Molecular Therapy
Nucleic Acids
Original Article

Deciphering a TB-related DNA methylation biomarker and constructing a TB diagnostic classifier

Mengyuan Lyu,[1,2,6] Jian Zhou,[2,3,6] Lin Jiao,[1,2] Yili Wang,[1,2] Yanbing Zhou,[1,2] Hongli Lai,[2] Wei Xu,[4,5] and Binwu Ying[1,2]

[1]Department of Laboratory Medicine, West China Hospital, Sichuan University, No. 37, Guoxue Alley, Chengdu, Sichuan 610041, China; [2]West China School of Medicine, Sichuan University, Chengdu, Sichuan 610041, China; [3]Department of Thoracic Surgery, West China Hospital, Sichuan University, Chengdu, Sichuan 610041, China; [4]Department of Biostatistics, Princess Margaret Cancer Centre, University Health Network, 10-511, 610 University Avenue, Toronto, ON M5G 2M9 Canada; [5]Dalla Lana School of Public Health, University of Toronto, Toronto, ON M5T 3M7 Canada

We systemically identified tuberculosis (TB)-related DNA methylation biomarkers and further constructed classifiers for TB diagnosis. TB-related DNA methylation datasets were searched through October 3, 2020. Limma and DMRcate were employed to identify differentially methylated probes (DMPs) and regions (DMRs). Machine learning methods were used to construct classifiers. The performance of the classifiers was evaluated in discovery datasets and a prospective independent cohort. Eighty-nine DMPs and 24 DMRs were identified based on 67 TB patients and 45 healthy controls from 4 datasets. Nine and three DMRs were selected by elastic net regression and logistic regression, respectively. Among the selected DMRs, two regions (chr3: 195635643–195636243 and chr6: 29691631–29692475) were differentially methylated in the independent cohort ($p = 4.19 \times 10^{-5}$ and 0.024, respectively). Among the ten classifiers, the 3-DMR logistic regression classifier exhibited the strongest performance. The sensitivity, specificity, and area under the curve were, respectively, 79.1%, 84.4%, and 0.888 in the discovery datasets and 64.5%, 90.3%, and 0.838 in the independent cohort. The differential diagnostic ability of this classifier was also assessed. Collectively, these data showed that DNA methylation might be a promising TB diagnostic biomarker. The 3-DMR logistic regression classifier is a potential clinical tool for TB diagnosis, and further validation is needed.

INTRODUCTION

The complex nature of *Mycobacterium tuberculosis* (MTB) has greatly contributed to the continuous effects of tuberculosis (TB), the world's top infectious killer, on human populations for thousands of years.[1] Globally, in 2019, approximately 10 million people fell ill with TB, and 1.4 million people died of TB.[2] Luckily, TB is a preventable and curable illness, and the key to thwarting the TB epidemic is early diagnosis.[3] The detection power of etiological tests (Xpert MTB/RIF, culture, etc.) largely depends on sample quality.[4] Traditional host immune response assays (interferon-γ [IFN-γ] release assays, tuberculin tests, etc.) work mainly by measuring the products of the host immune response. Most of these products are downstream molecules in various biological pathways, suggesting that these molecules are regulated by various factors.[5] In addition to the limited performance of the available detection methods, certain features of TB also increase the difficulty of its clinical diagnosis. For example, incipient TB is characterized by an indeterminate period of asymptomatic infection or absence of typical clinical symptoms and imaging features, which limit the effectiveness of TB diagnosis.[6] Therefore, new and more powerful detection tools are urgently needed. Current evidence has indicated that the detection of non-sputum blood biomarkers of TB is a preferred approach for clinical diagnosis and progression surveillance.[7,8]

TB is an infectious disease whose pathogenesis involves dynamic interactions among the host, MTB, and the environment. The circulatory system serves as a site for cellular communication and dynamic exchange of various cellular factors and chemokines, and thus peripheral blood has been considered the preferred sample type for studying and diagnosing infectious diseases.[9] To date, peripheral blood biomarkers of TB at the genomic, transcriptomic, epigenetic, and proteomic levels have received enormous attention, and numerous markers have been identified.[7,9–11] Of note, epigenetics can bridge the gaps between the host, MTB, and the environment.[12,13] Therefore, epigenetic biomarkers in the peripheral blood may have great potential in TB diagnosis and progression surveillance.

Received 9 June 2021; accepted 16 November 2021;
https://doi.org/10.1016/j.omtn.2021.11.014.

[6]These authors contributed equally

Correspondence: Wei Xu, PhD, Department of Biostatistics, Princess Margaret Cancer Centre, University Health Network, 10-511, 610 University Avenue, Toronto, ON M5G 2M9, Canada.
E-mail: wei.xu@uhnres.utoronto.ca
Correspondence: Binwu Ying, PhD, MD, Department of Laboratory Medicine, West China Hospital, Sichuan University, No. 37, Guoxue Alley, Chengdu, Sichuan 610041, China.
E-mail: yingbinwu@scu.edu.cn

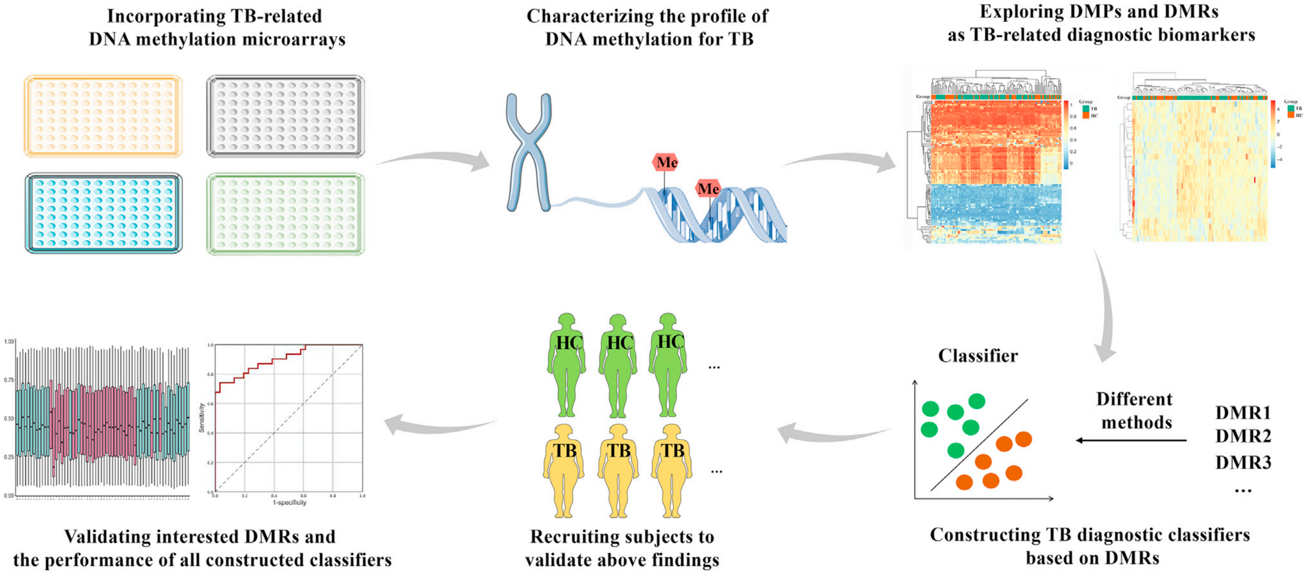
Incorporating TB-related
DNA methylation microarrays
Characterizing the profile of
DNA methylation for TB
Me
Me
Exploring DMPs and DMRs
as TB-related diagnostic biomarkers
Classifier
Different
methods
DMR1
DMR2
DMR3
...
Constructing TB diagnostic classifiers
based on DMRs
HC
HC
HC
TB
TB
TB
Recruiting subjects to
validate above findings
Validating interested DMRs and
the performance of all constructed classifiers

Figure 1. The flow chart of the study protocol

Abbreviations: TB, tuberculosis; DMPs, differentially methylated probes; DMRs, differentially methylated regions; HC, healthy control.

DNA methylation, the most widely studied epigenetic modification, refers to the formation of 5-methylcytosine through the transfer of a methyl group to the carbon-5 position of the cytosine base by DNA methyltransferase enzymes.[14] DNA methylation is responsible for regulating gene expression, chromatin structure, and alternative splicing.[15,16] In human somatic cells, methylation occurs in approximately 90% of the cytosines in CpG sites; most importantly, this ratio varies among different tissue types, cell types, and disease states.[15] As the advantages of DNA methylation include critical roles in biological processes, disease/tissue/cell-type-specific patterns, and inherent stability, specific DNA methylation profiles have been mapped for many diseases, such as breast cancer, adiposity, and others.[17,18] Some pathogens, such as human immunodeficiency virus and hepatitis B virus, have been verified to alter the DNA methylation landscape of the host;[19,20] MTB has also been reported to have such activity. DiNardo et al.[21] found that the methylation level of mycobacterial immunity-related genes was upregulated, leading to a downregulation of the host immune response against MTB. These findings indicated that DNA methylation is important in TB development and may serve as a promising biomarker. However, systematic work to explore the value of DNA methylation in TB has thus far been less common.

Toward this end, we (1) summarized available datasets to systematically describe the DNA methylation characteristics of TB and identify potential biomarkers; (2) applied machine learning methods to construct TB diagnostic classifiers by triaging a parsimonious list of the most promising targets; and (3) performed validation studies in an independent cohort to ensure that these molecules are robust as TB-related biomarkers. This study aimed to establish TB-related methylation profiles and facilitate TB diagnosis, as well as decipher the underlying connection between TB and DNA methylation.

RESULTS

The flow chart of the study protocol is shown in Figure 1.

