**Deep learning prediction of chemotherapy response in cancer patients using multi-omics and histological image data**

**A. Research Goal**

The efficacy of chemotherapies is highly variable across cancers and patients. Accurate prediction of an individual patient’s response to different chemotherapies prior to treatment could radically improve clinical practice and improve proper selection of therapeutic agents. The primary goal of this study is to predict response to chemotherapy drugs in cancer patients using molecular data and image data. As advanced deep learning algorithms can handle ultra-high dimensional data of different types in the generation of predictive models, these computational approaches will be used to synthesize genetic, epigenetic, and histological image data in a well-characterized, large cancer data set with pre- and post-treatment tumor pathology information. Our expectation is that the combination of genome-wide genetic data, genome-wide DNA methylation data, transcriptome (miRNA and mRNA) data and high resolution histological images—all measured prior to treatment—collectively carries signal for determining tumor response to chemotherapy agents. We anticipate that a two-phased feature selection procedure (Sure Independence Screening, followed by Gini impurity index selection), followed by conditional generative adversarial networks (CGANs), convolutional neural networks (CNNs) and self-attention generative adversarial networks (SAGANs) to produce predictive models will accurately predict tumor size response to chemotherapies.

**B. Specific Aims**

**Specific Aim 1: To develop an analysis pipeline**

**Specific Aim 2**: **To predict chemotherapy response on tumor size using pre-treatment histological image data and genome-wide multi-omics data (genome-wide DNA methylation, DNA mutation, miRNA and mRNA data).**

**Specific Aim 3: Validation**

**C. Background**

Paragraph on mortality and morbidity of cancers

Paragraph on chemotherapies used for cancer treatment

Paragraph on current clinical challenges to chemotherapy treatment and the variation in response to treatment

Paragraph on the need for prediction of chemotherapy response

Paragraph on TCGA project

Paragraph on molecular data

Paragraph on image data

Paragraph on feature selection

Paragraph on ‘traditional’ machine learning classification algorithms

Paragraph on deep learning classification algorithms

**D. Significance**

**E. Preliminary Studies**

**Predict DNA methylation with conditional generative adversarial networks to histological image data**

**Multi-Omics data provide better prediction to chemotherapy response than single-omics data**

In this study, we investigated the performance of deep-learning algorithms applied to multi-omics genomic data from the TCGA project. Drug response status including complete response (n=815), partial response (n=102), stable disease (n=182) and clinical progressive (n=892) across 32 cancers were evaluated. For computational efficiency and classifier performance, we proposed an approach using Sure Independence Screening (SIS) and Gini impurity index for feature selection from ultrahigh dimensional genomic feature set followed by conditional generative adversarial networks (CGAN) to produce predictive models for chemotherapy response, internally validated by 10-fold cross-validation. mRNA, miRNA, and methylation datasets were used both independently and jointly. For the miRNA-based prediction, we identified 23 highly informative features, including miR-141, miR-200c, miR-205, miR-9 and miRR-338, which, in combination, produced an AUC=0.64 for discriminating complete and partial response from stable disease and clinical progressive endpoints. Using mRNA-seq data, we identified 264 highly informative features including *MYCBP*, *KLF15*, *IGIP*, and *GRIA3*. The mRNA-based predictive model yielded an average AUC=0.71. The DNA methylation model attained a higher level of performance with AUC=0.81 (95% CI: 0.78-0.84). Combining mRNA-seq, miRNA, and methylation data improved the predictive performance with accuracy of 84.2% and an AUC=0.86 (95% CI: 0.82-0.90). Functional enrichment analysis for mRNA and DNA methylation selected features showed enrichment for the antioxidant response pathway and basal transcription factors associated with the platinum drug resistance (PDR) pathway. Importantly, our chemotherapy response classifier substantially outperforms a predictive model focused on mRNA and methylation features within PDR pathway (hsa01524 in KEGG) AUC=0.62, demonstrating that taking a broad approach using genome-wide multi-omics data dramatically improves discrimination. In summary, this work demonstrates that applying deep-learning algorithms to multi-omics data can generate informative predictive models for chemotherapy response across numerous cancers.

