Targeting Cancer via Signaling Pathways: A Novel Approach to the Discovery of Gene CCDC191's Double-agent Function using Differential Gene Expression, Heat Map Analysis through AI Deep Learning, and Mathematical Modeling

Annie Ostojic
Society for Science and the Public – Regeneron STS Finalist, United States of America

Abstract: Over the past couple of years, my family has experienced multiple encounters with cancer, some unfortunately, ending in death. This inspired me to tackle cancer using an analytical, computational biology approach. A need exists to study gene functions in pathways to meet a changing medical industry of personalized medicine and cancer treatments relative to gene expression patterns. In my computational biology and bioinformatics research project, I utilized a public access database to study gene CCDC191. Scientists do not know much about this gene other than its location and coiled-coil 3D structure; its function is unknown. Through an artificial intelligence deep learning model and reverse engineering mathematical modeling, I created a general process for determining the function of genes with unknown functions. Specifically, I found that the gene CCDC191 is a double agent which means it can lead to either the loss of controlled cell death or uncontrolled cell growth. Both can contribute to cancer development depending upon the type of cancer or situation. This study presents new insights into gene CCDC191, and it provides a replicable methodology which incorporates AI deep learning image classification and reverse engineering mathematical modeling to determine gene functions in pathways and cancer connectedness.

1 Introduction

Knowledge of gene functions and interactions are necessary to meet a changing medical industry of personalized medicine. “Some genes function through a dozen signaling pathways that regulate three core cellular processes: cell fate determination, cell survival, and genome maintenance” [4]. Over 21,000 protein-coding genes exist in the genome with many that remain unstudied in their functions. Determination of gene functions as tumor suppressors or gain of function oncogenes is vital to their correlations to pathways in development of various cancers. Cancer growth is described as “an increase in the ratio of cell birth to cell death causing a selective growth advantage to the cell in which it resides” [4]. Research has determined that some genes are “double agents” with both tumor suppressor and oncogenic functions depending upon the cancer they are associated with [3]. In addition, elevated signaling of Phosphoinositide 3-kinase (Pi3K) is “considered a hallmark of cancer” [2].

“New methods are needed in four areas to realize the potential of personalized medicine: (i) processing large-scale robust genomic data; (ii) interpreting the functional effect and the impact of genomic variation; (iii) integrating systems data to relate complex genetic interactions with phenotypes; and (iv) translating these discoveries into medical practice” [1]. Consequently, this research addresses (i) and (ii) by examining gene CCDC191’s (KIAA1407’s) function in 33 cancer types from The Cancer Genome Atlas (TCGA). From these results, the study analyzes gene expression of CCDC191 and its impact correlations. The research narrows its focus to human breast cancer in one part of the Pi3K signaling pathway. As a gene with an unstudied function, the study of CCDC191 plays an important role in providing molecular biologists with information for further gene analyses as well as personalized medicine with targeted cancer treatment studies. This type of in-silico study saves research funds and time by isolating a pathway function of gene CCDC191 for further investigation in a wet lab setting. To improve accuracy in gene heat map analyses and to expedite the pathway analysis process, this in-silico study created a replicable and novel AI deep learning model. In addition, gene interactions were analyzed further with mathematical modeling.

2 Materials and Methods

Materials:

This research project was fully completed at home with the utilization of a laptop computer, desktop computer, and server (Fig. 1). There was no university support for this research project. The databases accessed were all free, open-access, public databases (such as Reactome and The Cancer Genome Atlas – TCGA). The server was installed at home to process, manipulate, and work with the databases.

Methods:

A computational and bioinformatics study of gene CCDC191 was conducted to determine its functionality in pathway interactions, tumor suppressor or oncogenic relationships, and associated diseases/cancers. This "in silico" research started with a general approach and focused in on a more in-depth analysis of CCDC191's function. This study analyzes CCDC191's gene expression in all cancers and patient survival analyses and narrows the focus to correlations of enriched gene expressions of CCDC191 and other genes in the constitutive signaling by aberrant Pi3K pathway for breast cancer. Heat map analyses include an artificial intelligence deep learning component and reverse engineering mathematical models to examine gene correlations.

