**Genome-wide DNA methylation analysis uncovers novel epigenetic changes in human atrial fibrillation**

1Center for Precision Medicine Research, Marshfield Clinic Research Institute, Marshfield, Wisconsin, USA

2Biomedical Informatics Research Center, Marshfield Clinic Research Institute, Marshfield, Wisconsin, USA

Abstract

Atrial fibrillation is a complex disease caused by genetic and epigenetic interaction. DNA methylation is critical for dysregulated gene expression in atrial fibrillation. However, the genome-wide DNA methylation landscape and abnormal functional DNA methylation regions remains unclear. In this study, we applied genome-wide DNA methylation array (EPIC) which covering 868,564 probes and 99% human reference genes to investigate the methylation profile of atrial fibrillation to 48 individuals from Marshfield Clinic Personalized Medicine Research Project. We identified xxx significant differential methylation regions which are located in *PAX2*, xx, xxx genes. The results of our study suggest that methylation signatures in the blood compartment are useful to be as valuable biomarkers for predicting atrial fibrillation. In addition, DNA methylation analysis provided an approach to validate genetic findings identified from genome-wide association study (GWAS).

Background

Genome-wide DNA methylation study to heart samples from AF identified 417 differential methylated CpG sites in a small Chinese Han population1.

**Result**

**Identification of PBMC derived differential DNA methylated CpGs of atrial fibrillation**

In our study, we included 865,918 probes with 411 control probes initially. Methylation probe detection failure ratio ranges from 7.6x10-5 to 2.2x10-4 which is qualified enough for further statistical and bioinformatics analysis (Table S0). We removed 120,390 probes whose detection p-value is not significant (N=1,057, P>0.01) or including SNPs within probes (N=98,544) or high-missing ratio probes (N=712) or the probes have multiple genomic hits (N=11) and 17,075 probes within chrX and chrY. Eventually, 745,528 probes were pass the quality control and went through the further analysis. BMIQ algorithm was applied for the inter-normalization as our previous study2. Through our dedicate study design and protocol, we did not observe batch effect with surrogate variable analysis (Figure S1).

We identified 10,731 (FDR<0.05) and 36 () significant DMCs between the PBMC derived from AF and health control (Table S2). We also tried blood cell-type composition adjustment was conducted to remove cell type heterogeneity confounding as previous study3. We estimated and corrected the influence from the different cell-type contribution from CD8-T, CD4-T, NK, B-cell, Monocytes, Granulocyte (Table S3). By blood cell corrected methylation data, we identified 29,167 DMCs with cell-type corrected methylation signals (Table S4). We observed 6 and 45 significant DMCs in old-sub-cohort and young-sub cohort (Table S8 and S9). Meanwhile, we observed 4,377 significant DMCs between old case and young control indicating the influences from the aging (Table S7). We didn’t observe signiant DMCs between young case and old case, young control and old control which is consistent with atrial fimbriation frequently occurred in older population. We didn’t observed any significant DMCs between case-before and case-after under multiple-test correction mode (q-value<0.05) since the limited sample size. However, we observed 19 DMCs between case-before and control-before (Table S10). In order to identify aging related DMCs, we compared methylation data between younger and older groups and found 25 significant DMCs (Table S6).

In order to avoid non-stable results from differential methylation CpGs, we conducted differential methylation regions analysis to Identify of PBMC derived differential DNA methylated regions (DMR) in atrial fibrillation. We identified 66 significant DMRs (Table S9) between atrial fibrillation and normal samples and 7 DMRs between old-case and old-control sub-cohort (Table S10). 3 DMRs located in *XXYLT1* (13p29), *CELF5* (19p13.3), and *PPP2R2C* (4p16.1) were found between case-before and case-after. We didn’t identify significant DMRs in older sub-cohort and younger sub-cohort which might be caused by limited sample size. We mapped these DMRs to nearby GWAS significant hits and found that xxx, xx, xx have been identified in previous studies. We also identified 57 DMRs between overall younger and older groups that showed aging related DNA methylation changes (Table S7).

**Identification of functional epigenetic modules with atrial fibrillation methylation dataset**

**Methylation age estimation reveals accelerated aging of atrial fibrillation**

Correlation between DNA methylation and disease characteristics

Discussion

We conducted copy number variation calling to our methylation data. However, we didn’t identify significant CNV regions. Usually, cell-type composition should be adjusted for PBMC based DNA methylation study. However, inaccurate cell type estimation will create second-hand bias for the study. In our study, we observed cell-type corrected methylation signal identified more DMCs (N= 29,167) compared with raw-normalized methylation signals (N= 2,614) and only 30.6% of DMCs from raw methylation signals are repeated in DMCs identified by cell-type corrected methylation signals. Since all the previous cell-type composition estimation are based on methylation 450K and the performance to methylation EPIC (850K) array have never been evaluated.

