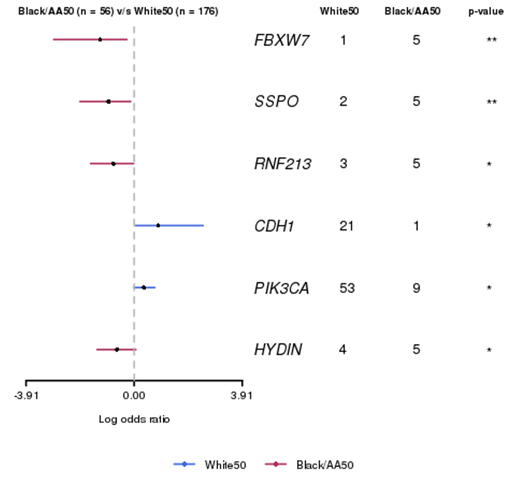

Figure 5a

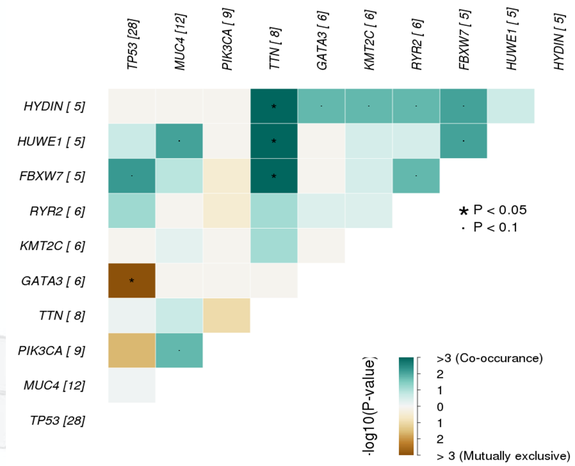

Figure 5b

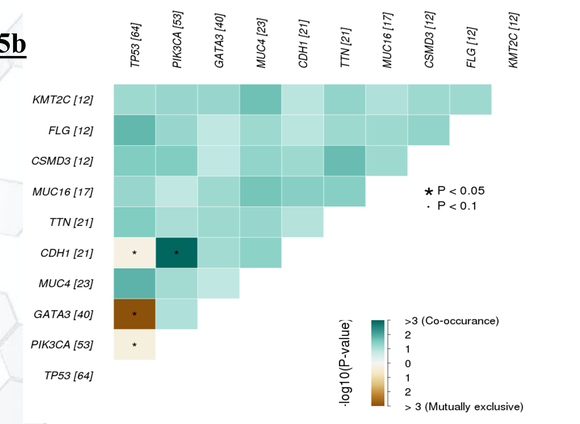

Figure 1: W50 patients (n=11), shown in blue, survived better than B/AA50 (n=3) patients, in red.
Figure 2: Master regulator transcription factor HES4 is in significant ($p < 0.0001$) correlation with miR-663b, miR-548ag-1, etc. Brown miR nodes expressed more in B/AA50 than in W50.
Figure 3: Processes over-represented among genes with elevated expression in B/AA50 relative to W50.
Figure 4: 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.
Figure 5: Co-occurrence plots of mutations found in (a) B/AA50, and (b) W50, respectively.

Summary and Conclusions

Cloud-based interrogations of a variety of BrCA and related data have tremendous potential. In this instance:

- 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.
- Mutations of HUWE1, HYDIN, and FBXW7 are more frequently mutated in B/AA50, and co-occur with mutations of TTN.