Fig. 1: desktop computer with installed floor server

Project procedures included:

1. Utilization of the PubMed database for research papers
2. Collation of CCDC191 information regarding structure, cell location, etc. from Ensembl, GeneCards, UniProt, ModBase, and The Human Protein Atlas
3. Utilization of The Cancer Genome Atlas (TCGA) to analyze:
    a. CCDC191's Gene Expression in 33 Cancer types
    b. CCDC191's Gene Expression in Cancer Subgroups/Progression
4. Analysis of patient survival defined by gene expression levels of CCDC191 using ggplot
5. Analysis of differential gene expression for select cancer groups
6. Utilization of Reactome, R Studio, and KEGG for pathway analyses
7. Creation of heat map to visualize enriched gene expression patterns in selected pathway
    a. Correlation between heat map data and actual pathway
8. Building, training, testing, & deploying AI deep learning image classification of heat map
9. Application of reverse engineering mathematical models to heat map results for other gene correlations with CCDC191, Mathematical Models include:
    a. Linear Additive Model for a 4-node gene network
    b. Boolean Function Model
    c. Boolean Function Binary Truth Tables for a Target Gene

3 Results

TCGA Cancer Database Analyses (ggplot and R) and Results

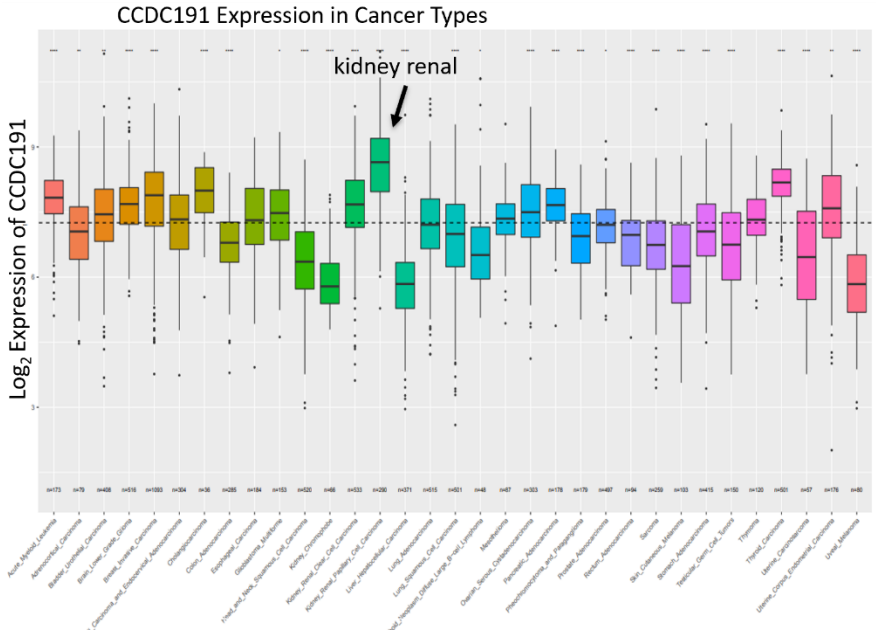
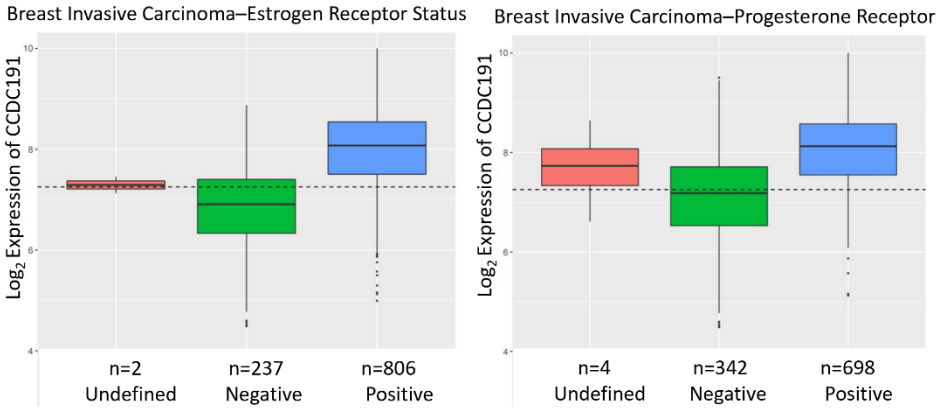