Method

Bisulfite-converted atrial fibrillation patient and health control DNA samples were prepared and quantified using a NanoDrop scanning spectrophotometer (Thermo, Wilmington, DE). For each sample, 500 ng of whole-genome bisulfite-converted DNA was denatured, fragmented, amplified and methylation signals were detected by Illumina Infinium HumanMethylation850K (EPIC) BeadChip (Illumina, San Diego, CA). Standard DNA methylation 850K analysis pipeline was implemented to conduct the methylation microarray analysis. Genome Studio (Illumina) was used to generate signal intensities and detected p-values with internal control normalization (ICN) and background subtraction (BS). Quality control and normalization were conducted with R package lumi4. Probes with SNPs located in ChrX and ChrY were removed before further analysis. In addition, probes with a detection of p-value 0.01 exceeding 5% of the samples were also filtered out while other probes less than 5% of samples were labeled as missing value (NA) to avoid further bias in the following statistic and bio-informatics analysis. Then, the overall signal intensity, the distribution of M-values and the number of significantly detected sites were used to measure the quality of the beadchip. Obvious outlier samples and probes were removed before the differential methylation loci identification. Color-bias adjustment and quantile normalization (QN) were performed on signal intensities with the package of ‘‘lumi’’. Finally, beta-mixture quantile normalization (BMIQ) to the b-values were conducted to adjust the bias caused by different types of probes (type I and type II). Statistical analysis PCA and Hierarchical cluster analysis were applied to show the correlation between the samples. Two RA samples were filtered since they were obviously different with others in the PCA analysis. Differentially methylated loci were identified by paired t-test based on the beta value of the normalization data. The q-value was applied for DMR and DMC identification (FDR<0.05) for multiple test correction. Association between clinical characteristics and differential methylation loci were conducted with linear regression with the significant threshold of 0.005. Gene ontology analysis was conducted with DAVID bioinformatics resources5. The interactions among the differential methylated genes were inferred by STRING11.06. Reactome pathway analysis was conducted by R package [ReactomePA](http://bioconductor.org/packages/ReactomePA)7. Human references of GRCh37/hg19 were used in the bioinformatics analysis and result representation. All methods and analyses were performed in R (version 3.5.2) and BiocManager (version 3.8). The data were deposited in the Gene Expression Omnibus (GEO accession: GSExxx). False discover rate (FDR, q-value) from Benjamini-Hochberg procedure was applied for significant DMC identification in for genome-wide DMC multiple comparison test8.

as shown in (Andrew E Jaffe and Rafael A Irizarry 2014), biological findings in blood samples can often be confounded with cell type composition. In order to estimate the confounding levels between phenotype and cell type composition, the function estimateCellCountsdepending on the package [FlowSorted.Blood.450k](http://bioconductor.org/packages/release/bioc/html/FlowSorted.Blood.450k.html)estimates the cell type composition of blood samples by using a modified version of the algorithm described in (E Andres Houseman et al. 2012)3. The function takes as input a RGChannelSet and returns a cell counts vector for each samples:

**Gene Set Enrichment Analysis**

The mutational landscape of ccRCC was explored through gene set enrichment analysis (GSEA), which was performed on those genes with p < 0.05 in the Fisher’s exact test of mutation burden (described above). This was performed via the Enrichr tool (Chen et al., 2013; Kuleshov et al., 2016) focusing on disease links

through the Jensen Diseases database, which compiles evidence of gene–disease associations through the analysis of existing literature on genetic studies. A sensitivity analysis was performed at this stage by subsampling 71,000 random subsets of variable size (30–100 genes) from the prioritized gene set that were then

evaluated in EnrichR, recording the proportion of times the terms were significant in the adjusted tests.

Result

1. Zhao, G. *et al.* Genome-wide DNA methylation analysis in permanent atrial fibrillation. *Mol Med Rep* **16**, 5505-5514 (2017).

2. Guo, S. *et al.* Genome-wide DNA methylation patterns in CD4+ T cells from Chinese Han patients with rheumatoid arthritis. *Mod Rheumatol* **27**, 441-447 (2017).

3. Houseman, E.A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).

4. Du, P., Kibbe, W.A. & Lin, S.M. lumi: a pipeline for processing Illumina microarray. *Bioinformatics* **24**, 1547-8 (2008).

5. Huang da, W., Sherman, B.T. & Lempicki, R.A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* **4**, 44-57 (2009).

6. Szklarczyk, D. *et al.* STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* **47**, D607-D613 (2019).

7. Yu, G. & He, Q.Y. ReactomePA: an R/Bioconductor package for reactome pathway analysis and visualization. *Mol Biosyst* **12**, 477-9 (2016).

8. Reiner, A., Yekutieli, D. & Benjamini, Y. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19**, 368-75 (2003).