**Title:**

**Research Goal**

The overarching goal of this study is to evaluate the epigenetic contribution for the phenotypes existed in the Marshfield Clinic Personalized Medicine Research Project (PMRP). Traditional genome-wide association studies have identified large number of susceptibility variants, however, majority of the significant variants are non-coding variants and these variants can only explain quite limited heritability. Epigenetic variants become one of the most important components of the missing heritability. In order to investigate these epigenetic contributions to human phenotypes, classic twin analyses could be conducted; however, the sample size usually is quite limited. Here, we hypothesize that epigenetic involvement in the disease or phenotype can be inferred by the significant association between the phenotypes and genetic variations in epigenetic factors. In this study, we will discover the epigenetic contributions to PMRP phenotypes with PheWAS study design and validate the PheWAS finding with genome-wide DNA methylation assay to all the PMRP phenotypes.

**Specific Aims**

**Specific Aim 1:** To identify a set of phenotypes coded in the EMR for 10,000 Marshfield Clinic patients (N=10,124) associated with one or multiple epigenetics factors. The hypothesis is that we can accurately extract phenotypic information from the EHR to generate a set of case/control and quantitative traits from which we can evaluate signals identified from the initial study on the same phenotypes. To reduce the false positive association, we will applied PMRP phase 2 dataset (N=8,400) as the validation dataset.

**Specific Aim 2:** To construct the phenotype networks by the sharing the common epigenetic susceptibilities. Phenotypes with similar susceptibilities variants might share similar therapy strategy. With the phenotype network, we can map the therapy approach or drug targets of the unknown disease to known disease.

**Specific Aim 3**: To illustrate the relationship between traits heritability with the strength of the association with epigenetics factors. We hypothesize the low heritability phenotypes will have higher epigenetic contributions and therefore have stronger association with genetic variance in epigenetic factors.

**Specific Aim 4**: To generate the intermediate phenotypes: genome-wide DNA methylation level for PMRP samples.

To systemically build the genome-wide DNA methylation profile changes to a widely phenotypes which have disclosed genetic architectures. We will build the representative genome-wide DNA methylation levels to PMRP phenotypes with bisulfite sequencing to LINE-1/SINE methylation levels. Genome-wide DNA methylation difference will be compared among all the PMRP phenotypes.

**Background**

In the past decades, population genetics based genome-wide association study has been unprecedentedly developed to identify the susceptibility genes for human complex disease. Thousands of susceptibility variants or genes were identified GWAS or candidate genes based fine mapping strategy. However, my previous study showed the prediction ability was severely limited with significant SNPs identified by GWAS study even for some high familial risk disease. Recently, some evidences shown genetic risk score would help to have improved distinguish ability between genetic disease and normal individuals, however, the effect were still limited. One the other side, majority of the GWAS identified significant SNPs were non-coding variants, which are locating in intergenic, intron or UTR regions. These evidences demonstrate we need pay more attention to human epigenetic variants and these variants might play quite important roles in the susceptibility and pathology of human disease.

Among all the epigenetics variants, DNA methylation is the most common investigated factors since the technique is relatively matured and the biological function of DNA methylation is deeply understudied. In order to investigate these epigenetic variants, Genome-wide epigenetic association study (EWAS) or genome-wide DNA methylation association study have been conducted in some mediate or low heritability human disease, such as rheumatoid arthritis, type 2 diabetes, obesity and human cancers. EWAS study could provide the fundamental evidence that whether human epigenetics play important roles for the pathology of the disease and then more case-control study can be conducted to find exact epigenetic variants in disease-origin tissues, which is determined by the truth that the epigenetic profiles have strong tissue-specificity. However, It is difficult to be extend to more diseases since the cost of EWAS is almost 10 times higher compared with GWAS study. What’s more, GWAS study could apply linkage disequilibrium to select tag-SNPs to cover the whole genomic regions, however, EWAS cannot rely this mechanism. Current, even by the most latest DNA methylation 850K array, we can only cover 3.01% total CpGs in human genome. In this study, we will evaluate the roles of epigenetic variants in more than 6,221 human phenotypes by investigate the genomic variants in the whole panel of epigenetic factors which including about 250 human epigenetic modification factors, such as DNMT1, DNMT3A/3B, TET1/2, DOT1L and so on.

The PMRP is a biobank of 20,000 participants with stored DNA, plasma and sera with accompanying electronic health records, the individuals of which have been drawn from a genetically homogeneous population in Central Wisconsin. In 2012-2013, Dr. Schrodi contributed externally-derived funds and worked with the MCRF core laboratory, Terrie Kitchner and a team of researchers at the University of Michigan to select 10,000 PMRP individuals and subject their DNAs to the Illumina exome beadchip. This genotyping array was designed to have excellent coverage of the exome down to fairly rare allele frequencies. As this genotyping array covers rare variants and the population studied is an extended kinship, this sample set is well-suited for linkage analyses to identify novel disease genes. Additionally, given that this sample set also directly targets the exome, the sample set is also highly germane for a compound heterozygosity scan.

