Name(s)
Bryan H. Chiang

Project Number
S0807

Project Title
Illuminating Gene Dysregulation in Cancer: Deep Learning Identification of Disrupted Transcription Factor Binding Sites

Abstract

Objectives/Goals
Over 90% of mutations associated with cancer lie in the regulatory regions of the genome, driving tumor development by disrupting transcription factor binding - the "on and off switches" of key cell life, growth, and death mechanisms. The purpose of my project was to develop a comprehensive deep learning and statistical framework to pinpoint and characterize sites of irregular transcription factor binding in cancer.

Methods/Materials
Integrating over 50 million DNA sequences with corresponding chromatin accessibility and gene expression data from the ENCODE database, I first constructed high-capacity deep convolutional neural networks (CNNs) to accurately identify genome-wide regions of transcription factor binding. Next, I rigorously screened 1,500 regulatory breast cancer variants regions from the GRASP and HoneyBadger databases for statistically significant regions of differential binding across the previously uncharacterized healthy MCF-10A and cancerous T47-D breast epithelial cell lines, using data provided by the CCLE and GEO. To highlight putative misregulated genes and processes, I performed regulatory gene set enrichment analyses with GREAT. Lastly, I explored downstream roles of the putative dysregulated genes through Ingenuity Pathway Analysis (IPA).

Results
My networks had an average auROC curve score of 98.7%, high sensitivities and specificities (> 90%), and low false positive and false negative rates (< 10%) when evaluated in unseen celltypes. My networks outperformed current state-of-the-art methods by over 15% (auROC). Known binding changes for MYC, SP1, and BRCA1 were confirmed, and more than 300 unique disrupted binding sites across 8 cancer-associated transcription factors were identified ($p<0.05$). I found over 240 putative dysregulated genes and dozens of protein interactions, canonical pathways, and disease functions relevant to cancer progression.

Conclusions/Discussion
To the best of my knowledge, this is the first instance of leveraging deep learning to locate regions of genetic dysregulation in true cancer tissue. My results give us great insight into the key molecular components and mechanisms underlying cancer development that can be further validated through in vitro and in vivo experimentation. My framework also aids the development of clinical applications such as targeted drug therapies, prognostic biomarkers based on abnormal binding patterns, and base reversal technologies.

Summary Statement
I devised a novel high-capacity, integrative deep learning framework to discover over 300 disrupted transcription factor binding sites in cancer, characterizing hundreds of downstream genes and pathways possibly linked to tumor development.

Help Received
Mentored by Irene Kaplow. Questions on deep learning, statistical concepts, and software usage answered by Johnny Israeli, Anshul Kundaje, Daniel Kim, Jin Lee, Vincent Gardeux, Devon Ryan, and Kevin Blighe. Dongwon Lee gave me models to benchmark. Project sponsored by Ms. Nicole Della-Santina.