Bioinformatics, 32(12), 2016, 1773–1778
doi: 10.1093/bioinformatics/btw089
Advance Access Publication Date: 15 February 2016
Original Paper

Genome analysis

Computationally expanding infinium HumanMethylation450 BeadChip array data to reveal distinct DNA methylation patterns of rheumatoid arthritis

Shicai Fan[1,2], Chengzhe Li[1], Rizi Ai[2], Mengchi Wang[2], Gary S. Firestein[3,*] and Wei Wang[2,*]

[1]School of Automation Engineering, University of Electronic Science and Technology of China, Chengdu, China, [2]Department of Chemistry and Biochemistry and [3]Division of Rheumatology, Allergy and Immunology, University of California San Diego, La Jolla, CA, USA

*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on November 17, 2015; revised on January 9, 2016; accepted on February 11, 2016

Abstract

Motivation: DNA methylation signatures in rheumatoid arthritis (RA) have been identified in fibroblast-like synoviocytes (FLS) with Illumina HumanMethylation450 array. Since <2% of CpG sites are covered by the Illumina 450K array and whole genome bisulfite sequencing is still too expensive for many samples, computationally predicting DNA methylation levels based on 450K data would be valuable to discover more RA-related genes.

Results: We developed a computational model that is trained on 14 tissues with both whole genome bisulfite sequencing and 450K array data. This model integrates information derived from the similarity of local methylation pattern between tissues, the methylation information of flanking CpG sites and the methylation tendency of flanking DNA sequences. The predicted and measured methylation values were highly correlated with a Pearson correlation coefficient of 0.9 in leave-one-tissue-out cross-validations. Importantly, the majority (76%) of the top 10% differentially methylated loci among the 14 tissues was correctly detected using the predicted methylation values. Applying this model to 450K data of RA, osteoarthritis and normal FLS, we successfully expanded the coverage of CpG sites 18.5-fold and accounts for about 30% of all the CpGs in the human genome. By integrative omics study, we identified genes and pathways tightly related to RA pathogenesis, among which 12 genes were supported by triple evidences, including 6 genes already known to perform specific roles in RA and 6 genes as new potential therapeutic targets.

Availability and implementation: The source code, required data for prediction, and demo data for test are freely available at: http://wanglab.ucsd.edu/star/LR450K/.

Contact: wei-wang@ucsd.edu or gfirestein@ucsd.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Rheumatoid arthritis (RA) is an autoimmune disease marked by synovial hyperplasia and invasion into cartilage and bone (Firestein, 2003). Fibroblast-like synoviocytes (FLS), which form the inner lining of the synovium, display an aggressive phenotype in RA that persists in long-term culture. The mechanism that contributes to functional alterations in RA FLS is only partially understood.

Recent studies have applied Illumina HumanMethylation450 BeadChip array to reveal distinct DNA methylation patterns that distinguish RA samples from osteoarthritis (OA) and normal (NL) FLS. Integrative analysis on the differentially methylated loci (DML), the associated genes and relevant pathways has provided insightful clues for identification of therapeutic targets (Ai *et al.*, 2015; Ekwall *et al.*, 2015; Nakano *et al.*, 2013; Whitaker *et al.*, 2013, 2015). A significant limitation of the Illumina 450K array is that it only covers less than 2% of CpG loci in the entire human genome. Expanding the coverage of CpG loci in defining the DNA methylation pattern of RA is becoming an urgent need.

Previously, computational methods have been developed to predict methylation values based on DNA sequences, histone modifications or integration of DNA methylation data from different techniques. Methods using only DNA sequence features can learn the methylation tendency for a CpG site based on its surrounding sequence compositions in a specific tissue (Bock *et al.*, 2006; Fang *et al.*, 2006; Feltus *et al.*, 2003; Feng *et al.*, 2014), but the trained model cannot be transferred to predict methylation levels in different tissues because the input sequence features remain the same and cell-type specific methylation obviously vary from one cell type to another. Histone modification features can reflect cell specificity and incorporating histone modifications allows prediction of cell-specific DNA methylation patterns (Bock *et al.*, 2009; Fan *et al.*, 2008; Zheng *et al.*, 2013). Using these models require histone modification data to be available in the cell type of interest, which is not always the case. Recently computational methods were developed to infer DNA methylation levels in the entire genome using MeDIP-seq and MRE-seq data (Stevens, *et al.*, 2013). Since neither histone modification data nor MeDIP-seq/MRE-seq data are available for RA, a new computational method to expand the coverage of Illumina 450K array is no doubt valuable to further refine the DNA methylation signature of RA.

In this study, we developed a prediction model integrating cell-type specific 450K array data and common DNA-sequence features. The model was trained on 14 cell types/tissues that have both 450K array and whole genome bisulfite sequencing (WGBS) data as local DNA methylation patterns are similar between similar tissues (Byun *et al.*, 2009; Fan and Zhang, 2009), and the methylation status of a CpG in one tissue is correlated with or affected by its flanking CpG sites and sequence compositions (Lister *et al.*, 2009; Stadler *et al.*, 2011). The model aims to capture such local similarity of DNA methylation patterns across cell types/tissues. We performed cross validations to confirm the success of this method, with an average prediction correlation coefficient around 0.9, accuracy over 0.9, and AUC close to 0.9. Particularly, 70–80% of differential methylation loci were correctly retrieved using the predicted methylation levels.