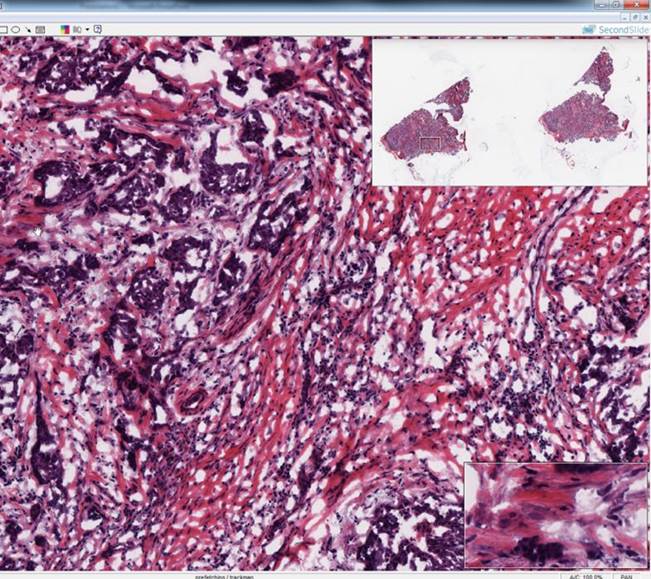
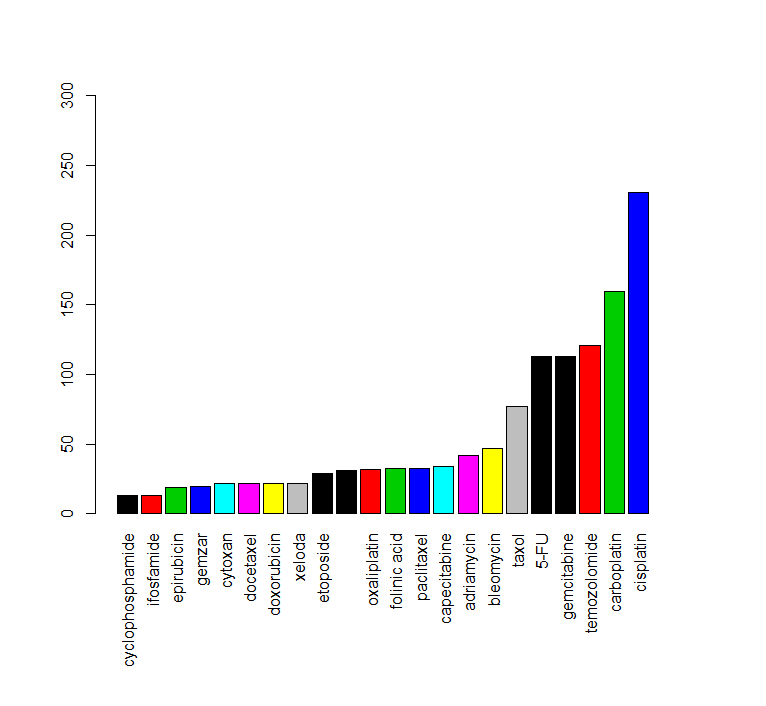
**F. Research Design and Methodology**

***Molecular data, image data collection and chemotherapy drug response definition.***

We collected 9,756 genome-wide DNA methylation data (methylation 450K microarray data), 12,759 mRNA-seq data and 11,667 miRNA-seq data and 11,766 histological diagnostic image data across 33 cancer types in TCGA project [1]. Meanwhile, we collected all 3,320 complete drug response records and eventually we identified 1,991 individuals with clear drug response status including complete response (n=815), partial response (n=102), stable disease (n=182) and clinical progressive (n=892) across 33 cancers.

