ISB-CGC

MCPHS UNIVERSITY

A tale of two cohorts:
Transcriptomics and epigenomic analysis in breast cancer

Boris Aguilar[1,2], Kawther Abdilleh[1,3], George Acquaah-Mensah[4]

Institute for Systems Biology-Cancer Genomics Cloud (ISB-CGC)[1], Institute for Systems Biology[2], General Dynamics Information Technology[3],
Massachusetts College of Pharmacy and Health Sciences, SOP-Worcester[4]

Abstract

- Breast Cancers are among the most common forms of cancers impacting women with over 1 million diagnoses every year worldwide. They are characterized by distinct clinical outcomes, morphological and molecular features.
- Breast cancers found in certain ethnic groups tend to be more aggressive than in others. According to data in the North American Association of Central Cancer Registries, relative to White women, Black women have a higher incidence rate prior to 40 years, and a lower incidence rate after 50 years.
- This study aimed to:
  - Use a cloud-based multi-omics data analysis approach to identify novel associations between molecular features in Breast Cancer
  - Find significant genomic and epigenomic differences between two cohorts: White and Black/African American participants younger than 50 years of age (W50 and BAA50).

Methods

- A novel multi-omics cloud-based approach was used to analyze genomic and methylation data on the Google Cloud Platform through the ISB-CGC, one of the National Cancer Institute's (NCI) Cloud Resources. We analyzed The Cancer Genome Atlas (TCGA) data stored in Google BigQuery tables hosted by ISB-CGC.
- Master Regulator genes were identified using the TCGA Gene Expression Compendium and the ARACNe and VIPER algorithms.
- The R package Methylmix was used to aggregate methylation beta values for each gene and to identify genes with gene expressions that are negatively correlated with methylation for the two cohorts.
- The AMARETTO package was used to identify activator and repressor driver genes in regulatory modules based on copy number variation, gene expression and DNA methylation data. We identified eight modules across both cohorts. We also applied the CommunityAMARETTO algorithm to identify driver and target genes unique to the individual cohorts.