Figure 2 (left upper) depicts CCDC191 gene expression in 33 cancers. The center dotted line across the graph marks average gene expressions across all 33 cancers. The results in this analysis were inconclusive for a correlation with CCDC191 due to the averaging of all cancers. The analysis was then run to check gene CCDC191 correlations with various cancer stages and receptors in individual cancers types (Figs. 3&4).



Figures 3 & 4 (left lower pictures) show a **higher** expression of CCDC191 in Estrogen Receptor Positive (ER+) tumors of breast cancer as well as a **higher** expression of CCDC191 in Progesterone Receptor Positive (PR+) tumors of breast cancer.

Patient Survival Analysis (ggplot) and Results

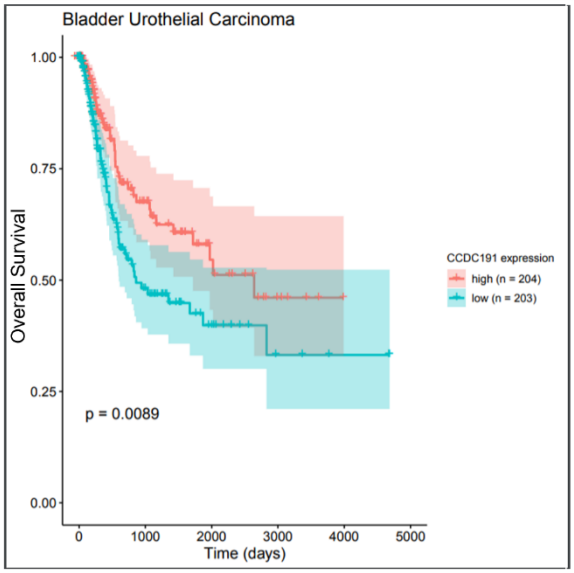


Figure 5 (left) depicts the patient survival analysis of CCDC191 in bladder urothelial carcinoma patients. The hypothesis was that patients with low CCDC191 expression would have a different prognosis than patients with high CCDC191 expression. The results from this graphical analysis revealed that:

1. Bladder urothelial carcinoma patients with high CCDC191 expression **have significantly better prognosis** ($p \leq 0.05$).
2. Expression level of CCDC191 can be useful in projecting survival of bladder urothelial cancer patients.
3. Since results of a survival analysis do not provide causative information, it **can not** be concluded that CCDC191's expression is causative of bladder cancer. Additional testing beyond a survival analysis is necessary to examine a possible causative relationship.

Heatmap: Pathway Enrichment Analysis (Reactome) for BRCA ER+/PR+ and Results

Figure 6 (left) shows results from the pathway enrichment heatmap analysis for CCDC191 expression of breast cancer ER+ and PR+ samples.

