Differential Expression in Cancer Driving Genes by Mutations in Metabolic Enzymes

Vinicius Wagner, Antonio Muscarella

Princeton University
ReMatch+ Summer 2019
Singh Lab

August 5, 2019

Abstract

The study of metabolism in cancer has gained traction for its potential in the field of cancer genomics to provide new diagnostic treatments and a better understanding of the ramifications of metabolic mutations in cancer development. Using differential gene expression analysis(DGE), a statistical technique often utilized in the realm of bioinformatics, we observe relationships between the RNA counts of patients affected with certain metabolic enzyme mutations and the expression of Cancer Geome Consesus(CGC) genes. Our preliminary results show the usefulness of this method, which is supported by previous research. The method is replicated across three different cancer types with mutations in the metabolic enzyme isocitrate dehydrogenase 1(IDH1) being the target of analysis. This will be useful in better understanding cancer development in the larger picture of personalized medicine in cancer treatment.

1 Introduction

In 2018 there were over 1.7 million new cancer cases in the United States, resulting in an estimated $147.3 billion in medical care expenditures[1]. With the number of cases projected to rise in the coming years, the need for new approaches to studying cancer is more pressing than ever. There is a growing attention towards the role that metabolic pathway mutations play in the development of different cancer types. Mutations in metabolism have been known to be a hallmark of cancer development and identifying these relationships has been a target for research in the past decades.

For instance, it was shown that mutations by cancer in the metabolic enzyme IDH1 are directly linked to the catalysis of α-ketoglutarate into the oncometabolite 2-hydroxyglutarate(2HG)[4]. The abundance of the latter product in patients is associated with a higher risk of malignant brain tumor development. This work paved the way for continued research into the ramifications of IDH mutations in the development of cancer[5][6]. This metabolic profiling of cancer cells was also performed in breast tumors, where 2HG was associated with MYC-pathway activation in breast cancer(BRCA)[7].

The prevalence of metabolic alterations in different types of cancers has already been demonstrated; we are interested in looking into additional metabolic relationships. Thus the pursuit of potentially new relationships between metabolites and cancer-driving genes shows promise and is the guiding motivation behind our work.

2 Methodology

2.1 Data Sources

Our research is primarily computational in nature and involves the analysis of vast amounts of patient data within The Cancer Genome Atlas(TCGA) provided by the National Institute of Health. We first subset this data set into a specific cancer type and retrieve its associated mutation annotation format(MAF) file, which contains all of the mutations found within the group of patients, among other data. Currently we are working with the data associated with LGG, BRCA, and prostate adenocarcinoma(PRAD) with different metabolic enzymes. We then divide these data sets into two populations, one containing a mutation in a chosen metabolic gene (e.g. IDH1) and the other not containing such mutation. The count data for each patient is then retrieved at which time a differential expression analysis is performed.

2.2 Differential Expression Analysis

Differential analysis of gene expression data has been shown to be effective at modeling the roles that certain mutated genes play in the development of cancer. In the realm of bioinformatics, differential expression analysis involves statistical analyses performed upon normalized count data[2]. This count data is usually in the form of RNA transcript counts that identify which genomic loci are expressed in an individual patient[3]. In our research, we use this technique to provide insight into the effects that mutated metabolic genes have upon the over and under expression of certain Cancer Genome Census(CGC) genes.

We can visualize the obtained expression data in a variety of ways, most notably with the help of the R-package edgeR[3] which is primarily used for data analysis in bioinformatics. One such preliminary visualization we obtained is shown in Figure 1, where the Fold Change(FC) describes the ratio in expression level in a certain gene between a healthy and an affected population. From those results we can single out the most positively expressed gene, sushi domain containing 2(SUSD2). There has been work surrounding this gene that highlight its role in tumorigenesis in patients with BRCA[8].

3 Results

A multitude of studies show that the most overexpressed CGC genes genes within the three observed cancer types were linked to the development and progression of cancer. Our selection of the following three cancer types is based on their documented effects on metabolic enzyme mutations and cancer-driving gene expression. The results of our DGE across the three cancer types are shown in Figure 1.

3.1 LGG Results

Low-grade glioma is a type of slow-growing tumor which develops from astrocytes and oligodendrocytes, both found in the human brain. The result from a clustering of patient data in mutated vs. nonmutated IDH1 is shown in Figure 2.

The most differentially overexpressed genes 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[10].

3.2 BRCA Results

Although mutations in BRCA1 and BRCA2 are more commonly linked to the development of breast cancer, there is evidence that metabolic alterations can also play a role in this cancer type[11].

The most differentially overexpressed genes were found to be NRIP3 and SUSD2. The overexpression of both of these genes has been linked to BRCA development[7][10].

3.3 PRAD Results

The metabolic mutation landscape of PRAD is less known than LGG and BRCA by comparison. However, because of the nature of prostatic cells to accumulate zinc and the subsequent inhibitions within the citric acid cycle(a metabolic pathway), a growing interest has surrounded the role of metabolism in PRAD[12].

APLP1 and ADH1C are the most differentially expressed genes gathered from the TCGA dataset in PRAD. Research shows that the APLP family of genes may be linked to the development of a variety of cancer types[13].