Different DNA methylation patterns in TB patients and healthy controls

Altogether, 4 datasets, including 67 TB patients and 45 healthy controls (HCs), met the inclusion criteria (Table 1). A total of 363,416 probes were retained for analysis after filtering and adjusting for batch effects. In total, 89 differentially methylated probes (DMPs) were found between TB patients and HCs (Figure 2A). Mapping of the 89 DMPs onto genomic features showed that the most probes targeted intronic features (29%, 26/89), followed by exons (21%, 19/89) and intergenic features (20%, 18/89) (Figure 2C). Among the 89 DMPs, 68.5% (61/89) sites were hypermethylated and 31.5% (28/89) were hypomethylated in TB patients compared with HCs (Figure 2E).

Further analysis was performed to identify ethnicity-specific and sample type-specific DMPs (Figure S1).

We then shifted focus to the methylation regions with the strongest prospective ability to regulate gene expression. Twenty-seven differentially methylated regions (DMRs) covering 310 CpG sites were identified (Figure 2B). Three regions were excluded because only one CpG site was present in each region. Of the 24 DMRs, genomic annotation showed that 42% (10/24) were located in exon regions, and 25% (6/24) were located in 5′ UTRs (Figure 2D). The locations of DMRs exhibited a chromosomal bias, and nearly half of the DMRs (10/24) were distributed on chromosome 6. Of all DMRs, only one (chr3: 50336343–50337494) mapped to two genes

Table 1. The detailed information of included methylation arrays

		Characteristics of included subjects				Sample numbers		
GEO ID	Platform	Race	Mean age (years)	Sex (male/female, %)	HIV status	TB	HC	Sample type
GSE118469	GPL13534	Asian	>18	100	0	15	6	PBMC
GSE104287	GPL13534	Caucasian	>18	62.5	0	32	16	PBMC and NK cell
GSE72338	GPL13534	African	>18	63.2	0	17[a]	20	monocyte and neutrophil
GSE107917	GPL23976	Asian	N/A	N/A	N/A	3	3	whole blood

GEO, Gene Expression Omnibus; HIV, human immunodeficiency virus; TB, tuberculosis; HC, healthy control; PBMC, peripheral blood mononuclear cell; N/A, not applicable; NK, natural killer.

[a] One sample (GSM1860484) was not included due to the filter conditions used in raw data procession.

(HYAL3 and NAT6). According to the mean fold change (FC) of the β value, TB patients showed significant hypermethylation compared with HCs in 22 regions and hypomethylation in 2 regions (Table 2; Figure 2F).

Principal-component analysis was conducted to explore the difference in 24 DMRs among ethnicity-specific and sample type-specific subgroups (Figure S2).

Function enrichment of DMRs

Gene ontology (GO) enrichment analysis and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis were performed for 25 DMR-associated genes. GO analysis indicated that DMR-associated genes contributed to many immune-related biological functions, such as immune cell activation and regulation, cellular response to IFN-γ, and cytotoxicity (Figure S3A). KEGG analysis suggested that these 25 genes mainly participated in antigen processing and presentation and pathogen infection-related pathways (cytomegalovirus, papillomavirus, etc.) (Figure S3B). A protein-protein interaction (PPI) network was built to visualize the interactions among the proteins encoded by these 25 genes. HLA-F, ZBTB22, SIN3A, and GABBR1 were considered hub genes in the constructed network (Figure S3C).

Construction of diagnostic classifiers based on the validated DMRs

Variables were selected by logistic regression and elastic net regression. A total of six machine learning methods were applied to construct classifiers based on the selected variables.

Through binary univariate and multivariate logistic regression, three DMRs (chr11: 65315205–65315625, chr3: 195635643–195636243, and chr6: 29691631–29692475) were finally selected as classifiers (Figures 3A and 3B). For these three DMRs, the logistic regression classifier yielded a sensitivity of 79.1%, specificity of 84.4%, and area under the curve (AUC) of 0.888 (95% confidence interval [CI]: 0.831–0.945) (Figures 3C and 3D). This classifier also showed a net benefit in performance regardless of the risk threshold selected (Figure 3E). The highest AUC was 0.999 (95% CI: 0.997–1.000) for the random tree classifier, followed by that for the extreme gradient boosting (XGBoost) (AUC = 0.972; 95% CI: 0.948–0.995) and k-nearest neighbor (KNN) (AUC = 0.945; 95% CI: 0.908–0.982) classifiers. The lowest AUC was observed for the support vector machine (SVM) classifier (AUC = 0.833; 95% CI: 0.758–0.907).

To avoid overfitting, the largest λ at which the mean squared error was within 1 standard error of the minimum (λ = 0.068) was used in the process of variable selection by elastic net regression (Figure 4A). In addition to the above three DMRs, six DMRs (chr15: 75743753–75744225, chr6: 30458519–30458601, chr6: 33244976–33246390, chr6: 31627090–31627313, chr6: 33283789–33284168, and chr6: 31937968–31938372) (Figure 4B) were selected by elastic net regression. For these nine DMRs, the elastic net regression classifier reached a sensitivity of 82.1%, specificity of 86.7%, and AUC of 0.918 (95% CI: 0.871–0.966) (Figure 4C). The highest AUC was found for the random tree classifier (AUC = 1.000, 95% CI: 1.000–1.000), followed by the XGBoost (AUC = 0.997; 95% CI: 0.990–1.000) and KNN (AUC = 0.919, 95% CI: 0.869–0.969) classifiers.

The optimal hyperparameters used in each classifier are provided in Table S2.

Validation of DMRs by region-specific multiple sequencing

The methylation levels of the above nine DMRs were further tested. Altogether, 62 samples from 31 TB patients and 31 HCs were collected in a prospective clinical cohort.

The regions of chr3: 195635643–195636243 and chr6: 29691631–29692475 were found to be differentially methylated ($p = 4.19 \times 10^{-5}$ and 0.024, respectively) (Figure 5A). However, no meaningful findings were observed in the other seven regions (original data are provided in Data S1).

Given that the data are generated in different ways (array and sequencing), the cutoff value of the classifier based on microarray datasets might not be suitable for sequencing data; therefore, new thresholds generated by sequencing data were used. Among all the classifiers constructed by different DMRs and modeling methods, the 3-DMR logistic regression classifier exhibited the highest AUC (0.838; 95% CI: 0.737–0.938). The specificity of this classifier

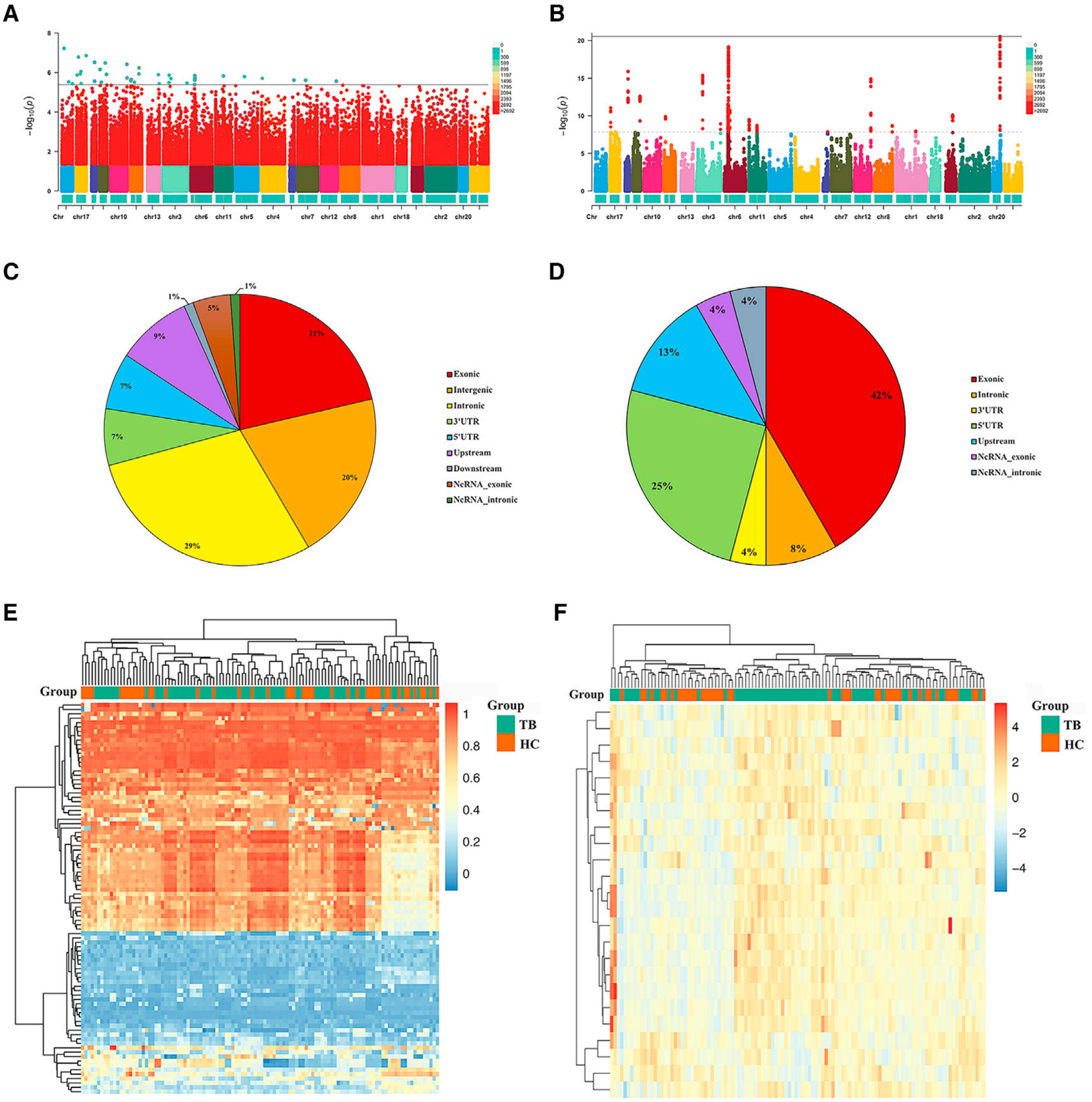

Figure 2. Differently methylated probes and differentially methylated regions between tuberculosis patients and healthy controls

(A and B) The Manhattan plot of all methylated probes and regions, respectively. The ordinate and abscissa represented the $-\log_{10}$ p value and chromosomes, respectively. The different colors above the abscissa show the probe numbers in corresponding chromosomes and the horizontal line indicates the threshold value of p. (C and D) The types of differentially methylated probe-related genes and differentially methylated region-related genes, respectively. ncRNA, non-coding RNA. (E and F) The heatmap of differentially methylated probes and differentially methylated regions between tuberculosis patients and healthy controls, respectively. The method of correlation was used for row clustering.

increased to 90.3%, while the sensitivity declined slightly but still reached 64.5% (Table 3; Figure 5B).

