Statement of Research Interests

Shicheng Guo, Ph.D.

Human complex diseases dynamically and progressively driven by sequences of genetic-epigenetic-environment interactions. Each of these components contributes to bring phenotypes from a susceptibility stage to clinical presentation. The most exciting ideas of precision medicine require scientists to consider numerous risk/susceptibility factors as a system to understand the pathogenesis of human diseases. Taking rheumatoid arthritis as an example, individuals carrying HLA-DRB1*04:01 together with triggering inflammation and citrullination caused by factors such as smoking is thought to initiate an autoimmune reaction and eventually generate circulating cell-free DNA methylation signals in patient plasma samples, caused by synovial cell apoptosis. In the past eight years, I have been working on genetic and epigenetic variants in multiple diseases and authored more than 50 SCI publications, 17 of which are first/co-first/co-corresponding author publications. My experience in both computational and molecular approaches has prepared me to conduct comprehensive research studies in precision medicine. Precision medicine requires precise, accurate and comprehensive clinical information. The extensive and longitudinal data collected by the Marshfield Clinic will provide an excellent base for conducting my research. The Marshfield Clinic Research Institute's Personalized Medicine Research Project (PMRP) contains exome-wide genotyping data on almost 20,000 samples. With my expertise in computation and epigenetics, I propose to 1) generate genome-wide DNA methylation profiles for PMRP and other related samples to conduct epigenetic-wide association studies (EWAS) and investigations designed to discover methylation-based diagnostic, prognostic and pharmaco-epigenomics (PeGx) biomarkers. 2) develop a novel research collaboration system and medical decision support system based on deep learning by integrating EHR, genetic and epigenetic information.

Research Background

I have conducted a wide array of genetic and epigenetic studies in my career, heavily relying on computational approaches. My experience with these studies has given me the ability to investigate clinically important research questions using novel and multi-faceted approaches.

Genetic susceptibility to the development of autoimmune disease. Early in my career, I investigated genetic variants involved in systemic sclerosis (SSc) and rheumatoid arthritis in the Chinese Han population. Applying a multiple-candidate pre-selection method (SNP and CNV screens), I identified multiple susceptibility genes, such as an important CNV within *HLA-DQA1* and *APOBEC3A/3B* for SSc, *CFH* for age-related macular degeneration, and *FOXE1* for thyroid cancer. I also conducted a large association study interrogating genetic variants in miRNA for human cancer and identified miR-4293 as being significantly associated with non-small cell lung cancer, and miR-196a2/miR-499 involved in esophageal squamous cell carcinoma. These findings have provided much needed molecular insight into the role of miRNA regulation and genetic variants involved in these cancer etiologies.

Epigenomic Research and Epigenetic Variations in Diagnosis and Prognosis Models for Complex Diseases. Starting from 2009, I investigated the epigenetics of human diseases with a particular focus on DNA methylation. I participated in several large projects to build a model of the epigenomic architecture for human cells and tissues under normal and disease conditions. Notable work includes evaluating the genomic methylation profiles (methylomes) for normal human blood cells, animal model 'silk', CD4+ T-cells of patients with rheumatoid arthritis, pancreatic cancer cells, and hepatocellular carcinoma cells with different methylation methods, such as BS-seq and MBD-seq. Concurrently, I identified a large number of methylation-based markers with diagnostic and prognostic implications for lung cancer, bladder cancer, and pancreatic cancer. Since DNA methylation has different patterns for different tissue types, we proposed a predictive model to map the origin of cell-free DNA fragments based on tissue-specific methylation signals. This model provides a potentially non-invasive approach for the diagnosis of solid cancers. This work has been published in *Nature Genetics* in 2017.

Development of Functional Principal Component and Randomized Sparse Clustering Algorithm for Medical Image Analysis. Medical imaging data is an important aspect of electronic health records (EHR) and has the potential to enhance the diagnosis of disease, the prediction of clinical outcomes and the characterization of disease progression. However, the growing data dimensions pose great methodological and computational challenges for the representation and selection of features in image cluster analysis, including both dimension deduction and feature selection. To address these challenges, my colleagues and I extended the functional principal component analysis (FPCA) from one dimension to two dimensions to capture the space variation of image signals. Additionally, we applied randomized algorithms to remove the irrelevant features for data clustering. We applied this method to the liver and kidney cancer histology image data from the TCGA project and demonstrated that the randomized feature selection method coupled with functional principal component analysis substantially outperforms the current sparse clustering algorithms in image cluster analysis.