We previously used 450K array data to identify a characteristic DNA methylation pattern in RA FLS, which are pathogenic cells that form the lining of the joint (Firestein, 2003; Whitaker *et al.*, 2013, 2015). These data implicated genes and pathways in the pathogenesis of RA, especially related to immunity, cell adhesion and matrix regulation. Applying this model to the original 28 FLS 450K array data, we expanded the CpG coverage to 8 555 846 sites, which is over 18-fold greater than the number of CpG sites covered by 450K array. Using the predicted methylation sites, we found 3874 genes differentially methylated between RA and OA/NL (referred as differentially methylated genes, DMGs). Combing these DMGs, genes differentially expressed between RA and OA/NL (DEGs), and RA-associated genes from Genome-Wide Association Study (GWAS) studies, we found 11 enriched KEGG pathways. Most of these pathways are related to immune system and expand upon with those found by 450K array data only. Twelve genes were supported by three-way evidences of DML, DMG and GWAS, among which half are related to RA such as HLA-DQA1, LBH and ELMO1 (Castro *et al.*, 2001; Ekwall *et al.*, 2015; Whitaker *et al.*, 2015).

2 Methods and results

2.1 Prediction model

We proposed the following strategy to predict the methylation levels of CpGs outside the Illumina 450K covered sites. This model is based on the following assumptions:

1) The methylation level of a CpG site in two cell types or tissues is similar if their flanking methylation patterns are similar (local methylation pattern);

2) The methylation level of a CpG site is related to its adjacent upstream and downstream CpG sites in the same cell type or tissue (neighbor CpG methylation levels);

3) The methylation level of a CpG site is related to its flanking DNA sequence composition.

Let variables x_1, x_2 and x_3 represent local methylation pattern, neighbor CpG methylation levels and methylation information derived from flanking DNA sequence composition, respectively. We constructed a logistic regression model $\beta = f(x_1, x_2, x_3)$ to predict the methylation values of a given CpG site (Fig. 1).

In Figure 1A, methylation patterns of tissue 1 to tissue n were measured by both WGBS and 450K array. In order to get the methylation value of the CpG locus l not covered by 450K array, we compared its local 450K methylation pattern of region R_1 with the n tissues, and selected the WGBS methylation value of locus l from the tissue which has the most similar local methylation pattern. The local methylation pattern of region R_1 was represented by the 450K methylation values of the closest 10 CpG loci, 5 upstream and downstream in the genome. The similarity was measured by the Pearson correlation coefficient between the local methylation patterns of different tissues. As the methylation level of locus l would be weakly correlated with CpG loci far away from it, we only considered the locus l within 5 kbp of a CpG site covered by 450K array.

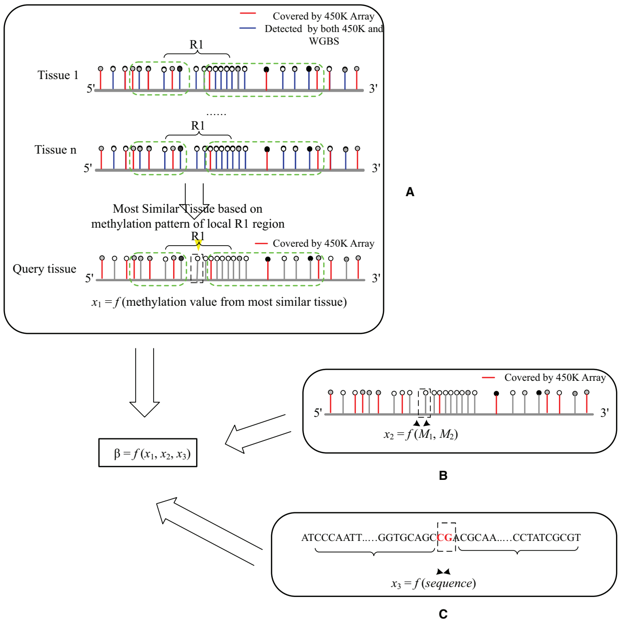

Covered by 450K Array
Detected by both 450K and WGBS
R1
Tissue 1
5'
3'
......
R1
Tissue n
5'
3'
Most Similar Tissue based on methylation pattern of local R1 region
R1
Covered by 450K Array
Query tissue
5'
3'
$x_1 = f$(methylation value from most similar tissue)
A
Covered by 450K Array
5'
3'
$x_2 = f(M_1, M_2)$
B
$\beta = f(x_1, x_2, x_3)$
ATCCCAATT.....GGTGCAGCCGACGCAA.....CCTATCGCGT
$x_3 = f(sequence)$
C

Fig. 1. The work flow of the computational expansion strategy (Color version of this figure is available at *Bioinformatics* online.)

