Statement of Research Interests

Shicheng Guo, Ph.D.

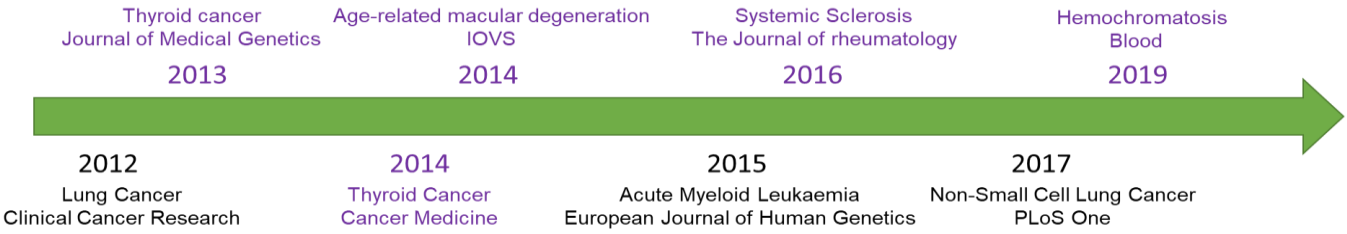
Human complex diseases are dynamically and progressively driven by a series of genetic-epigenetic-environmental interactions. These interactions contribute to bringing phenotypes from a susceptibility stage to clinical presentation. The most exciting ideas of precision medicine involve the consideration of numerous genetic susceptibility variants as potential baseline risk factors and track the epigenetic changes in pathogenesis of human diseases over time. Taking rheumatoid arthritis as an example, individuals carrying HLA-DRB1*04:01 in combination with inflammatory processes and citrullination caused by factors such as smoking, are thought to initiate an autoimmune process. Early molecular and epigenetic biomarkers such as rheumatoid factor (RF), anticyclic citrullinated peptide (anti-CCP) and circulating cell-free DNA methylation signals caused by synovial cell apoptosis precede the manifestation of symptomology. I have a strong interest in delineating these types of genetic-epigenetic-environmental interactions to elucidate the etiologies of complex diseases. In the past 10 years, I have been working to identify genetic and epigenetic variants in multiple human complex diseases. This work has resulted in >50 peer-reviewed publications, 17 of which are first/co-first/co-corresponding author publications. My extensive experience in conducting human genetics, epigenetics research combined with my bioinformatics skills in analyzing and interpreting data has prepared me to conduct comprehensive research studies in precision medicine. My future research will focus on 1) identifying and implementing detection of circulating cell-free DNA-based genetic and epigenetic biomarkers for early diagnosis, real-time monitoring and prognosis prediction; 2) functional assessment of the genetic and epigenetic variants identified by GWAS, PheWAS, EWAS studies; and 3) developing pharmaco-epigenetic biomarkers (PeGx) which will aid in individualizing therapies..

Research Background

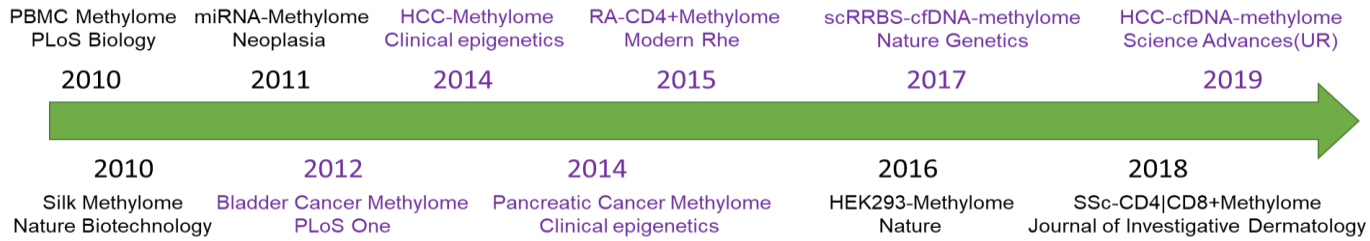
I have conducted a wide array of genetic and epigenetic studies which have utilized both my computational and bench science skills. My experience from these studies has given me the ability to investigate clinically important research questions using novel and multi-faceted approaches.