***DNA methylation, RNA-seq, miRNA-seq, DNA mutation and histological image data structure.***

Genome-wide DNA methylation was measured my Illumina methylation 450K microarray. All the beta-values were extracted by the unique ID for the 1,991 individuals with complete drug response status. Transcriptomic data were measured by RNA-seq assay and gene-based counts identified by HTSeq [2] and fragments per kilobase of transcript per million mapped reads upper quartile (FPKM-UQ, [2]) for all the gene were collected. Genome-wide miRNA data were generated by miRNA-seq and the miRNA expression were produced by BCGSC miRNA profiling pipeline [3]. Cancer sample DNA mutation were called by the comparison between cancer tissues and matched normal tissues or blood cells with 4 different algorithms including MuTect2 [4], SomaticSniper [5], MuSE [6], VarScan2 [7]. Digital images of glass slides were scanned by using digital slide scanner (Aperio's ScanScope AT). Digital images in SVS format can be viewed with Aperio's ImageScope viewer.

***G. External Collaborators***

None. The assembled research team has all of the necessary expertise to conduct the proposed study.

***H. Timeline***

Timeline for Dr. Shicheng Guo, Dr. Steven J. Schrodi and Dr. Fereshteh Bashiri are attached as appendix 1.

***I. Budget Justification***

**BUDGET PERIOD: 28/06/2019 – 28/06/2020 (12 Months)**

**PERSONNEL**

**Shicheng Guo, Ph.D. – Principal Investigator:** Dr. Guo will coordinate all efforts for this project including directing all research aims, prepare molecular data (miRNA, mRNA, DNA methylation, DNA mutation and drug response), interpreting predation results, providing progress reports, and publishing results. Dr. Guo will devote x C/M to the project.

**Steve J. Schrodi, PhD – Co-Investigator:** Dr. Schrodi will assist Dr. Guo in all aspects for this project including directing all research aims, interpreting results, providing progress reports, and publishing results. Dr. Schrodi will devote x C/M effort to the project which will be cost shared.

**Fereshteh Bashiri, PhD – Co-Investigator**: Dr. Bashiri will assist Dr. Guo in several aspects of this project including image data analysis, image feature extraction, conditional generative adversarial network (CGAN) training, interpreting results, providing progress reports, and publishing results. Dr. Bashiri will devote x C/M effort to the project.

**J. Budget**

**K. Biographical Sketch**

**L. Literature Cited**

1. Cancer Genome Atlas Research, N., et al., *The Cancer Genome Atlas Pan-Cancer analysis project.* Nat Genet, 2013. **45**(10): p. 1113-20.

2. Anders, S., P.T. Pyl, and W. Huber, *HTSeq--a Python framework to work with high-throughput sequencing data.* Bioinformatics, 2015. **31**(2): p. 166-9.

3. Chu, A., et al., *Large-scale profiling of microRNAs for The Cancer Genome Atlas.* Nucleic Acids Res, 2016. **44**(1): p. e3.

4. Cibulskis, K., et al., *Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples.* Nat Biotechnol, 2013. **31**(3): p. 213-9.

5. Larson, D.E., et al., *SomaticSniper: identification of somatic point mutations in whole genome sequencing data.* Bioinformatics, 2012. **28**(3): p. 311-7.

6. Fan, Y., et al., *MuSE: accounting for tumor heterogeneity using a sample-specific error model improves sensitivity and specificity in mutation calling from sequencing data.* Genome Biol, 2016. **17**(1): p. 178.

7. Reble, E., et al., *VarScan2 analysis of de novo variants in monozygotic twins discordant for schizophrenia.* Psychiatr Genet, 2017. **27**(2): p. 62-70.

8. Kather, J.N., et al., *Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer.* Nat Med, 2019.

9. Lengauer, C., K.W. Kinzler, and B. Vogelstein, *DNA methylation and genetic instability in colorectal cancer cells.* Proc Natl Acad Sci U S A, 1997. **94**(6): p. 2545-50.

10. Jin, C., et al., *Left Atrial Appendage Segmentation Using Fully Convolutional Neural Networks and Modified Three-Dimensional Conditional Random Fields.* IEEE J Biomed Health Inform, 2018. **22**(6): p. 1906-1916.