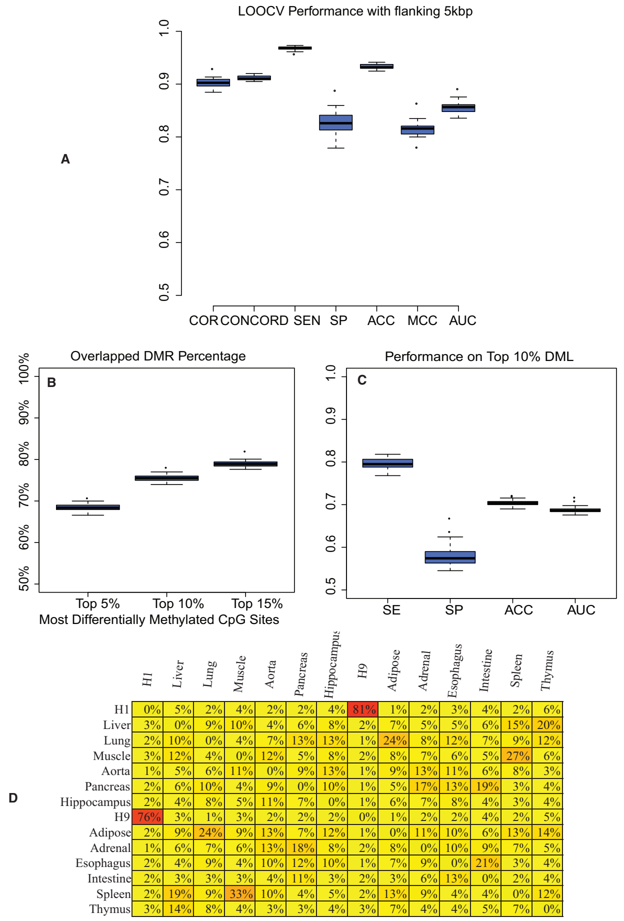
In Figure 1B, x_2 of CpG locus l is the weighted450K methylation values of its closest upstream and downstream CpGs. The normalized weights of upstream and downstream CpG are inversely proportional to its genomic distance to CpG locus l. To model the relationship between DNA methylation and sequence composition, we trained a sub-model before merging the three variables into the final model to avoid the large number of sequence feature dominating the final model (Fig. 1C). We extracted 362 features including NpN ratio, NpN content (N represents any nucleotide, i.e. A/G/T/C) and 1- to 4-mers occurrence frequencies (Fig. 1C), where NpN content = (#N + #N)/len(Sequence), and NpN ratio = (#NpN × len(Sequence)))/(#N×#N). Then we performed feature selection using random forest. For each chromosome of each tissue, 3-fold cross-validation was repeated 10 times, and we recorded the 50 most frequently selected sequence features. Support vector regression was used to construct the sub model to obtain the methylation information x_3 as it showed better prediction results based on DNA-sequence features (Bock *et al.*, 2006; Fan *et al.*, 2008; Fang *et al.*, 2006).

2.2 The model showed superior performance in leave-one-tissue-out cross validations

We retrieved 14 tissues or cell lines that have both WGBS and 450K array data, including adipose, adrenal, aorta, esophagus, H1, H9, hippocampus, intestine, liver, lung, muscle, pancreas, spleen, thymus, generated by the NIH epigenomics roadmap project (Bernstein *et al.*, 2010). The methylation proportion values of WGBS data and beta values of 450K array were downloaded from the GEO Database directly. Both WGBS data and 450K array data were quantile normalized. The performance of our strategy was assessed by leave-one-tissue-out cross validation on all the 22 autosomes. The evaluation metrics included Pearson correlation coefficient, Concordance (the percent of CpGs with a methylation proportion difference <0.25 (Harris *et al.*, 2010), Sensitivity (SE), Specificity (SP), Accuracy (ACC), Matthew's correlation coefficient (MCC) and AUC (Area Under ROC Curve). For calculating SE, SP, ACC and MCC, we defined the methylation status as +1 if the methylation value is larger than 0.5, and the methylation status as −1 otherwise.

In leave-one-tissue-out cross validation, we trained the prediction model on the remaining 13 tissues, and evaluated the prediction performance on the left-out tissue. Figure 2A shows that the Pearson correlation coefficients between the predicted and measured values were 0.9025 ± 0.0093, indicating that the predicted methylation values are overall similar to the WGBS measurements. The scatter plots between predicted and measured WGBS values of the 22 chromosomes (Supplementary Figure S1 in Supplementary Materials) showed that the majority of the dots distributed along the diagonal line. Consistently, high concordance (0.9103 ± 0.0040), SE (0.9684 ± 0.0039), SP (0.8260 ± 0.0224), ACC (0.9318 ± 0.0045), MCC (0.8160 ± 0.0167) and AUC (0.8565 ± 0.124) also demonstrated a satisfactory performance of our proposed strategy.

The most challenging but also the most meaningful prediction is to correctly identify DMLs across tissues. To define DMLs, we calculated the methylation value variation of each CpG site across the 14 tissues. CpG sites were sorted according to their methylation variation. The CpG locus with larger variations would be regarded to be more differentially methylated. We called DMLs using both measured WGBS and predicted DNA methylation values. The overlapping of measured and predicted DMLs among the 14 tissues are shown in Figure 2B. In the most distinguished 5, 10 and 15% of DMLs called by the WGBS data, 68.38 ± 0.93%, 75.51 ± 0.95% and 78.91 ± 0.92% were correctly identified by DML called by the predicted DNA methylation values, respectively. When focusing on the CpG loci corresponding to the top 10% DMLs, we got the average prediction ACCof 0.7033 ± 0.0073 and AUC of 0.6856 ± 0.0092, respectively (Fig. 2C).