Current Research

My research in Center for Precision Medicine Research is composed of several studies to elucidate disease genetics with certain non-traditional statistical methods and to identify epigenetic variant-based diagnostic and prognostic biomarkers. The details for these projects are as follows:

A Gene-Based Recessive Diplotype Exome Scan to Identify Disease Genes for 10 PMRP Phenotypes. This project started from late 2017 and supported by Dr. Schrodi's MCRI grant. We planed to map disease genes in iron overload disorders, obesity, type 2 diabtetes, rheumatoid arthritis, psoriatic arthritis, multiple sclerosis, premature myocardial infraction, and obsessive compulsive disorder using a novel approach of gene-based recessive diplotype analysis based on exome-wide genotype data (PMRP). Standard analyses applied to genome-wide association data are well-designed to detect additive effects of moderate strength. However, the power for standard GWAS analyses to identify effects from recessive diplotypes is not typically high. With this approach applied to iron overload, a strong association signal was identified between the fibroblast growth factor-encoding gene, *FGF6*, and hemochromatosis in the central Wisconsin population. Our functional validation result showed Fgf-6 regulates iron homeostasis and induces transcriptional regulation of hepcidin. Moreover, specific *FGF6* variants identified in the study were shown to differentially impact iron metabolism. Using the recessive diplotype approach, we revealed a novel susceptibility hemochromatosis gene and has extended our understanding of the mechanisms involved in iron metabolism. Furthermore, significant novel findings have also been made in type 2 diabetes, premature myocardial infarction and rheumatoid arthritis. For obesity, we have determined the most promising candidate genes and submitted it to our collaborators for functional interrogation. Corresponding molecular and cellular validation have designed with knock down and up-regulation experiments. Meanwhile, I am working on the bioinformatics manuscript of "A R Packge for Compound Heterozygote Scanning to Identify Novel Disease Genes based on Exome-wide Genotype Data".

Epigenetic silencing mediated by hyper-methylation in esophageal squamous cell carcinoma (ESCC) and cholangiocarcinoma (CCA). This project started in early 2017 and collabrate with Dr. Schrodi and Dr. Wang. Based on TCGA dataset, I identified the most signficant top 200 hypemethylated regions with optimized functional differential DNA methylation region identification algorithm (fDMRI) for ESCC and CCA. In the new developed prioritization algorithm genome-wide TFBS assay (ChIP-seq), eQTL assay and RNA-seq from GTEx were integrated and given different weights so that functional DMRs could be identified. With this method, we identified the most important 50 functionally abnormal methylation regions. Further biological validation demonstrated that our method could identify novel interactions between DNA methylation and other genomic functional elements. For example, we found epigenetic silencing of ZNF132 mediated by methylation-sensitive Sp1 binding promotes cancer progression in ESCC. We will be continue to validate the remaining fDMRs and discover the mechanisms of these interaction in the diseases. Meanwhile, Dr. Schrodi and I will integrate GWAS study to investigate the interaction between genetics and epigenetics in the pathology of ESCC and CCA.

Proposed Research Program

While human genetics opened the road to precision medicine, genetics is not enough for the achievement of precision medicine. Individual variability from genetics, epigenetics and environment should be concurrently considered in disease treatment and prevention. In addtion, our understanding of complex disease pathogenesis is limited as the majority of the heritability is missing in results from traditional GWAS studies. Epigenetic variation, non-additive but complex modes of inheritance, misdiagnosis, and rare SNP and CNV effects, are the most promising solutions to the missing heritability problem. With diverse interdisciplinary training in human genetics, biostatistics, computational biology and epidemiology, my research will focus on 1) identifying genetic and epigenetic biomarkers for diagnosis, prognosis, pharmacogenomics and race disparity prediction. 2) identify interactions between genetic and epigenetic variants and elucidate their roles in complex diseases. 3) Implement low-cost large-scale methylation analytical method in PMRP cohort (eg. Fecal-seq, RAD-seq, cfMeDIP–seq) to conduct epigenetic-wide association study (EWAS) with PheWAS style.

Phenome-wide Association Study of Genetic Variation in Epigenetic Factors to Test the Role of Epigenetics in Human Complex Disease. This project is a collaboration with Dr. Steven Schrodi and Dr. Mark Craven as part of the Computation and Informatics in Biology and Medicine (CIBM) training program. Human complex disease is generated by the interaction between genetics, epigenetics and the environment. While the rationale for genetic association studies have been supported by different fundamental observations such as heritability estimates from twin studies, there is no fundamental research to illustrate whether epigenetic changes are involved in disease heritability, although we know that epigenetic elements are an important interface between genetics and the environment. In this study, I hypothesize that genetic variants in epigenetic genes are a proxy to infer the epigenetic involvement in phenotypes. We will apply a phenome-wide association study (PheWAS) approach to test the association between a panel of epigenetic factors against 6,221 clinical traits within the Marshfield Clinic Personalized Medicine Research Project (PMRP) dataset. This will enable us to identify all the significant phenotypes whose pathology are potentially driven by epigenetic changes and apply the measurement of genome-wide DNA methylation levels in the corresponding phenotypes to validate the above findings. This project will also feature a collaboration with Dr. Scott J Hebbring. We will share the DNA methylation dataset with Dr. Hebbring so that he will investigate the relationship between aging, telomeres length and genome-wide DNA methylation. To date, there has not been this type of study conducted and the results of this work will provide insight into epigenetic architecture underlying important clinical traits.