11. de Vos, B.D., et al., *ConvNet-Based Localization of Anatomical Structures in 3-D Medical Images.* IEEE Trans Med Imaging, 2017. **36**(7): p. 1470-1481.

12. Han Zhang, I.G., Dimitris Metaxas, Augustus Odena, *Self-Attention Generative Adversarial Networks.* arXiv:1805.08318, 2018.

The primary goal of this study is to use ultra-high dimensional genetic, epigenetic and image data to predict response to chemotherapy drugs in cancer patients using advanced deep learning algorithms. Deep learning has become one of the most important tools for precision medicine research and provides a highly efficacious approach for personalized drug response prediction, early disease diagnosis and disease subtype classification based on ultra-high dimensional genetic, epigenetic and image data.

We hypothesize that histological image data and molecular features driving regulatory mechanisms contain information useful in the prediction of chemotherapy response. This study is designed to investigate machine learning approaches using both image data and genome-wide molecular data to predict chemotherapy drug response across several cancers. Of particular interest will be to use deep learning algorithms such as conditional generative adversarial neural networks (CGANs), convolutional neural networks (CNNs), and self-attention generative adversarial networks (SAGANs). The multi-omics data is highly comprehensive and spans genetics (genome-wide DNA mutations), transcriptional data (genome-wide miRNA and mRNA data), and epigenetics (genome-wide DNA methylation).

contains information to infer molecular characteristics, such as genome-wide DNA methylation and clinical outcomes such as chemotherapy response. The goal of this study is to investigate the approach to apply multiple deep learning approaches such as conditional convolutional neural networks (CGAN), convolution-al neural networks (CNNs) and [self-attention generative adversarial networks](https://arxiv.org/abs/1805.08318) (SAGAN) to human genetic (DNA mutation), transcriptional data(miRNA and mRNA) and epigenetics data (DNA methylation) as well as histological diagnostic image data to predict chemotherapy drug response and human molecular profile and characteristics. We will also compare the deep learning algorithm to multi-Omics-image data with traditional machine learning method such as Naïve Bayes (NB), Hidden Markov Models (HMMs), Hidden Semi-Markov Models (HSMM), Conditional Random Fields (CRFs), Support Vector Machine (SVM) and Random Forest Models (RF).

with traditional machine learning method and deep learning (neural networks, such as conditional convolutional neural networks). Recently, Dr. Kather, in his Nature Medicine paper [8], demonstrated that histological image data could provide powerful prediction to microsatellite instability (MSI). Meanwhile, DNA methylation capacity have been shown striking correlation with genetic instability [9]. We plan to investigate the prediction ability of histological image data to genome-wide DNA methylation level by 9,756 paired DNA methylation and histological image data with the most powerful deep learning method: conditional convolutional neural networks (CGAN)[10], convolutional neural networks (CNNs)[11] and [self-attention generative adversarial networks](https://arxiv.org/abs/1805.08318) (SAGAN) [12].

with traditional machine learning method and conditional convolutional neural networks (CGAN). Chemotherapy drug response prediction become one of most important tasks to achieve precision medicine. However, previous evidence showed the prediction accuracy is very low with biomarker panel or single-omics data. More and more studies were trying to apply multi-omics data for the chemotherapy drug response. We plan to integrate multi-Omics data and H&E histological image data to predict chemotherapy drug response within 1,991 well recorded drug response cross 33 cancers with Sure Independence Screening (SIS) and Gini impurity index (SII) for feature selection from ultrahigh dimensional genomic feature set followed by conditional generative adversarial networks (CGAN)[10], convolutional neural networks (CNNs)[11], [self-attention generative adversarial networks](https://arxiv.org/abs/1805.08318) (SAGAN) [12]

Currently, we don’t have any existed external collaboration in this research. Dr. Mark Craven at Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison and Dr. Momiao Xiong at Department of Biostatistics, University of Texas Health Science Center in Houston (Uthealth) may join the discussion to the research if necessary.