	H1	*Liver*	*Lung*	*Muscle*	*Aorta*	*Pancreas*	*Hippocampus*	*H9*	*Adipose*	*Adrenal*	*Esophagus*	*Intestine*	*Spleen*	*Thymus*
H1	0%	5%	2%	4%	2%	2%	4%	81%	1%	2%	3%	4%	2%	6%
Liver	3%	0%	9%	10%	4%	6%	8%	2%	7%	5%	5%	6%	15%	20%
Lung	2%	10%	0%	4%	7%	13%	13%	1%	24%	8%	12%	7%	9%	12%
Muscle	3%	12%	4%	0%	12%	5%	8%	2%	8%	7%	6%	5%	27%	6%
Aorta	1%	5%	6%	11%	0%	9%	13%	1%	9%	13%	11%	6%	8%	3%
Pancreas	2%	6%	10%	4%	9%	0%	10%	1%	5%	17%	13%	19%	3%	4%
Hippocampus	2%	4%	8%	5%	11%	7%	0%	1%	6%	7%	8%	4%	3%	4%
H9	76%	3%	1%	3%	2%	2%	2%	0%	1%	2%	2%	4%	2%	5%
Adipose	2%	9%	24%	9%	13%	7%	12%	1%	0%	11%	10%	6%	13%	14%
Adrenal	1%	6%	7%	6%	13%	18%	8%	2%	8%	0%	10%	9%	7%	5%
Esophagus	2%	4%	9%	4%	10%	12%	10%	1%	7%	9%	0%	21%	3%	4%
Intestine	2%	3%	3%	3%	4%	11%	3%	2%	3%	6%	13%	0%	2%	4%
Spleen	2%	19%	9%	33%	10%	4%	5%	2%	13%	9%	4%	4%	0%	12%
Thymus	3%	14%	8%	4%	3%	3%	4%	3%	7%	4%	4%	5%	7%	0%

Fig. 2. Evaluating the model performance using leave-one-tissue-out cross validations. (A) Pearson correlation, Concordance, SE, SP, ACC, MCC and AUC on 22 autosomal chromosomes.Prediction performance of detecting DML are in (b) and (c). (B) The overlap between the DMLs called based on WGBS and predicted DNA methylation data in the 14 tissues. (C) The SE, SP, ACC and AUC of the prediction model for the top 10% DMLs called by the WGBS data. (D) The percentage of the most predictive tissue

We also investigated whether one tissue was always selected for predicting another tissue. Figure 2D shows that the most predictive tissues/cell line for the query tissue/cell was not dominant by a single tissue/cell line, except H1 and H9. More than 75% of CpG loci of H1 and H9 shared the most similar local methylation patterns, which is not surprising because both H1 and H9 are human embryonic stem cell lines. The selected percentages of other tissues are relatively evenly distributed, which indicates no biased selection of predictive WGBS data in the model based on tissue/cell similarity.

2.3 Predicted RA methylation patterns

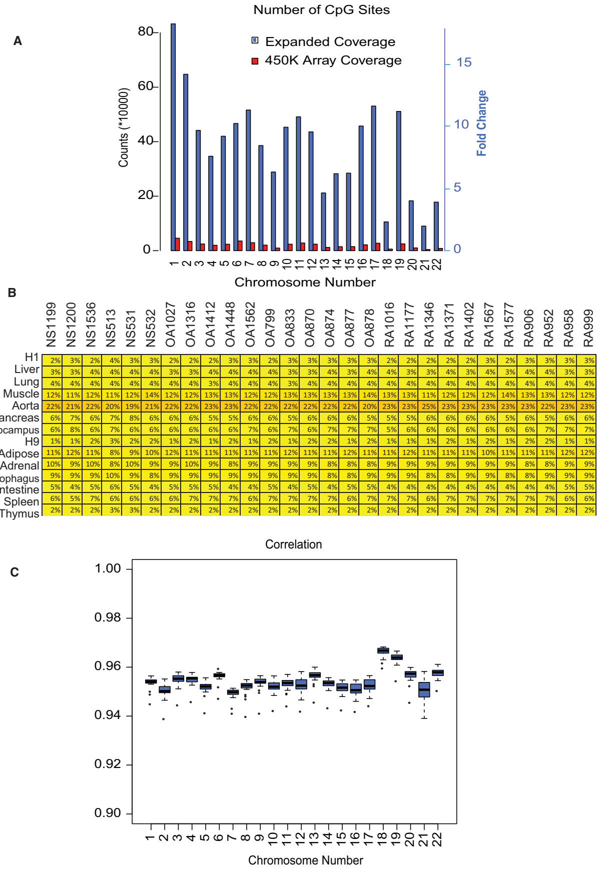
Illumina 450K array data on 11 RA, 11 OA and 6 NL samples were generated in our previous studies (Whitaker *et al.*, 2013). Applying the prediction model to FLS samples, we predicted methylation values for the CpG sites within 5 kbp of any 450K CpG. The overlapped CpG sites in the 14 tissues/cell lines and 28 FLS samples after array quality filtering were 462 105. The number of predicted CpG sites was 8 555 846, which is 18.5 times of the sites covered by 450K array and about 30% of all the CpGs in the human genome. The expanded number of CpGs in each of the 22 autosomes is shown in Figure 3A. Consistent with the cross validations, the predictive tissues for each RA/OA/NL tissue based on the local methylation pattern were wide spread over the 14 tissues (Fig. 3B), which further confirmed that the selection was not biased. For the validation purpose, we also predicted the methylation levels of CpGs covered in 450K array using our model and then calculated the correlations between the predicted and measured values (Fig. 3C). All the Pearson correlation coefficients were around 0.95, confirming the success of the prediction model.

