Polymorphism in maternal HLA-DRB5 is associated with the risk of preeclampsia in Chinese population

Chenhong Xu
Shenzhen University Medical School
Lu Zhou
Maternal and Child Healthcare Hospital of Shenzhen City, Southern Medical University
Qiongfang Fang
Shenzhen University Medical School
Yinglin Liu
Maternal and Child Healthcare Hospital of Shenzhen City, Southern Medical University
Jielin Yang
Shenzhen University Medical School
Lijun Luo
Shenzhen University Medical School
Sichun Li
Shenzhen University Medical School
Peiyu Guo
Shenzhen University Medical School
Yifei Niu
Shenzhen University Medical School
Wenxin Deng
Shenzhen University Medical School
Xueqing Wu
Shenzhen University General Hospital, Shenzhen University Health Science Center
Yueming Hu
Shenzhen University Medical School
Ming-an Sun
Yangzhou University
Dong Ni
Shenzhen University Medical School
Yejun Wang (✉ wangyj@szu.edu.cn)
Shenzhen University Medical School

Research Article

Keywords: Preeclampsia, Genome-Wide Association Study, HLA-DRB5, Multi-gene Model, Early Screening

Posted Date: April 11th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2746165/v1

Abstract

Background: Preeclampsia is an important clinical syndrome occurring during pregnancy. It shows genetic disposition, and the genetic risk has large ethnic heterogeneity. The study was designed to investigate the genetic risk of preeclampsia in Chinese pregnancies, and to apply it in early screening of the disease.

Methods: We performed a genome-wide association study to screen candidate risk loci associated with preeclampsia in Chinese people, and validated them with an independent cohort of enlarged size. We also trained prediction models using the genotypes of newly identified risk loci to screen the pregnancies with high preeclampsia risk.

Results: A segment in chromosome 6 covering *HLA-DQB1*, *HLA-DRB5* and other immune-related genes shows the most significant association, and three loci in *HLA-DRB5* were confirmed with an enlarged validation cohort. One of the validated loci, rs147440497, forms an amino acid change by the nucleotide polymorphism, which further causes a conformational change in the antigen-binding domain of HLA-DRB5 protein. With the genotypes of risk genetic loci and other demographic features, a machine-learning model was trained, which can predict Chinese preeclampsia pregnancies accurately, with a cross-validated recalling rate of 0.63 at a false positive rate of 8%.

Conclusion: We identified a novel gene from maternal genome, *HLA-DRB5*, the polymorphism in which is associated with preeclampsia. The genotypes of risk SNP loci can also be used for prediction of preeclampsia risk in Chinese population accurately.

Background

Preeclampsia (PE) is an important clinical syndrome that occurs during pregnancy, which affecting 2–8% of women worldwide. It is one of the leading causes of maternal and perinatal morbidity and mortality [1–3]. PE is characterized by new onset of hypertension and proteinuria or other syndromes at ≥ 20 weeks of gestation impairing multiple maternal and fetal systems simultaneously [2, 3]. The exact mechanisms underlying the pathogenesis of PE remain unclear. It is widely believed that incomplete spiral arterial remodeling in the uterus is the main pathophysiological basis of PE [4, 5]. The former leads to placental ischemia, which further aggravates the imbalance between proangiogenic factors (placental and vascular endothelial growth factors) and antiangiogenic factors (soluble fms-like tyrosine kinase-1 and soluble endoglin), and causes damage to the maternal vascular endothelium [4, 5]. As a result, pregnancies with PE show various symptoms of vascular endothelium injury such as proteinuria, headache, and hepatic obstruction [6]. PE can be divided into two types according to the time of onset, with early onset referring to the syndrome occurring before 34 weeks and late onset occurring after 34 weeks of pregnancy [7]. Around 80–90% of the PE cases are late-onset in most countries [7]. Till present, there remains a lack of an effective therapy for PE other than to end the pregnancy [8].

Clinical studies demonstrated that low-dose aspirin could be applied for the prevention of PE among women at a high risk of developing the disease [9–12]. Drugs are not preferred for low-risk pregnancies, and therefore it appears important to screen the high-risk population during the first trimester or even at an earlier time [13, 14]. However, till now, there remains a lack of effective screening methods and corresponding markers for early detection of PE, although some factors have been found with the capability to increase the early detection rate, such as maternal age at the first birth, previous history, body mass index (BMI), and blood pressure fluctuation [15–18]. The effect remains to be improved for screening PE with a single clinical feature, serological marker, or uterine artery blood flow test [19–20]. Recent studies showed that free RNA detection from peripheral blood of pregnant women could screen pregnancies with high risk of PE independently and accurately, and yet its clinical applicability needs to be further evaluated [21, 22].

Genetic biomarkers are more stable than RNA molecules and therefore also prospective for early screening of PE. Although the genetic basis remains unclear, associations have been found between PE and polymorphisms involving multiple genes. Genome-Wide Association Studies (GWAS) have identified a large number of susceptibility gene loci of PE, which are involved in different biological processes such as immunity, vascular resistance control, coagulation, cell signaling pathways and metabolism [23–26]. The associations often show apparent ethnic specificity [27]. Johnson et al conducted a GWAS study involving 538 PE cases and 540 healthy pregnancy controls in Australian Caucasian women, and reported three single-nucleotide polymorphisms (SNPs) that were significantly associated with the occurrence of PE [24]. The SNPs reside in an intergenic region less than 15 kb downstream from the 3' terminus of the inhibin, beta B (*INHBB*) gene on 2q14.2 [24]. Zhao et al conducted a GWAS study in Iowa white women and found four other PE-associated SNPs located in the PSG family cluster (chr19:47.918–48.465 Mb) and the immediately flanking region (± 10kb) [28]. The group further performed a case-control study involving 1,070 Afro Caribbean (21 PE cases and 1049 controls), 723 Hispanic (62 cases and 661 controls) and 1,257 European (50 cases and 1,207 controls), and identified a set of potential candidate PE-associated risk SNP loci across populations [29]. McGinnis et al performed a GWAS study on PE offspring and identified a significant susceptibility locus (rs4769613) that was nearby the gene *FLT1* encoding FMS like tyrosine kinase 1 [25]. Zhou et al selected multiple genetic risk loci that have been reported to be significantly associated with PE in other multiethnic cohorts, performed genotype testing in Chinese PE and healthy pregnancies, and found very few risk alleles or genotypes consistent with previous reports [30, 31]. Some risk loci even showed opposite effects, i.e., being protective to the Chinese population for the risk alleles or genotypes reported for other ethnicity, further demonstrating the ethnic specificity of PE genetic susceptibility.

Because of the large population, late marriage and late childbirth, PE has become a severe health burden in China. However, there is still a lack of GWAS study on Chinese PE pregnancies, which could facilitate identification of novel risk genetic alleles and genotypes specific to Chinese people. In addition, there could be differences in pathogenesis and genetic risks between early- and late-onset PE. Late-onset PE was reported with closer associations with genetic factors [23, 24]. However, most of the GWAS studies did not differentiate the two PE types. In this study, we performed an exome sequencing based study in

Chinese Han pregnancies to identify late-onset PE-associated risk alleles and genotypes. A validated cohort was also recruited with enlarged size of cases and controls, to validate the candidate risk loci and identify possible risk differences between late- and early-onset PE patients.

Materials And Methods

1. Discovery cohort

The subjects in this study were pregnant women from Shenzhen Maternity & Child Healthcare Hospital, who signed the informed consent form before they were recruited to the project. The diagnostic criteria of preeclampsia are as follows: new onset of hypertension during pregnancy (systolic blood pressure ≥ 140 mmHg or diastolic blood pressure ≥ 90 mmHg, confirmed repeatedly at intervals ≥ 6 hours), accompanied by proteinuria (24-hour urine sample with protein ≥ 300 mg) or other disorders impairing the function of maternal organs (e.g., liver, kidney or nerve system) or hematological system, and / or uterine and placental dysfunction [13]. Samples of preeclampsia or healthy controls were collected from the pregnancies admitted to Shenzhen Maternity & Child Healthcare Hospital from July 2014 to May 2018. The pregnant women in the control group had no preeclampsia or previous medical history. The subjects are all from the Han ethnic group and lived in the same area during the study period. A subset of the preeclampsia cases and healthy pregnant controls were selected as the discovery cohort for GWAS analysis with further rigorous screening criteria. All cases in the cohort were late onset. All the subjects were excluded from multiple pregnancy, fetal malformation, hypertension, diabetes, chronic infection, autoimmune disease, thyroid disease, chronic nephropathy, rheumatoid arthritis or systemic lupus erythematous. The pregnancies too old (>40-year) or too young (<20 year) were excluded. Age and BMI distribution were strictly controlled and maintained similar between the preeclampsia and healthy groups. This study was approved by the Ethics Review Committee for Human Research of Shenzhen Maternity & Child Healthcare Hospital (SFYLS (2019) NO. 055).