Results

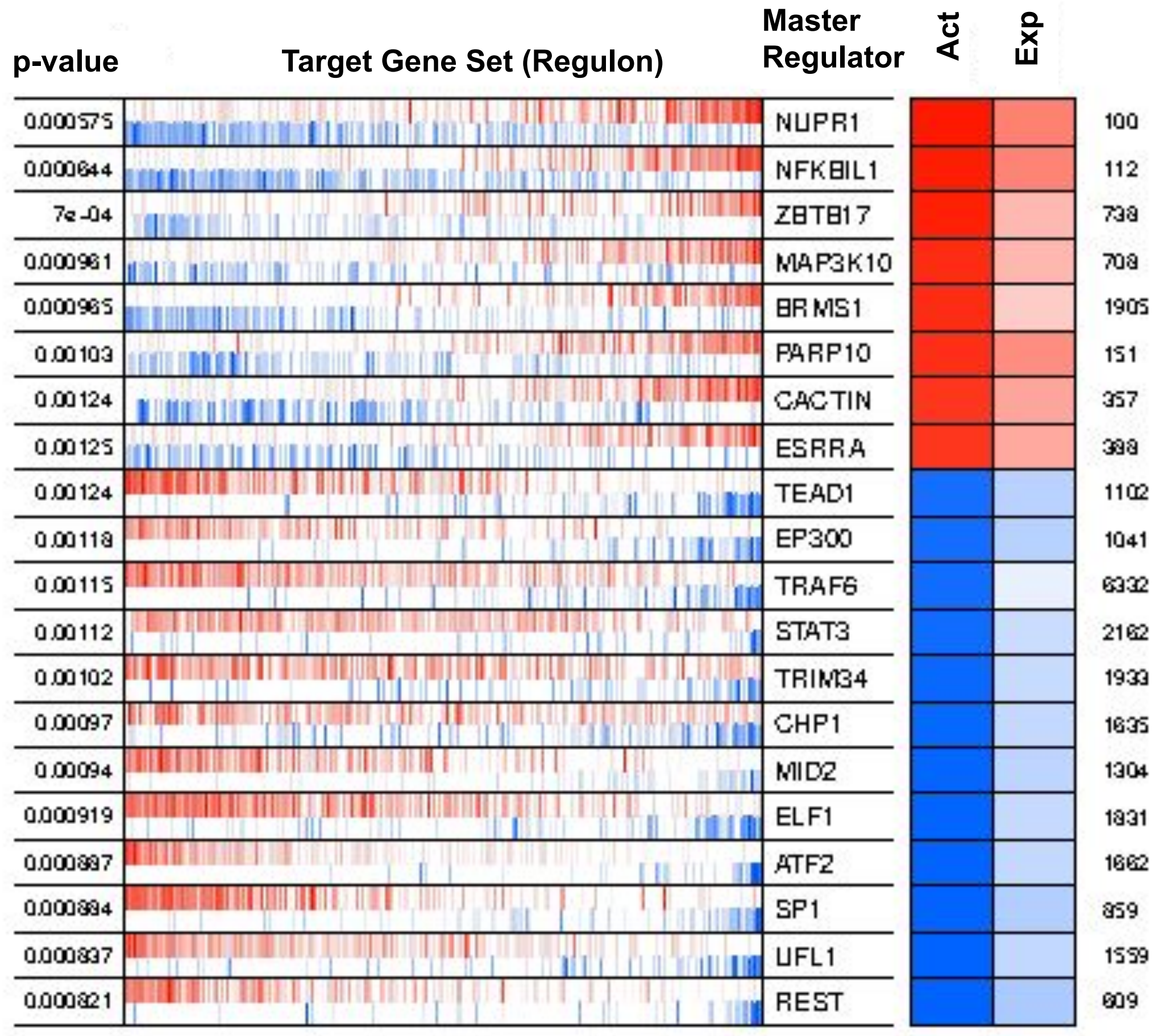


Figure 1: The top 20 differentially active transcription factors with their associated enriched regulons.

The differential activity is between W50 and BAA50. Blue columns represent gene products with suppressed expression in BAA50; red columns represent gene products with elevated expression in BAA50 relative to W50. Each regulon is identified by its regulator as noted in column "Master Regulator", expression level and the inferred protein activity of each regulator is captured under the columns "Exp" and "Act", respectively. Red hues indicate high levels; blue hues indicate low levels.

gene name	p(Fisher exact)	Correlation
OTUB1	2.14E-04	-0.3145
PRRC1	1.06E-03	-0.3081
FUZ	1.10E-03	0.0432
DNAH5	3.07E-03	-0.3768
ZNF667	3.16E-03	-0.7130
ZNF91	3.20E-03	-0.2515
EHHADH	4.39E-03	-0.4539
CAT	4.55E-03	-0.324
KLHL3	0.11E-02	-0.296