Dr. Schrodi and Dr. Guo have recently performed several different types of studies on this initial sample set of 10,000 individuals, including compound heterozygote test, IBD scanning test. We initially developed phenotype algorithms for 13 diseases (cases and controls) using combinations of laboratory test results and ICD-9/ICD-10 codes. These algorithms were implemented by Brent Olson. These 13 diseases are taken as preliminary data to prepare the code, power analysis and PheWAS pre-analysis. For the full-size of the phenotype definition, it will be much easier since a routine ICD9 based case-control classify algorithms have been widely use in PMRP association research (details see below).

Previously, we have found that, for numerous diseases, simply using diagnostic codes was insufficiently accurate in identifying cases and controls. Additionally, we explored the impact of this misclassification on genetic association studies, demonstrating a strong inflation in type I and II error rates for marginal levels of misclassification.22 This was the impetus behind supplementing ICD-9 codes with laboratory test data or other medical measurements where available. Subsequently, we investigated both the shared chromosomal regions in those diseases and subjected those data to exome-wide compound heterozygosity scans. These efforts have produced several highly interesting findings with statistical support. That said, replication of these findings in the 8,400 sample set is critical for determining which of our initial findings are true positives. Doing so, will not only generate novel genes involved in the pathogenesis of these diseases but will also validate the identity-by-descent and compound heterozygosity approaches. Neither Dr. Schrodi nor Dr. Maadooliat have research funds. Therefore we cannot support a programmer analyst applying the previously implemented algorithms to the new set of 8,400 PMRP samples to determine which of these individuals has each of these 13 diseases and which can serve as controls.

**Significance**

PheWAS based association between 6,221 PMRP phenotypes with a while panel of epigenetic factors will identify novel susceptibility genes to improve the understanding to the corresponding phenotypes. Meanwhile, it will provide novel therapy target to these phenotypes. What’s more, it will infer the epigenetic contribution for each phenotypes, which provided the fundamental evidence to the necessary for the further genome-wide epigenetic association study to identify novel diagnosis and prognosis biomarkers. On the other side, the genome-wide DNA methylation levels created in this study will also provide a novel analysis to investigate the relationship between LINE-1 methylation with genetic variants, which might bring insights to the genetic variations associated with aging or longevity.

**Preliminary Studies**

Dr. Schrodi and Dr. Guo have completed several preliminary aspects of the study, including 1) power calculations of statistical test of compound heterozygosity under a wide spectrum of the parameter space in comparison with orthodox statistical tests used in GWAS, 2) quality control analyses using principal components analysis on the 8,400 PMRP samples with exome beadchip genotype data, 3) the development of a preliminary software tool that performs basic analyses of identity-by-descent shared chromosomal region analyses and visualization across the genome, 4) haplotype phasing of all individuals from the initial phase of PMRP samples genotyped on the exome beadchip (n=10,000) and subsequent exhaustive determination of pairwise identity-by-descent regions, 5) conducting an exome-wide compound heterozygote scan for iron overload in the initial phase of PMRP samples, and 6) conducting an exome-wide scan of shared chromosomal regions in rheumatoid arthritis cases compared to controls using the initial phase of PMRP samples.

The phenome was defined by patient EHR data as Dr. Hebbring previously applied. Briefly, ICD9 codes were used to define cases and controls at varying levels of phenotypic resolution using a roll-up strategy (e.g. ICD9 720.89 ✧ 720.8\*✧720.\*). Patients coded for any one specific code (e.g., ICD9 720.89) became “cases” for that code, whereas those not coded for the specific code or related codes (e.g., ICD9 720\*) became “controls.” For common ICD9 codes (≥300 individuals), cases were defined by those coded two or more times (“rule-of-two”); those coded only once were not considered a case or a control. For less frequent ICD9 codes (<300 individuals), all individuals coded for that ICD9 code were designated as a case. As requested by Marshfield Clinic’s Institutional Review Board, case status was not defined for rare ICD9 codes (<9 individuals) to protect patient privacy as PMRP participants originate from a very specific region in Central Wisconsin.

***Power calculations***

**Research Design and Methodology**

To address Specific Aim 1, we will work with a programmer analyst to apply the previously developed phenotyping algorithms (applied to the set of 10,000 exome genotyped PMRP individuals several years ago) to the set of 8,400 PMRP individuals recently genotyped on the Illumina exome beadchip.

**Timeline**

Year 1: Obtain Marshfield IRB approval for the study. Partition exome data for the PMRP into data types (exome variants, GWAS-significant SNPs, AIMs). Define the PMRP cohort to be studied by applying exclusion criteria and performing the PCA analysis on the exome beadchip data to remove outliers. Identify cases and controls for all diseases with aid from Dr. Smith. Send plasma from 50% of the samples to Eve Technologies to run the Human Cytokine Array/Chemokine Array 42-Plex (HD42) and the Human Metabolic Hormone Array 9-Plex. Send plasma from 50% of the samples to Antigen Discovery for interrogation on the 800 antigen protein microarray platform.

Year 2:

**Budget**

**Budget Justification**

**Literature Cited**

Shicheng Guo is supported by an NLM training grant to the Computation and Informatics in Biology and Medicine Training Program (NLM5T15LM007359).