To facilitate the application of these findings to clinical practice, an online classifier based on the 3-DMRs logistic regression was designed at https://mengyuan.shinyapps.io/TB_DNAmethylation/. For simplicity, in this online tool, region-related genes were employed to represent the corresponding DMRs. LTBP3, TNK2-AS1, and HLA-F represent the regions of chr11: 65315205–65315625, chr3: 195635643–195636243, and chr6: 29691631–29692475, respectively.

Further evaluation of the 3-DMR logistic regression classifier in different situations

Considering the serious consequences of TB spread, 35 TB patients and 32 participants injected with Bacillus Calmette-Guerin (BCG) were combined into the same group to decrease the likelihood of missed diagnosis. However, differential diagnosis of TB patients

Table 2. Identified top different methylation regions between TB patients and HCs

	Genetic position[a]					
Chromosome	Start	End	Gene symbol	CpG site number	Mean beta fold change	Adjusted p value
Chr1	160068509	160068681	IGSF8	6	1.22×10^{-2}	1.06×10^{-8}
Chr3[b,c]	195635643	195636243	TNK2-AS1	2	9.36×10^{-3}	1.17×10^{-9}
Chr3	50336343	50337494	HYAL3; NAT6	17	7.01×10^{-4}	4.13×10^{-16}
Chr6	33160067	33160976	COL11A2	19	4.60×10^{-3}	3.96×10^{-18}
Chr6[c]	30458519	30458601	HLA-E	3	-3.76×10^{-2}	5.74×10^{-9}
Chr6	44190729	44191600	SLC29A1	10	6.24×10^{-4}	1.88×10^{-11}
Chr6[c]	33244976	33246390	B3GALT4	37	1.19×10^{-2}	7.05×10^{-20}
Chr6[c]	31627090	31627313	C6orf47	6	1.04×10^{-2}	6.81×10^{-9}
Chr6[c]	33283789	33284168	ZBTB22	10	8.26×10^{-3}	3.96×10^{-9}
Chr6	29600108	29600468	GABBR1	10	2.66×10^{-3}	3.30×10^{-9}
Chr6[c]	31937968	31938372	DXO	6	4.98×10^{-3}	7.91×10^{-10}
Chr6[b,c]	29691631	29692475	HLA-F	21	-1.10×10^{-2}	2.75×10^{-12}
Chr6	30711586	30712559	IER3	23	5.67×10^{-5}	3.03×10^{-11}
Chr8	144328914	144329279	ZFP41	7	4.18×10^{-3}	2.05×10^{-9}
Chr11[b,c]	65315205	65315625	LTBP3	4	2.28×10^{-2}	3.63×10^{-9}
Chr11	2019930	2020560	H19	15	1.65×10^{-2}	3.12×10^{-10}
Chr11	66034896	66035392	KLC2	14	3.86×10^{-3}	1.98×10^{-9}
Chr12	133065912	133066762	FBRSL1	18	5.40×10^{-3}	1.22×10^{-15}
Chr15[c]	75743753	75744225	SIN3A	8	6.00×10^{-3}	8.06×10^{-11}
Chr16	11350112	11350371	SOCS1	6	6.65×10^{-3}	1.28×10^{-10}
Chr17	7210796	7211307	EIF5A	6	6.65×10^{-3}	8.61×10^{-12}
Chr19	55850629	55851365	KMT5C	12	3.31×10^{-3}	2.55×10^{-13}
Chr20	57426538	57427973	GNAS	38	1.59×10^{-2}	2.88×10^{-21}
Chr22	38851318	38852154	KCNJ4	9	1.06×10^{-2}	1.26×10^{-16}

[a] The positions were obtained according to hg19, GRCh37 (Genome Reference Consortium Human Reference 37).
[b] These three different methylation regions were selected by logistic regression into models.
[c] These nine different methylation regions were selected by elastic net regression into models.

and BCG recipients is needed. Although the 3-DMR logistic regression classifier exhibited moderate AUC (0.689; 95% CI: 0.563–0.816), it could distinguish TB patients from BCG participants with a specificity of 82.9%.

The differential diagnosis performance of the 3-DMR logistic regression classifier was also evaluated. The details of disease controls (DCs) are shown in Table S3. The 3-DMR logistic regression classifier showed a strong ability to distinguish TB from malaria (sensitivity, 61.2%; specificity, 87.5%; AUC = 0.778; 95% CI: 0.598–0.959) and systemic inflammatory response syndrome (sensitivity, 100%; specificity, 94.0%; AUC = 0.955; 95% CI: 0.907–1.000). The sensitivity, specificity, and AUC were 100%, 92.5%, and 0.965 (95% CI: 0.929–1.000) when using this classifier to differentiate sepsis patients (GEO: GSE138074) from TB patients, while they were 100%, 100%, and 1.000 (95% CI: 1.000–1.000) when distinguishing sepsis patients (GEO: GSE58651 or GEO: GSE155952) from TB patients. However, this classifier failed to efficiently differentiate TB patients from patients with subclinical parasitemia (sensitivity, 100%; specificity, 41.8%; AUC = 0.631; 95% CI: 0.417–0.844).

The differential diagnosis ability of the 3-DMR logistic regression classifier is shown in Figure 6.

DISCUSSION

This work represents a comprehensive analysis of TB-related DNA methylation biomarkers. By integrating the available datasets, we identified TB-related targets (89 DMPs and 24 DMRs). With respect to the likelihood and degree of influencing gene expression, 24 DMRs were treated as subsequent candidates. Logistic regression and elastic net regression were used to select potential biomarkers for further validation in an independent cohort. Based on the selected candidates, six different methods were applied to construct classifiers, and the 3-DMR logistic regression classifier outperformed the others. The good performance of this classifier in both discovery and

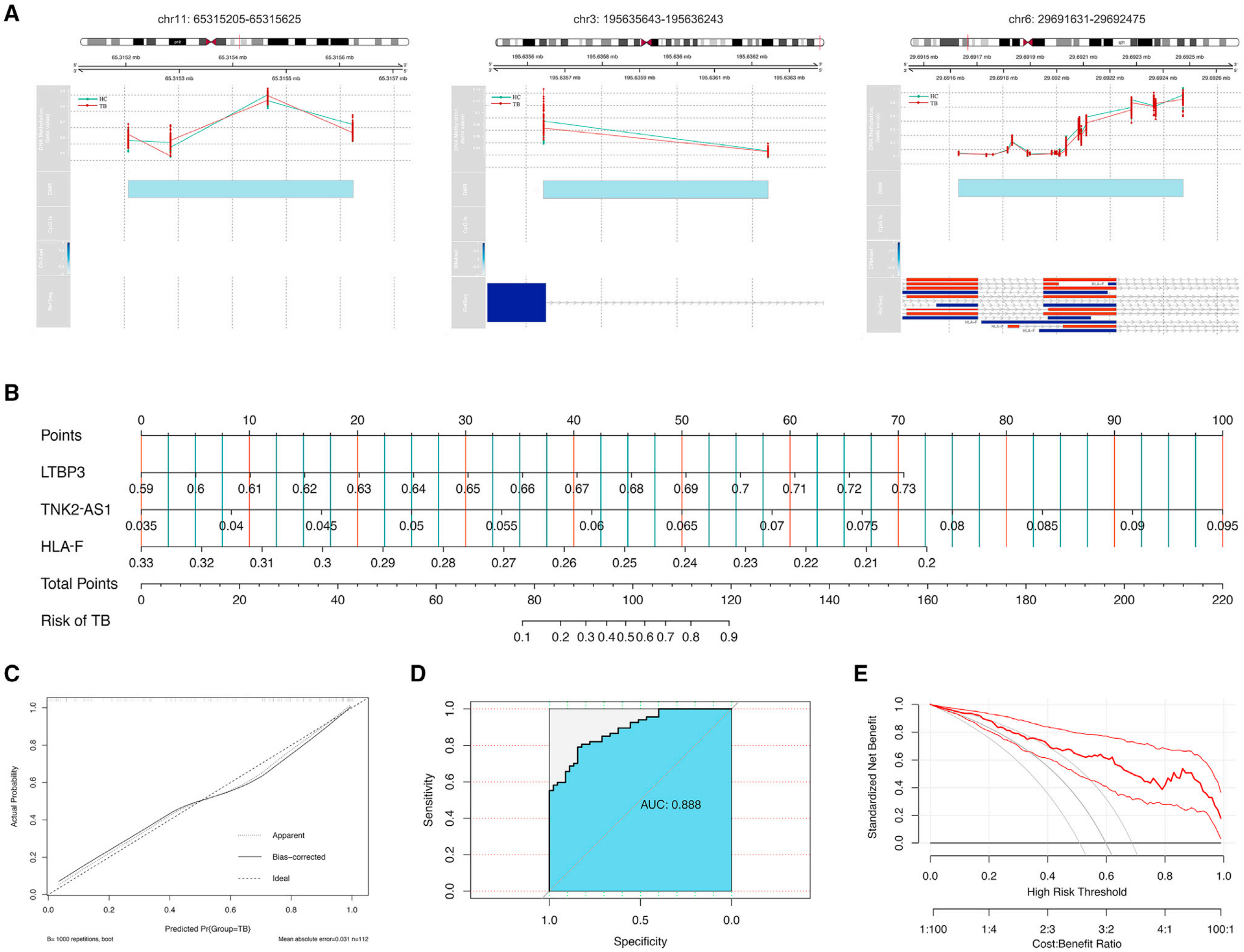

Figure 3. The included variables, nomogram, and performance of the 3-DMR logistic regression classifier