2）Validation cohort

A total of 399 parturient were recruited from the outpatient department of high-risk pregnancy in Shenzhen Maternity & Child Healthcare Hospital from June 2014 to May 2018, including 135 cases of preeclampsia and 264 controls. The diagnostic criteria of preeclampsia were the same as above. The pregnant women in the control group had no preeclampsia or previous medical history. All the subjects were Han Chinese and lived in the same area during the study period. All the subjects signed the informed consent, and the study was approved by the Ethics Review Committee for Human Research of Shenzhen Maternal and Child Health Hospital.

3）DNA extraction, exome selection and sequencing

Peripheral blood samples were collected from the subjects of the discovery cohort, the mononuclear cells were isolated and the genome DNA was extracted with a standard DNA extraction procedure. The genome DNA was sheared to produce small fragments, Illumina specific adapters were added and sequencing libraries were built. The exome-targeted DNA fragments were captured and enriched using a SureSelect V5-post capture kit and the corresponding target enrichment workflow (Agilent, USA). The captured fragments were amplified and the products were loaded on a HiSeq X Ten sequencer, generating paired-end 101-bp reads (Macrogen, Korea).

4、Genome- and Transcriptome-wide preeclampsia association analysis

The raw sequencing reads were preprocessed by quality control, adapter removal, and filtering of the repeats and low-quality sequences. BWA was used to map the cleaned reads to human genome reference hg19 (NCBI builder GRCh37) (http://bio-bwa.sourceforge.net/bwa.shtml). After imputation and pre-genotyping, GWAS analysis was performed to the loci with minor allele counts not smaller than 10. SAIGE was applied to perform GWAS analysis [32]. Briefly, for each locus, a null logistic mixed model was fitted with the genotype data and non-genetic covariates (e.g., age, BMI, etc), followed by a univariate correlation test. Cochran's Q test was used to estimate effect-value heterogeneity. Regional correlation scatter plots were drawn using LocusZoom [33]. Stepwise conditional analysis and repeated correlation tests were performed in the range of ±1 Mb for major loci until no significant correlation was measured.

To observe possible effect of genetic polymorphisms on preeclampsia at transcriptome level, we combined the preeclampsia GWAS and mQTL data from 49 GTEx tissues using S-PrediXcan [34]. The mQTL data were obtained from PredictDB, which included eQTL annotating the expression information and sQTL curating splicing-related information. S-MultiXcan multivariate regression was performed to the whole-tissue S-PrediXcan data jointly [34]. Bonferroni thresholds ($b < 0.05/n$ genes) were set to identify statistically significant TWAS genes.

5、Genotype verification of risk loci

Genome DNA was extracted from the peripheral blood monocular cells of donors from the validation cohort. Sanger sequencing or SNaPshot (ABI, USA) was used to genotype the candidate genetic loci associated with preeclampsia. In this research, the main loci validated included the three candidate risk loci within *HLA-DRB5*, which were close to each other. Primers were designed in the conserved flanking regions on both sides, followed by PCR amplification, cloning and sequencing. The sequences generate were aligned to the reference *HLA-DRB5* gene sequence, and the genotypes were determined for the target loci.

6、Gene expression analysis

The bulky and single-cell RNA-seq (or microarray) data for placentas and peripheral blood mononuclear cells of preeclampsia and/or healthy donors were annotated from literature and publically available database (Supplemental Table S1). For bulky RNA-seq datasets, the cleaned reads after quality control were mapped to human reference genome with TopHat2 [35], quantified for genes by FeatureCounts [36], and normalized and compared with edgeR [37]. An R package Limma (version 3.50.3) was used for microarray-based gene expression quantification and comparison [38]. The single-cell gene expression matrix were loaded into, preprocessed and clustered by R package Seurat (version 4.1.1) [39]. The identity of cell clusters was discerned by specific expression of classical markers, i.e., *CD3* for T cells, *CD19* / *CD27* / *CD38* for B cells, *CD14* / *CD52* / *CD83* / *CD86* for DCs, *CD14* / *CD163*, *CD209* / *CD53* / *CSF1R* for macrophages, *ALAS2* / *HBA1* / *HBB* / *HBG1* for erythrocytes, *DKK1*/ *IGFBP1* / *PRL* for decidual cells, *PARP1* for villous cytotrophoblasts (VCTs), *HLA-G* / *PAPPA2* for extravillous trophoblast cells (EVCTs), *ECM1* for villous stromal cells (VSCs), *CGA* / *CYP19A1* and *GH2* for syncytiotrophoblasts (STs), and *CD34* / *CD44* for mesenchymal stem cells (MSCs), etc.

7□Structure modeling

RoseTTAFold was used to predict the tertiary structure of protein sequences derived from the annotated *HLA-DRB5* transcripts with polymorphous nucleotide composition at rs147440497 [40]. The 3D structure of most reliable models were visualized and compared with PyMol (https://pymol.org/2/).

8□Machine-learning models predicting pregnancies with high PE risk

PE risk prediction models were trained with the using the genotype and/or clinical features of the validation cohort. Five-fold cross validation was adopted to train and evaluate the performance of models. Briefly, the PE cases and healthy controls were randomly divided into five folds with equal size respectively, with each combination of four folds using as a training dataset and the rest one as the testing dataset. The genotype of each risk loci was used as features and represented by 2 bits of 0 and 1. The cases were also stratified according to the age and BMI respectively. Four groups were classified based on the distribution of age or BMI, i.e., <= the lower quartile, <= mean, <= the upper quartile and > the upper quartile, and the age and BMI were also represented with 2 bits of 0 and 1, respectively. Logistic regression models, SVM and Naïve Bayes models, and Random Forest models were trained with the *glm* R function, ‘e1071’ and ‘randomForest’ R packages, respectively (http://cran.r-project.org/). For the SVM models, the hyper-parameters, including the kernel, *gamma* and *cost*, were optimized were optimized using grid search based on 10-fold cross-validation. A voting-based ensembler was also built based on the prediction results of individual models.

9. Performance assessment of the prediction models

The model performance was evaluated based on prediction results of 5-fold cross-validated testing datasets. The parameters for performance assessment, including Specificity (*Sp*), Sensitivity (*Sn*), Accuracy (*Acc*), Precision (*Pre*), Receiver Operating Characteristic (ROC) curve, the area under ROC curve (*AUCroc*), Matthews Correlation Coefficient (*MCC*) and F1-Score, were well defined in Zhou et al, 2018[31] or elsewhere. The average 5-fold average performance was calculated and assessed.

10. Statistics

Statistic analysis was performed with the methods indicated in context. Unless specified, the significance threshold was preset as $p < 0.05$.

11. Codes availability

The training and testing datasets, codes for the machine learning models to predict preeclampsia risk, and the optimized hyper-parameters and parameters were curated in the website: http://61.160.194.165:3080/PE/models/PERPer2.

Results

1. Research design and clinical characteristics of cohorts

The study involves two cohorts composed by Chinese people of Han ethnicity. The first group, discovery cohort, are used to identify possible risk genetic loci and genotypes for late-onset PE by exome sequencing based Genome-Wide Association Study (GWAS), while the other one, validation cohort, are recruited to verify the candidate genetic risks identified by GWAS with an expanded size of subjects (Fig. 1). The genetic predisposition is also compared between early-onset and late-onset PEs (Fig. 1).

A total of 42 subjects were collected in the discovery cohort, including 22 late-onset PE patients, and 20 healthy controls. Confounding factors that are highly correlated with PE, such as age, BMI, etc., are not differentially distributed between groups (Table 1). None of the subjects had complications or underlying diseases that may affect PE genetic research. The blood pressure and albuminuria levels of PE cases were significantly higher than those of the control group, since they are the main diagnostic standards for PE. The fetal birth weight and neonatal score (Apgar score) of PE cases were significantly lower than those of control, while the incidence of intrauterine growth retardation (IUGR) was significantly increased in the PE group, further suggesting fetal development could be influenced by PE apparently (Table 1).

In the validation cohort, 399 subjects were enrolled, including 135 PE patients and 264 controls. The PE group included 64 late-onset and 71 early-onset cases. Age, BMI, and blood pressure were not significantly different between early-onset and late-onset subgroups, but were significantly larger than those of the control group (Table 2). Fetal birth weight and the Apgar score of newborns showed a gradual increase between early-onset PE, late-onset PE, and healthy controls, which may reflect the different effect of the disease time or types of PE on the fetus, with early-onset PE showing larger impact than late-onset PE (Table 2). The possibility could not be excluded yet that the difference was actually caused by the different birth time of the fetuses from the pregnant women with early- late-onset PE. The rate of IUGR in early-onset PE was significantly higher than that in late-onset PE, further confirming that different time or types of PE have different effect on the fetuses (Table 2). The incidence of maternal GDM, HELLP, and other important underlying diseases and complicating gynecological and obstetric diseases, were generally higher in the PE group than in the control group, but there was no significant difference between the two types of PE (Table 2). Taking together, the characteristics of the validation cohort suggested a more severe impact of early-onset PE on the fetus than late-onset PE.