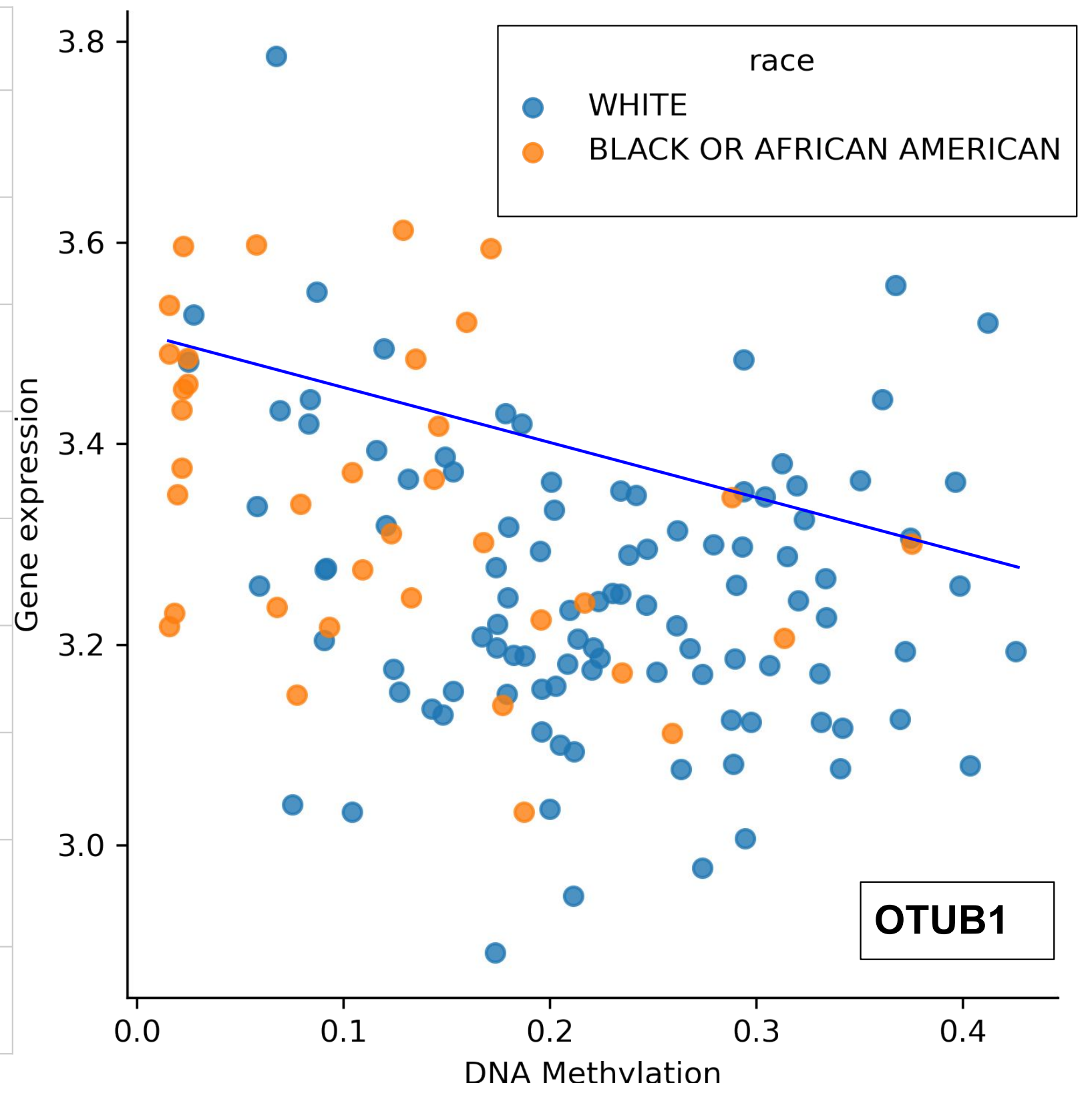
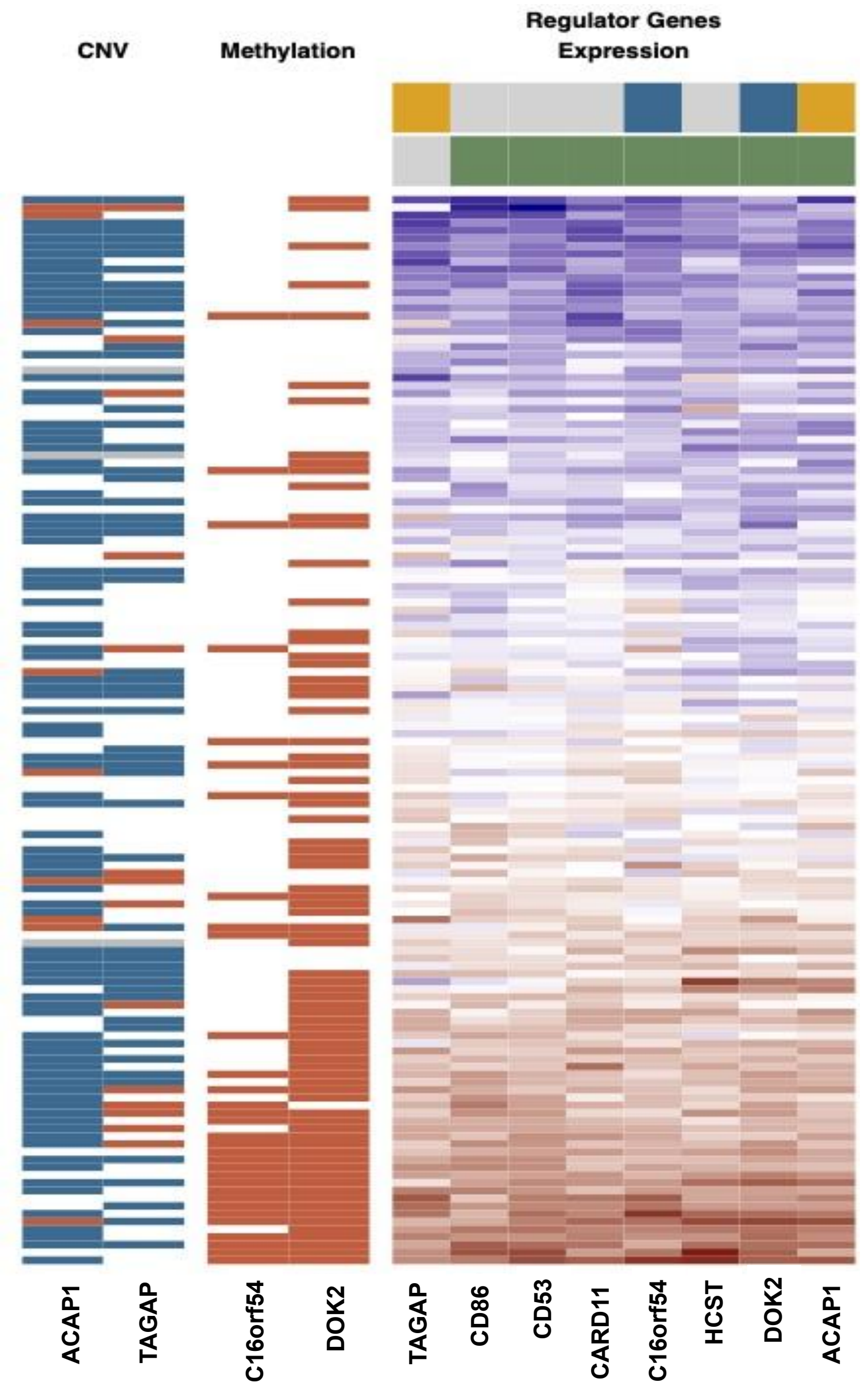