Results for CCDC191:

ERBB4 highly correlated

FGFR3 and FOXO4 possibly correlated

AI Deep Learning Image Classification of Heat Map – Analysis and Results

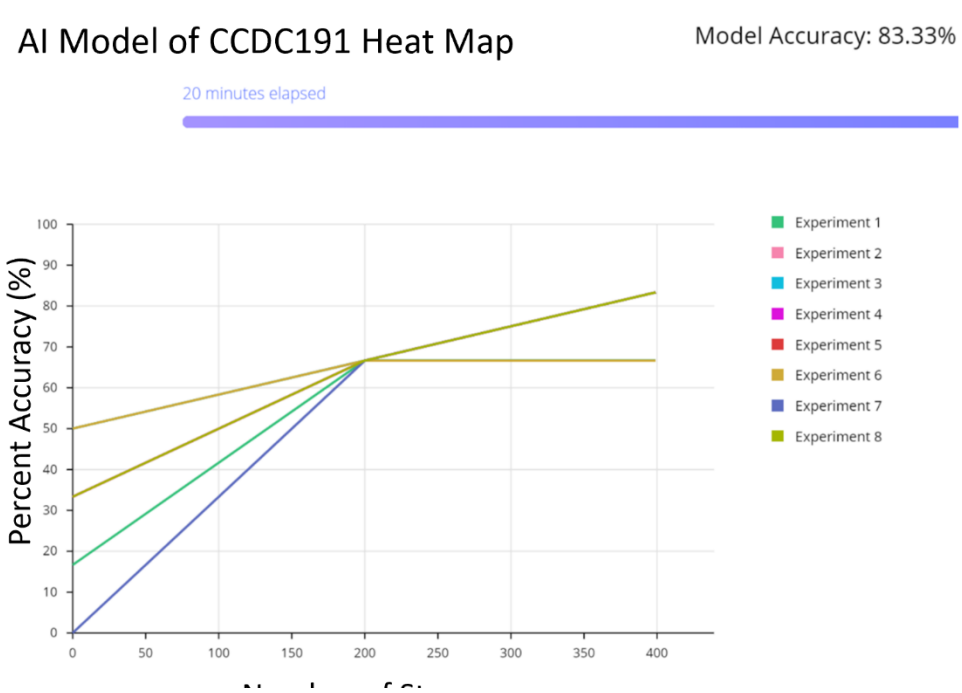


Figure 6 (above): AI Deep Learning Model of CCDC191 Heat Map after Normalizing Images Pixel-to-pixel. Used Nanonets and Python to create image classification AI model

- Model was built, trained, tested, and deployed. Used 75 heat map images to build model and 25 *different* images to test model.
- Three layers in model: 1. input, 2.Nanonets, 3. output
- Model **83.33%** accuracy
- Analyzed Receiver Operating Characteristics (ROC) graphs (not shown) to determine model's specific category accuracy
- Genes which are "not targeted or correlated" as well as "possibly correlated" genes are accurately distinguished 100% of time
- Model shows some difficulty determining "target gene"

Applying Reverse Engineering Mathematical Models to Heat Map Results for other gene correlations with