(A) Customizing visualizations of included variables in 3-DMR logistic regression classifiers. The upper two plots are the genomic coordinates of the targeted differentially methylated regions, followed by a line chart that shows the mean beta value of all probes in the corresponding region among tuberculosis patients (red line) and healthy controls (green line). Next, the genomic annotations, including CpG island locations and DNaseI hypersensitive sites, were plotted. The data of CpG island locations and DNaseI hypersensitive sites were obtained from Wu et al.[22] and UCSC Genome Browser. Finally, RefSeq tracks were added. (B) Nomogram of the 3-DMR logistic regression classifier. For simplicity, region-related genes were employed to represent the corresponding differentially methylated regions. LTBP3, TNK2-AS1, and HLA-F represent the regions of chr11: 65315205–65315625, chr3: 195635643–195636243, and chr6: 29691631–29692475, respectively. (C) Calibration curve of the 3-DMR logistic regression classifier. (D) Receiver operating characteristic of the 3-DMR logistic regression classifier. (E) Decision curve analysis curve of the 3-DMR logistic regression classifier. The bold red curve shows the benefit net of this classifier at different risk thresholds, while the curves on both sides represent its 95% confidence interval.

validation cohorts emphasize the tremendous potential of DNA methylation as a TB diagnostic biomarker.

Over the past few decades, clinicians have expressed a preference for obtaining the most direct evidence possible to diagnose diseases. Therefore, tissues or body fluid from lesion sites are considered ideal sample types; and, certainly, analyses based on these samples are regarded as reference methods. However, a high level of invasiveness, challenging techniques, and limited applicable populations currently constrain the use of such samples for disease diagnosis. In this context, peripheral blood has emerged as a focus for researchers. The peripheral blood is the environment in which various cytokines and immune cells interact in infectious diseases. Therefore, the circulatory system can be treated as a source of markers for the host immune response[23] and offers information about when and how disease progresses. The exploration of diagnostic biomarkers in the peripheral blood has encompassed genomics, transcriptomics, and epigenetics studies. However, genomics and transcriptomics do not provide an appropriate balance between molecular stability and flexible monitoring of disease progression. DNA methylation perfectly offsets the above two limitations. Thanks to technological developments, many choices for detection methods are available, which also

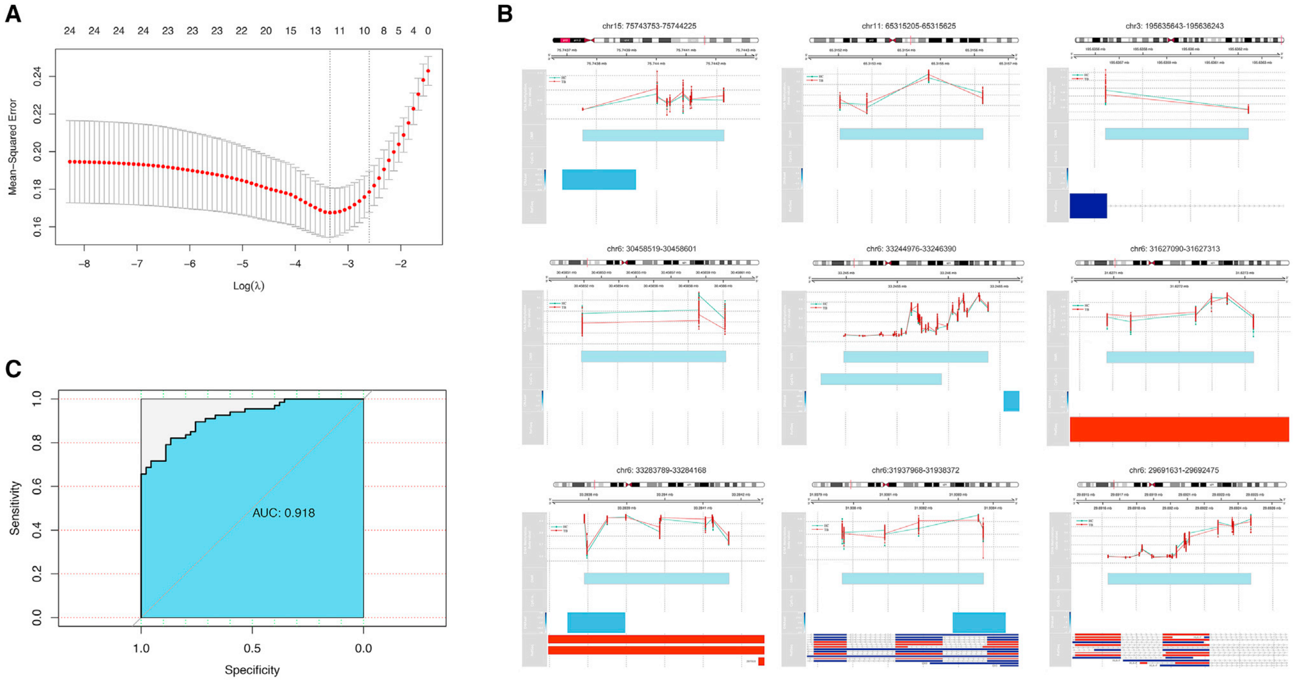

Figure 4. Optimal hyperparameter selection, included variables and receiver operating characteristics of the 9-DMR elastic net regression classifier

(A) The cross-validation plot for the optimal λ selection. The red dots represent the target mean squared error corresponding to each λ. The left dotted line shows the value of λ when MSE was at its minimum, while the right dotted line shows the maximum value of λ when MSE was within 1 standard error of the minimum. (B) Customizing visualizations of included variables in 9-DMR elastic net regression classifier. The upper two plots are the genomic coordinates of the targeted differentially methylated regions, followed by the line chart which shows the mean beta value of all probes in the corresponding region among tuberculosis patients (red line) and healthy controls (green line). Next, the genomic annotations, including CpG island locations and DNAseI hypersensitive sites, were plotted. The data of CpG island locations and DNAseI hypersensitive sites were obtained by the publication of Wu et al.[22] and UCSC Genome Browser. Finally, RefSeq tracks were added. (C) The receiver operating characteristics of the 9-DMR elastic net regression classifier.

contributes greatly to the clinical application of DNA methylation analysis. These advantages have prompted the deep exploration of DNA methylation in infectious diseases.

To clarify how peripheral blood biomarkers are regulated and determine their functions in targeted disease, it is essential to demonstrate their roles as biomarkers of disease progression. The association between DNA methylation and TB was reported as early as 1980.[24] Although the question of whether the change in host genomic methylation is caused by MTB itself or secondary to inflammation remains unanswered, some progress has been made to decode the functions of these changes in DNA methylation. By summarizing and comparing the characteristics of different methylation sites before and after host infection with MTB, scholars found that significantly changed candidates were mainly located in promoter regions. Methylation in the promoter region can lead to the failure of gene transcription initiation and gene silencing the by hampering transcription factor (TF) binding to specific motifs.[25,26] Changes in gene activity and expression will further affect a series of signaling pathways and biological functions, including immune cell regulation, cytokine regulation, IFN-γ signaling pathways, and other TB-related immune activities,[21,27] consistent with the results of this study. As mentioned above, methylation is able to interfere with TF binding, while TFs are also capable of regulating methylation status. TF occupancy theory suggests that TFs can modulate gene methylation levels by competing with DNA methyltransferase binding at promoter sequences.[26] Pacis et al.[28] cultured MTB-infected human dendritic cells and collected time-dependent data on DNA methylation, gene expression, and chromatin accessibility patterns. The resulting data verified that immune-related TFs could regulate methylation levels by binding to *cis*-acting elements. Notably, not all TFs are negatively regulated by DNA methylation, and positive correlations between them have also been reported.[29] Collectively, the existing data show that DNA methylation has crucial roles in the pathogenesis and development of TB. Mapping TB-specific methylation and further selecting promising sites may open a new horizon for TB elimination to some extent.