2.4 Identification of RA-related genes and pathways

We collected the original 450K data and the expanded CpG methylation values from our predictions, based on which we aimed to identify the DMGs between 11 RA, 11 OA and 6 NL samples. Welch's *t*-test was used to calculate the *P*-values of CpGs located in promoter regions [(TSS-2500 bp, TSS + 500 bp)] by comparing RA versus OA, RA versus NL and RA versus (OA + NL). For each CpG locus, the lowest *P*-value in the three pairs of comparison tests was selected. For each promoter, the Fisher's combined test was used to evaluate whether a gene is differentially methylated. Then *P*-values were adjusted to *q*-values. 3874 genes with *q*-value < 0.05 and mean difference of DNA methylation >0.1 were selected as DMGs.

Next, we performed integration analysis from the expanded DNA methylation data, gene expression data and GWAS studies to identify gens whose relevance to RA are supported by multiple lines of evidences. For Differentially Expressed Genes (DEGs) in RA, we downloaded the microarray data of 9 RA, 11 OA and 11 NL FLS samples from GEO database (Del Rey *et al.*, 2012) (GEO ID GSE29746). Using the same processing method in our previous work (Whitaker *et al.*, 2015), we took genes >2-fold change in expression and *P*-value < 0.05(Welch's *t*-test), and found 2947 DEGs. Furthermore, we collected GWAS genes from reference Hindorff *et al.* (2009) and a recent meta-analysis of over 100 000 cases and controls (Okada *et al.*, 2014).

There were 484 genes supported by two evidences (Fig. 4A). We first analyzed the GO terms of these 484 genes. Totally, there were 8 enriched GO Molecular Function (MF) terms, 25 cellular component (CC) terms and 132 biological process (BP) terms. Some enriched GO termed related to RA are listed in Table 1 and others are given in the Supplementary Materials. We next analyzed the enrichment of KEGG pathways (Kanehisa and Goto, 2000) in these genes. The enrichment *P*-value was calculated with hypergeometric distribution and then adjusted to *q*-values. Pathways with *q*-value < 0.05 were considered to as significantly enriched. This way we found 11 enriched pathways (Fig. 4B). Among them, five are related to human



	NS1199	NS1200	NS1536	NS513	NS531	NS532	OA1027	OA1316	OA1412	OA1448	OA1562	OA799	OA833	OA870
H1	2%	3%	2%	4%	3%	3%	2%	2%	2%	3%	3%	2%	3%	3%
Liver	3%	3%	4%	4%	4%	4%	3%	3%	4%	4%	4%	4%	3%	3%
Lung	4%	4%	4%	4%	4%	4%	4%	4%	3%	4%	4%	4%	3%	4%
Muscle	12%	11%	12%	11%	12%	14%	12%	12%	13%	13%	12%	12%	13%	13%
Aorta	22%	21%	22%	20%	19%	21%	22%	22%	23%	23%	22%	22%	22%	22%
Pancreas	6%	7%	6%	7%	8%	6%	6%	6%	5%	5%	6%	6%	5%	6%
Hippocampus	6%	8%	6%	7%	6%	6%	6%	6%	6%	6%	7%	6%	7%	6%
H9	1%	1%	2%	3%	2%	2%	1%	2%	1%	2%	1%	2%	1%	2%
Adipose	11%	12%	11%	8%	9%	10%	12%	11%	11%	11%	11%	11%	12%	11%
Adrenal	10%	9%	10%	8%	10%	9%	9%	10%	9%	8%	9%	9%	9%	9%
Esophagus	9%	9%	9%	10%	9%	8%	9%	9%	9%	9%	9%	9%	9%	9%
Intestine	5%	4%	5%	6%	5%	4%	5%	5%	5%	4%	4%	5%	4%	5%
Spleen	6%	5%	7%	6%	6%	6%	6%	7%	7%	7%	6%	7%	7%	7%
Thymus	2%	2%	2%	3%	3%	2%	2%	2%	2%	2%	2%	2%	2%	2%

	OA874	OA877	OA878	RA1016	RA1177	RA1346	RA1371	RA1402	RA1567	RA1577	RA906	RA952	RA958	RA999
H1	3%	3%	3%	2%	2%	2%	2%	2%	3%	2%	3%	3%	3%	3%
Liver	4%	3%	4%	4%	3%	3%	4%	3%	4%	3%	4%	4%	3%	3%
Lung	4%	4%	4%	4%	4%	4%	4%	4%	4%	4%	4%	4%	4%	4%
Muscle	13%	13%	14%	13%	13%	11%	12%	12%	12%	14%	13%	13%	12%	12%
Aorta	22%	22%	20%	23%	23%	25%	23%	23%	23%	23%	23%	22%	23%	23%
Pancreas	6%	6%	5%	5%	5%	6%	6%	6%	6%	5%	6%	5%	6%	6%
Hippocampus	7%	7%	7%	5%	6%	6%	6%	6%	6%	6%	6%	7%	7%	7%
H9	2%	1%	2%	1%	1%	1%	2%	1%	2%	1%	2%	2%	1%	1%
Adipose	11%	12%	11%	11%	12%	11%	11%	11%	10%	11%	11%	11%	12%	12%
Adrenal	8%	8%	8%	9%	9%	9%	9%	9%	9%	9%	8%	8%	9%	9%
Esophagus	8%	8%	9%	9%	9%	8%	8%	9%	9%	8%	8%	9%	9%	9%
Intestine	4%	5%	5%	4%	4%	4%	4%	4%	4%	4%	4%	4%	5%	5%
Spleen	7%	6%	7%	7%	7%	6%	7%	7%	7%	7%	7%	7%	6%	6%
Thymus	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%	2%

Fig. 3. Prediction results on the RA data. (A) The expanded CpG sites of the RA data on the 22 automal chromosomes. (B) The percentage of predictive tissue for each FLS samples. (C) The correlations between predicted methylation values and detected methylation values with 450K array (Color version of this figure is available at *Bioinformatics* online.)