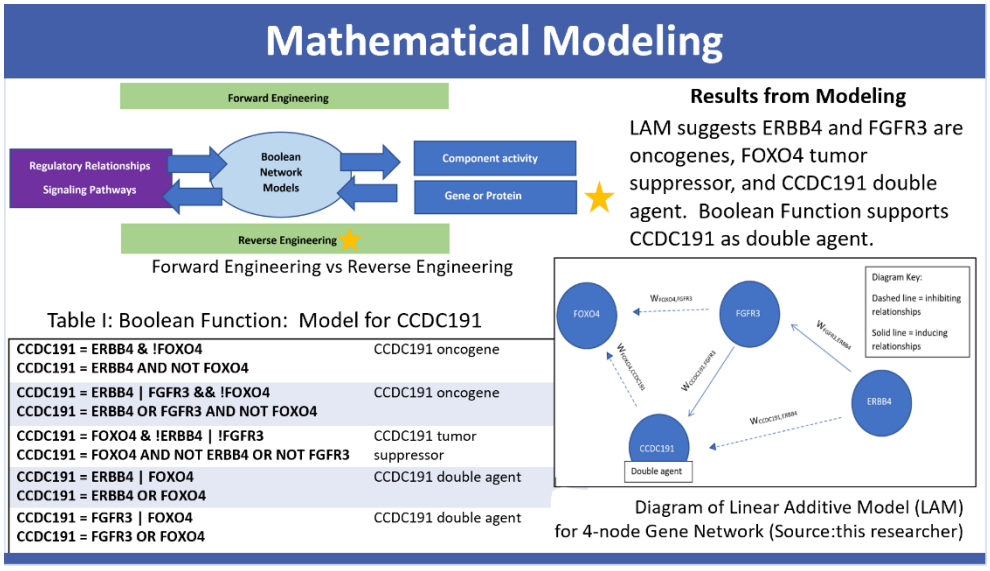


Table I and Figures on left support CCDC191's double agent functions as an oncogene or tumor suppressor when expressed with ERBB4, FGFR3. and FOXO4

4 Discussion

This study addresses a need for rapid, reliable gene function identification with improved methodology. It specifically identifies the function of unknown gene CCDC191 as a double agent tumor suppressor and oncogene found through reverse engineered mathematical modeling from a pathway enrichment heatmap. The pathway in this study, the Constitutive Signaling by Aberrant Pi3K Pathway, utilized data from breast cancer patients. Through three reverse engineering mathematical models (Linear Additive Model, Boolean function, Boolean truth table), a correlation was shown to exist between CCDC191 and another pathway gene, ERBB4. The heatmap had been generated through a narrowing series of bioinformatic analyses of TCGA data. Artificial intelligence deep learning was incorporated into the heat map analysis to enhance accuracy and pixel-to-pixel depth in the investigation. Combining AI with bioinformatic heat map analyses supported the project's goal for faster, more reliable results.

This project's overall methodology to narrow and funnel its large database research toward a focused pathway-gene correlation including AI deep learning and mathematical modeling of the gene enrichment pathway is novel in its approach. Such bioinformatic studies are good precursors to guide wet lab gene function analyses, and it is advisable that the results be viewed with follow-up lab studies. With changes in the medical industry of personalized medicine and cancer treatments based upon gene expression, rapid and reliable gene function determinations are necessary. Since earlier bioinformatic processes have been unable to produce results quickly enough for a changing medical industry of personalized medicine, many gene functions remain unknown to date. This research addresses a need for a replicable bioinformatic method that rapidly and accurately analyzes gene functions in pathways.

References

1. Fernald, G. H., Capriotti, E., Daneshjou, R., Karczewski, K. J., & Altman, R. B. (2011). Bioinformatics challenges for personalized medicine. *Bioinformatics*, 27(13), 1741-1748.

2. Fruman, D. A., Chiu, H., Hopkins, B. D., Bagrodia, S., Cantley, L. C., & Abraham, R. T. (2017). The PI3K pathway in human disease. *Cell*, 170(4), 605-635.

3. Shen, L., Shi, Q., & Wang, W. (2018). Double agents: genes with both oncogenic and tumor-suppressor functions. Oncogenesis, 7(3), 25, doi:10.1038/s41389-018-0034-x

4. Vogelstein, B., Papadopoulos, N., Velculescu, V. E., Zhou, S., Diaz, L. A., & Kinzler, K. W. (2013). Cancer genome landscapes. *Science*, 339(127), 1546-1558.

Bioinformatic Databases and Resources

The results in this research paper are based upon data and information generated from the following research networks:

TCGA: https://www.cancer.gov/tcga

Reactome: Reactome. [http://www.reactome.org]

KEGG: https://www.genome.jp/kegg/

Ensembl: website (http://www.ensembl.org)

GeneLoc: https://genecards.weizmann.ac.il/geneloc/index.shtml

Human Protein Atlas: https://www.proteinatlas.org/

ModBase: https://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi

BioGPS: http://biogps.org/#goto=welcome