The understanding of how DNA methylation functions increases its potential as a promising biomarker. Over time, a number of TB-specific DNA methylation biomarkers, from the global DNA methylation level to detailed methylated sites or regions, have been reported. Maruthai et al.[30] took the global DNA methylation level as a TB-specific biomarker and reported an AUC of 0.81. Their group also

Table 3. The performance of built classifiers

3-DMR classifier							9-DMR classifier				
Variable selection method		Logistic regression					Elastic net regression				
Modeling method		Logistic regression	Support vector machine	K-nearest neighbor	Random tree	XGBoost	Elastic net regression	Support vector machine	K-nearest neighbor	Random tree	XGBoost
Discovery set	AUC (95% CI)	0.888 (0.831–0.945)	0.833 (0.758–0.907)	0.945 (0.908–0.982)	0.999 (0.997–1.000)	0.972 (0.948–0.995)	0.918 (0.871–0.966)	0.849 (0.780–0.918)	0.919 (0.869–0.969)	1.000 (1.000–1.000)	0.997 (0.990–1.000)
	Sensitivity	79.1%	76.1%	65.7%	100%	92.5%	82.1%	91.0%	86.6%	100%	97.0%
	Specificity	84.4%	77.8%	100%	93.3%	88.9%	86.7%	66.7%	86.7%	100%	100%
Validation set	AUC (95% CI)	0.838 (0.737–0.938)	0.778 (0.663–0.894)	0.500 (0.355–0.645)	0.750 (0.625–0.875)	0.772 (0.652–0.891)	0.536 (0.389–0.638)	0.506 (0.357–0.654)	0.516 (0.371–0.662)	0.512 (0.367–0.657)	0.532 (0.389–0.675)
	Sensitivity	64.5%	61.3%	–	93.5%	80.6%	61.3%	77.4%	71%	58.1%	87.1%
	Specificity	90.3%	87.1%	–	54.8%	64.5%	58.1%	45.2%	38.7%	45.2%	22.6%

DMR, different methylation region; XGBoost, extreme gradient boosting; AUC, area under curve; CI, confidence interval.

focused on the VDR promoter methylation level and the median DNA methylation level of Alu repetitive elements and found that these two indicators exhibited good performance for diagnosing pediatric TB (AUC = 0.977 and 0.969, respectively).[31,32] However, these results were obtained from relatively small cohorts (76 and 68 children). The team of Das and Chen reported some differentially methylated sites in TLR2, SNX26, and other genes but did not analyze the diagnostic values of these sites.[33–35] Wang et al.[36] demonstrated meaningfully methylated sites in key genes in the vitamin D metabolic pathway and further assessed the diagnostic capacity of these genes by calculating the cumulative methylation level of CpG sites on these genes via four different methods. The AUCs of candidate genes varied from 0.578 to 0.794. Esterhuyse et al.[9] incorporated different CpG sites by machine learning to distinguish active TB from latent TB infection and obtained a model with an AUC of 0.74. However, these methylation data were generated from only monocytes and granulocytes, which might restrict the applicability of the model. The high AUCs of these candidate loci underscores the potential of DNA methylation as a TB diagnostic biomarker. However, these biomarkers were identified in relatively small cohorts or specific cell types. Moreover, current studies tend to use a single CpG site or CpG sites in a single gene as a biomarker, which implies limited diagnostic power and generalizability. Comprehensive and systematic work is particularly necessary to avoid the above problems and further develop the diagnostic potential of DNA methylation. Therefore, we first combined all available datasets to properly expand the sample size, cover more ethnic populations, and target more sample types (whole blood, peripheral blood mononuclear cells, etc.). Through rigorous data analysis, the high population coverage and many different sample types ensure the wide applicability of biomarkers. Compared with CpG sites, methylated regions are more likely to regulate gene expression to a greater extent and thus were chosen as biomarkers in this work. In addition, each single biomarker has its own limitations, which will be particularly magnified in complex and changeable clinical settings. Constructing models to incorporate different biomarkers can resolve these questions effectively and maximize the underrated power of TB-specific DNA methylation biomarkers. We tested various combinations of variable selection methods and modeling methods to build different classifiers, which offered us a better opportunity to select a high-performing classifier. Such efforts were finally realized in the 3-DMR logistic regression classifier, which exhibited superior performance in both the discovery and validation cohorts. Fewer included variables, excellent capacity, strong generalizability, and availability as a convenient online tool enable this classifier to be used in real clinical applications.

Understanding why this 3-DMR logistic regression classifier possesses such capacity and differential diagnostic ability is crucial to its optimal utilization. The regions of chr3: 195635643–195636243 and chr6: 29691631–29692475 occupied a dominant position in this classifier. The region of chr3: 195635643–195636243 is at the position of TNK2-AS1. TNK2-AS1 is closely related to cell proliferation, invasion and apoptosis in cancer.[37,38] TNK2-AS1 acts as a microRNA sponge and, to date, miR-4319 and miR-150-5p have been reported as

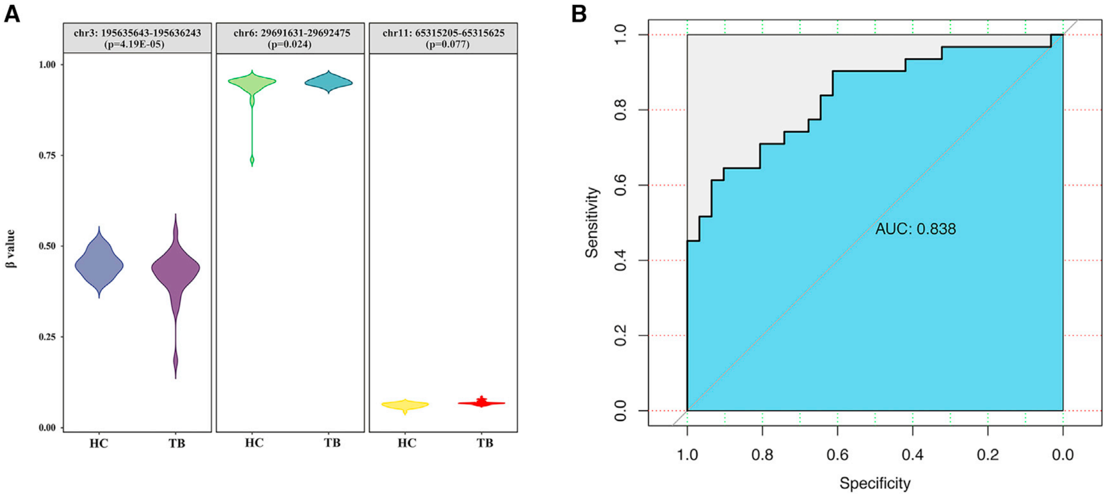

Figure 5. The methylation levels of three differentially methylated regions and the performance of the 3-DMR logistic regression classifier in an independent cohort

(A) A violin plot of the methylation levels of three differentially methylated regions in an independent cohort. TB, tuberculosis; HC, healthy control. (B) The receiver operating characteristics of the 3-DMR logistic regression classifier in an independent cohort.

targets of TNK2-AS1.[38,39] Our previous study suggested that miR-150-5p was a promising biomarker for TB diagnosis,[40] and mounting evidence demonstrates that miR-150-5p contributes greatly to immune cell differentiation and the capacity of effector CD8+ T cells to kill infected cells.[41] Herein, we hypothesized that the interaction between MTB and the host influences TNK2-AS1 expression by regulating its methylation level, while altered TNK2-AS1 expression triggers a series of downstream pathways through a ceRNA mechanism.

The region of chr6: 29691631–29692475 is located within the promoter region of HLA-F. In the late 1980s, Geraghty et al.[42] first reported HLA-F as a nonclassical class I antigen in the human leukocyte antigen (HLA) family. Due to their prominent role in antigen presentation and immune regulation, many members of the HLA family have received ongoing attention from researchers,[43] but HLA-F has been less well studied. However, the current evidence suggests that HLA-F is an underappreciated player in immune regulation.[44] Changes in HLA-F at the genome, transcription, or epigenetic level can fuel a series of immune alterations and are thus involved in cancer, infection, autoimmunity, and other disorders.[45] In our work, TB patients exhibited decreased methylation levels in the promoter region of HLA-F, also indicating elevated HLA-F mRNA expression. Increased HLA-F mRNA expression had positive impacts on its interaction with immune receptors (KIR3DL1, KIR3DS1, etc.), cytokine production (CCL4, IFN-γ, etc.), and antigen presentation, which have already been verified in other infectious diseases. Lunemann et al.[46] documented that through binding to KIR3DS1, increased HLA-F promoted natural killer cell cytolysis and cytokine production and thus suppressed hepatitis C virus replication. This was also observed in HIV-1 infection.[47] Based on the above, we speculated that the host responded to MTB infection by downregulating the methylation level of the promoter region in HLA-F, elevating HLA-F mRNA expression and then promoting immune cell activation to kill the bacteria. The CpG sites in the HLA-F promoter region may serve as promising targets to enhance host immunity and achieve precise MTB clearance.