2. Identification of risk genetic loci and genotypes of Chinese late-onset preeclampsia cases

Exome sequencing was performed on the discovery cohort composed of patients with late-onset PEs and healthy controls. The Manhattan plot reveals a concentrated segment of chromosome 6 (6p21.32), which is clearly associated with PE (Fig. 2A). With the preset statistical significance threshold ($p < 0.001$), 139 risk genetic loci distributed within 22 protein-encoding genes were recalled, for which the genotypes are associated with late-onset PE in Chinese people significantly (Supplemental Dataset S1). The loci are enriched in 6p21.32 most strikingly, especially in two genes, *HLA-DQB1* and *HLA-DRB5* (Fig. 2B; Supplemental Dataset S1). The significant SNPs in *HLA-DQB1* show high correlation between each other but low correlation with those in *HLA-DBR5* or other nearby genes (Fig. 2B). Similarly, there is low linkage between the significant SNPs within and outside the *HLA-DRB5* gene (Supplemental Fig. S1). The recombination rates are generally low between *HLA-DQB1* or *HLA-DRB5* and other loci (Fig. 2B; Supplemental Fig. S1). The results suggest that *HLA-DQB1* and *HLA-DRB5* could be associated with PE incidence independently at gene level, while the SNPs within individual genes could potentially contribute to the disease synergistically.

Furthermore, we combined GTEx gene expression data and transcriptome-wide disease association study (TWAS) approaches, and identified three genes significantly associated with late-onset PE, i.e., *HLA-DQB1*, *CFAP61*, and *RARS2* (Fig. 2C; Supplemental Dataset S2). Among them, *HLA-DQB1* is most significant and typical (Fig. 2C). *HLA-DRB5* was not included in the original model, and therefore could not be analyzed for its possible association with PE at transcriptome level.

We compared the preeclampsia-associated loci with the significant GWAS results of other cohorts. However, only one common locus was identified, i.e., rs4762 in *AGT* gene, which was reported with significant association with preeclampsia in cohorts of multiple ethnics previously. It only shows low significance in the current Chinese GWAS cohort before multi-testing correction ($p = 0.035$). Despite the potential contribution of small size of the GWAS cohort to the relatively low power, it could also reflect the heterogeneity of preeclampsia genetic risk in different ethnics [30, 31].

3. Validation of three novel preeclampsia risk loci within *HLA-DRB5* gene

Based on the results of GWAS and TWAS analysis, we identified a risk genetic region 6p21.32 that showed a strong association with Chinese late-onset PE and the polymorphic genetic loci were mostly enriched in genes *HLA-DQB1* and *HLA-DRB5*. To further confirm the association of the region and genes with PE, we included more PE and control samples in a validation cohort, resolving the nucleotide and genotype composition for candidate risk polymorphic loci within the region. However, the main association loci within *HLA-DQB1* and some loci in *HLA-DRB5* are not suitable for design of sequencing or genotyping primers attributed to the poor conservation of flanking sequences. Eventually, only three risk loci within *HLA-DRB5* were detected successfully for the base composition and genotypes in the complete validation cohort (Table 3).

For all the three loci, the results in the validation cohort are consistent with the GWAS results. The proportion of minor alleles for rs147440497 (GRCh37.p13 - Chr6: 32489888) increases significantly in the PE population, and the odd ratio reaches 2.0 (95% CI: 1.5-2.8) (Table 3; Fig. 3A). For rs141378803

(GRCh37.p13 - chr6: 32490016) and rs149025589 (GRCh37.p13 - chr6: 32490000), the minor alleles in healthy controls even shift to major ones in PE population, with composition odd ratios of 2.4 (95% CI: 1.8-3.3) and 2.3 (95% CI: 1.7-3.1) between PE and control, respectively (Table 3; Fig. 3A). The proportion of homozygous genotype consisting of minor alleles in control for each locus is significantly higher in the PE population, and the recessive heritability related odd ratios of rs147440497, rs141378803, and rs149025589 are 6.0 (95% CI: 2.9-12.5) and 3.6 (95% CI: 2.2-5.7) and 4.0 (2.5-6.4), respectively (Table 3; Fig. 3B).

The PE cases in the validation cohort could be further divided into late-onset and early-onset subgroups. In order to observe whether the PE genetic associations are specific to late-onset cases, the association analysis was repeated to the different types of PE for the three candidate loci. However, for both allele and genotype, the composition distributions for all the three loci show significant risk associations with either type of PE, and there is no significant difference between the two PE subgroups (Table 3; Fig. 3A-B).

The SNPs in *HLA-DRB5*, especially rs14740497, show some correlation with other SNPs in the same gene based on the GWAS data (Supplemental Fig. S1). Therefore, we further explored the possible linkage of genotypes among the three validated sites within *HLA-DRB5*. The three sites were genotyped with Sanger sequencing using the same set of primers for the validation cohort, and therefore they could be haplotyped. For rs147440497, rs149025589 and rs141378803, seven combinatorial forms for the nucleotide composition could be identified except for ‘TCG’, though three haplotypes, i.e., ‘ATA’, ‘TTA’ and ‘TTG’, are rarely detected (Fig. 3C). Haplotypes ‘TCA’ and ‘ACA’ are enriched while ‘ATG’ is depleted in PE cases significantly (Fig. 3C; χ2 tests, $p < 0.05$). Most of the haplotypes show similar proportion between early-onset and late-onset PE cases, except for ‘ACA’, which is with marginally significant higher proportion in early-onset PE cases (Fig. 3C; χ2 tests, $p = 0.057$).

4. Possible function of *HLA-DRB5* on preeclampsia incidence

To infer the possible functional relevance on PE incidence, gene expression of *HLA-DRB5* was examined in the bulky RNA-seq data obtained from placentas of 5 Chinese PE and 5 healthy pregnancies. The gene is detected with low expression in both types of tissues and with slightly higher expression in controls though not being significant statistically (Fig. 4A). The results are consistent with a microarray-based gene expression profiling experiment involving the placentas of 15 PE and 10 healthy control pregnancies (Fig. 4B). Interestingly, the transcriptional level of *HLA-DRB5* is significantly higher in late-onset PE cases than in early-onset cases (Fig. 4C). The peripheral blood mononuclear cells (PBMCs) of the pregnancies express *HLA-DRB5* with increased abundance, but not differentially between PE cases and healthy controls (Fig. 4D). *HLA-DQB1* is expressed in a pattern similar to *HLA-DRB5* in the placentas and PBMCs of PE and healthy pregnancies (Supplemental Fig. S2).

A single-cell RNA-seq dataset was further explored to observe the cell types expressing *HLA-DRB5*, which comprises placentas of two PE and two healthy Chinese pregnancies. The gene is expressed with low level in both PE and control placentas, but preferentially in dendritic cells (DCs), a C1Q-expressed macrophage subset and mesenchymal stem cells (MSCs), which also show the most abundant expression of *HLA-DQB1* (Fig. 4E; Supplemental Fig. S3A). No difference is found between PE and control, for the expression level of either *HLA-DRB5* or *HLA-DQB1* (Fig. 4E; Supplemental Fig. S3A). In the peripheral blood mononuclear cells of healthy human subjects, *HLA-DRB5* and *HLA-DQB1* are also mainly expressed in subsets of myeloid cells and B cells (Supplemental Fig. S3B-C). Both *HLA-DRB5* and *HLA-DQB1* were also found with preferential expression in macrophages and DCs in human placentas and other tissues, based on online single-cell type analysis with the Human Protein Atlas (https://www.proteinatlas.org). These cell types where the genes are mainly expressed are major antigen presentation cells (APCs), and therefore *HLA-DRB5* and *HLA-DQB1* may contribute to PE incidence by antigen presentation processes.

Without typical difference in expression level, *HLA-DRB5* could still contribute to PE pathogenicity or progress by the differential protein structure and function caused by the single nucleotide variations. Among the 3 PE risk sites of *HLA-DRB5*, rs141378803 and rs149025589 are located in the same intron dispersed by 16 nucleotides (Supplemental Fig. S4). The third site, rs147440497, is located at the 164th nucleotide and within the second exon of the corresponding transcript. The variation from T to A at the site causes a change of the encoding amino acid from Phe to Tyr (F55Y) (Supplemental Fig. S4). The exon where rs147440497 is located and the upstream sequences are known to encode an extracellular domain, which is important for antigen recognition and presentation (Supplemental Fig. S4). Structural modeling analysis demonstrated that the F55Y substitution causes a change in an important turn angle, which further lead to a striking conformation change of the extracellular domain for antigen recognition and presentation (Fig. 4F).