Identification of Genetic Susceptibility Genes to Human Complex Diseases.

Early in my career, leveraging GWAS and candidate gene studies, I identified multiple susceptibility genes, CFH for age-related macular degeneration and FOXE1 for thyroid cancer. I also conducted a series of association studies investigating the association between genetic variants in epigenetic factors such as miRNAs and human cancers. In these studies, miR-4293 was demonstrated to be significantly associated with lung cancer, and miR-196a2/miR-499 was clearly involved in esophageal cancer. Furthermore, I generated the first genome-wide copy number architecture for systemic sclerosis (SSc) in 2016. Applied whole-genome aCGH microarray in a systemic sclerosis (SSc) discovery population, I identified five statistic-significant and potential functional related systemic sclerosis (SSc) copy number variations (CNVs) and two CNVs including HLA-DQA1and APOBEC3A/3B were successfully validated with independent cohort. Recently, I implemented a novel approach termed exome-wide gene-based recessive diplotype scanning. Compared with traditional methods, our method demonstrated higher power to identify recessive compound heterozygotes. The study identified a susceptibility gene, FGF6, for hemochromatosis. By conducting an evolutionary analysis, protein-protein network analysis, and examining both molecular and cellular evidence, we demonstrated FGF6 plays an important role in iron metabolism which may be involved in multiple conditions underlying diseases associated with iron metabolism. The work was accepted in Blood in 2019. The following flowchart shows the timeline for publication of outcomes of my research efforts (Purple indicates 1st author).



Epigenomic Research in Diagnosis and Prognosis Models for Complex Diseases. Beginning in 2009, I have investigated the epigenetics of human diseases with a particular focus on DNA methylation. I participated in several large projects to build a model depicting the epigenomic architecture for human cells and tissues under normal and disease conditions. Notable work includes the evaluation of the WGBS-based and MBD-seq-based methylomes for normal human blood cells, animal model 'silk', CD4+ T-cells of patients with rheumatoid arthritis, pancreatic cancer cells, and hepatocellular carcinoma cells. 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, my colleagues and I proposed a predictive model to map the origin of cell-free DNA fragments based on tissue-specific methylation signals. This model provides a non-invasive approach for improving the diagnosis of solid cancers. This work was published in *Nature Genetics* in 2017.



Current Research

My research in Center for Precision Medicine Research consists of several studies to elucidate genetics of disease 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 15 PMRP phenotypes. Motivated by successfully showing that the fibroblast growth factor *FGF6* is involved in iron metabolism, as a Co-Investigator I am currently conducting exome-wide scans for recessive diplotype effects in morbid obesity, type 2 diabetes, rheumatoid arthritis, psoriatic arthritis, multiple sclerosis, premature myocardial infarction and obsessive-compulsive disorder using exome-chip genotype data on 19,000 central Wisconsin samples. My analyses have produced signficant signals for type 2 diabetes (*MTNR1B* and *KIF2C*) and morbid obesity (*SPTBN5*). I am collaborating with other investigators to conduct biological validation of these targets and preliminary data shows highly interesting findings from knockdown and up-regulation experiments. I am working on a manuscript describing these results. Additionally, I am writing a manuscript describing the R package I have written for compound heterozygote scanning to identify disease genes (to be submitted to *Bioinformatics*).