Aside from its diagnostic ability, generalizability, detection time, cost, and feasibility in economically poor areas must also be considered in the transition of this classifier to clinical practice. In terms of generalizability, data generated from different sample types and ethnicities were included to increase the generalizability of the classifiers. The limited sample size in subgroups of different ethnicities restricted the reliability of related findings, whereas relevant results were presented in this paper to provide potential clues. For sample type subgroups, principal-component analysis indicated that 24 regions in 5 subgroups exhibited similar methylation patterns. This suggests that the 3-DMR logistic regression classifier has the potential to be used in these five sample types. The excellent performance of the 3-DMR logistic regression classifier in whole blood samples from an independent cohort supports the above speculation. Although different blood cells have their own epigenetic patterns, they are all found in the circulation and thus may be stimulated by the same factors and communicate frequently with each other, leading to similarities in their epigenetic patterns.[48,49] In terms of cost and detection time, the use of whole blood as samples allows a shorter detection time and lowers costs because it does not require complex cell sorting, gradient density centrifugation, or other sample processes. However, sample processing accounts for a limited part of the total detection time and costs, while detection assays greatly influence detection time and cost. Assays based on bisulfite conversion, restriction enzymes, and affinity enrichment are currently applied to analyze specific methylated regions and, among these, bisulfite conversion is the most widely used.[50] Polymerase chain reaction (PCR),

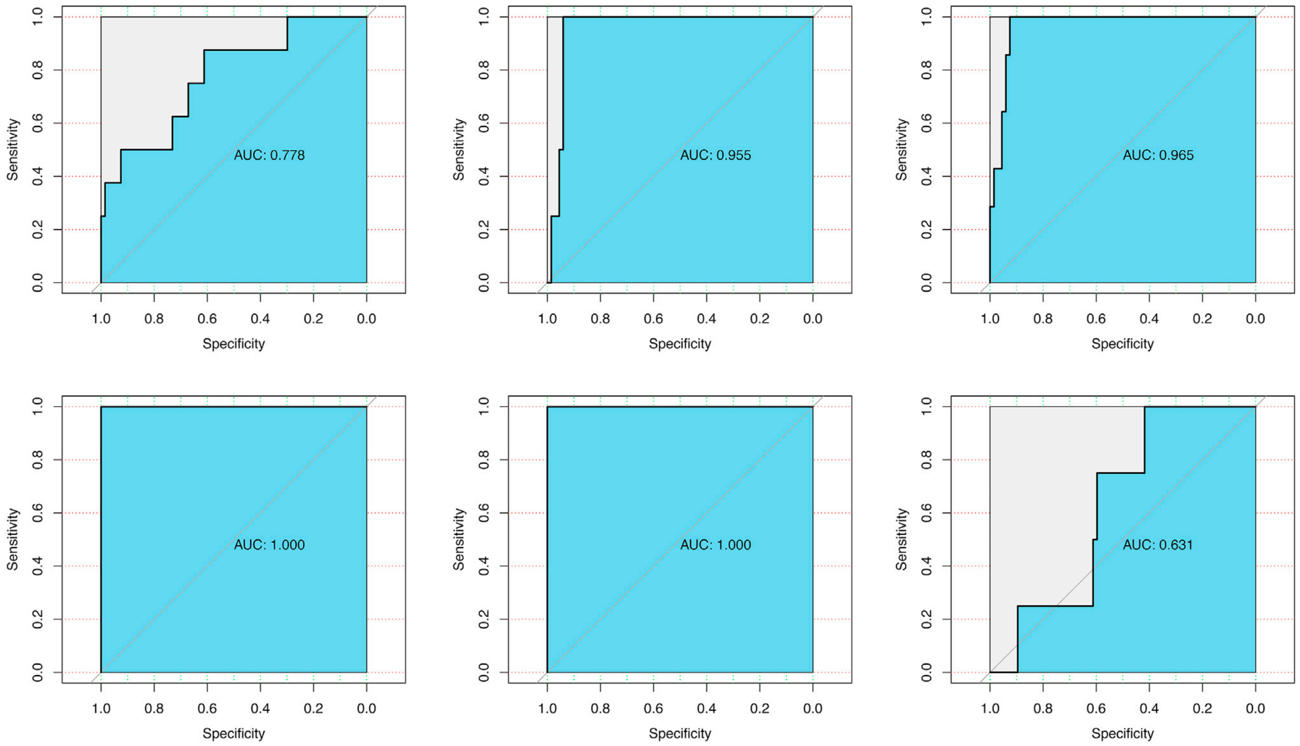

Figure 6. Receiver operating characteristic curves of the 3-DMR logistic regression classifier when distinguishing tuberculosis patients from other disease controls

(A) The receiver operating characteristic curve of the 3-DMR logistic regression classifier when distinguishing tuberculosis patients from malaria cases. (B) The receiver operating characteristic curve of the 3-DMR logistic regression classifier when distinguishing cases with systemic inflammatory response syndrome from tuberculosis patients. (C) The receiver operating characteristic curve of the 3-DMR logistic regression classifier when distinguishing sepsis patients (GEO: GSE138074) from tuberculosis patients. (D) The receiver operating characteristic curve of the 3-DMR logistic regression classifier when distinguishing sepsis patients (GEO: GSE58651) from tuberculosis patients. (E) The receiver operating characteristic curve of the 3-DMR logistic regression classifier when distinguishing sepsis patients (GEO: GSE155952) from tuberculosis patients. (F) The receiver operating characteristic curve of the 3-DMR logistic regression classifier when distinguishing tuberculosis patients from patients with subclinical parasitemia.

microarrays, and sequencing are utilized to read methylation information after bisulfite conversion.[51] PCR is suitable for primary health facilities due to its low cost, short detection time, and simple operation; however, limited sensitivity and poor ability for methylated region detection should also be taken seriously. Microarrays and sequencing overcome the above shortcomings of PCR but at the expense of high costs and rigorous laboratory conditions. When choosing different detection methods, it should be considered that these methods had different sensitivities and thus different captured efficiencies of CpG sites in targeted regions, which was also observed in this study. Unfortunately, we have not proposed a standardized process to address this problem. In addition, the differential diagnostic ability of the targeted classifier was evaluated in relatively small populations, and more reliable evidence is needed. Herein, much work remains to be done before applying a novel diagnostic pattern in clinical practice.

Collectively, the available evidence suggests that DNA methylation might be a kind of biomarker for TB diagnosis. Among all classifiers tested here, the 3-DMR logistic regression classifier presented excellent performance in both the discovery and validation datasets. This classifier might provide insights into how DNA methylation biomarkers could fit into future TB diagnosis and allow TB patients to be shielded from disease progress by timely diagnosis. Further decoding the profile of DNA methylation could offer crucial hints related to TB development, a research direction that requires an increased emphasis on TB.

MATERIALS AND METHODS

Data preparation

DNA methylation array datasets were searched in the NCBI Gene Expression Omnibus database (GEO) and European Bioinformatics Institute ArrayExpress from their inception to October 3, 2020. The search terms used were (“tuberculosis” OR “TB”) AND (“methylation” OR “methylate”), with restrictions on species (*Homo sapiens*) and sample type (peripheral blood or its components). Considering the influence of complex body interactions on methylation patterns,[52] datasets generated based on *in vitro* infection cell models, such as GEO: GSE83379, were excluded. Datasets without raw IDAT or

CEL files, such as GEO: GSE50835, were also excluded. The search strategy for DCs is described in Data S1.

Two authors (Lyu and Zhou) were responsible for reviewing the eligible datasets and extracting the design information of each included dataset. The extracted content included the demographic characteristics of the subjects (age, sex, race, etc.), number and type of samples, detection platform used, etc. Any discrepancy was resolved by discussion.

Raw data processing

To decrease the impact of using different probe filtering conditions and data processing methods, the raw files for the included datasets were downloaded and processed in the same way. The Minfi v.1.30.0 package[53] was applied to handle the data. The p value of the probe was calculated to assess the reliability of the probe signal, and a probe was filtered out if its p value was more than 0.01 in any sample.[53] Samples with poor quality, such as GEO: GSM1860483, were excluded according to the mean p values of all probes for each sample. In addition, probes with single-nucleotide polymorphisms at CpG sites or on sex chromosomes were excluded. Preprocessing, background noise reduction, and normalization were also performed.

DMP and DMR analysis

The batch effects among different datasets were adjusted by the sva v.3.32.1 package.[54] Both beta and M values were calculated to assess the methylation level of each CpG site in each sample. Based on the M value matrix,[55] the Limma package v.3.40.6[56] was used to assess DMPs with the standard of $|FC| > 1.5$ and $p < 0.05$.

For DMRs, the DMRcate package v.1.20.0[57] was employed to identify methylated regions and perform differential methylation analysis. Methylated regions were identified based on a Gaussian kernel, and DMRs were found by tunable kernel smoothing of the differential methylation signal.[57] The region methylation level was assessed on the basis of the mean beta value of all CpG sites in the corresponding region. Based on 363,416 probes, methylated regions with a false discovery rate < 0.05 were regarded as meaningful regions. The differential analysis was also performed in different ethnicities and sample types. DMR annotation was carried out by an online tool, wANNOVAR: http://wannovar.wglab.org.

Functional analysis

GO enrichment analysis and KEGG pathway enrichment analysis were performed for DMR-related genes. A PPI network was constructed to identify the hub genes among DMR-associated genes by an online tool, STRING: https://string-db.org/. The node score suggested the importance of this node in the whole PPI network. The interactions of the combined score were set at 0.4.

Diagnostic classifier development

Incorporating DMRs into classifiers might open the door to new methods for clinical diagnosis of TB. Logistic regression and elastic net regression were used to select classifier variables and construct classifiers. All DMRs underwent binary univariate and multivariable logistic regression. DMRs with $p < 0.05$ in the final multivariable logistic regression were incorporated. For elastic net regression, the key parameter λ was selected by K-fold cross-validation. In addition to the above two methods, SVM, KNN, random tree, and XGBoost were also used to construct classifiers, and optimal hyperparameters of these modeling methods were chosen by grid search or cross-validation. Grid search was realized by GridSearchCV, which evaluates and compares all scores by adjusting a series of parameter values and outputs the optimal values of the parameters that generated the highest score. Cross-validation divides all data into K parts, calculates the average verification accuracy, and selects the optimal parameter when the verification accuracy reached the best value. The Youden index was calculated to determine suitable cutoff values. Sensitivity, specificity, and AUC were used to assess the diagnostic capacity of the classifiers. Receiver operating characteristic curves were plotted to visually present classifier performance.

Validation cohort recruitment and sample preparation

A total of 62 Chinese participants admitted to West China Hospital of Sichuan University between January 2019 and October 2020 were enrolled. TB patients met the following criteria: (1) diagnosed with TB according to the Diagnostic Criteria for Tuberculosis (WS 288-2008);[58] (2) age older than 18 years; and (3) free of other lung diseases (lung cancer, chronic obstructive pulmonary disease, etc.), liver diseases (hepatitis, hepatocellular carcinoma, etc.), metabolic diseases (diabetes mellitus, hyperlipidemia, etc.), and autoimmune diseases. Pregnant women were excluded unless specifically indicated. HCs with negative results on TB-related examinations and no history of TB were recruited. The TB and HC groups were age and sex matched. All participants signed written informed consents.

