Genome-wide DNA Methylation profiling in Personalized Medicine Research Project (PMRP)

SP Code GUO30119

Abstract

DNA methylation have been widely observed to be abnormal in human cancers. However, diagnosis and prognosis biomarker panel have not been well-established. In order to accelerate biomarkers into clinical application, I proposed two biotechnology methods MBD-seq and MSD-SNuPET for methylation-based biomarker identification and validation. Methylation-binding domain-sequencing (MBD-seq) was applied to identify cancer differential methylation regions and these DMR were validated and the best biomarker panel was tested in different clinical scenarios: case-control study and double-blind clinical study. By this method, I identified DNA methylation biomarkers for bladder cancer including 5 diagnostic biomarker (VAX1, KCNV1, TAL1, PROX1 and CFTR) and 2 recurrence associated biomarker (VAX1 and LMX1A) and 2 differentiation related biomarker (ECEL1 and TMEM26). In the second study, I proposed low cost DNA methylation dection method: methylation status determined single nucleotide primer extension technique (MSD-SNuPET) to valiate the differential methylation loci identified from public methylation microarray data (such as HM450K array). I applied this method to validate the prediction model build by TCGA lung cancer dataset and built a five-gene diagnosis panel. The panel showed very high prediction performance in different machine learning models such as support vector machine (SVM), random forest (RF) and logistic regression (sensitivity=78%, specificity=97% and AUC=0.91).

Phenome-wide Association Study of Genetic Variation in Epigenetic Factors to Test the Role of Epigenetics in Human Complex Disease

Human complex disease is generated by the interaction between genetics, epigenetics and environment. While the rationale for genetic association studies have been supported by different fundamental observations such as heritability estimation from twin studies, there is no fundamental research to illustrate whether epigenetic changes are involved in disease phenotypes, although we know that epigenetic elements are an importance interface between the interaction of genetics and environment. In this study, I hypothesize that genetic variants in epigenetic genes are a proxy to infer the epigenetic involvement in phenotypes. In this study, 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.

In the past decade, population-based genome-wide association studies (GWAS) have been developed to identify the susceptibility genes for human complex disease. These studies have successfully identified thousands of disease-related susceptibility variants. However, these variants can only explain a small part of the total heritability and clinical use of these findings to distinguish between disease and normal individuals is currently limited. Additionally, the majority of the GWAS identified significant SNPs are non-coding variants (i.e., located in intergenic, intronic or UTR regions). These evidences indicate epigenetic variants might play important roles in disease etiology.

In order to investigate the roles epigenetic variants in human disease, genome-wide epigenetic association study (EWAS), especially genome-wide DNA methylation association study, have been conducted in human diseases, such as rheumatoid arthritis (RA), type 2 diabetes, obesity and human cancers. However, extending these studies to a large number of diseases can be prohibitively expensive. Moreover, even the most high-throughput DNA methylation microarray only covers about 3.01% of the total CpGs in the genome which significantly increases the difficulty of EWAS since they cannot rely on linkage disequilibrium to select tag-SNPs. Therefore, assessment of whether epigenetic mechanisms are involved in the phenotypes represents an important scientific advance. In this study, we will evaluate the roles of epigenetic variants in more than 6,221 clinical phenotypes by investigating the genetic variants in a panel of ~250 important epigenetic genes, such as DNMT1, DNMT3A/3B, TET1/2, and DOT1L, and test for the association with full statistical approach to infer the epigenetic contribution to these phenotypes.

To achieve our objectives, we will conduct a two-staged association study. In the discovery stage, we will conduct a phenome-wide association study (PheWAS) to test the set of variants within the panel of epigenetic factors against the 6,221 clinical traits within the Marshfield Clinic PMRP phase I dataset (10,124 samples). In the validation stage, PMPR phase II dataset (8,258 samples) will be taken as a repeat dataset. Genome-wide exome genotype data is available on all 18,382 samples. All variants within the epigenetic factors and variants in high linkage disequilibrium with variants within those factors will be extracted. Additionally, all variants found to be correlated with the expression of the epigenetic factors (trans-eQTL) (in public databases) will be included in the analysis. My preliminary analysis shows that 5,873 SNPs satisfy these criteria. Using imputed genetic data on the measured genotypes (using the Beagle software package) will further increase this set of variants to ~50k. The phenotype information on all samples will be extracted from the ICD9/ICD10 data within the Marshfield Clinic electronic health record.