4 Discussion

In this work, we successfully developed a pipeline to link mutations in metabolic enzymes to changes in expression in CGC genes. Supported by previous research and data on The Human Protein Atlas, we show that this type of DGE is a powerful tool in accurately highlighting over- and under-expressed CGC genes. Many of our preliminary results highlight relationships between metabolic enzymes and cancer recently uncovered in experimental work.

At this stage in our work we did not take into consideration some important confounding factors that could potentially have adverse effects on our results: Copy Number Variation (CNV) and biological variation among patients. CNV is a phenomenon where the number of repeats in some loci of the genome can vary among individuals in a population. High CNV status of any given gene is correlated with higher expression of that gene, which may confound DGE results. In order to account for this, we will utilize the CNV data provided by the TCGA and incorporate it as a covariate in our statistical model.

Biological variations between patients are important considerations since they are independent of cancer sub-type or metabolic mutations. In certain cancer types, the TCGA has data pertaining to control(non-cancerous) tissue in order to account for variations that do not arise from the cancer itself. In future work, we must include these additional covariates to refine our results.

Additionally, we aim to investigate the relationships between certain functional sites within mutated proteins (as opposed to the mutation as a whole) and the expression of CGC genes. The Singh Lab has developed software tools to perform such analyses, such as CanBind[14]. Going forward we can incorporate CanBind into our results to investigate the impact that mutations in specific sites in a metabolic enzyme can have upon the expression of cancer-driving genes.

References

[1] Mariotto AB, Yabroff KR, Shao Y, Feuer EJ, Brown ML. *Projections of the cost of cancer care in the United States: 2010–2020.*
J Natl Cancer Inst 2011;103(2): 117–28

[2] Friederike Dündar, Luce Skrabanek, Paul Zumbo *Introduction to differential gene expression analysis using RNA-seq.*
Applied Bioinformatics Core — Weill Cornell Medical College, 2015-2018

[3] Mark D. Robinson, Davis J. McCarthy2, and Gordon K. Smyth *edgeR: a Bioconductor package for differential expression analysis of digital gene*

expression data.
BIOINFORMATICS,
doi:10.1093/bioinformatics/btp616, 2010

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

[5] Wei Xu, Hui Yang, et. al *Oncometabolite 2-Hydroxyglutarate Is a Competitive Inhibitor of α-Ketoglutarate-Dependent Dioxygenases.*
Cancer Cell, Volume 19, Issue 1, Pages 17-30, 2011

[6] Chao Lu, Patrick S. Ward, et. al, *IDH mutation impairs histone demethylation and results in a block to cell differentiation.*
Nature, 2012 Feb 15;483(7390):474-8. doi: 10.1038/nature10860.

[7] 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

[8] Watson AP1, Evans RL, Egland KA. *Multiple functions of sushi domain containing 2 (SUSD2) in breast tumorigenesis..*
Mol Cancer Res. 2013 Jan;11(1):74-85. doi: 10.1158/1541-7786.MCR-12-0501-T. Epub 2012 Nov 6.

[9] Hur H1, Lee JY, Yun HJ, Park BW, Kim MH. *Analysis of HOX gene expression patterns in human breast cancer.*
Mol Biotechnol. 2014 Jan;56(1):64-71. doi: 10.1007/s12033-013-9682-4.

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

[11] Annapoorna Sreedhar, Yunfeng Zhao *Dysregulated metabolic enzymes and metabolic reprogramming in cancer cells*
Biomed Rep. 2018 Jan; 8(1): 3–10. doi: 10.3892/br.2017.1022

[12] Eric Eidelman, Jeffrey Twum-Ampofo, Jamal Ansari, and Mohummad Minhaj Siddiqui textit-The Metabolic Phenotype of Prostate Cancer
Front Oncol. 2017; 7: 131. doi: 10.3389/fonc.2017.00131

[13] 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.

[14] Dario Ghersi, Mona Singh *Interaction-based discovery of functionally important genes in cancers*
Nucleic Acids Research, Volume 42, Issue 3, 1 February 2014, Page e18, https://doi.org/10.1093/nar/gkt1305