EDTA-treated whole blood (3.0 mL) was collected from each subject, and genomic DNA was extracted (Kuang Yuan Diagnostics Q1001, China) according to the manufacturer's protocol. DNA purity was measured by spectrometry at 260/280 nm. The extracted DNA was stored at −80°C.

The protocol of this study was approved by the Clinical Trials and Biomedical Ethics Committee of West China Hospital, Sichuan University (registration number in the Chinese Clinical Trial Registry: ChiCTR1900028670).

Region-specific multiple sequencing and data analysis

To verify DMRs found in array analysis and test classifier performance and generalization, region-specific multiple sequencing was performed. Ultrasound was used to fragment the genomic double-stranded DNA into pieces of 300 bp. After purification with 2× magnetic beads, DNA was bisulfite-treated with an EZ DNA Methylation-Gold Kit (Zymo Research D5005, USA). Bisulfite-converted DNA samples further underwent phosphorylation modification by T4 polynucleotide kinase (Thermo Scientific EK0031, USA) and the addition of 5′ adapters. Bisulfite-specific PCR was conducted to amplify regions of interest (primers used in this step are listed in Table S1) using EpiTaq HS (TaKaRa R110A, Japan). Then, Taq (TaKaRa R001WZ, Japan) was used for a second PCR to add sample barcodes and sequencing adapters. PCR amplicons were visualized by gel electrophoresis, purified by 1.2× magnetic beads and sequenced on an Illumina Sequencer.

Raw data were subjected to quality control and read trimming. Bismark was applied to align trimmed reads to the reference genome sequence and calculate the percentage of 5-methylated cytosine (5-mC). According to DMRfinder, the methylation degree of each region was assessed by the following formula: methylation fraction of a region = sum of methylated counts at CpG sites within a region/sum of methylated and unmethylated counts at CpG sites within a region. Differences between TB patients and HCs were evaluated by Limma package v.3.40.6.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.omtn.2021.11.014.

ACKNOWLEDGMENTS

This research was supported by the National Science & Technology Pillar Program during the 13th Five-year Plan Period (grant no. 2018ZX10715003), the 1·3·5 project for disciplines of excellence–Clinical Research Incubation Project, West China Hospital, Sichuan University (grant no. 2020HXFH015) and the Technology Department of Sichuan Province (grant no. 2020YJ0049).

AUTHOR CONTRIBUTIONS

M.L., X.W., and B.Y. designed this study. W.X. and B.Y. gave administrative support. M.L., J.Z., L.J., W.X., and B.Y. collected data and conducted analysis. M.L., J.Z., L.J., Y.W., Y.Z., and H.L. were responsible for data interpretation. All authors wrote and revised this paper. All authors agreed to approve this work and were accountable for this work.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

1. Knight, G.M., McQuaid, C.F., Dodd, P.J., and Houben, R. (2019). Global burden of latent multidrug-resistant tuberculosis: trends and estimates based on mathematical modelling. Lancet Infect. Dis. *19*, 903–912.
2. Global, W.H.O. (2020). Tuberculosis report 2020. https://apps.who.int/iris/bitstream/handle/10665/336069/9789240013131-eng.pdf.
3. Rangaka, M.X., Cavalcante, S.C., Marais, B.J., Thim, S., Martinson, N.A., Swaminathan, S., and Chaisson, R.E. (2015). Controlling the seedbeds of tuberculosis: diagnosis and treatment of tuberculosis infection. Lancet *386*, 2344–2353.
4. Yang, Q., Chen, Q., Zhang, M., Cai, Y., Yang, F., Zhang, J., Deng, G., Ye, T., Deng, Q., Li, G., et al. (2020). Identification of eight-protein biosignature for diagnosis of tuberculosis. Thorax *75*, 576–583.
5. Lu, L.L., Smith, M.T., Yu, K.K.Q., Luedemann, C., Suscovich, T.J., Grace, P.S., Cain, A., Yu, W.H., McKitrick, T.R., Lauffenburger, D., et al. (2019). IFN-γ-independent immune markers of *Mycobacterium tuberculosis* exposure. Nat. Med. *25*, 977–987.
6. Lewinsohn, D.M., Leonard, M.K., LoBue, P.A., Cohn, D.L., Daley, C.L., Desmond, E., Keane, J., Lewinsohn, D.A., Loeffler, A.M., Mazurek, G.H., et al. (2017). Official American Thoracic Society/Infectious Diseases Society of America/Centers for Disease Control and Prevention clinical practice guidelines: diagnosis of tuberculosis in adults and children. Clin. Infect. Dis. *64*, 111–115.
7. Penn-Nicholson, A., Hraha, T., Thompson, E.G., Sterling, D., Mbandi, S.K., Wall, K.M., Fisher, M., Suliman, S., Shankar, S., Hanekom, W.A., et al. (2019). Discovery and validation of a prognostic proteomic signature for tuberculosis progression: a prospective cohort study. PLoS Med. *16*, e1002781.
8. WHO (2014). High-priority target product profiles for new tuberculosis diagnostics: report of a consensus meeting. https://apps.who.int/iris/bitstream/handle/10665/135617/WHO_HTM_TB_2014.18_eng.pdf?sequence=1.
9. Esterhuyse, M.M., Weiner, J., 3rd, Caron, E., Loxton, A.G., Iannaccone, M., Wagman, C., Saikali, P., Stanley, K., Wolski, W.E., Mollenkopf, H.J., et al. (2015). Epigenetics and proteomics join transcriptomics in the quest for tuberculosis biomarkers. mBio *6*, 01187–01115.
10. Gupta, R.K., Turner, C.T., Venturini, C., Esmail, H., Rangaka, M.X., Copas, A., Lipman, M., Abubakar, I., and Noursadeghi, M. (2020). Concise whole blood transcriptional signatures for incipient tuberculosis: a systematic review and patient-level pooled meta-analysis. Lancet Respir. Med. *8*, 395–406.
11. Cubillos-Angulo, J.M., Arriaga, M.B., Silva, E.C., Müller, B.L.A., Ramalho, D.M.P., Fukutani, K.F., Miranda, P.F.C., Moreira, A.S.R., Ruffino-Netto, A., Lapa, E.S.J.R., et al. (2019). Polymorphisms in TLR4 and TNFA and risk of *Mycobacterium tuberculosis* infection and development of active disease in contacts of tuberculosis cases in Brazil: a prospective cohort study. Clin. Infect. Dis. *69*, 1027–1035.
12. Gómez-Díaz, E., Jordà, M., Peinado, M.A., and Rivero, A. (2012). Epigenetics of host-pathogen interactions: the road ahead and the road behind. PLoS Pathog. *8*, e1003007.
13. Berdasco, M., and Esteller, M. (2019). Clinical epigenetics: seizing opportunities for translation. Nat. Rev. Genet. *20*, 109–127.
14. Skvortsova, K., Stirzaker, C., and Taberlay, P. (2019). The DNA methylation landscape in cancer. Essays Biochem. *63*, 797–811.
15. Vymetalkova, V., Vodicka, P., Vodenkova, S., Alonso, S., and Schneider-Stock, R. (2019). DNA methylation and chromatin modifiers in colorectal cancer. Mol. Aspects Med. *69*, 73–92.
16. Singh, S., Narayanan, S.P., Biswas, K., Gupta, A., Ahuja, N., Yadav, S., Panday, R.K., Samaiya, A., Sharan, S.K., and Shukla, S. (2017). Intragenic DNA methylation and BORIS-mediated cancer-specific splicing contribute to the Warburg effect. Proc. Natl. Acad. Sci. U S A *114*, 11440–11445.
17. Xu, Z., Sandler, D.P., and Taylor, J.A. (2020). Blood DNA methylation and breast cancer: a prospective case-cohort analysis in the sister study. J. Natl. Cancer Inst. *112*, 87–94.
18. Agha, G., Houseman, E.A., Kelsey, K.T., Eaton, C.B., Buka, S.L., and Loucks, E.B. (2015). Adiposity is associated with DNA methylation profile in adipose tissue. Int. J. Epidemiol. *44*, 1277–1287.
19. Bogoi, R.N., de Pablo, A., Valencia, E., Martín-Carbonero, L., Moreno, V., Vilchez-Rueda, H.H., Asensi, V., Rodriguez, R., Toledano, V., and Rodés, B. (2018). Expression profiling of chromatin-modifying enzymes and global DNA methylation in CD4+ T cells from patients with chronic HIV infection at different HIV control and progression states. Clin. Epigenetics *10*, 20.
20. Okamoto, Y., Shinjo, K., Shimizu, Y., Sano, T., Yamao, K., Gao, W., Fujii, M., Osada, H., Sekido, Y., Murakami, S., et al. (2014). Hepatitis virus infection affects DNA methylation in mice with humanized livers. Gastroenterology *146*, 562–572.
21. DiNardo, A.R., Rajapakshe, K., Nishiguchi, T., Grimm, S.L., Mtetwa, G., Dlamini, Q., Kahari, J., Mahapatra, S., Kay, A., Maphalala, G., et al. (2020). DNA hypermethylation during tuberculosis dampens host immune responsiveness. J. Clin. Invest. *130*, 3113–3123.
22. Wu, H., Caffo, B., Jaffee, H.A., Irizarry, R.A., and Feinberg, A.P. (2010). Redefining CpG islands using hidden Markov models. Biostatistics *11*, 499–514.
23. Blankley, S., Berry, M.P., Graham, C.M., Bloom, C.I., Lipman, M., and O'Garra, A. (2014). The application of transcriptional blood signatures to enhance our understanding of the host response to infection: the example of tuberculosis. Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci. *369*, 20130427.