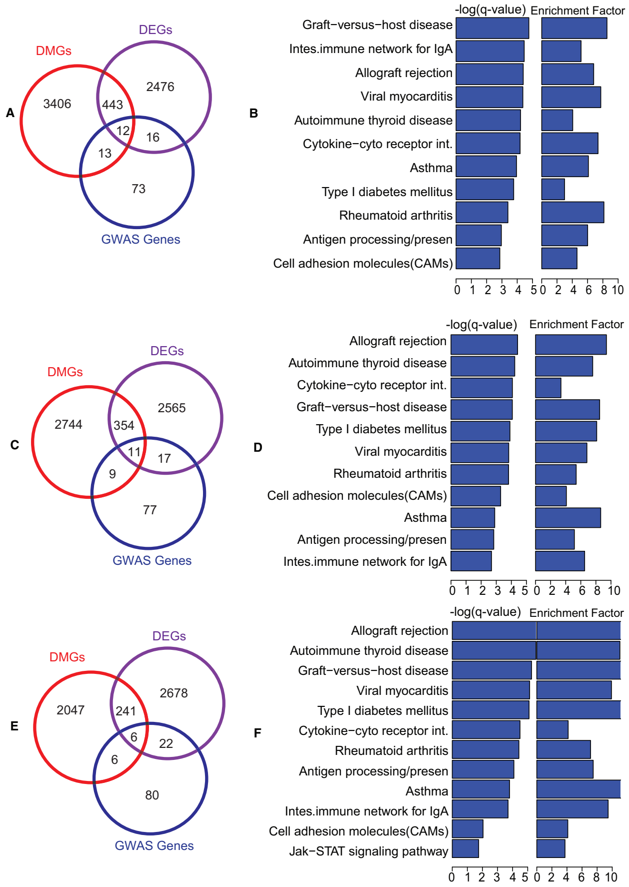


Fig. 4. Integrated Omics analysis on the RA data. (A and B) are results based on the 450K and the expanded CpG loci; (Cand D) are results based on only expanded CpG loci. (E and F) are results based on DMGs not found using 450K array data analysis (Color version of this figure is available at *Bioinformatics* online.)

Table 1. Selected enriched GO terms related to RA

GO terms	GO type	Enrichment	*q*-value
MHC class II protein complex	CC	>5	9.35E-03
Integral component of lumenal side of endoplasmic reticulum membrane	CC	>5	5.35E-03
Positive regulation of T cell activation	BP	4.81	5.03E-05
Inflammatory response	BP	2.76	2.44E-02
Immune response-regulating signaling pathway	BP	2.75	5.97E-04
Cytokine-mediated signaling pathway	BP	2.72	9.25E-03
Response to cytokine	BP	2.42	3.02E-03
Immune response	BP	2.23	1.68E-06
Immune system process	BP	2.06	1.42E-08
Signal transduction	BP	1.63	8.63E-09
Cytokine receptor binding	MF	3.37	8.70E-03

immune diseases: Graft-versus-host disease, Allograft rejection, Autoimmune thyroid disease, Asthma, RA, which covers 62.5% of all the 8 annotated immune disease related pathwaysin human; two of them are immune system pathways: Intestinal immune network for IgA production, Antigen processing and presentation; two are signaling pathways: cell adhesion molecules, Cytokine-cytokine receptor interaction. These genes and pathways are highly relevant to the pathogenesis of RA, which are consistent with our previous results based on only 450K array data (Whitaker *et al.*, 2015). Despite increasing the number of CpGs by 18-fold, differentially methylated pathways remained highly relevant to RA pathogenesis. Biologic validation of the newly identified genes will be important, but the fact that they are consistent with previously identified RA-associated genes suggests that the results are not random.

There were 12 genes supported by all three evidences, including IL23R, LBH, CASP8, HLA-DQA1, OLIG3, HLA-G, IRF5, ELMO1, TRHDE, SLCO1C1, PLD4, AIRE. Five of them overlapped with the seven triple-evidenced genes identified from 450K array data only analysis. More importantly, six of them have reported association with RA. HLA-DQA1 has known roles in RA (Castro *et al.*, 2001), LBH is a regulator of cell cycle in RA FLS and also a potential RA therapeutic target (Ekwall *et al.*, 2015) and ELMO1 contributes to the pathogenesis of RA as a regulator of FLS migration and invasion (Whitaker *et al.*, 2015); The *AIRE* gene was identified as a genetic risk factor for RA in a GWAS study(Garcia-Lozano *et al.*, 2013; Shao *et al.*, 2014; Terao *et al.*, 2011); IRF5 confers susceptibility to RA and influences its erosive phenotype (Dawidowicz *et al.*, 2011); GWAS replication study confirmed the association of PDE3A–SLCO1C1 with anti-TNF therapy response in RA in reference Acosta-Colman *et al.* (2013).