DNA Hypermethylation Mediated Epigenetic Silenced Diagnostic and Prognostic Biomarkers in Human Cancers. This project started in early 2017 as a collaboration with Dr. Steven Schrodi and Dr. Minghua Wang. Leveraging my well-trained bioinformatics and data-mining skills, I have conducted a systematic analysis integrating well-known public data including the TCGA project, ENCODE project, Roadmap project, GTEx project, ExAC dataset, and UK-biobank dataset to identify the most interesting differential DNA methylation regions using the 'differential DNA methylation region identification algorithm' (fDMRI) for human cancers. I applied this method and identified ~50 fDMRs for esophageal cancer (ESCC) and cholangiocarcinoma. These regions are likely to have high utility for use as novel diagnostic and prognostic cancer biomarkers. Currently, biological validation has demonstrated that our method identifies 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. Results were published in *Cell Death &* Disease in 2018. We will continue to validate the remaining fDMRs and discover the mechanisms surrounding these epigenetic factors in cancers. In another study, I proposed a novel low-cost non-invasive cell-free DNA methylation based diagnosis approach for liver cancer. In this study, I demonstrated the measurement of long-range methylation could be applied in low-pass cell-free WGBS at as low as 5-million reads to reflect liver disease status of chronic hepatitis, cirrhosis and HCC. Moreover, DNA hypo-methylation in HBV integration regions was shown to be a promising novel biomarker for early HCC detection.The manuscript was submitted this month and now under review in Science Advances.

Phenome-wide association study of genetic variation in epigenetic factors. This project is a collaboration with Dr. Steven Schrodi (Marshfield Clinic and Laboratory of Genetics, UW-Madison) and Dr. Mark Craven (Department of Biostatistics and Medical Informatics, UW-Madison) that is currently in the initial stages as part of the CIBM training program supported by UW-Madison and National Library of Medicine (NLM). 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 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. Such studies have not been previously conducted and the results of this work will provide insight into epigenetic architecture underlying important clinical traits. If offered the position, I anticipate transitioning this study to a K-award mechanism.

Proposed Research Program

While human genetics opened the road to precision medicine, genetics is not sufficient to drive a revolution in individualized clinical treatment. Individual variability caused by genetics, epigenetics and environment should be concurrently considered in disease subtype, diagnosis, treatment and prevention. In addition, our understanding of the pathogenesis of complex disease is currently limited since the variants identified by traditional GWAS studies can only explain very limited proportions of the overall heritability. Epigenetic variation, non-additive complex modes of inheritance, misdiagnosis, and rare SNP and CNV effects are the most promising solutions to the missing heritability problem. In addition, recently, a series of powerful and low-cost functional assessment and identification tools have

been developed such as WES, ATAT-seq, scRRBS, BSPP, CUT&Tag, Fecal-seq, RAD-seq, cfMeDIP-seq which offer an excellent opportunity to investigate the role of genetic, epigenetic and interactions in the development of complex diseases. With interdisciplinary training in human genetics, biostatistics, computational biology and epidemiology, my research will take advantage of my knowledge of these advanced technologies, which will be applied to the following three fields: 1) developing and implementing new epigenetic approaches to cancer susceptibility, early diagnosis and prognosis.; 2) determining the function of epigenetic variants associated with disease using integrated GWAS, EWAS and PheWAS results.; And 3) developing epigenetic biomarkers based pharmaco-epigenomics (PeGx) approaches that complement PGx and improve the personalized application of therapeutics.

Development and implementation of novel epigenetic approaches to human cancer diagnosis and prognosis. DNA methylation is one of the most promising 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 germline genetic variation (completely stable) and mRNA (highly dynamic). DNA methylation is involved in transcriptional regulation and therefore plays critical roles in cell differentiation, development and disease. Given its regulatory 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, such as *SEPT9* and *SHOX2* which have been approved by FDA for colon cancer and early lung cancer screening. However, DNA methylation diagnostic biomarkers for other cancer types remain to be defined. In my previous publication *(Nature Genetics*, 2017), I demonstrated that non-invasive cell-free DNA methylation haplotype-based tissue-of-origin prediction could be effectively applied to cancer diagnosis. However, the study had a small sample size (N=134) and limited cancer types (N=2 includes lung cancer and colon cancer). In this project, I will make full use of the abundant cancer samples in pre-built biobanks in which plasma samples were collected before the clinical diagnosis which provides me an opportunity to identify early biomarkers. I will integrate genetic (cancer risk alleles which have been generated for PMRP samples, somatic mutations), and other informative variables, such as cell-free DNA fragment distribution, metabolites in plasma to develop a multi-omics prediction platform with deep learning (artificial neural networks) to generate AI-based models for early cancer diagnosis. Following sufficient preliminary results, I will seek NIH funding for the expansion of this study.