Taken together, the results indicate that *HLA-DRB5* potentially functions in the pathogenicity or progress of PE by working differentially in the antigen presentation process through the local protein structure changes endued by the single nucleotide polymorphisms. *HLA-DQB1* may contribute to PE by similar mechanisms.

5. A multi-gene model predicts Chinese women with high preeclampsia risk accurately

The genotypes of the three genetic loci in *HLA-DRB5* were tested for the capability to predict the PE risk of Chinese women. A Support Vector Machine (SVM) model was trained only with the three loci of single nucleotide polymorphisms, and the Chinese validation cohort was predicted with the model. The 5-fold cross-validated average Area Under the Receiver Operating Characteristic Curve (ROC-AUC) reaches 0.65 while the accuracy reaches 0.69 (Fig. 5A). The model performs as well as those developed previously that are based on both the genotypes of 5 or 8 PE risk loci and clinical factors [31]. We also included the 8 previously identified loci and the age and BMI at pregnancy, and trained different regression or machine learning models. The AUC of ROCs is close to each other, averagely 0.83, 0.82, 0.85 and 0.84 for SVM, Logistic Regression (LR), Random Forest (RF) and Naïve Bayes (NB) models, respectively (Fig. 5B). A voting-based stacking model was built with the four primary models, but its performance is not better than individual ones (Fig. 5B). Other performance items such as Sensitivity (*Sn*), Specificity (*Sp*), Accuracy (*Acc*), Precision (*Pre*), F1-score and Mathews Correlation Coefficient (*MCC*) were also evaluated and compared. Generally, the SVM model shows best performance, outperforming the LR, RF and NB ones (Fig. 5C). The model can reach a sensitivity of 0.63 at a high specificity of 0.92 to predict the cases with high PE risk in Chinese pregnancy population (Fig. 5C).

Discussion

A GWAS study identified a genomic segment associated with the incidence of preeclampsia in Chinese population, which is located in 6p12.32 of maternal genome and covers immune-related genes including *HLA-DRB5* and *HLA-DQB1*. Genotyping analysis in a validation cohort confirmed the association of three SNP loci in *HLA-DRB5*, rs141378803, rs149025589 and rs147440497, with preeclampsia. Structure modeling demonstrated a striking local conformation change in HLA-DRB5 due to the nucleotide polymorphism of rs147440497. Genotypes of the three SNP loci in *HLA-DRB5* largely improved the performance of models predicting the pregnancies with high preeclampsia risk in Chinese population.

The genotype polymorphism was reported to be associated with preeclampsia incidence for *HLA-DQB1* previously [41, 42]. The association of *HLA-DQB1* was mainly reported in East Asian, and occasionally in West Asian population [43]. We confirmed the observation and broadened it to the adjacent gene *HLA-DRB5* newly using the GWAS cohort in this study (Fig. 2A-B). Risk SNPs located in *HLA-DRB5* were further validated in the validation cohort (Fig. 3A-B). Both *HLA-DRB5* and *HLA-DQB1* encode HLA II molecules and play important roles in antigen presentation [44]. Using bulky and single-cell RNA sequencing datasets from the placenta and peripheral blood samples of preeclampsia and healthy donors, we confirmed the expression of the two genes in major APCs preferentially (Fig. 4E). Considering the hypothesis of immunity-related pathogenicity [45, 46], *HLA-DRB5* and *HLA-DQB1* could potentially contribute to the incidence or progress of preeclampsia.

Different hypotheses have been proposed to explain late- and early-onset preeclampsia.[7] Previous GWAS studies seldom differentiated the different preeclampsia types. We attempted to investigate the genetic risk in cases of early-onset preeclampsia with the GWAS cohort, and then compare the candidate risk loci between late- and early-onset cases with the validation cohort (Fig. 1). We cannot exclude the possibility that there are other loci either not validated or not identified that contribute to late- and early-onset preeclampsia differentially. According to our results, the three risk sites in *HLA-DRB5* identified from exclusively late-onset preeclampsia cases of the GWAS cohort, interestingly, all showed the similar association with the incidence of early-onset and late-onset preeclampsia in the validation cohort (Fig. 3A-B). The immune processes involving *HLA-DRB5* could contribute to both types of preeclampsia. However, we also noticed higher expression of *HLA-DRB5* (and *HLA-DQB1*) in the placentas of late-onset preeclampsia cases (Fig. 4C; Supplemental Fig. S2C). It remains unclear whether the genes function differentially or it only reflects different immune statuses for the two types of preeclampsia.

Besides *HLA-DRB5* and *HLA-DQB1*, the risk loci and genotypes also show large heterogeneity between Chinese and other ethnic groups. The heterogeneity is not likely false interpretation due to the small size of the GWAS cohort, since the size mainly influence the power to detect significantly associated loci [47] while the Chinese cohort-specific risk loci are statistically significant and some of them were validated with an enlarged cohort independently. It is possible that more intersects of risk loci could be identified with other cohorts if the power was increased with enlarged cohort size; however, the most significantly associated SNPs and genes could still show large heterogeneity. The ethnic heterogeneity of genetic risks in preeclampsia has been reported frequently [27, 30, 31], though the mechanisms remain unknown.

The SNPs identified in this study could also be used as biomarkers for preeclampsia. Recently, other biomarkers have also been identified for preeclampsia, e.g., fetal cfRNAs in peripheral blood of pregnancies, which show a prospect for early screening of pregnancies with high risk of preeclampsia [21, 22]. In this study, we identified novel risk loci in Chinese preeclampsia cases and developed polygenetic models that can recall the preeclampsia cases with high accuracy (Fig. 5). Hopefully, combination of the multi-different biomarkers, clinical factors and screening approaches would facilitate early detection of the pregnancies with high preeclampsia risks precisely.

An SNP locus in *HLA-DRB5* (rs147440497) associated with preeclampsia is located in the extracellular region, and the F-to-Y mutation causes a striking conformation change of the antigen-binding domain [48] (Fig. 4F). The antigen-binding domain of HLA-DRB5 contains two sub-domains while F55 or Y55 is located at the connecting site. F55 forms a narrow angle connecting the two anti-paralleled sub-domains at its two sides, whereas the turning angle of Y55 becomes wide, leading to disappear of the anti-paralleled structure of subdomains (Fig. 4F). Therefore, we tentatively hypothesize that the polymorphism of *HLA-DRB5*, and possibly *HLA-DQB1*, could influence the capability of recognition and presentation of sperm antigens, thereby causing the extent difference of maternal immune responses against the placenta and the incidence of preeclampsia. The association is required to be confirmed with larger size of cohorts in future, and more experiments should be designed to validate the hypothesis and to investigate the specific molecular mechanisms.

We found a consolidated association between the three SNP loci in *HLA-DRB5* and preeclampsia in Chinese population, according to multiple lines of evidence. However, the study also has some limitations. One of major concerns could be the small size of GWAS cohort. We tried to select the cases carefully with confounding factors excluded or controlled. For example, besides the recruited cases of exclusively late-onset preeclampsia without basic or other gynecological and obstetric diseases, the subject number and distribution of age and BMI were similar between PE and control groups. A validation cohort was recruited to ensure the accuracy of positive observations from the discovery cohort. However, the low power caused by inborn property of small cohort size could not be overcome, and therefore more genetic loci with lower but significant association with preeclampsia remain to be identified.

Conclusions

The polymorphism of *HLA-DRB5* is associated with the incidence of preeclampsia in Chinese pregnancies. The genotypes of rs141378803, rs149025589 and rs147440497 in *HLA-DRB5* can facilitate prediction of the women with high risk of preeclampsia.

Abbreviations

DCs: dendritic cells; EVCTs: extravillous trophoblast cells; MSCs: mesenchymal stem cells; PE: preeclampsia; STs: syncytiotrophoblasts; VCTs: villous cytotrophoblasts; VSCs: villous stromal cells.

Declarations

Ethics approval and consent to participate

This study was approved by the Ethics Review Committee for Human Research of Shenzhen Maternity & Child Healthcare Hospital (SFYLS (2019) NO. 055). All methods were carried out in accordance with relevant guidelines and regulations. All the subjects signed the informed consent.

Consent for publication

Not applicable.

Availability of data and materials

The GWAS sequencing data were submitted into and stored in the National Genomics Data Center (NGDC) under the accession of HRA002503 (https://ngdc.cncb.ac.cn/).