24. Kostromin, A.P. (1980). Dynamics of splenic DNA methylation in guinea pigs with experimental tuberculosis. Probl. Tuberk. 67–70.

25. Kathirvel, M., and Mahadevan, S. (2016). The role of epigenetics in tuberculosis infection. Epigenomics *8*, 537–549.

26. Domcke, S., Bardet, A.F., Adrian Ginno, P., Hartl, D., Burger, L., and Schübeler, D. (2015). Competition between DNA methylation and transcription factors determines binding of NRF1. Nature *528*, 575–579.

27. Wu, L.J., Xin, Z.D., Huang, Y.C., Zhou, W.J., Zhang, J.Y., Hu, X.J., Zhuang, J., and Ying, B.W. (2019). Methylation chip screening and verification of differential genes related to tuberculosis infection. J. Sichuan Univ. Med. Sci. Ed. *50*, 234–240.

28. Pacis, A., Mailhot-Léonard, F., Tailleux, L., Randolph, H.E., Yotova, V., Dumaine, A., Grenier, J.C., and Barreiro, L.B. (2019). Gene activation precedes DNA demethylation in response to infection in human dendritic cells. Proc. Natl. Acad. Sci. U S A *116*, 6938–6943.

29. Yin, Y., Morgunova, E., Jolma, A., Kaasinen, E., Sahu, B., Khund-Sayeed, S., Das, P.K., Kivioja, T., Dave, K., Zhong, F., et al. (2017). Impact of cytosine methylation on DNA binding specificities of human transcription factors. Science *356*, eaaj2239.

30. Maruthai, K., Kalaiarasan, E., Joseph, N.M., Parija, S.C., and Mahadevan, S. (2018). Assessment of global DNA methylation in children with tuberculosis disease. Int. J. Mycobacteriol. *7*, 338–342.

31. Maruthai, K., Sankar, S., and Subramanian, M. (2020). Methylation status of VDR gene and its association with vitamin D status and VDR gene expression in pediatric tuberculosis disease. Immunol. Invest. 1–16, https://doi.org/10.1080/08820139.2020.1810702.

32. Maruthai, K., and Subramanian, M. (2018). Methylation status of alu repetitive elements in children with tuberculosis disease. Int. J. mycobacteriology *7*, 242–246.

33. Das, J., Verma, D., Gustafsson, M., and Lerm, M. (2019). Identification of DNA methylation patterns predisposing for an efficient response to BCG vaccination in healthy BCG-naïve subjects. Epigenetics *14*, 589–601.

34. Chen, Y.C., Hsiao, C.C., Chen, T.W., Wu, C.C., Chao, T.Y., Leung, S.Y., Eng, H.L., Lee, C.P., Wang, T.Y., and Lin, M.C. (2020). Whole genome DNA methylation analysis of active pulmonary tuberculosis disease identifies novel epigenotypes: PARP9/miR-505/RASGRP4/GNG12 gene methylation and clinical phenotypes. Int. J. Mol. Sci. *21*, 3180.

35. Chen, Y.C., Hsiao, C.C., Chen, C.J., Chao, T.Y., Leung, S.Y., Liu, S.F., Wang, C.C., Wang, T.Y., Chang, J.C., Wu, C.C., et al. (2014). Aberrant Toll-like receptor 2 promoter methylation in blood cells from patients with pulmonary tuberculosis. J. Infect. *69*, 546–557.

36. Wang, M., Kong, W., He, B., Li, Z., Song, H., Shi, P., and Wang, J. (2018). Vitamin D and the promoter methylation of its metabolic pathway genes in association with the risk and prognosis of tuberculosis. Clin. Epigenetics *10*, 118.

37. Wang, Y., Han, D., Pan, L., and Sun, J. (2018). The positive feedback between lncRNA TNK2-AS1 and STAT3 enhances angiogenesis in non-small cell lung cancer. Biochem. Biophys. Res. Commun. *507*, 185–192.

38. Yao, W., Yan, Q., Du, X., and Hou, J. (2021). TNK2-AS1 upregulated by YY1 boosts the course of osteosarcoma through targeting miR-4319/WDR1. Cancer Sci. *112*, 893–905.

39. Cai, T., Cui, X., Zhang, K., Zhang, A., Liu, B., and Mu, J.J. (2019). LncRNA TNK2-AS1 regulated ox-LDL-stimulated HASMC proliferation and migration via modulating VEGFA and FGF1 expression by sponging miR-150-5p. J. Cell. Mol. Med. *23*, 7289–7298.

40. Lyu, M., Cheng, Y., Zhou, J., Chong, W., Wang, Y., Xu, W., and Ying, B. (2021). Systematic evaluation, verification and comparison of tuberculosis-related non-coding RNA diagnostic panels. J. Cell. Mol. Med. *25*, 184–202.

41. Chen, Y.C., Lee, C.P., Hsiao, C.C., Hsu, P.Y., Wang, T.Y., Wu, C.C., Chao, T.Y., Leung, S.Y., Chang, Y.P., and Lin, M.C. (2020). MicroRNA-23a-3p down-regulation in active pulmonary tuberculosis patients with high bacterial burden inhibits mononuclear cell function and phagocytosis through TLR4/TNF-α/TGF-β1/IL-10 signaling via targeting IRF1/SP1. Int. J. Mol. Sci. *21*, 8587.

42. Geraghty, D.E., Wei, X.H., Orr, H.T., and Koller, B.H. (1990). Human leukocyte antigen F (HLA-F). An expressed HLA gene composed of a class I coding sequence linked to a novel transcribed repetitive element. J. Exp. Med. *171*, 1–18.

43. Sabbatino, F., Liguori, L., Polcaro, G., Salvato, I., Caramori, G., Salzano, F.A., Casolaro, V., Stellato, C., Col, J.D., and Pepe, S. (2020). Role of human leukocyte antigen system as a predictive biomarker for checkpoint-based immunotherapy in cancer patients. Int. J. Mol. Sci. *21*, 7295.

44. Dulberger, C.L., McMurtrey, C.P., Hölzemer, A., Neu, K.E., Liu, V., Steinbach, A.M., Garcia-Beltran, W.F., Sulak, M., Jabri, B., Lynch, V.J., et al. (2017). Human leukocyte antigen F presents peptides and regulates immunity through interactions with NK cell receptors. Immunity *46*, 1018–1029.e1017.

45. Lin, A., and Yan, W.H. (2019). The emerging roles of human leukocyte antigen-F in immune modulation and viral infection. Front. Immunol. *10*, 964.

46. Lunemann, S., Schöbel, A., Kah, J., Fittje, P., Hölzemer, A., Langeneckert, A.E., Hess, L.U., Poch, T., Martrus, G., Garcia-Beltran, W.F., et al. (2018). Interactions between KIR3DS1 and HLA-F activate natural killer cells to control HCV replication in cell culture. Gastroenterology *155*, 1366–1371.e1363.

47. Garcia-Beltran, W.F., Hölzemer, A., Martrus, G., Chung, A.W., Pacheco, Y., Simoneau, C.R., Rucevic, M., Lamothe-Molina, P.A., Pertel, T., Kim, T.E., et al. (2016). Open conformers of HLA-F are high-affinity ligands of the activating NK-cell receptor KIR3DS1. Nat. Immunol. *17*, 1067–1074.

48. Schlums, H., Cichocki, F., Tesi, B., Theorell, J., Beziat, V., Holmes, T.D., Han, H., Chiang, S.C., Foley, B., Mattsson, K., et al. (2015). Cytomegalovirus infection drives adaptive epigenetic diversification of NK cells with altered signaling and effector function. Immunity *42*, 443–456.

49. Lau, C.M., Adams, N.M., Geary, C.D., Weizman, O.E., Rapp, M., Pritykin, Y., Leslie, C.S., and Sun, J.C. (2018). Epigenetic control of innate and adaptive immune memory. Nat. Immunol. *19*, 963–972.

50. Pajares, M.J., Palanca-Ballester, C., Urtasun, R., Alemany-Cosme, E., Lahoz, A., and Sandoval, J. (2021). Methods for analysis of specific DNA methylation status. Methods *187*, 3–12.

51. Morselli, M., Farrell, C., Rubbi, L., Fehling, H.L., Henkhaus, R., and Pellegrini, M. (2021). Targeted bisulfite sequencing for biomarker discovery. Methods *187*, 13–27.

52. Kaelin, W.G., Jr., and McKnight, S.L. (2013). Influence of metabolism on epigenetics and disease. Cell *153*, 56–69.

53. Aryee, M.J., Jaffe, A.E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A.P., Hansen, K.D., and Irizarry, R.A. (2014). Minfi: a flexible and comprehensive bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics *30*, 1363–1369.

54. Leek, J.T., and Storey, J.D. (2007). Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet, 3:e161.

55. Du, P., Zhang, X., Huang, C.C., Jafari, N., Kibbe, W.A., Hou, L., and Lin, S.M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics *11*, 587.

56. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47.

57. Peters, T.J., Buckley, M.J., Statham, A.L., Pidsley, R., Samaras, K., Lord, R.V., Clark, S.J., and Molloy, P.L. (2015). De novo identification of differentially methylated regions in the human genome. Epigenetics Chromatin *8*, 6.

58. NHFPC (2017). Diagnostic Criteria for Tuberculosis (WS 288—2008), http://www.moh.gov.cn/zhuz/s9491/201712/a452586fd21d4018b0ebc00b89c06254.shtml.