Functional assessment of human genetic and epigenomic variants using computational and biological approaches. Recently, GWAS, EWAS and PheWAS have identified hundreds of significant disease-associated genetic and epigenetic loci. However, there is an enormous gap between the statistically-significant associations and understanding their causal role, if any, in human phenotypes. Functional assessment of genetic and epigenomic variants will provide the opportunity to advance our understanding of the functional significance of these loci and how they contribute to disease risk and the pathogenesis of complex diseases. As the of multi-dimensional data including genetic, gene expression, epigenetic, eQTL, pathway and drugbank databases increases, computational approaches could provide a solution to identify the disease causal variants and to design the most efficient functional discovery approach for these variants. In this study, computational assessment was conducted leveraging publically available databases and tools such as ENCODE, GTEx, ExAC, FANTOM, RegulomeDB, BioGPS, STRING, Reactome and KEGG pathway. In my previous publication (*Blood*, 2019), I have demonstrated that comprehensive computational and evolutionary analyses provides a highly efficient functional, protein structure-related, interactive network prediction and cell-of-origin determination. These findings are very useful for understanding the pathogenic role of candidate genes and variants. To date, I have conducted these analyses on 15 phenotypes within 20,000 PMRP samples and identified several highly interesting candidate genes, such as *MTNR1B* in T2D, and *SPTBN5* for morbid obesity and the full functional predictions have been completed. In the next stage, I aim to validate these discoveries with indepedent dataset and biological functional experiences. Take *MTNR1B* as an example: In the previous GWAS study, *MTNR1B* was found to be associated with T2D since promoter SNPs were found to be an eQTL which influence the expression of *MTNR1B* and mediate early insulin secretion (Lyssenko, *Nature Genetics*. 2009). However, I found non-synonymous diplotype-recessive exomic variants located in *MTNR1B* functional domain: Melatonin receptor signature and rhodopsin-like G protein-coupled receptors might cause functional silencing of *MTNR1B*. Genome-wide RNAi screening found *MTNR1B* knockdown would have depressed secretory pathway (Jeremy, *Nature Cell Biology*, 2012) in HeLa cells. Hence, I will utilize human clonal β cells to validate the function of exonic variants with wild-type/mutation-type knock-in and insulin-releasing tracking with GFP-tagging. Our working hypothesis is that *MTNR1B* exonic variants play a critical role in insulin transmembrane release and could be an important target for gene therapy to patients with these deleterious mutations.

Developing pharmaco-epigenomic (PeGx) biomarkers to improve PGx-based treatment. Genetic variation in drug-metabolizing enzymes and transporter (DMET) genes modifying drug response is the main focus of pharmacogenetics (PGx). Multiple studies have shown that SNPs, even on a genome-wide scale, are only partially informative for the prediction of drug response (typical AUC<0.8). Hence, transcriptional regulation of DMET genes may also play an important role in drug response, drug resistance and adverse drug events. In my previous studies, I have developed and implemented multiple epigenetic approaches to investigate how DNA methylation plays a role in drug response. My hypothesis is that epigenetic status could significantly improve prediction of drug response. In this study, I plan to apply epigenetic approaches to identify blood-based DNA methylation biomarkers for Methotrexate (MTX) is one of the chemotherapy drugs for cancer treatment and also the gold standard drug in rheumatoid arthritis (RA) which is widely used by 350 million RA patients. Although several polymorphisms have been identified to be associated with MTX efficacy and toxicity, the outcome prediction performance is very disappointing. In this study, I will build a prediction system based on genetics and epigenetics of 66 very important pharmacogenes. We can apply the research strategy in other drugs whose samples are easy to be collected in School of Public Health, Harvard. My aim is to advance precision medicine through the incorporation of DNA methylation-based PeGx.