Competing Interests

The authors declare no conflict of interest.

Funding

The study was supported by a Shenzhen University Multidisciplinary Team Project (no. 843-000033), a Natural Science Fund of Shenzhen (JCYJ20190808165205582), a Shenzhen Maternal and Child Healthcare Hospital 2018 Internal Research Fund and Doctoral Research Fund Project (FYA2018009). C.X. was supported by an Undergraduate Training Program of Guangdong Province for Innovation and Entrepreneurship (no. 202210590102). All funders had no role in the study design, data collection, data analysis, data interpretation, or writing of the manuscript.

Authors' contributions

Y.W. and D.N. conceived and supervised the project. C.X., L.Z., Q.F., J.Y., L.L., S.L., P.G., Y.N., W.D., Y.H., Y.W. performed the data analysis. L.Z. and Y.L. contributed the materials. S.L., Y.H. and Y.W. contributed algorithms and tools. C.X., L.Z., Q.F., L.L., S.L., P.G., W.D., D.N., Y.W. wrote the first draft. M.A.S., X.W., D.N. and Y.W. revised the manuscript. All the authors approved the final manuscript.

Acknowledgements

There is no acknowledgement.

References

1. Steegers EA, Von Dadelszen P, Duvekot JJ, Pijnenborg R. Pre-eclampsia. Lancet. 2010; 376(9741): 631-644.
2. Chaiworapongsa T, Chaemsaithong P, Yeo L, Romero R. Pre-eclampsia part 1: current understanding of its pathophysiology. Nat Rev Nephrol. 2014; 10(8): 466-480.
3. Johnston AN, Batts TL, Langohr IM, Moeller C, Liu CC, Sones JL. The BPH/5 mouse model of superimposed preeclampsia is not a model of HELLP syndrome. Biology (Basel). 2021; 10(11): 1179.
4. Fisher SJ. Why is placentation abnormal in preeclampsia? Am J Obstet Gynecol. 2015; 213(4 Suppl): S115-S122.
5. Cerdeira AS, Karumanchi SA. Angiogenic factors in preeclampsia and related disorders. Cold Spring Harb Perspect Med. 2012; 2(11): a006585.
6. Rana S, Lemoine E, Granger JP, Karumanchi SA. Preeclampsia: pathophysiology, challenges, and perspectives. Circ Res. 2019; 124(7): 1094-1112.
7. Robillard PY, Dekker G, Scioscia M, Saito S. Progress in the understanding of the pathophysiology of immunologic maladaptation related to early-onset preeclampsia and metabolic syndrome related to late-onset preeclampsia. Am J Obstet Gynecol. 2022; 226(2S): S867-S875.
8. Bokslag A, Van Weissenbruch M, Mol BW, De Groot CJ. Preeclampsia; short and long-term consequences for mother and neonate. Early Hum Dev. 2016; 102: 47-50.
9. US Preventive Services Task Force, Davidson KW, Barry MJ, et al. Aspirin use to prevent preeclampsia and related morbidity and mortality: US preventive services task force recommendation statement. JAMA. 2021; 326(12): 1186-1191.
10. Rolnik DL, Wright D, Poon LC, et al. Aspirin versus placebo in pregnancies at high risk for preterm preeclampsia. N Engl J Med. 2017; 377(7): 613-622.
11. Gu W, Lin J, Hou YY, et al. Effects of low-dose aspirin on the prevention of preeclampsia and pregnancy outcomes: A randomized controlled trial from Shanghai, China. Eur J Obstet Gynecol Reprod Biol. 2020; 248: 156-163.
12. Henderson JT, Vesco KK, Senger CA, Thomas RG, Redmond N. Aspirin use to prevent preeclampsia and related morbidity and mortality: Updated evidence report and systematic rreview for the US preventive services task force. JAMA. 2021; 326(12): 1192-1206.
13. Chappell LC, Cluver CA, Kingdom J, Tong S. Pre-eclampsia. Lancet. 2021; 398(10297): 341-354.
14. Ma'ayeh M, Costantine MM. Prevention of preeclampsia. Semin Fetal Neonatal Med. 2020; 25(5): 101123.
15. Henderson JT, Thompson JH, Burda BU, Cantor A. Preeclampsia screening: evidence rreport and systematic review for the US preventive services task force. JAMA. 2017; 317(16): 1668-1683.
16. Goffin SM, Derraik JGB, Groom KM, Cutfield WS. Maternal pre-eclampsia and long-term offspring health: Is there a shadow cast? Pregnancy Hypertens. 2018; 12: 11-15.

17. Mula R, Meler E, Albaiges G, Rodriguez I. Strategies for the prediction of late preeclampsia. J Matern Fetal Neonatal Med. 2019; 32(22): 3729-3733.

18. McCarthy FP, Ryan RM, Chappell LC. Prospective biomarkers in preterm preeclampsia: A review. Pregnancy Hypertens. 2018; 14: 72-78.

19. ACOG Committee on Practice Bulletins. ACOG Practice bulletin no. 134: fetal growth restriction. Obstet Gynecol. 2013; 121(5): 1122-1133.

20. Zhou L, Sun H, Cheng R, Fan X, Lai S, Deng C. ELABELA, as a potential diagnostic biomarker of preeclampsia, regulates abnormally shallow placentation via APJ. Am J Physiol Endocrinol Metab. 2019; 316(5): E773-E781.

21. Moufarrej MN, Vorperian SK, Wong RJ, et al. Early prediction of preeclampsia in pregnancy with cell-free RNA. Nature. 2022; 602(7898): 689-694.

22. Rasmussen M, Reddy M, Nolan R, et al. RNA profiles reveal signatures of future health and disease in pregnancy. Nature. 2022; 601(7893): 422-427.

23. Nejatizadeh A, Stobdan T, Malhotra N, Pasha MA. The genetic aspects of pre-eclampsia: achievements and limitations. Biochem Genet. 2008; 46(7-8): 451-479.

24. Johnson MP, Brennecke SP, East CE, et al. Genome-wide association scan identifies a risk locus for preeclampsia on 2q14, near the inhibin, beta B gene. PLoS One. 2012; 7(3): e33666.

25. McGinnis R, Steinthorsdottir V, Williams NO, et al. Variants in the fetal genome near FLT1 are associated with risk of preeclampsia. Nat Genet. 2017; 49(8): 1255-1260.

26. Gray KJ, Saxena R, Karumanchi SA. Genetic predisposition to preeclampsia is conferred by fetal DNA variants near FLT1, a gene involved in the regulation of angiogenesis. Am J Obstet Gynecol. 2018; 218(2): 211-218.

27. Johnson JD, Louis JM. Does race or ethnicity play a role in the origin, pathophysiology, and outcomes of preeclampsia? An expert review of the literature. Am J Obstet Gynecol. 2020; S0002-9378(20)30769-9.

28. Zhao L, Triche EW, Walsh KM, et al. Genome-wide association study identifies a maternal copy-number deletion in PSG11 enriched among preeclampsia patients. BMC Pregnancy Childbirth. 2012; 212: 61.

29. Zhao L, Bracken MB, DeWan AT. Genome-wide association study of pre-eclampsia detects novel maternal single nucleotide polymorphisms and copy-number variants in subsets of the Hyperglycemia and Adverse Pregnancy Outcome (HAPO) study cohort. Ann Hum Genet. 2013; 77(4): 277-287.

30. Zhou L, Cheng L, He Y, Gu Y, Wang Y, Wang C. Association of gene polymorphisms of FV, FII, MTHFR, SERPINE1, CTLA4, IL10, and TNFalpha with pre-eclampsia in Chinese women. Inflamm Res. 2016; 65(9): 717-724.

31. Zhou L, Hui X, Yuan H, Liu Y, Wang Y. Combination of genetic markers and age effectively facilitates the identification of people with high risk of preeclampsia in the Han Chinese population. Biomed Res Int. 2018; 2018: 4808046.

