Differential Expression Analysis of Cancer Driving Genes

Vinicius Wagner, Lewis-Sigler Institute for Integrative Genomics, vwagner@princeton.edu

SUMMER RESEARCH COLLOQUIUM
PRINCETON 2019

Background

Metabolism has long been known to play an active role in the development of many human cancers. Novel computational techniques have been developed to analyze the large-scale cancer data sets gathered by the The Cancer Genome Atlas (TCGA), the primary source of data for our project. One such technique, differential gene expression analysis (DGE), statistically links experimental conditions with changes in gene expression. Studies using this technique have been informative in understanding cancer, but to our knowledge, little work in linking metabolic relationships has been done.

Research Questions

We are interested in using DGE to link observed cancer somatic mutations with changes in gene expression. This project seeks to demonstrate the capabilities of DGE as a method to supplement experimental work as well as reveal potentially untapped or overlooked avenues of cancer genomics research.

Methods and Materials

Differential Expression:

- Normalized RNASeq count data from TCGA.
- 33 Cancer types, primary focus on low-grade glioma(LGG), breast cancer(BRCA), and prostate adenocarcinoma(PRAD).

Data Visualization:

- EdgeR bioinformatics package.
- Other packages have been developed by lab.

Limitations:

- Copy Number Variations
- Tissue Matched Samples

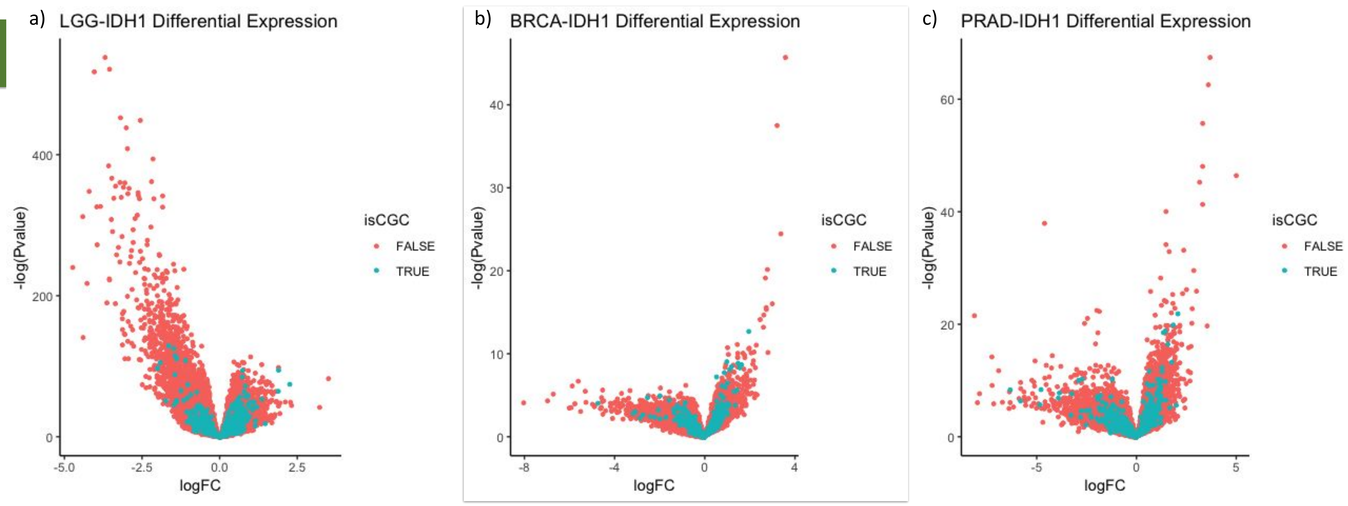

Figure 1: The log Fold Change(FC) describes the ratio between the counts of patients with and without mutations in IDH1. Shown are the results from 3 cancer types and highlighted are genes within the Cancer Genome Consensus(CGC).

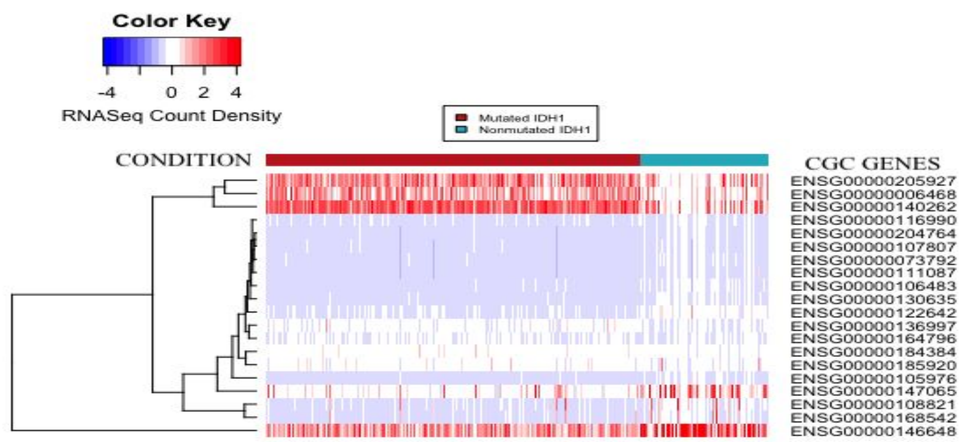

Figure 2: The LGG dataset is split between individuals affected with a mutation in IDH1. The density (in relative units) of their respective counts is shown, along with the corresponding CGC gene.

Results

We successfully developed this pipeline to analyze the links from metabolic enzymes to expression in CGC genes.

LGG: The most differentially overexpressed genes(Figure 1a, Figure 2) were found to be OLIG2, ETV1, and TCG12. *All* of these genes are highly linked to the development of LGG as shown in The Human Protein Atlas and relevant research[1][2].

BRCA: The most differentially overexpressed genes(Figure 1b) were found to be NRIP3 and SUSD2, the overexpression of which is linked to BRCA development[3].

PRAD: APLP1 and ADH1C are the most differentially expressed genes gathered from the TCGA dataset in PRAD. Research shows that the APLP family of genes is linked to the dvelopment a variety of cancer types[4].

Conclusion

Conclusion/Discussion:

- Supported by previous research and data on THPA, DGE is shown to be a powerful tool in accurately highlighting over- and underexpressed CGC genes.

Future Directions:

- Investigate mutations in other metabolic enzymes.
- DGE on data sets filtered by enzyme activation sites.
- Consider covariates in data analysis such as CNV and for interpatient heterogeneity.

Acknowledgments

Special thanks to the Office of Undergraduate Research (OUR) for funding this project, the ReMatch+ program and its advisors as well as Antonio Muscarella for hostinng the project and guiding me through the Summer.

References

[1]: Uhlen M et al, 2017. **A pathology atlas of the human cancer transcriptome.** *Science.* PubMed: 28818916 DOI: 10.1126/science.aan2507

[2]: Lenny Dang, David W. White, et. al., **Cancer-associated IDH1 mutations produce 2-hydroxyglutarate.** Nature volume 462, pages 739[744 (10 December 2009)

[3]: Atsushi Terunuma, et. al, **MYC-driven accumulation of 2-hydroxyglutarate is associated with breast cancer prognosis.** The Journal of Clinical Investigation, 124(1):398-412, doi: 10.1172/JCI71180, 2014

[4]: Poomy Pandey, Bailee Silker, et. al, **Amyloid precursor protein and amyloid precursor-like protein 2 in cancer.** Oncotarget. 2016 Apr 12; 7(15): 19430–19444.