Among the remaining six genes, HLA-G was reported to be a candidate biomarker for prognosis and disease activity in early RA patients (Rizzo *et al.*, 2013); IL23R, which plays a role in Th17 cell differentiation, was a controversial gene: one group found association of two IL23R SNPs with RA in Hungarian population (Farago *et al.*, 2008; Szabo *et al.*, 2013), while in a Spanish study no association was detected (Orozco *et al.*, 2007); OLIG3 was reported for association with susceptibility and severity in an inception cohort in Morgan *et al.* (2010), while a later study found it was not associated with the severity of joint destruction in RA (Knevel *et al.*, 2012); For TRHDE, there was only one work reporting that it might be a RA susceptibility genes based on Korean RA samples (Freudenberg *et al.*, 2011); For PLD4, there was no report showing its direct relation with RA.

To make sure that these results were not dominated by the 450K array data, we repeated the above analyses using only the expanded CpG loci not covered by 450K array. We called DMGs identified this way as eDMGs (Fig. 4C and D). There were 391 genes supported by two evidences and the enriched pathways are shown in Figure 4D. The 11 enriched pathways are the same as those analyzed with CpG loci including 450K array data. There were 11 genes supported by all three evidences including LBH, CASP8, HLA-DQA1, OLIG3, HLA-G, IRF5, ELMO1, TRHDE, SLCO1C1, PLD4, AIRE (only IL23R was missed and its role in RA is controversial). These results suggested that the expanded methylation data alone is informative of RA pathogenesis.

Furthermore, we investigated the DMGs only identified by the expanded data but not by 450K array data. We repeated the above analyses using eDMGs exclusively identified from the predicted data. There were 275 genes supported by two evidences (Fig. 4E). The 12 enriched pathways are shown in Figure 4F, among which 11 are the same as those found in the above two analyses and the additional Jak-STAT signaling pathway is known to be important for RA.

3 Discussion

Illumina 450K BeadChip array is a useful way to investigate the different DNA methylation patterns in RA and many diseases. In fact, tens of thousands of Illumina 450K array data have been generated on precious disease samples. Despite the invaluable insights generated by these 450K array data, the small coverage of the CpG sites limited the scope of the investigated DNA methylation patterns in RA and other diseases. Our computational strategy to predict DNA methylation based on 450K data alone opens a new avenue of reprocessing the existing data that were previously generated by 450K array to uncover new disease-related genes before these samples are re-analyzed using WGBS. When applying to a new sample, our model only requires input of 450K array data and avoids the need of histone modification or MeDIP-seq/MRE-seq data that are not always available, which significantly expands its applicability.

Illumina recently released a new Infinium MethylationEPIC Array which is to replace the current methylation450K array (referred to as the 850K array), which covers about 3% of CpGs of the human genome. It is straightforward to apply our expanding algorithm to significantly expand the coverage of the 850K array once enough 850K array data are available.

Our method significantly expanded the coverage of CpG sites, which is 18.5 times of the CpGs covered by 450K array and accounts for about 30% of all the CpGs in the human genome. The current model is trained on 14 tissues that have both WGBS and 450K array data. The performance of the model is expected to be further improved when more tissues/cells are included in training the model. Importantly, our model can successfully predict DML and its performance was confirmed by both leave-one-tissue-out cross validations and identification of RA-related genes/pathways. The 12 triple-evidenced genes with predicted DNA methylation data, a significant increase from 7 based on 450K array data, include 6 genes with reported functions in RA and 6 genes as potential therapeutic targets. We expect the similar applications to other diseases would greatly facilitate discovery of new drug targets and understanding of disease mechanisms.

Funding

This project was supported by the National Natural Science Foundation of China under [61503061] (SCF), China Scholarship Council, and grants from the Rheumatology Research Foundation, the Arthritis Foundation, and the National Institute of Arthritis and Musculoskeletal and Skin Diseases [1R01 AR065466], and [CIRM RB5-07012].

Conflict of Interest: none declared.

References

Acosta-Colman,I. *et al*. (2013) GWAS replication study confirms the association of PDE3A-SLCO1C1 with anti-TNF therapy response in rheumatoid arthritis. *Pharmacogenomics*, **14**, 727–734.

Ai,R. *et al*. (2015) DNA methylome signature in synoviocytes from patients with early rheumatoid arthritis compared to synoviocytes from patients with longstanding rheumatoid arthritis. *Arthritis Rheumatol*., **67**, 1978–1980.

Bernstein,B.E. *et al*. (2010) The NIH Roadmap Epigenomics Mapping Consortium. *Nat. Biotechnol*., **28**, 1045–1048.

Bock,C. *et al*. (2009) EpiGRAPH: user-friendly software for statistical analysis and prediction of (epi)genomic data. *Genome Biol*., **10**, R14.

Bock,C. *et al*. (2006) CpG island methylation in human lymphocytes is highly correlated with DNA sequence, repeats, and predicted DNA structure. *PLoS Genet*., **2**, e26.

Byun,H.M. *et al*. (2009) Epigenetic profiling of somatic tissues from human autopsy specimens identifies tissue- and individual-specific DNA methylation patterns. *Hum. Mol. Genet*., **18**, 4808–4817.