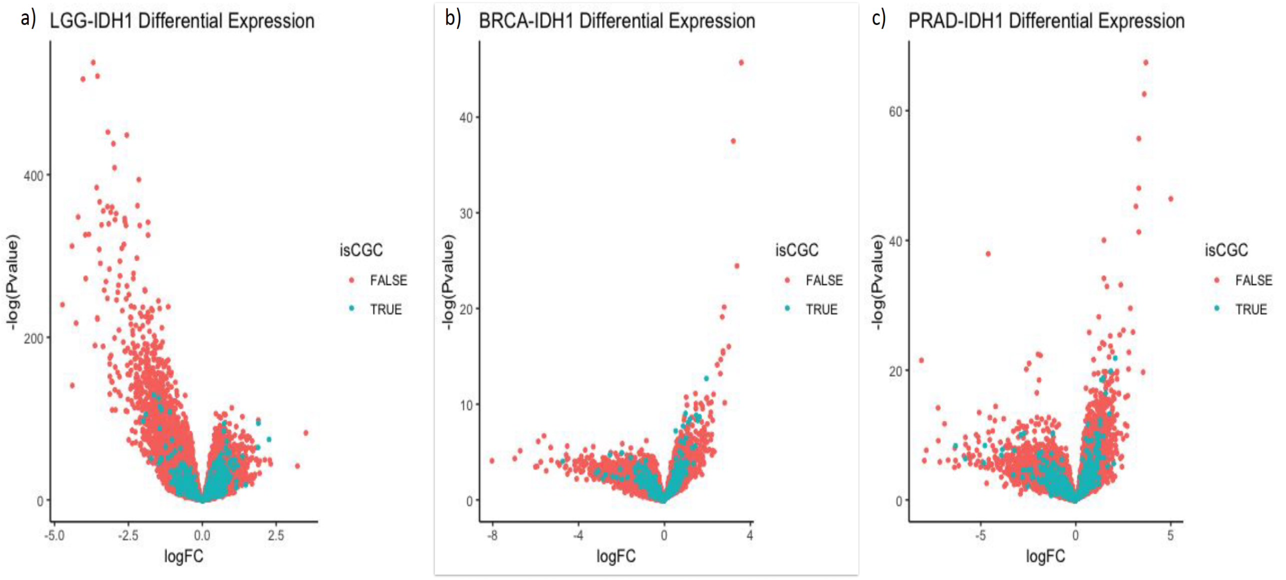

Figure 1: Shown are the differential gene expression data from the three analyzed cancer types. The data are represented as the log Fold Change (x-axis) by the negative-log p-values(y-axis). The data set is color split between the populations containing a mutation in a CGC gene(blue) and not containing such mutation(red) .

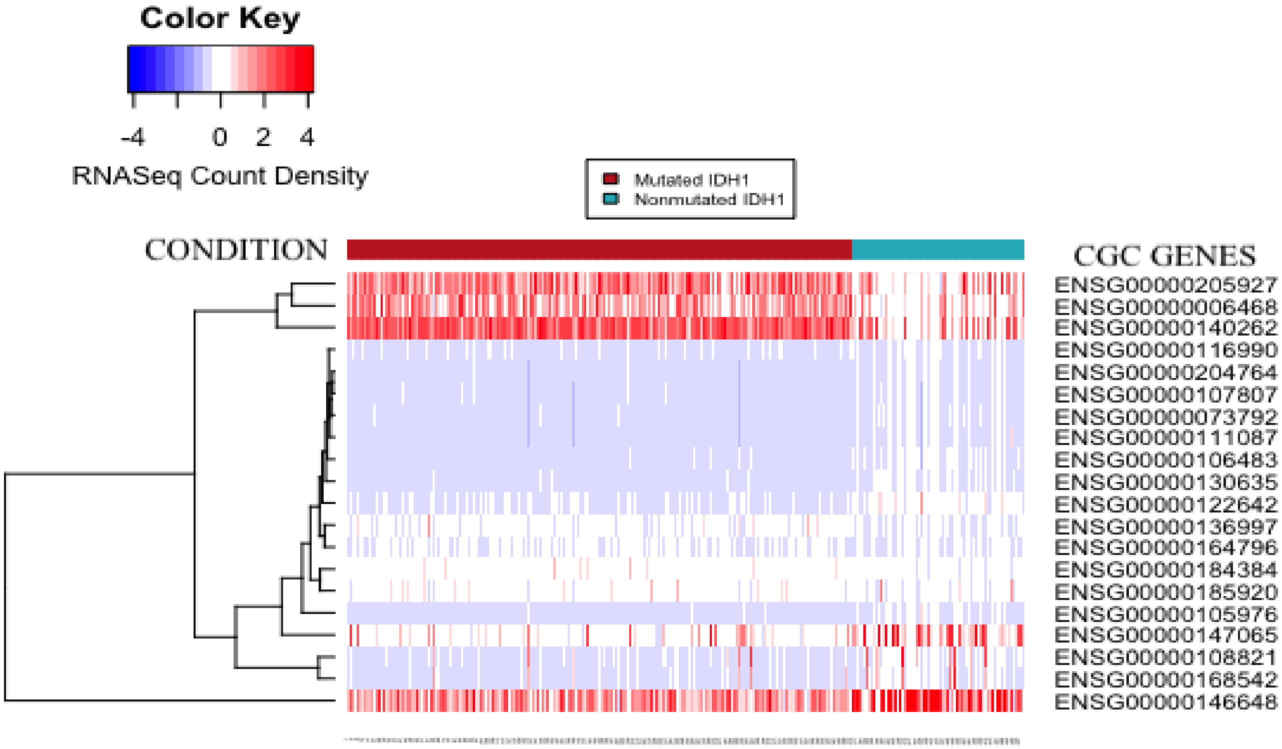

Figure 2: Shown is a clustering of CGC genes from the LGG patient dataset, split among patients who show a mutation in IDH1(red) and do not present the mutation(blue). The heatmap coloring corresponds to the RNASeq count density(in relative units).