32. Zhou W, Nielsen JB, Fritsche LG, et al. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat Genet. 2018; 50(9): 1335-1341.
33. Pruim RJ, Welch RP, Sanna S, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010; 26(18): 2336-7.
34. Barbeira AN, Dickinson SP, Bonazzola R, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun. 2018; 9(1): 1825.
35. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013; 14: R36.
36. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014; 30(7): 923-30.
37. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010; 26(1): 139-40.
38. Sun MA, Shao X, Wang Y. Microarray data analysis for transcriptome profiling. Methods Mol Biol. 2018; 1751: 17-33.
39. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. Cell. 2021; 184(13): 3573-3587.
40. Baek M, DiMaio F, Anishchenko I, et al. Accurate prediction of protein structures and interactions using a three-track neural network. Science. 2021; 373(6557): 871-876.
41. Mao Y, Zhang Z, Fan L, et al. HLA-DQA1, -DQB1 polymorphism distribution in Chinese women with pregnancy induced hypertension in Shanghai area. Chin Med J (Engl). 1998; 111(2): 163-5.
42. Honda K, Takakuwa K, Hataya I, Yasuda M, Kurabayashi T, Tanaka K. HLA-DQB1 and HLA-DPB1 genotypes in severe preeclampsia. Obstet Gynecol. 2000; 96(3): 385-9.
43. Mohammadi M, Farazmandfar T, Shahbazi M. Relationship between human leukocyte antigen (HLA)-DQA1*0102/HLA-DQB1*0602 polymorphism and preeclampsia. Int J Reprod Biomed. 2017; 15(9): 569-574.
44. Roche PA, Furuta K. The ins and outs of MHC class II-mediated antigen processing and presentation. Nat Rev Immunol. 2015; 15(4): 203-16.
45. Aneman I, Pienaar D, Suvakov S, Simic TP, Garovic VD, McClements L. Mechanisms of key innate immune cells in early- and late-onset preeclampsia. Front Immunol. 2020; 11: 1864.
46. Collier ARY, Smith LA, Karumanchi SA. Review of the immune mechanisms of preeclampsia and the potential of immune modulating therapy. Hum Immunol. 2021; 82(5): 362-370.
47. Sham PC, Purcell SM. Statistical power and significance testing in large-scale genetic studies. Nat Rev Genet. 2014; 15: 335–346.
48. Li Y, Li H, Martin R, Mariuzza RA. Structural basis for the binding of an immunodominant peptide from myelin basic protein in different registers by two HLA-DR2 proteins. J Mol Biol. 2000; 304: 177-

88.

Tables

Table 1. Clinical characteristics of the PE and control groups in GWAS cohort

Characteristic	Case (22)	Control (20)	*p*-value
Age	28.27±4.89	27.61±4.08	0.65
BMI	23.75±4.44	22.46±4.51	0.38
SBP	164.05±19.45	120.60±8.04	1.36E-11
DBP	100.23±13.49	74.91±5.49	1.24E-09
Weight of fetus	2739.05±527.09	3329.00±423.52	0.0004
Urine protein	2+	-	8.34E-12
Apgar score	9.88±0.22	10.00	0.02
IUGR	4/22	0/20	0.10
GDM	0/22	0/20	NA
HELLP	0/22	0/20	NA
Other basic and maternity diseases	0/22	0/20	NA

Note: The blood pressure on admission was recorded.

Table 2. Clinical characteristics of the PE and control groups in validation cohort

Characteristic	**PE (135)**	**Control (264)**	***p*-value**
Age (years)	31.76±5.03	28.99±3.72	8.57E-11
BMI	32.04±20.25	19.66±5.49	1.52E-17
SBP (mmHg)	167.59±18.89	121.96±8.63	3.13E-125
DBP (mmHg)	104.64±12.19	75.84±7.04	9.97E-112
Weight of fetus (g)	1940.98±819.86	3331.74±484.17	1.72E-78
Urine protein	2+	-	7.06E-103
Apgar score	9.46±1.13	9.95±0.62	1.72E-08
IUGR	48/135	0/264	0.000
GDM	23/135	1/264	0.000
HELLP	6/135	0/264	0.001
Other basic and maternity diseases	70/135	53/264	0.000

Characteristic	**Later PE (64)**	**Early PE (71)**	***p*-value**
Age	31.78 ±4.83	31.73 ±5.18	0.96
BMI	33.45±18.55	34.66±19.58	0.73
SBP	166.38±19.03	168.69±18.69	0.48
DBP	103.59±11.26	105.58±12.90	0.35
Weight of fetus	2563.57±588.93	1367.42±484.94	1.00E-23
Urine protein	2+	2+	0.15
Apgar score	9.86±0.30	9.04±1.09	1.58E-08
IUGR	15/64	33/71	0.01
GDM	14/64	9/71	0.18
HELLP	2/64	4/71	0.68
Other basic and maternity diseases	32/64	38/71	0.73

Note: The blood pressure on admission was recorded.

Table 3. Contribution of SNP genotypes and alleles to PE risk in the validation cohort

SNP	Genotype	PE (E-L)#[1]	Ctrl#	OR[2]	OR_E[3]	OR_L[4]
rs147440497	TT	28 (14-14)	11	**6.0** (2.9-12.5)	**5.7** (2.4-13.1)	**6.4** (2.8-15.0)
	AT	40 (22-18)	92			
	AA	67 (35-32)	161			
rs149025589	CC	63 (37-26)	47	**4.0** (2.5-6.4)	**4.0** (2.3-7.0)	**3.1** (1.7-5.7)
	TC	25 (11-14)	93			
	TT	45 (22-23)	117			
rs141378803	AA	58 (32-26)	45	**3.6** (2.2-5.7)	**3.9** (2.2-6.9)	**3.2** (1.8-5.9)
	AG	26 (14-12)	72			
	GG	51 (25-26)	141			

SNP	Allele	PE (E-L)#	Ctrl#	OR	OR_E	OR_L
rs147440497	T	96 (50-46)	114	**2.0** (1.5-2.8)	**2.0** (1.3-3.0)	**2.0** (1.3-3.1)
	A	174 (92-82)	414			
rs149025589	C	151 (85-66)	187	**2.3** (1.7-3.1)	**2.7** (1.8-4.0)	**1.9** (1.3-2.9)
	T	115 (55-60)	327			
rs141378803	A	142 (78-64)	162	**2.4** (1.8-3.3)	**2.7** (1.8-3.9)	**2.2** (1.5-3.2)
	G	128 (64-64)	354			

Note: [1] The number of early-onset (E) and late-onset (L) cases was recorded in the parentheses separated by a hyphen. [2] Average odd ratio (OR) between PE and control (Ctrl) was shown in bold and the 95% confidential intervals were shown in the parentheses. [3] OR_E: OR between PE and control for the early-onset cases. [4] OR_L: OR between PE and control for the late-onset cases.