Castro,F. *et al*. (2001) Tumour necrosis factor microsatellites and HLA-DRB1*, HLA-DQA1*, and HLA-DQB1* alleles in Peruvian patients with rheumatoid arthritis. *Ann. Rheum. Dis*., **60**, 791–795.

Dawidowicz,K. *et al*. (2011) The interferon regulatory factor 5 gene confers susceptibility to rheumatoid arthritis and influences its erosive phenotype. *Ann. Rheum. Dis*., **70**, 117–121.

Del Rey,M.J. *et al*. (2012) Transcriptome analysis reveals specific changes in osteoarthritis synovial fibroblasts. *Ann. Rheum. Dis*., **71**, 275–280.

Ekwall,A.K. *et al*. (2015) The Rheumatoid Arthritis Risk Gene LBH Regulates Growth in Fibroblast-like Synoviocytes, *Arthritis. Rheumatol*., **67**, 1193–1202.

Fan,S. *et al*. (2008) Histone methylation marks play important roles in predicting the methylation status of CpG islands. *Biochem. Biophys. Res. Commun*., **374**, 559–564.

Fan,S. and Zhang,X. (2009) CpG island methylation pattern in different human tissues and its correlation with gene expression. *Biochem. Biophys. Res. Commun*., **383**, 421–425.

Fang,F. *et al*. (2006) Predicting methylation status of CpG islands in the human brain. *Bioinformatics*, **22**, 2204–2209.

Farago,B. *et al*. (2008) Functional variants of interleukin-23 receptor gene confer risk for rheumatoid arthritis but not for systemic sclerosis. *Ann. Rheum. Dis*., **67**, 248–250.

Feltus,F.A. *et al*. (2003) Predicting aberrant CpG island methylation. *Proc. Natl. Acad. Sci. USA*, **100**, 12253–12258.

Feng,P. *et al*. (2014) Prediction of CpG island methylation status by integrating DNA physicochemical properties. *Genomics*, **104**, 229–233.

Firestein,G.S. (2003) Evolving concepts of rheumatoid arthritis. *Nature*, **423**, 356–361.

Freudenberg,J. *et al*. (2011) Genome-wide association study of rheumatoid arthritis in Koreans: population-specific loci as well as overlap with European susceptibility loci. *Arthritis. Rheum*., **63**, 884–893.

Garcia-Lozano,J.R. *et al*. (2013) Association of the AIRE gene with susceptibility to rheumatoid arthritis in a European population: a case control study. *Arthritis Res. Ther*., **15**, R11

Harris,R.A. *et al*. (2010) Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat. Biotechnol*., **28**, 1097–1105.

Hindorff,L.A. *et al*. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad. Sci. USA*, **106**, 9362–9367.

Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*., **28**, 27–30.

Knevel,R. *et al*. (2012) Studying associations between variants in TRAF1-C5 and TNFAIP3-OLIG3 and the progression of joint destruction in rheumatoid arthritis in multiple cohorts. *Ann. Rheum. Dis*., **71**, 1753–1755.

Lister,R. *et al*. (2009) Human DNA methylomes at base resolution show widespread epigenomic differences. *Nature*, **462**, 315–322.

Morgan,A.W. *et al*. (2010) Evaluation of the rheumatoid arthritis susceptibility loci HLA-DRB1, PTPN22, OLIG3/TNFAIP3, STAT4 and TRAF1/C5 in an inception cohort. *Arthritis Res. Ther*., **12**, R57.

Nakano,K. *et al*. (2013) DNA methylome signature in rheumatoid arthritis. *Ann. Rheum. Dis*., **72**, 110–117.

Okada,Y. *et al*. (2014) Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, **506**, 376–381.

Orozco,G. *et al*. (2007) Investigation of the IL23R gene in a Spanish rheumatoid arthritis cohort. *Hum. Immunol*., **68**, 681–684.

Rizzo,R. *et al*. (2013) HLA-G may predict the disease course in patients with early rheumatoid arthritis. *Hum. Immunol*., **74**, 425–432.

Shao,S. *et al*. (2014) Association of AIRE polymorphisms with genetic susceptibility to rheumatoid arthritis in a Chinese population. *Inflammation*, **37**, 495–499.

Stadler,M.B. *et al*. (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature*, **480**, 490–495.

Stevens,M. *et al*. (2013) Estimating absolute methylation levels at single-CpG resolution from methylation enrichment and restriction enzyme sequencing methods. *Genome Res*., **23**, 1541–1553.

Szabo,M. *et al*. (2013) Marked diversity of IL23R gene haplotype variants in rheumatoid arthritis comparing with Crohn's disease and ankylosing spondylitis. *Mol. Biol. Rep*., **40**, 359–363.

Terao,C. *et al*. (2011) The human AIRE gene at chromosome 21q22 is a genetic determinant for the predisposition to rheumatoid arthritis in Japanese population. *Hum. Mol. Genet*., **20**, 2680–2685.

Whitaker,J.W. *et al*. (2015) Integrative omics analysis of rheumatoid arthritis identifies non-obvious therapeutic targets. *PLoS One*, **10**, e0124254.

Whitaker,J.W. *et al*. (2013) An imprinted rheumatoid arthritis methylome signature reflects pathogenic phenotype. *Genome Med*., **5**, 40.

Zheng,H. *et al*. (2013) CpGIMethPred: computational model for predicting methylation status of CpG islands in human genome. *BMC Med. Genomics*, **6(Suppl 1)**, S13.