We will apply different models of inheritance in the association testing of epigenetic variants with the phenotypes (dominant, recessive, additive, gene-based weighted burden, and compound heterozygote models will be evaluated). Significant variant/gene and phenotypes identified in the discovery dataset will be re-tested in the independent validation dataset. Further bioinformatics validation will also be considered, such as 1) gene expression specificity patterns of significant genes in corresponding tissue by the public gene expression database such as GTEx, BLUEPRINT, FANTOM5, and the TCGA data. 2) apply existed eQTL data, such as GTEx-eQTL and the Chicago eQTL database, to illustrate the function mechanisms of the significant SNVs identified by our study in disease susceptibility and pathology.

Furthermore, I will generate the genome-wide DNA methylation level data to the significant associated phenotypes for PMRP samples with different strategies, including: 1) genome-wide DNA methylation representation by LINE-1, SINE-1 and Alu methylation status. 2) illumina methylation 450K microarray and 3) Reduced representation bisulfite sequencing (RRBS). The first method have quite low cost and can be applied in 100 phenotypes and 1000 samples. The second method is more expensive can be applied only to validated phenotypes. RRBS method can be applied for the phenotypes with the most solid findings since it can provide single base resolution for DNA methylation. In this stage, we will also consider conducting a DNA methylation assay to certain solid tissues since there is tissue specificity of the DNA methylation profiles. Tissue-of-origin DNA methylation profile of the significant associated phenotypes will be bisufilte sequenced with RRBS to evaluate genome-wide DNA methylation and differential DNA methylation regions (DMR).

At the conclusion of this study, we will have a multi-omic dataset, including genotype and genome-wide DNA methylation level (LINE-1, SINE-1, Alu), for multiple phenotypes. Therefore, we will try to construct a genetic and epigenetic feature-based multi-class prediction to distinguish different phenotypes. We will compare the performance of different machine learning methods, such as random forest, support vector machine and neural networks. We will investigate different features (allele, compound heterozygosity or genetic risk score) and use a feature selection strategy to achieve the better prediction model.

With this project, we will also investigate the relationship between phenotype heritability from UK biobank and with association status between the phenotype and epigenetics factors. We hypothsize the low heritability phenotypes maybe be have more signficant associations. In our preliminary analysis, we indeed found that lower heritability phenotypes will have more significant associated epigenetics factor genetics variants through the comparison among rheumatoid arthritis (h2=0.6), obsessive compulsive disorders (h2=0.29) and iron overload (h2=0.25-0.4). Meanwhile, Since we will generate genome-wide DNA methylation level for parts of PMRP samples. We can also take genome-wide methylation level as an intermediate phenotype and conduct a genome-wide association to identify genetic variants to be associated with baseline level of genome-wide DNA methylation adjusted by gender and age to identify genome-wide methylation related genetic variants.

Overall, this project will provide a most compressive association study between epigenetic factors over 6,000 human phenotypes and provide the fundamental clues for the necessary of the genome-wide epigenetic association study (EWAS) for specific phenotypes. We will also provide a multi-class phenotype prediction model based on genome-wide genetic and epigenetic features to improve the disease distinguish or diagnosis ability.

**Preliminary Data**

Power analysis to determine the minimum sample size for each phenotype

We collected LINE-1 methylation targeted methylation sequencing data from my previous published paper and investigated the variation between different normal samples including esophagus (N=91) [1], bile duct (N=98).

1. Jiang, D., et al., *Epigenetic silencing of ZNF132 mediated by methylation-sensitive Sp1 binding promotes cancer progression in esophageal squamous cell carcinoma.* Cell Death Dis, 2018. **10**(1): p. 1.