Figures

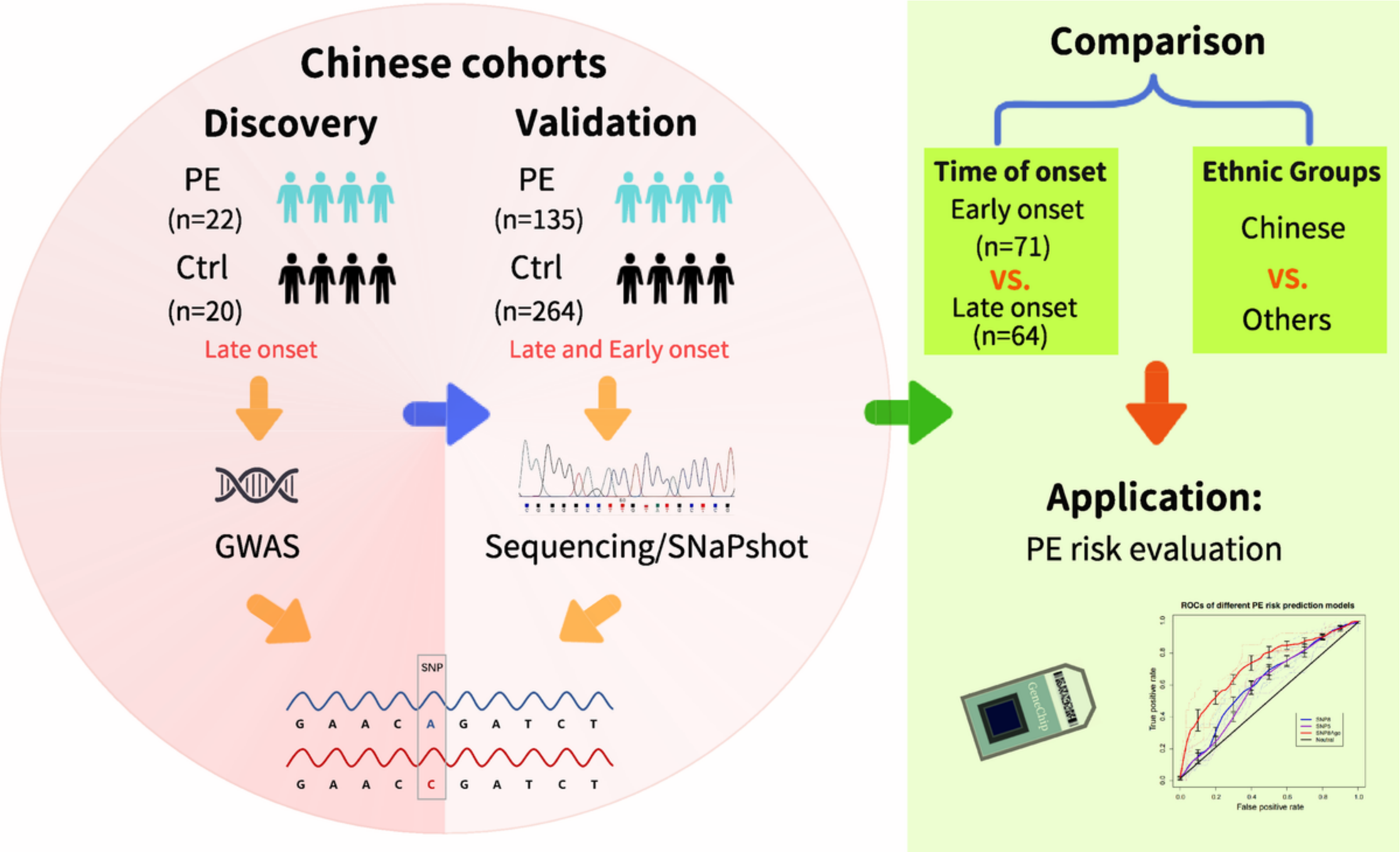

Figure 1

Research scheme and the subject composition of cohorts. Two cohorts are used in the study, i.e., discovery cohort and validation cohort. Both of the cohorts are Chinese women, comprising pregnancies with PE (PE) and healthy controls (Ctrl). The PE cases in discovery cohort are all late-onset, while the cases in validation cohort contain both late- and early- onset ones. The discovery cohort is used to perform GWAS analysis and to screen candidate risk loci and genotypes of Chinese PE, followed by verification of them with the validation cohorts using sequencing or SNaPshot based genotyping. The identified Chinese-associated PE risk alleles and genotypes are further compared between the two types of PE, i.e., early- and late- onset cases. They are also compared between different ethnic groups. Finally, machine-learning models are trained using the PE-associated alleles and their genotypes to screen the pregnancies with high PE risks.

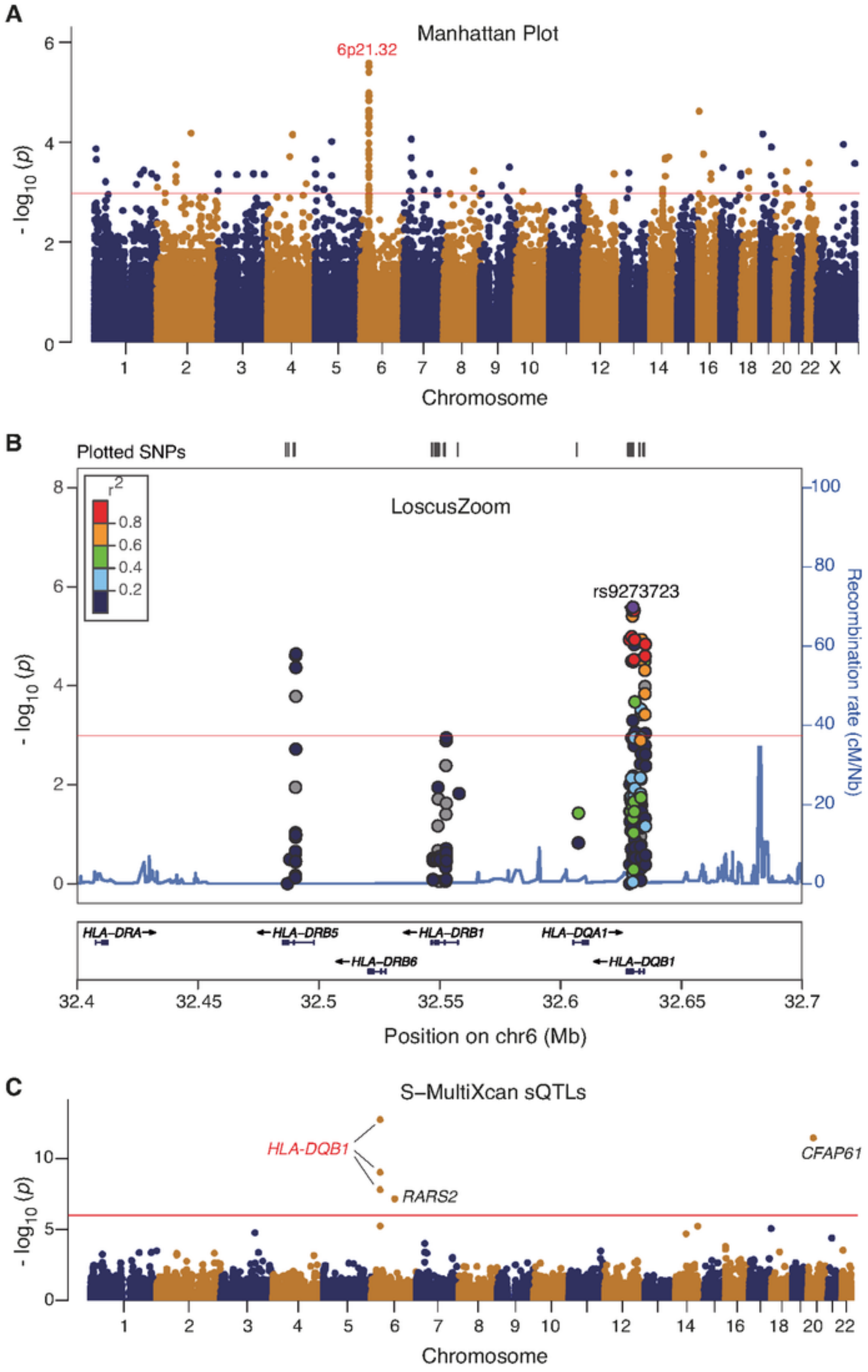

Figure 2

Distribution of Chinese PE-associated risk loci. (A) PE-associated risk loci identified form GWAS. **(B)** Regional plot of a novel risk locus (6p21.32) for PE. Estimated recombination rates (from 1000 Genomes) are plotted in blue. The SNPs within the locus are plotted and color-coded to represent the correlation with the most significant SNP (rs9273723) in *HLA-DQB1*. The reference SNP rs9273723 is plotted in purple. Pairwise r^2 values are from 1000 Genomes East Asian data (Nov 2014 release). Genes and the gene models from UCSC genome browser are noted. LocusZoom [33] was used to generate the plot. **(C)** Distribution of the PE-associated risk loci identified from TWAS. The red lines in **(A)** – **(C)** indicate the cutoffs of statistical significance.

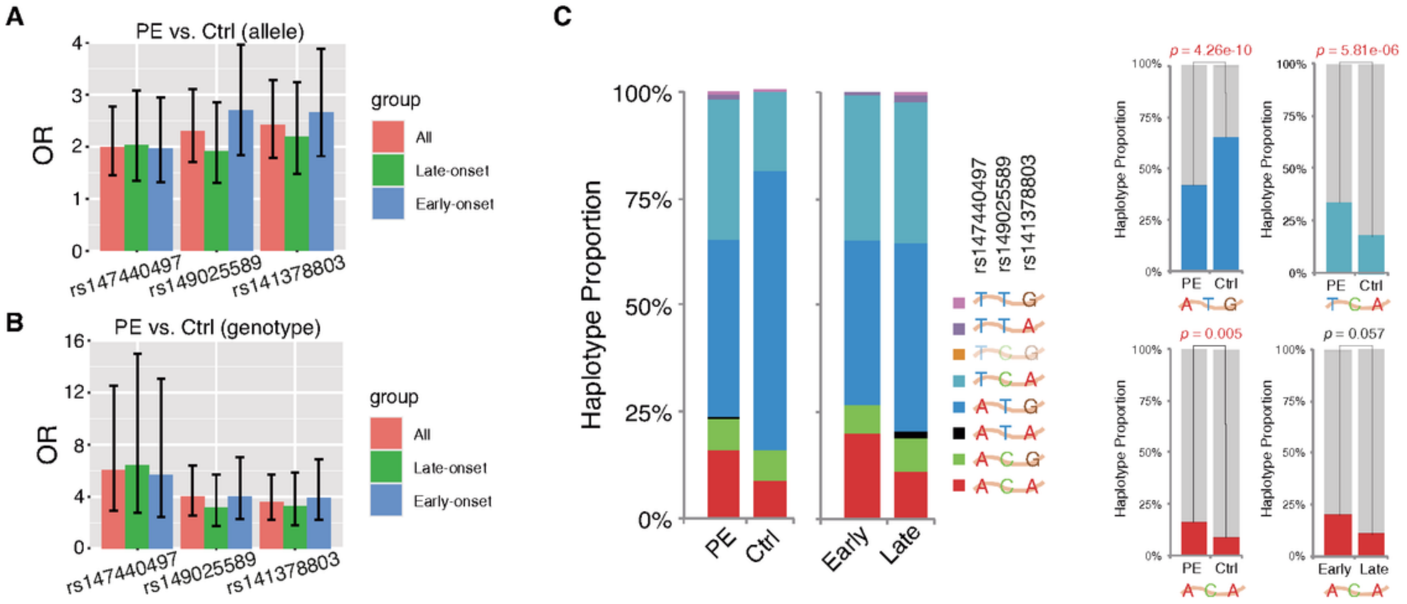

Figure 3

Validation of three Chinese PE risk sites located in *HLA-DRB5*gene. (A) Odd ratios of the minor alleles between PE and healthy pregnancies in the validation cohort. **(B)** Recessive heritability related odd ratios of genotypes between PE and healthy pregnancies in the validation cohort. The odd ratios were calculated between all PE cases (all), late-onset cases or early-onset cases and healthy controls. **(C)** Haplotyping the three SNPs in *HLA-DRB5* and comparison of the haplotypes between PE and healthy controls or between different PE subgroups. Chi-square tests were performed to compare the composition of haplotypes between groups or subgroups. Significance level was preset as $p < 0.05$.