Medical Decision Support and Research Collaboration System Integrating EHR, Genetic and Epigeneitc Information. Modern health care is accelerating quickly since the synergistic development of electrical health record system, human genetics and epigenetics. Every year, FDA approves a large number of novel disease screening assays, diagnostic panels, outcome prediction approaches and personalized medications. Meanwhile, the guideline of the evolution of disease managment is rapidly accelerating. Multi-omics information including genetics, epigenetics and phenomes will be integrated by EHR system. One the other side, translational research informatics (TRI) system will be signficantly changed from clinical information management (CIM), translational study management (TSM), research collaboration system (RCS) to biorepository management systems (BMS). I plan to develop a comprehensive system to combine these elements especially for autoimmnue diseases, such as rheumatoid arthritis.

Aim 1. Develop Novel Research Collaboration System and Medical Decision Support System for Autoimmnue Disease. As the better understanding to the pathology of human diseases, EHR information is also evolving quickly. However, corresponding research collaboration system and medical decision support system are always lack of the timely update, such as disease identification algorithm (DIA), disease subtype identification algorithm (SIA), personalized medication selection algorithm (PMSA), outcome prediction algorithm (OPA). In this project, I will propose a novel pipeline to apply deep learning algorithm to the informations from computerized provider order entry (CPOE) system including codified data (such as ICD codes), natural language information from narrative EHR data and research information from science data management system (SDMS) to provide better medical decision support to the diagnosis, prognosis and drug selection. Take rheumatoid arthritis as an example, all the information from the different systems will be integrated including ICD9/10, anti-CCP, *PTPN22*, *PADI4*, *HLA-DRB1*, electronic prescriptions, ANA and DAS28-ESR and then multiple heuristic algorithms including deep learning, random forest, Bayesian nework and support vector machine will be applied in DIA, SIA, PMSA and OPA. These systems will be dramatically increase the efficiency of the health care provider and decease the cost of the service.

Aim 2. Identify Novel Diagnostic, Prognostic Biomarkers and Pharmaco-Epigenomics for Complex Disease. DNA methylation has been demonstrated to be one of the most promsing diagnostic, prognostic and pharmaco-epigenomics biomarkers for human complex disease. This may be attributable to the fact that DNA methylation is partially stable and partially dynamic, compared to genetic variation (completely stable) and mRNA (highly dynamic). DNA methylation is involved in transcriptional regulation and alternative splicing and therefore plays critical roles in differentiation, development and disease. Given its regulation roles, DNA methylation changes usual occur earlier than other classes of molecular variation. A large number of DNA methylation-based diagnostic and prognostic biomarkers have been identified and *SEPT8* and *SHOX2* methylation has been approved by FDA for colon cancer and earlty lung cancer screening. However, the majority of these projects are focused only on cancers. With my expertise on DNA methylation, I will extend my previous work on other human complex diseases, especially on autoimmnue diseases and inflammatory diseases in which cell death releases the DNA into blood and can therefore be taken as a tissue-of-origin and disease status biomarker. I will apply multiple DNA methylation assays such as BS-seq, RRBS, BSPP, MH450K, MBD-seq to identify the most promising DNA methylation signals and apply Bayesian networks and other machine learning methods to build the most stable diagnosis or prognosis models. In these models I will integrate established risk factors. Taking rheumatoid arthritis as an example, rheumatoid factor (RF), anti-CCP, genetic risk variants (*HLA-DRB1*), smoking status will be considered. I expect such prediction model based on genetic, epigenetic and environmental factors to have powerful prediction performance.

Aim 3. Genetics and Epigenetics Interaction and Precision Medicine in PMRP cohort. Increasingly, inexpensive gentopying or sequencing technologies are now being introduced to identify genetic variants. We already have genotype data for more than 20,000 samples in PMRP cohort. However, epigenetic data for PMRP has not yet been generated—a situation which impacts the progression of precision medicine. (Descirbe what you want to do here. The hypothesis that you want to address, then talk about the approach) MRCI will be the perfect place to initiate this work. In order to start this project, I will implement low-cost large-scale methylation analysis method in PMRP cohort such as Fecal-seq, RAD-seq, cfMeDIP–seq, MBD-seq so that we can conduct EWAS with PheWAS style and therefore MRCI will be leading in the second generation of EWAS research. Meanwhile, aften we generate the genome-wide DNA methylation dataset, we can initiate some other related project such as 1) Interaction between DNA methylation and genetic variations and the application in precision medicine including precision subtypes, pharmacogenomics and pharmacoepigenomics. 2) DNA methylation mediated eQTL analysis in complex disease based on PMRP cohort. 3) Genome-wide association study to identify genetic variants associated with overall methylation levels. 4) relationship between aging, DNA methylation and telomere length in PMRP cohort.