Figure 2: Left: a List of genes for which cohort (W50 and BAA50) is significantly associated (p-values < 0.1, Fisher exact test) with methylation state generated by Methylmix. The genes are also negatively correlated to methylation beta values. **Right:** scatter plot of of gene expression and methylation beta values for the gene **OTUB1** and all the participants of the cohort. The activity of the deubiquitination enzyme OTUB1 is known to have an impact on immunosuppression in breast cancer.

Results



- Identified modules of activator and repressor driver genes and their targets based on multi-omics data integration including transcriptomic, copy number and epigenomic (methylation) data.
- In addition, using the CommunityAMARETTO algorithm, we Identified unique targets governed by CNV alterations between the cohorts (figure not shown):
  - The **ABAT** gene exhibited copy number amplifications in BAA50 patients relative to the W50
  - The **APC** gene exhibits copy number deletions in W50 patients relative to BAA50

Figure 3: Inferring modules of cancer driver genes from multi-omics data AMARETTO identified driver genes after integrating copy number variation, methylation and gene expression data, as shown in the heatmap.

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 identified a set of genes that are regulated via DNA methylation and that are suppressed in WII50 (relative to BAAII50) including **OTUB1**, DHODH, RBX1, RAB11B, KIFC3, NME3, etc.
- Using a multi-omics approach, we identified modules of regulatory genes and their targets both between and across the cohorts.

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