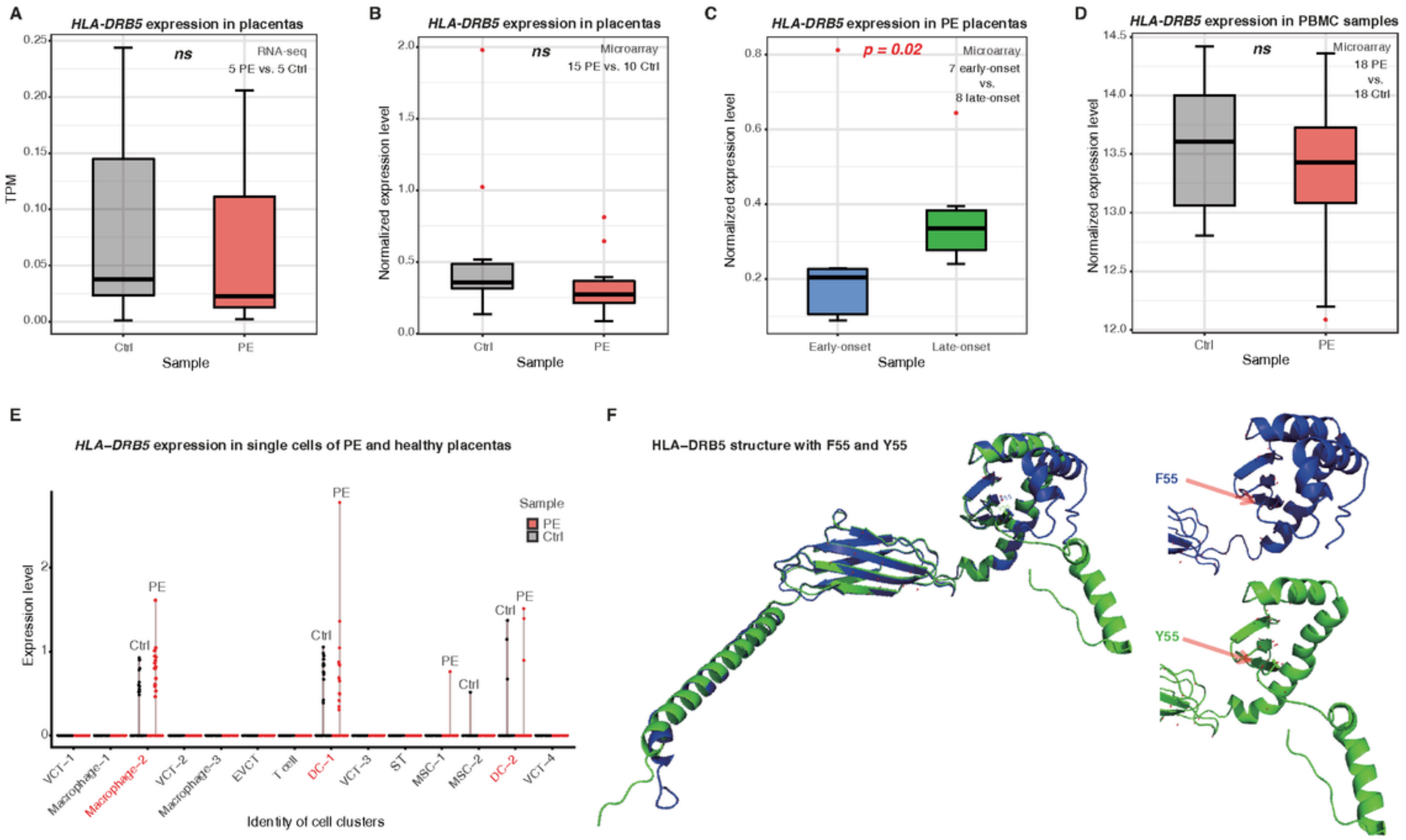

Figure 4

Expression and protein structure of *HLA-DRB5*gene. (A) – (D) *HLA-DRB5* expression in placentas or PBMC samples of PE and healthy subjects. RNA-seq or microarray-based techniques were used to profile the gene expression, as indicated in the figures. The number of samples was also indicated. Gene expression between groups was compared using Mann–Whitney U tests. Significance was predefined as $p < 0.05$. TPM, transcripts per kilobase million; *ns*, not significant. **(E)** Violin diagram of *HLA-DRB5* gene expression in single-cell clusters of PE ($n = 2$) and control ($n = 2$) placentas. The identity of cell clusters was determined according to the specifically expressed marker genes. **(F)** Structure modeling and comparison of HLA-DRB5 protein with F55 and Y55. Structure alignment was shown in the left, and the differential segmental structures were shown in the right individually with the differentiated residues indicated with arrows.

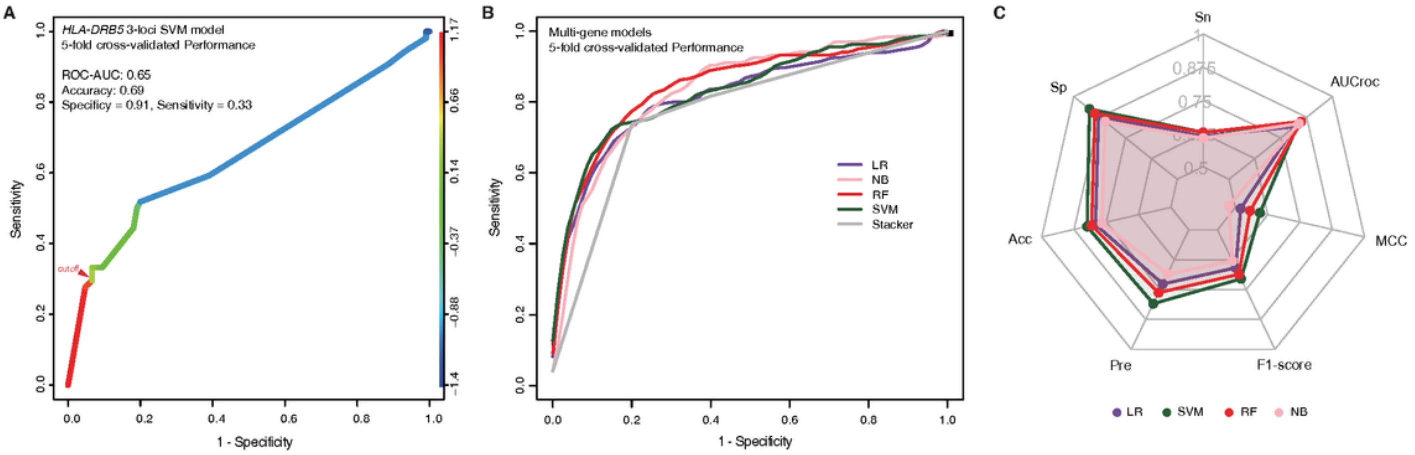

Figure 5

Performance of genotype-based models predicting PE risk. (A) The average ROC curve of the 5-fold cross-validated SVM models. The cutoff was indicated with an arrow after optimization with a specificity larger than 0.90. **(B)** The average 5-fold cross-validated ROC curves of different models. 'Stacker' was a voting-based ensembler of the other four models. **(C)** Performance comparison of the machine-learning models based on 5-fold cross-validation assessment.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- DatasetS1.xlsx
- DatasetS2.xlsx
- FigureS1.pdf
- FigureS2.pdf
- FigureS3.pdf
- FigureS4.pdf
- TableS1.xlsx
- SupplementaryMaterials.docx